ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Αυτόματη ταξινόμηση κειμένων με χρήση Αυτό-Οργανούμενων Χαρτών & μεθόδων Μηχανικής Μάθησης

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ελένη Θ. Γιαννοπούλου

Αθήνα, Ιούνιος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Αυτόματη ταξινόμηση κειμένων με χρήση Αυτό-Οργανούμενων Χαρτών & μεθόδων Μηχανικής Μάθησης

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ελένη Θ. Γιαννοπούλου

Συμβουλευτική Επιτροπή: Μήτρου Νικόλαος
Βασιλείου Ιωάννης
Συκάς Ευστάθιος

Αθήνα, Ιούνιος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Αυτόματη ταξινόμηση κειμένων με χρήση Αυτό-Οργανούμενων Χαρτών & μεθόδων Μηχανικής Μάθησης

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

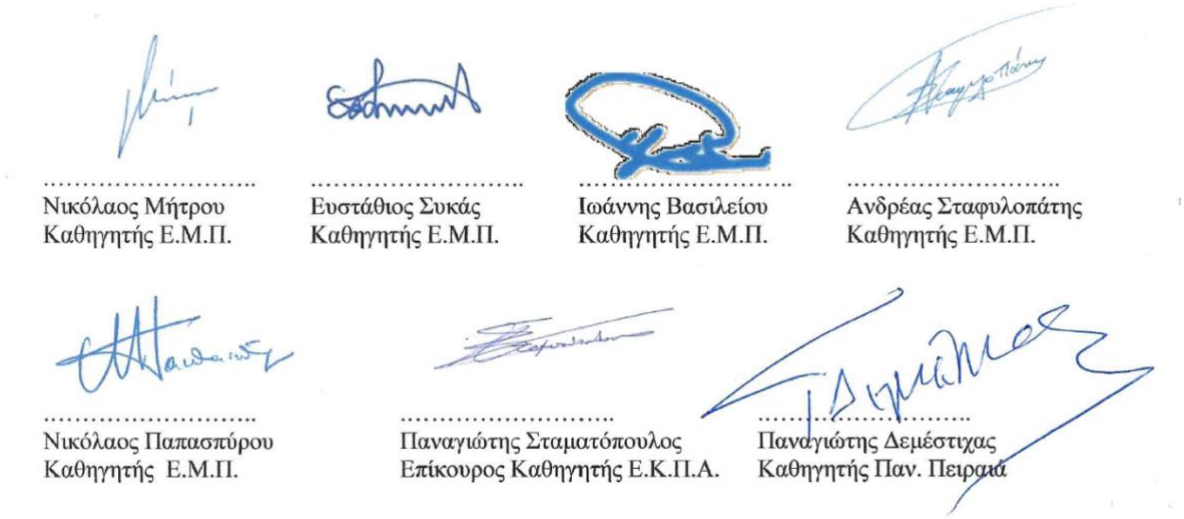
Ελένη Θ. Γιαννοπούλου

Συμβουλευτική Επιτροπή: Νικόλαος Μήτρου

Ευστάθιος Συκάς

Ιωάννης Βασιλείου

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 22α Ιουλίου 2021.



.....	.....	.....	.....
Νικόλαος Μήτρου	Ευστάθιος Συκάς	Ιωάννης Βασιλείου	Ανδρέας Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

.....	.....	.....
Νικόλαος Παπασπύρου	Παναγιώτης Σταματόπουλος	Παναγιώτης Δεμέστιχας
Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Κ.Π.Α.	Καθηγητής Παν. Πειραιά

. . . . . .

Ελένη Θ. Γιαννοπούλου
Διδάκτωρ Ε.Μ.Π

Περίληψη

Η αύξηση του Παγκόσμιου Ιστού τόσο ως προς το πλήθος των συνδεδεμένων κόμβων, όσο και ως προς τον όγκο των πληροφοριών που περιέχει έχει οδηγήσει σε δυσκολίες αποτελεσματικής αναζήτησης και ανάκτησης πληροφοριών από τους τελικούς χρήστες. Αντίστοιχα, σε μικρότερη κλίμακα, στα πλαίσια μιας Ψηφιακής Βιβλιοθήκης ή ενός Ιδρυματικού Αποθετηρίου, η αύξηση του όγκου των πληροφοριών τείνει να μειώσει την αποτελεσματικότητα αναζήτησης. Έτσι, δημιουργήθηκε η ανάγκη για την ανάπτυξη νέων τρόπων αναπαράστασης της διαθέσιμης πληροφορίας, πρόσβασης σε αυτήν και μετατροπής της εν τέλει σε γνώση. Ως καταλληλότερη τεχνολογία για την αποτελεσματική αναζήτηση και ανάκτηση πληροφορίας από κείμενα θεωρούνται οι τεχνικές Μηχανικής Μάθησης και πιο συγκεκριμένα τεχνικές που βασίζονται στην Μη Εποπτευόμενη και Βαθιά Μηχανική Μάθηση. Οι εν λόγω τεχνικές έχουν τη δυνατότητα να ανακαλύπτουν συναφή κείμενα με αυτόματο τρόπο χρησιμοποιώντας μέτρα ομοιότητας διανυσμάτων. Ειδικότερα, οι τεχνικές Μη Εποπτευόμενης Μηχανικής Μάθησης προκρίνονται, στη συγκεκριμένη περίπτωση, έναντι των αντίστοιχων τεχνικών Εποπτευόμενης Μηχανικής Μάθησης, καθώς οι τελευταίες απαιτούν ένα εκτεταμένο, σχολαστικά επισημασμένο σύνολο δεδομένων, που συνήθως δύσκολα είναι διαθέσιμο σε πραγματικές εφαρμογές. Η παρούσα διδακτορική Διατριβή εντάσσεται στο ευρύτερο ερευνητικό πεδίο της αυτόματης Εξαγωγής Πληροφορίας από Κείμενα με χρήση τεχνικών Μηχανικής Μάθησης και πραγματεύεται ανοικτά θέματα στην περιοχή αυτή. Συγκεκριμένα, στην παρούσα Διατριβή προσεγγίζεται το δημοφιλές πρόβλημα της αυτόματης εξαγωγής πληροφορίας ταξινόμησης από κείμενα, με μεθόδους/προσεγγίσεις οι οποίες χωρίζονται αδρά σε τέσσερις βασικές κατηγορίες: α) προσεγγίσεις εξαγωγής, β) προσεγγίσεις ανάθεσης, γ) μεικτές προσεγγίσεις και δ) προσεγγίσεις πρόβλεψης. Οι μέθοδοι εξαγωγής πληροφορίας από κείμενα παρουσιάζουν μεγάλη ποικιλομορφία και εφαρμόζονται σε ένα πλήθος πεδίων με ποικίλες εφαρμογές. Αφού παρουσιαστεί, αρχικά, ένα πλήθος διαφορετικών εφαρμογών, όπου οι μέθοδοι εξαγωγής πληροφορίας έχουν υιοθετηθεί με επιτυχία, εξετάζονται τα πλεονεκτήματα που προκύπτουν από την χρήση τέτοιων μεθόδων ειδικότερα στις Ψηφιακές Βιβλιοθήκες. Στη συνέχεια προσεγγίζεται το πρόβλημα της αυτόματης ταξινόμησης ενός συνόλου δεδομένων ειδήσεων, το οποίο μοντελοποιείται ως ένα πρόβλημα ταξινόμησης πολλαπλής ετικέτας. Σε αυτή την περίπτωση χρησιμοποιείται ένα Νευρωνικό Δίκτυο Μη Εποπτευόμενης Μηχανικής Μάθησης, οι Αυτό-Οργανούμενοι Χάρτες (Self-Organized Maps – SOM), ενώ προτείνεται μια απλή, αλλά αποτελεσματική διαδικασία που αντιμετωπίζει το πρόβλημα πολλαπλής ετικέτας ως ένα πρόβλημα ταξινόμησης πολλαπλών κλάσεων. Επιπλέον, προτείνεται ένας έξυπνος αλγόριθμος για την επιλογή ετικετών, με στόχο να δείξει ότι οι γειτονικοί κόμβοι στον Χάρτη επηρεάζουν την επιλογή των ετικετών για έναν συγκεκριμένο κόμβο. Τέλος, εφαρμόζεται μια ευρετική μέθοδος για την επιλογή του μεγέθους του SOM. Η εκτεταμένη πειραματική ανάλυση που πραγματοποιήθηκε έδειξε ότι η προτεινόμενη λύση βελτιώνει την αποτελεσματικότητα της ταξινόμησης, όχι μόνο όσον αφορά στην ακρίβεια, αλλά και στους υπολογιστικούς πόρους που απαιτούνται και στο χρόνο για την εκπαίδευση του Δικτύου. Στα πλαίσια της παρούσας Διατριβής πραγματοποιείται, επίσης, μια επισκόπηση των μεθόδων ταξινόμησης πολλαπλών κλάσεων, ενώ προτείνεται μια διαδικασία για την αυτόματη ταξινόμηση ηλεκτρονικών βιβλίων εξάγοντας πληροφορία από τους πίνακες περιεχομένων των βιβλίων. Στην περίπτωση αυτή χρησιμοποιήθηκε ένα νευρωνικό δίκτυο μη εποπτευόμενης μηχανικής μάθησης (SOM) και δύο αρχιτεκτονικές Νευρωνικών Δικτύων Βαθιάς Μάθησης κάτω από διαφορετικά σενάρια διαμόρφωσης. Στόχος της διαδικασίας αυτής ήταν η μελέτη ανάπτυξης ενός

συστήματος συστάσεων για την υποστήριξη φοιτητών και καθηγητών στον εντοπισμό σχετικών πηγών βάσει μιας λεπτομερούς θεματικής περιγραφής (π.χ. της περίληψης ή του πίνακα περιεχομένων ενός βιβλίου) αντί για μερικές λέξεις-κλειδιά με βάση την πειραματική ανάλυση που πραγματοποιήθηκε. Τέλος, στα πλαίσια της Διατριβής αυτής προτείνεται η δημιουργία μιας Πύλης Διασυνδεδεμένων Δεδομένων με χρήση τεχνολογιών Σημασιολογικού Ιστού, με στόχο την ενσωμάτωση των μηχανισμών αυτόματης εξαγωγής πληροφορίας ταξινόμησης και των αποτελεσμάτων αυτών και απώτερο σκοπό τον εμπλουτισμό μεταδεδομένων, έτσι ώστε να υποβοηθηθεί η αποτελεσματικότερη αναζήτηση και ανάκτηση πληροφοριών από τους τελικούς χρήστες στις συλλογές μιας Ψηφιακής Βιβλιοθήκης.

Λέξεις-κλειδιά: Αυτόματη Εξαγωγή Πληροφορίας, Τεχνητή Νοημοσύνη, Στατιστική Ανάλυση Φυσικής Γλώσσας, Αλγόριθμοι Ταξινόμησης, Ταξινόμηση Πολλαπλής Ετικέτας, Αυτό-Οργανούμενοι Χάρτες, Μη εποπτευόμενη Μηχανική Μάθηση, Νευρωνικά Δίκτυα Βαθιάς Μάθησης, Εμπλουτισμός Μεταδεδομένων, Ψηφιακές Βιβλιοθήκες, Ταξινόμηση Βιβλίων, Διανυσματικοποίηση Πίνακα Περιεχομένων.

Abstract

The growth of the World Wide Web both in terms of the number of connected nodes and the volume of information it contains has led to difficulties in effectively searching and retrieving information from end users. Similarly, on a smaller scale, in the context of a Digital Library or an Institutional Repository, increasing the volume of information tends to reduce search efficiency. Thus, the need arose for the development of new ways of representing the available information, accessing it and finally turning it into knowledge. Machine Learning techniques and more specifically techniques based on Unsupervised and Deep Neural Networks are considered as the most appropriate technology for the effective search and retrieval of information from texts. These techniques have the ability to discover related texts automatically using vector similarity measures. In particular, Unsupervised Learning techniques are, in this case, superior to the corresponding Supervised Learning techniques, as the latter require an extensive, meticulously labeled data set, which is usually difficult to develop in real applications. This doctoral dissertation is part of the broader research field of "Automatic Information Extraction from texts" using Machine Learning Techniques and deals with open topics in this area. Specifically, this dissertation approaches the popular problem of automatically extracting information from texts, with methods / approaches that are roughly divided into four main categories: a) extraction, b) assignment, c) hybrid and d) prediction methods. Methods of extracting information from texts are very diverse and are applied in a number of fields with a variety of applications. Having first introduced a number of different applications, where information extraction methods have been successfully adopted, the advantages of using such methods, especially in Digital Libraries, are examined. Then the problem of automatically classifying a news data set is approached, which is modeled as a multi-label classification problem. In this case, a Self-Organized Maps (SOM) Neural Network is used, while a simple but effective procedure is proposed that transforms the multi-label problem into a multi-class classification problem. In addition, an intelligent algorithm for selecting labels is proposed, in order to show that the neighboring nodes in the Map affect the selection of tags for a specific node. Finally, a heuristic method is used to select the size of the SOM. The extensive experimental analysis carried out showed that the proposed solution improves the efficiency of the classification, not only in terms of accuracy, but also in terms of computational resources and time required for the training of the Network. In the framework of this dissertation, an overview of multi-class classification methods is also carried out, while a process for the automatic classification of e-books by extracting information from the table of contents of the books is proposed. In this case an unsupervised machine learning neural network (SOM) and two Deep Neural Network architectures were used under different configuration scenarios. The aim of this process was to study the feasibility of developing a recommendations system to support students and teachers in identifying relevant sources based on a detailed thematic description (e.g., the summary or table of contents of a book) instead of some keywords based on the experimental analysis performed. Finally, in the context of this dissertation, the creation of a Linked Data portal using Semantic Web technologies is proposed, with the aim of integrating information extraction mechanisms and their results, with the ultimate goal of enriching metadata , in order to assist efficient search and retrieval of information from end users in the collections of a Digital Library.

Keywords: Information Extraction, Artificial Intelligence, Statistical Natural Language Processing, Classification algorithms, Multi-label Classification, Self-Organizing Maps, Unsupervised

Machine Learning, Deep Neural Networks, Metadata Enrichment, Digital Libraries, Book Classification, ToC Vectorization.

Ευχαριστίες

Η ολοκλήρωση αυτής της διδακτορικής Διατριβής ήταν μια μακρά και δύσκολη διαδικασία, η οποία απαιτούσε προσπάθεια και απόλυτη αφοσίωση. Παρά τις αντιξοότητες και τις δυσκολίες κατάφερα να αποκτήσω πολύτιμες γνώσεις στον τομέα της αυτόματης Εξαγωγής Πληροφορίας ταξινόμησης από κείμενα με χρήση τεχνικών Μηχανικής Μάθησης, καθώς μου δόθηκε η ευκαιρία να συμμετάσχω σε συναφή ερευνητικά έργα και να μελετήσω σημαντικά άρθρα σε αυτόν τον τομέα. Κανένα από τα παραπάνω δεν θα είχε επιτευχθεί χωρίς την πραγματική υποστήριξη πολλών ανθρώπων των οποίων η συμβολή στην έρευνά μου, με διάφορους τρόπους, ήταν σημαντική και αξίζει ιδιαίτερης αναφοράς.

Η ανά χείρας διδακτορική Διατριβή εκπονήθηκε στη Σχολή ΗΜΜΥ του Εθνικού Μετσόβιου Πολυτεχνείου υπό την επίβλεψη του καθηγητή κ. Νικολάου Μήτρου, στον οποίο θα ήθελα από τη θέση αυτή να εκφράσω την ευγνωμοσύνη μου για την ευκαιρία που μου έδωσε, για τον χρόνο που μου αφιέρωσε, καθώς και για την εμπιστοσύνη, την υποστήριξη, την καθοδήγηση και τις διαφωτιστικές υποδείξεις του. Θα ήθελα, επίσης, να ευχαριστήσω θερμά και τους καθηγητές κ. Ευστάθιο Συκά και κ. Ιωάννη Βασιλείου για την υποστήριξη και τις εποικοδομητικές συμβουλές τους καθόλη τη διάρκεια εκπόνησης της παρούσας Διατριβής, καθώς και τους καθηγητές κ. Νικόλαο Παπασπύρου, κ. Ανδρέα Σταφυλοπάτη, τον Επικ. Καθηγητή κ. Παναγιώτη Σταματόπουλο και τέλος τον Καθηγητή κ. Παναγιώτη Δεμέστιχα για την τιμή που μου έκαναν να συμμετέχουν στην επιτροπή αξιολόγησης της διδακτορικής μου Διατριβής.

Θερμές ευχαριστίες οφείλω να εκφράσω, επίσης στους Δημήτρη Σπανό, Σταμάτη Αρκουλή, Περικλή Σταύρου, Απόστολο Παλαιό, Ελένη Παπαδάτου και Νικόλαο Μακρή καθώς οι συζητήσεις μας, αλλά και το ευχάριστο περιβάλλον που επικρατούσε στην μεταξύ μας συνεργασία, με ενθάρρυναν να συνεχίσω και να φέρω εις πέρας την παρούσα επιστημονική προσπάθεια.

Ακόμα, αισθάνομαι την ανάγκη να ευχαριστήσω τους γονείς μου, Θεόδωρο και Αναστασία, και τον αδελφό μου, Εμμανουήλ, για την ηθική υποστήριξη καθόλη τη διάρκεια της εκπόνησης της παρούσας Διατριβής. Τέλος, θέλω να ευχαριστήσω το σύζυγό μου Βασίλη καθώς χωρίς την υποστήριξή, κατανόηση και καθοδήγησή του το ακαδημαϊκό αυτό εγχείρημα θα ήταν αδύνατο να ολοκληρωθεί. Είμαι για πάντα ευγνώμων για την υποστήριξη του όλα αυτά τα χρόνια.

Ελένη Θ. Γιαννοπούλου

Αθήνα, Ιούνιος 2021

Πίνακας Περιεχομένων

Ευρετήριο Εικόνων

Ευρετήριο Πινάκων

1. Εισαγωγή

Τα τελευταία χρόνια ο Παγκόσμιος Ιστός αυξάνεται με ιλιγγιώδεις ρυθμούς, τόσο ως προς το πλήθος των συνδεδεμένων κόμβων, όσο και ως προς τον όγκο των πληροφοριών που περιέχει. Αυτό το γεγονός έχει οδηγήσει στη μετατροπή της έννοιας του Παγκόσμιου Ιστού, ο οποίος σταδιακά αλλάζει τόσο σε πολυπλοκότητα όσο και στην υποκείμενη αρχιτεκτονική. Το όραμα του εμπνευστή του Παγκόσμιου Ιστού, Tim Berners–Lee ήταν ένα παγκόσμιο δίκτυο πληροφοριών, που θα είχε ως σκοπό την παροχή υψηλού επιπέδου υπηρεσιών πλοήγησης και αναζήτησης στους χρήστες του. Με την αύξηση όμως του όγκου της διαθέσιμης πληροφορίας, τα υπάρχοντα μέσα πλοήγησης και αναζήτησης δεν ήταν εύκολο να επιτύχουν τον στόχο τους. Έτσι, δημιουργήθηκε η ανάγκη για την ανάπτυξη νέων τρόπων αναπαράστασης και πρόσβασης στην διαθέσιμη πληροφορία αλλά και στη μετατροπή της πληροφορίας αυτής σε γνώση.

Η μετατροπή αυτή ακολουθήθηκε από ένα σύνολο κανόνων, πρακτικών αλλά και τεχνολογιών όπως Οντολογίες (Ontologies) και ελεγχόμενα λεξιλόγια (Controlled Vocabularies) ή Θησαυροί (Thesaurus), τόσο για την δημοσίευση αλλά και για τη διασύνδεση δομημένων δεδομένων στον Ιστό, τα οποία αποκαλούνται Διασυνδεδεμένα Ανοικτά Δεδομένα (Linked Open Data - LOD) [1]. Η προσπάθεια για τη δημιουργία του νέου αυτού Ιστού επικεντρώνεται στον τρόπο δημοσίευσης και «κατανάλωσης» της πληροφορίας έτσι ώστε η τελευταία να είναι κατανοητή από μηχανές (machine readable), οι οποίες θα είναι σε θέση να αντλήσουν συμπεράσματα από μετά-δεδομένα (metadata) τα οποία είναι συνδεδεμένα με πληροφορίες δημοσιευμένες στον Ιστό. Η τάση αυτή είναι πολλά υποσχόμενη και τα τελευταία χρόνια έχει υιοθετηθεί σε ένα πλήθος ερευνητικών και εμπορικών εφαρμογών, παρόλο που ακόμα δεν χρησιμοποιούνται στο έπακρο οι δυνατότητες που έχει να προσφέρει.

Παρόλο τον όγκο των δεδομένων που έχουν διαχυθεί και συνεχίζουν να διαχέονται με γρήγορους ρυθμούς στον Ιστό, αυτά διατίθενται σε διάφορα μορφότυπα (formats), τα οποία τις περισσότερες φορές είναι δύσκολα διαχειρίσιμα και πολλές φορές είναι σχεδόν αδύνατη η μακροχρόνια διατήρησή τους. Αυτό έχει ως άμεσο αποτέλεσμα την απώλεια τόσο της δομής όσο και της υποκείμενης σημασιολογίας αυτών. Ακόμα, είναι κοινός τόπος πως η αναζήτηση βασίζεται σε λέξεις-κλειδιά, η οποία έχει τον περιορισμό ότι η λέξη-κλειδί πρέπει να ταιριάζει επακριβώς με τη λέξη η οποία βρίσκεται μέσα στην πηγή δεδομένων. Όλα τα παραπάνω οδηγούν σε αδυναμία επιτυχούς αναζήτησης σε ανομοιογενείς και κατανεμημένες πηγές. Στόχος λοιπόν είναι η δόμηση των πληροφοριών με τέτοιο τρόπο έτσι ώστε να καθίσταται δυνατή η εξαγωγή συμπερασμάτων από τις υπάρχουσες πληροφορίες οδηγώντας έτσι σε ανακάλυψη νέας γνώσης με βάση την υπάρχουσα αλλά ταυτόχρονα και η κατανόηση της σημασίας των δεδομένων, έτσι ώστε η πληροφορία αυτή να είναι πλήρως κατανοητή και επεξεργάσιμη με αυτόματο τρόπο. Αυτό μπορεί να επιτευχθεί με τη δημιουργία εργαλείων επισημείωσης (annotation tools) αλλά και μηχανών συλλογισμού (inference engines). Επειδή όμως η γνώση που κρύβεται στην σημασιολογία των υποκείμενων δεδομένων πολλές φορές είναι δύσκολο να εξαχθεί, υπάρχει μεγάλο ερευνητικό ενδιαφέρον και προς αυτή την κατεύθυνση, καθώς έχουν δημοσιευθεί διάφορες προσεγγίσεις για την εξαγωγή μετά-δεδομένων από πηγές δεδομένων με χρήση τεχνικών μηχανικής μάθησης (machine learning) αλλά και επεξεργασίας φυσικής γλώσσας (natural language processing). Σκοπός των παραπάνω προσεγγίσεων αποτελεί η εξαγωγή χρήσιμης πληροφορίας από ένα δεδομένο τμήμα κειμένου.

Η Αυτόματη Εξαγωγή Πληροφορίας (Automatic Information Extraction) από κείμενα, ενώ αποτελεί μια ιδέα που εδώ και αρκετά χρόνια έχει απασχολήσει την ερευνητική κοινότητα, έχει διαμορφωθεί κάτω από διαφορετικό πρίσμα τον τελευταίο καιρό λόγω της επινόησης και της μεγάλης ανάπτυξης του Σημασιολογικού Ιστού. Οι μέθοδοι εξαγωγής πληροφορίας από κείμενα παρουσιάζουν μεγάλη ποικιλομορφία και μπορούν να ενσωματωθούν σε ένα πλήθος ερευνητικών πεδίων παρέχοντας εφαρμογές που έχουν ως στόχο την παροχή αξιοποιήσιμων υπηρεσιών προς τους τελικούς χρήστες. Με αυτό τον τρόπο είναι δυνατόν να προκύψει ένα πλήθος διαφορετικών εφαρμογών όπου οι τεχνικές εξαγωγής πληροφορίας μπορούν να υιοθετηθούν με επιτυχία, καθώς είναι ποικίλα τα πλεονεκτήματα που απορρέουν από την ενσωμάτωση τέτοιων μεθόδων για τους τελικούς χρήστες. Σε επόμενη παράγραφο του παρόντος, εξετάζονται ενδελεχώς οι δυνατότητες εφαρμογής των τεχνικών αυτών σε διάφορους τομείς, όπως στις Ψηφιακές Βιβλιοθήκες.

Η παρούσα διδακτορική Διατριβή πραγματεύεται την αυτόματη Εξαγωγή Πληροφορίας ταξινόμησης από κείμενα με χρήση τεχνικών Μηχανικής Μάθησης (ΜΜ), όπως οι Αυτό-Οργανούμενοι Χάρτες (Self-Organising Maps) και τα Νευρωνικά Δίκτυα Βαθιάς Μάθησης (Deep Neural Networks) για την αποτελεσματική αναζήτηση και ανάκτηση συναφών εγγράφων στα πλαίσια μιας Ψηφιακής Βιβλιοθήκης. Όπως είναι φανερό το συγκεκριμένο πρόβλημα σε συνδυασμό με τον όγκο τον διαθέσιμων δεδομένων που συνεχώς αυξάνονται αλλά και ο παράγοντας της υποκειμενικότητας στο τι θεωρείται συναφές έγγραφο, θέτει συγκεκριμένες προκλήσεις που αφορούν τόσο στην μοντελοποίηση του υπό εξέταση πεδίου όσο και στον τρόπο εφαρμογής τεχνικών εξαγωγής πληροφορίας σε γενικότερο πλαίσιο. Παρόλα αυτά, τα οφέλη που προκύπτουν είναι πολλά, πιο συγκεκριμένη αναφορά όμως θα γίνει σε επόμενη παράγραφο.

Η διάρθρωση του Κεφαλαίου που ακολουθεί έχει ως εξής. Αρχικά αναλύεται το πλαίσιο έρευνας της παρούσας Διατριβής, συγκεκριμένα γίνεται μια εισαγωγή στο πεδίο της Εξαγωγής Πληροφορίας από κείμενα ενώ στην επόμενη παράγραφο γίνεται μια σύντομη αναφορά στην αυτόματη ταξινόμηση πολλαπλής ετικέτας, ενός συνόλου δεδομένων ειδήσεων, με χρήση ενός Νευρωνικού Δικτύου μη εποπτευόμενης Μηχανικής Μάθησης, συγκεκριμένα των Αυτό-Οργανούμενων Χαρτών (Self-Organizing Maps - SOM). Τέλος, στην τρίτη υποενότητα γίνεται μια αναφορά στις τεχνικές αυτόματης Εξαγωγής Πληροφορίας από κείμενα με χρήση τεχνικών Αυτό-Οργανούμενων Χαρτών και δικτύων Βαθιάς Μάθησης και πώς αυτές μπορούν να εφαρμοστούν για την ταξινόμηση πολλαπλών κλάσεων, ενώ προτείνεται μια διαδικασία για την αυτόματη ταξινόμηση ηλεκτρονικών βιβλίων εξάγοντας πληροφορία από τους πίνακες περιεχομένων των βιβλίων.

1.1. Πλαίσιο Έρευνας

1.1.1. Αυτόματη Εξαγωγή Πληροφορίας από κείμενα

Η τρέχουσα ενότητα αποτελεί μια εισαγωγή στο θέμα της αυτόματης Εξαγωγής Πληροφορίας από κείμενα, των μεθόδων που χρησιμοποιούνται αλλά και των πλεονεκτημάτων ή/και των περιορισμών που οι μέθοδοι αυτές μπορεί να παρουσιάζουν. Στον τομέα των Ψηφιακών Βιβλιοθηκών όσο και στο πεδίο της επιστήμης των Πληροφοριών μια λέξη-κλειδί (keyword) καθορίζεται ως μια λέξη η οποία περιγράφει ευκρινώς και με ακρίβεια το θέμα ή ένα μέρος του θέματος που αναφέρεται σε ένα έγγραφο [5]. Για τον ίδιο λόγο οι βιβλιοθηκονόμοι χρησιμοποιούν θεματικές επικεφαλίδες (subject headings). Οι θεματικές επικεφαλίδες αποτελούν όρους ή φράσεις που περιγράφουν το περιεχόμενο ενός εγγράφου και επιπρόσθετα μπορούν να χρησιμοποιηθούν για να ομαδοποιήσουν και να οργανώσουν τις συλλογές μιας βιβλιοθήκης [20], [26]. Για παράδειγμα,

έγγραφα τα οποία σχετίζονται με το ίδιο θέμα μπορούν να ομαδοποιηθούν και να υπάρχει σε αυτά γρήγορη πρόσβαση με την επιλογή μιας συγκεκριμένης θεματικής επικεφαλίδας.

Όπως οι θεματικές επικεφαλίδες έτσι και οι λέξεις-κλειδιά, καθώς και σχετικοί με το θέμα όροι, μπορούν να επιλεγούν από ένα τυποποιημένο σύνολο περιγραφέων όπως ένα ελεγχόμενο λεξιλόγιο (controlled vocabulary). Ένας άλλος τρόπος για την ανάθεση των όρων είναι η εξαγωγή τους από το σώμα του κειμένου ή όπως κάνουν συνήθως οι βιβλιοθηκονόμοι να τα εξάγουν μόνο από τον τίτλο του κειμένου. Η έννοια της αυτόματης Εξαγωγής Πληροφορίας (Automatic Information Extraction) από κείμενα αναφέρεται στην εξαγωγή από κάποιο κείμενο μιας συγκεκριμένου μεγέθους λίστας όρων, στην οποία κάθε μια εγγραφή μπορεί να αποτελείται από δύο ή περισσότερες λέξεις [6], [18]. Στόχος αυτής της λίστας είναι να αποτυπώσει τα βασικά σημεία ή τις βασικές έννοιες ενός δεδομένου εγγράφου. Σε σχέση με την ευρετηρίαση πλήρους κειμένου (full-text indexing) όπου όλοι οι όροι που εμφανίζονται στο κείμενο αποθηκεύονται σε κάποια μονάδα αποθήκευσης όπως μια βάση δεδομένων, η ευρετηρίαση με βάση λέξεις ή φράσεις κλειδιά απαιτεί λιγότερο αποθηκευτικό χώρο και παρέχει μια συνοπτική επισκόπηση των θεματικών περιοχών από μια συλλογή ή ένα μεμονωμένο κείμενο.

Κατά τη δημοσίευση ενός κειμένου σε ένα επιστημονικό περιοδικό ζητείται από τους συγγραφείς, εκτός από το κείμενο της δημοσίευσης και μια λίστα με αντιπροσωπευτικές λέξεις ή φράσεις. Συνήθως οι λέξεις ή οι φράσεις αυτές είτε παρέχονται απ' ευθείας από τους συγγραφείς είτε από εξειδικευμένο προσωπικό το οποίο είναι υπεύθυνο για την κατάρτιση της λίστας αυτής. Η διαδικασία αυτή μπορεί να είναι επίπονη και χρονοβόρα και επίσης μειονεκτεί στο ότι είναι χειρωνακτική και άρα υπόκειται σε λάθη. Ακόμα είναι μια υποκειμενική διαδικασία στην οποία πολύ σημαντικό ρόλο παίζει ο ανθρώπινος παράγοντας διότι διαφορετικός άνθρωπος μπορεί να καταρτίσει διαφορετική λίστα φράσεων. Η παραπάνω περίπτωση αποτελεί ένα κλασικό πρόβλημα στο οποίο η εφαρμογή μεθόδων αυτόματης Εξαγωγής Πληροφορίας ταξινόμησης με χρήση τεχνικών Μηχανικής Μάθησης μπορεί να δώσει λύση.

Όλα τα ως άνω προβλήματα είναι δυνατόν να ξεπεραστούν εάν η παραπάνω διαδικασία αυτοματοποιηθεί με κάποιο τρόπο. Στο σημείο αυτό μπορούμε να πούμε ότι το πρόβλημα μπορεί να μετατραπεί από την εξαγωγή πληροφορίας από κείμενα με χειροκίνητο τρόπο, σε εξαγωγή πληροφορίας με αυτόματο ή ημιαυτόματο τρόπο. Η χρήση της πληροφορίας που εξάγεται μπορεί να υπηρετήσει πολλούς και διαφορετικούς σκοπούς. Η σύνοψη του θέματος ενός επιστημονικού περιοδικού (summarization) αποτελεί έναν από αυτούς. Σε αυτή την περίπτωση είναι δυνατόν να δημιουργηθεί η σύνοψη του υποκείμενου θέματος χρησιμοποιώντας τεχνικές εξαγωγής πληροφορίας, το αποτέλεσμα των οποίων θα τοποθετηθεί στην αρχική σελίδα του άρθρου ή γενικότερα του υπό επεξεργασία κειμένου, στη μορφή μιας «ειδικής» περίληψης [7]. Άμεσο πλεονέκτημα αυτού είναι να δύναται ο εκάστοτε χρήστης να αναγνωρίσει σε ελάχιστο χρόνο εάν το θέμα ενός δεδομένου κειμένου είναι μέσα στα ενδιαφέροντά του ή όχι.

Ένα ακόμα πλεονέκτημα των φράσεων-κλειδιών που προκύπτουν από την εξαγωγή είναι η πολύ-λειτουργικότητα (multi-functionality) που τις διακρίνει. Οι φράσεις-κλειδιά αλλά και τα σύνολα φράσεων μπορεί να χρησιμοποιηθούν για ποικίλες εφαρμογές επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP) όπως κατηγοριοποίηση κειμένου (text clustering) και ταξινόμηση (classification) [8]. Δύο ακόμα περιοχές όπου οι τεχνικές αυτόματης εξαγωγής πληροφορίας ταξινόμησης μπορούν να εφαρμοστούν με επιτυχία, αποτελούν η ανάκτηση πληροφοριών που βασίζεται στο περιεχόμενο (content-based retrieval) και η θεματική αναζήτηση (topic search). Τέλος,

όπως αναφέρθηκε παραπάνω, η πληροφορία που εξάγεται μπορεί να χρησιμοποιηθεί αποδοτικά κατά την αναζήτηση [9] ή και την πλοήγηση είτε στον Παγκόσμιο Ιστό είτε πιο συγκεκριμένα στα πιο στενά όρια μιας Ψηφιακής Βιβλιοθήκης. Ένας παράγοντας που παίζει πολύ σημαντικό ρόλο στη συγκεκριμένη περίπτωση είναι η ποιότητα της πληροφορίας που θα εξαχθεί από το κείμενο. Για το λόγο αυτό, είναι πολύ σημαντικό να έχουν εξαχθεί ποιοτικοί και ταυτόχρονα ακριβείς περιγραφείς οι οποίοι σε περίπτωση που η εξαγωγή αυτών δεν γίνεται με αυτόματο ή ημιαυτόματο τρόπο, θα παρέχονται από ανθρώπους ειδικούς στο αντικείμενο.

Οι λέξεις ή οι φράσεις που εξάγονται μπορούν ακόμα να χρησιμοποιηθούν για λόγους ευρετηρίασης σε ένα ευρετήριο. Σε αυτή την περίπτωση ένας χρήστης είναι δυνατόν να βρει ένα άρθρο που μπορεί να καλύπτει τις ανάγκες του μέσα σε ένα μικρό χρονικό διάστημα. Επίσης, είναι δυνατή η χρήση της λίστας αυτής κατά τη διάρκεια μιας αναζήτησης. Οι φράσεις αυτές μπορούν να αποτελέσουν την είσοδο μιας μηχανής αναζήτησης αντί για μεμονωμένες λέξεις κλειδιά έτσι ώστε η τελευταία να επιστρέψει πιο σχετικά ή/και πιο ακριβή αποτελέσματα. Τα αποτελέσματα της αναζήτησης μπορούν να βελτιστοποιηθούν περαιτέρω εάν η λίστα αυτή των φράσεων συνδυαστεί ή ορισμένες φράσεις που ανήκουν στη λίστα αντιστοιχηθούν με όρους που προέρχονται από κάποιο ελεγχόμενο λεξιλόγιο. Έτσι μπορεί να καταστεί δυνατή η ανάκτηση περισσότερο σχετικών ή πιο γενικών ή ακόμα και πιο ειδικών αποτελεσμάτων λόγω του ότι τα ελεγχόμενα λεξιλόγια ή οι θησαυροί ενσωματώνουν συνήθως ένα είδος ιεραρχίας. Απόρροια αυτού θα είναι μια μικρότερη σε μέγεθος και ταυτόχρονα πιο σχετική λίστα αποτελεσμάτων με περισσότερο ποιοτικά αποτελέσματα. Όπως είναι φανερό από τα παραπάνω, παρόλο που οι τεχνικές εξαγωγής πληροφορίας μπορεί να χρησιμοποιούνται για να επιτελέσουν διαφορετικούς σκοπούς, υπάρχει μια κοινή απαίτηση: μια μικρή λίστα από φράσεις οι οποίες θα αποτυπώνουν το βασικό νόημα ενός εγγράφου [6], [18]. Οι υπάρχουσες προσεγγίσεις εξαγωγής πληροφορίας από κείμενα μπορεί να χωριστούν σε δύο μεγάλες κατηγορίες [20], [26]. Μπορούν να θεωρηθούν είτε ως μια διαδικασία εξαγωγής (extraction) είτε ως μια διαδικασία ανάθεσης (assignment). Στην πρώτη περίπτωση, οι όροι και οι φράσεις που εμφανίζονται στο σώμα του κειμένου αναλύονται με σκοπό τελικά να επιλεγούν οι πιο αντιπροσωπευτικές. Στη δεύτερη περίπτωση, τα έγγραφα κατηγοριοποιούνται σε έναν προκαθορισμένο αριθμό ομάδων με βάση αυτούς τους όρους.

1.1.2. Αυτόματη ταξινόμηση πολλαπλής ετικέτας με χρήση Αυτό-Οργανούμενων χαρτών

Η εξέλιξη του Διαδικτύου δημιούργησε νέα πρότυπα στον τρόπο διάθεσης, ευρετηρίασης, αναζήτησης και ανάκτησης ηλεκτρονικών εγγράφων. Επιπλέον, ο αριθμός των ηλεκτρονικών εγγράφων που διατίθενται μέσω του διαδικτύου αυξάνεται συνεχώς, δημιουργώντας έτσι προκλήσεις σχετικά με τη διαθεσιμότητα, την ανάκτηση και τη μακροπρόθεσμη διατήρησή τους. Σε αυτό το πλαίσιο, η αυτόματη ταξινόμηση εγγράφων και πιο συγκεκριμένα η αυτόματη ταξινόμηση πολλαπλής ετικέτας (multi-label), κατά την οποία σε κάθε έγγραφο έχουν ανατεθεί περισσότερες από μία ετικέτες κάθε φορά, καθίσταται πολύ σημαντική. Επιπλέον, η εργασία ταξινόμησης πολλαπλής ετικέτας είναι πιο περίπλοκη σε σύγκριση με τη δυαδική ταξινόμηση (binary) ή την ταξινόμηση πολλαπλών κλάσεων (multi-class) καθώς η υπόθεση ότι οι ετικέτες που εκχωρούνται σε κάθε στοιχείο είναι στοχαστικά ανεξάρτητες δεν ισχύει πάντα [39]. Στην περίπτωση της αυτόματης ταξινόμησης πολλαπλής ετικέτας κάθε στοιχείο δεδομένων διαθέτει περισσότερες από μία ετικέτες κάθε φορά, ενώ στην ταξινόμηση πολλαπλών κλάσεων κάθε στοιχείο δεδομένων διαθέτει μόνο μια ετικέτα κάθε φορά. Η ταξινόμηση πολλαπλών κλάσεων διαφέρει σε σχέση με τη δυαδική ταξινόμηση καθώς στην πρώτη περίπτωση ο ταξινομητής έχει ως στόχο να αναθέσει το δείγμα σε μια από τις διαθέσιμες

κατηγορίες του συνόλου δεδομένων, ενώ στόχος της δυαδικής ταξινόμησης είναι το να διακρίνει εάν ένα δείγμα δοκιμής ανήκει σε μια συγκεκριμένη κατηγορία ή όχι.

Υπάρχουν διαφορετικές ερευνητικές εργασίες που επικεντρώνονται στο πρόβλημα ταξινόμησης πολλαπλών κλάσεων (multi-class) [4], ενώ οι περισσότερες από τις πρόσφατες έρευνες σχετικά με την ταξινόμηση πολλαπλής ετικέτας στοχεύουν στην αναγωγή της σε ένα σύνολο προβλημάτων δυαδικής ταξινόμησης [41]. Σε αυτόν τον μετασχηματισμό, για κάθε κατηγορία του συνόλου δεδομένων, δημιουργείται και εκπαιδεύεται ένας δυαδικός ταξινομητής, έτσι ώστε κάθε ένας από τους δυαδικούς ταξινομητές να μπορεί να διακρίνει εάν ένα δείγμα δοκιμής ανήκει σε μια συγκεκριμένη κατηγορία ή όχι. Δυστυχώς, αυτή η προσέγγιση θα αρκούσε μόνο υπό την προϋπόθεση ότι οι κατηγορίες είναι ανεξάρτητες [39]. Στο σύνολο δεδομένων Reuters, που χρησιμοποιήθηκε σε αυτή την περίπτωση, οι ετικέτες που ανατίθενται σε ένα αντικείμενο ειδήσεων σχετίζονται, και έτσι η παραπάνω μετατροπή δεν είναι εφικτή.

Παραδοσιακά, η αυτόματη ταξινόμηση ενός συνόλου ηλεκτρονικών εγγράφων σε ένα σύνολο διαφορετικών κατηγοριών, που υποδηλώνεται από διαφορετικές ετικέτες, είναι μια εποπτευόμενη εργασία ταξινόμησης [39]. Σε αυτήν την περίπτωση, ο ταξινομητής μαθαίνει με βάση το επισημειώμενο υποσύνολο δεδομένων που χρησιμοποιήθηκε κατά τη φάση της εκπαίδευσης και στη συνέχεια το υποσύνολο δεδομένων ελέγχου χρησιμοποιείται για την αξιολόγηση της απόδοσής του. Ωστόσο, αυτή η προσέγγιση απαιτεί ένα μεγάλο, σωστά επισημασμένο σύνολο δεδομένων και θεωρείται αναποτελεσματική με σύνολα δεδομένων που έχουν μια μη ισορροπημένη κατανομή των αντικειμένων μεταξύ των διαφορετικών κλάσεων. Αυτός ο περιορισμός επιβάλλει την επιλογή μη εποπτευόμενων προσεγγίσεων έναντι των εποπτευόμενων σε αυτή την περίπτωση.

Σε αντίθεση με τις παραδοσιακές προσεγγίσεις, οι Αυτό-Οργανούμενοι Χάρτες (SOM) είναι μια μη εποπτευόμενη μέθοδος Μηχανικής Μάθησης, η οποία μπορεί να συγκεντρώσει παρόμοια έγγραφα μαζί χρησιμοποιώντας μέτρα ομοιότητας διανυσμάτων, χωρίς προηγούμενη γνώση των κλάσεων. Σε αυτή την περίπτωση, οι ετικέτες των δεδομένων εκπαίδευσης θα χρησιμοποιηθούν μόνο για να καθοδηγήσουν την εκχώρηση ετικετών στο σύνολο δεδομένων ελέγχου μετά τη φάση της εκπαίδευσης. Μια άλλη διαφοροποίηση σχετικά με τις εποπτευόμενες τεχνικές, είναι ότι ένας εκπαιδευμένος χάρτης SOM περιέχει αναγνωρίσιμες αλλά μη οριοθετημένες ομάδες [39]. Στις ομάδες που δημιουργούνται, παρόμοια έγγραφα τείνουν να μοιράζονται κοινές ετικέτες και επομένως το SOM διατηρεί εγγενώς τη συσχέτιση μεταξύ των ετικετών, καθώς οι σχετικές ετικέτες μπορούν να βρεθούν είτε στον ίδιο είτε σε έναν γειτονικό κόμβο στον χάρτη.

1.1.3. Αυτόματη ταξινόμηση πολλαπλών κλάσεων με χρήση τεχνικών Αυτό-Οργανούμενων χαρτών και δικτύων Βαθιάς Μάθησης

Κατά την αναζήτηση πηγών, καθηγητές και φοιτητές χρησιμοποιούν συχνά την Ψηφιακή Βιβλιοθήκη του ιδρύματός τους δίνοντας ως είσοδο λέξεις-κλειδιά που πιστεύουν ότι είναι σχετικές. Ωστόσο, τα αποτελέσματα που επιστρέφονται από τα συστήματα βιβλιοθηκών είναι συχνά άσχετα, καθώς ενδέχεται να αντιστοιχούν σε ένα εντελώς διαφορετικό θέμα ή σχετικά βιβλία μπορεί να εξαιρούνται από τη λίστα αποτελεσμάτων, εάν δεν περιέχουν τον ακριβή όρο αναζήτησης. Αυτό το πρόβλημα γίνεται περισσότερο εμφανές καθώς το περιεχόμενο μιας Ψηφιακής Βιβλιοθήκης αυξάνεται σε όγκο. Τα αποτελέσματα ταξινομούνται με βάση την ομοιότητα του όρου αναζήτησης με όρους που εμφανίζονται στον τίτλο κάθε βιβλίου. Ωστόσο, η χρήση μιας πιο λεπτομερούς περιγραφής του θέματος που βρίσκεται υπό αναζήτηση (π.χ. περιγραφή μαθήματος ή ο πίνακας περιεχομένων ενός γνωστού βιβλίου) μπορεί να οδηγήσει σε πιο ακριβή αποτελέσματα σε

περίπτωση που υπάρχει ένα κατάλληλο σύστημα ταξινόμησης και ανάκτησης που να βασίζεται σε μέτρα ομοιότητας τέτοιων στοιχείων κειμένου. Στην παρούσα Διατριβή επιχειρούμε να μοντελοποιήσουμε το πρόβλημα της αναζήτησης παρόμοιων βιβλίων ως ένα πρόβλημα ταξινόμησης πολλαπλών κλάσεων χρησιμοποιώντας πληροφορία που έχουμε εξάγει από τους πίνακες περιεχομένων (ToC) των βιβλίων για την παροχή αποτελεσματικών προτάσεων σχετικά με συναφή βιβλία και συγγράμματα.

Η αυτόματη ταξινόμηση εγγράφων, και συγκεκριμένα η ταξινόμηση πολλαπλών κλάσεων, στην οποία κάθε έγγραφο μπορεί να αντιστοιχιστεί μόνο σε μία κατηγορία, είναι σημαντική σε διάφορους τομείς, συμπεριλαμβανομένων των Ψηφιακών Βιβλιοθηκών. Έχουν διεξαχθεί διάφορες ερευνητικές μελέτες που εστιάζουν στην ταξινόμηση πολλαπλών κλάσεων όπως αυτές που παρουσιάζονται στα [76]- [82], [84], [85]. Οι περισσότερες από τις προσεγγίσεις στοχεύουν στη μετατροπή του προβλήματος ταξινόμησης πολλαπλών κλάσεων σε ένα σύνολο προβλημάτων δυαδικής ταξινόμησης. Σε αυτόν τον μετασχηματισμό, για κάθε κατηγορία στο σύνολο, δημιουργείται και εκπαιδεύεται ένας δυαδικός ταξινομητής στο σύνολο δεδομένων έτσι ώστε κάθε ταξινομητής να μπορεί να διακρίνει εάν ένα δείγμα δοκιμής ανήκει σε μια συγκεκριμένη κατηγορία ή όχι. Δυστυχώς, αυτή η προσέγγιση επαρκεί μόνο με την υπόθεση ότι το σύνολο δεδομένων είναι πρακτικά αμετάβλητο, καθώς για κάθε νέα κατηγορία που προστίθεται, θα πρέπει να αναπτυχθεί ένας νέος ταξινομητής. Στο σύνολο δεδομένων Springer που επιλέχθηκε στην περίπτωσή μας, τα ηλεκτρονικά βιβλία μπορούν να προστεθούν ή να αφαιρεθούν και να εισαχθούν νέες κατηγορίες. Επομένως, η μετατροπή σε ένα σύνολο δυαδικών προβλημάτων ταξινόμησης δημιουργεί σοβαρές δυσκολίες.

Παραδοσιακά, η αυτόματη κατηγοριοποίηση ενός συνόλου ηλεκτρονικών εγγράφων σε ένα σύνολο κατηγοριών, οι οποίες επισημαίνονται με ετικέτες, είναι μια εποπτευόμενη εργασία ταξινόμησης [85], [97]. Σε αυτήν την περίπτωση, ο ταξινομητής μαθαίνει με βάση τα επισημασμένα δεδομένα εκπαίδευσης και στη συνέχεια, το υποσύνολο ελέγχου χρησιμοποιείται για την αξιολόγηση της απόδοσής του ταξινομητή. Ωστόσο, αυτή η προσέγγιση απαιτεί ένα μεγάλο, καλά επισημασμένο σύνολο δεδομένων και είναι αναποτελεσματική σε σύνολα δεδομένων που έχουν μια μη ισορροπημένη κατανομή δειγμάτων μεταξύ των κλάσεων. Λόγω αυτού του περιορισμού, επιλέγονται συνήθως μη εποπτευόμενες προσεγγίσεις και προσεγγίσεις που χρησιμοποιούν Νευρωνικά Δίκτυα Βαθιάς Μάθησης, έναντι των εποπτευόμενων προσεγγίσεων, σε περιπτώσεις που είτε δεν υπάρχει διαθέσιμο επισημασμένο σύνολο δεδομένων είτε υπάρχει ανισορροπία στην κατανομή των δειγμάτων μεταξύ των κλάσεων. Πρόσφατα, προ-εκπαιδευμένα γλωσσικά μοντέλα όπως οι Transformers και τα BERT, XLNet και GPT-2 έχουν προσελκύσει την προσοχή της ερευνητικής κοινότητας [91]-[95]. Τα συγκεκριμένα μοντέλα έχει αποδειχθεί ότι είναι αποτελεσματικά σε σύγκριση με μοντέλα που εκπαιδεύονται από το μηδέν. Επιπλέον, η μεταφορά μάθησης από εποπτευόμενα δίκτυα μπορεί να εφαρμοστεί και εδώ ως αντίδοτο στο πρόβλημα ανισορροπίας των δειγμάτων μεταξύ των κλάσεων.

1.2. Δομή και συνεισφορά διατριβής

Η παρούσα Διατριβή εντάσσεται στο ευρύτερο ερευνητικό πεδίο της αυτόματης Εξαγωγής Πληροφορίας ταξινόμησης από κείμενα και εξετάζει ανοιχτά θέματα τα οποία άπτονται της ταξινόμησης κειμένων με χρήση τεχνικών Μηχανικής Μάθησης και Νευρωνικών Δικτύων Βαθιάς Μάθησης. Ακριβέστερα, στο Κεφάλαιο 2 το ενδιαφέρον εστιάζεται στην ανάλυση των διαφορετικών τεχνικών εξαγωγής πληροφορίας από κείμενα με σκοπό την κατανόηση των τεχνικών που χρησιμοποιούνται αλλά και το πλαίσιο ενσωμάτωσης των προσεγγίσεων αυτών σε εφαρμογές

σημασιολογικού περιεχομένου. Το Κεφάλαιο 3 πραγματεύεται την επιλογή των κατάλληλων τεχνικών εξαγωγής για την αυτόματη ταξινόμηση ενός συνόλου δεδομένων πολλαπλών ετικετών μέσω μιας εκτεταμένης πειραματικής αξιολόγησης της χρήσης Αυτό-Οργανούμενων Χαρτών (Self-Organizing Maps). Στη συνέχεια, το Κεφάλαιο 4 πραγματεύεται το πρόβλημα της αυτόματης ταξινόμησης πολλαπλών κλάσεων (multi-class) μιας συλλογής ηλεκτρονικών βιβλίων (e-books) με βάση μια μεθοδολογία που βασίζεται στην τεχνητή νοημοσύνη εξάγοντας πληροφορία από τον πίνακα περιεχομένων. Το Κεφάλαιο 5 παρουσιάζει άλλα θέματα που έχουν εξεταστεί στην πορεία αυτής της Διατριβής και σχετίζονται με την εφαρμογή των τεχνικών αυτόματης Εξαγωγής Πληροφορίας ταξινόμησης για τον εμπλουτισμό μεταδεδομένων και την μετέπειτα εφαρμογή τους στα πλαίσια μιας Πύλης Διασυνδεδεμένων Δεδομένων με βάση την ανάλυση που πραγματοποιήθηκε στο Κεφάλαιο 2.

Αναλυτικότερα, στο **Κεφάλαιο 2** αναλύονται οι μέθοδοι Εξαγωγής Πληροφορίας από κείμενα, οι οποίες παρουσιάζουν μεγάλη ποικιλομορφία και εφαρμόζονται σε ένα πλήθος πεδίων και εφαρμογών. Στο συγκεκριμένο Κεφάλαιο αρχικά γίνεται μια εισαγωγή στο χώρο της εξαγωγής πληροφορίας από κείμενα και μια λεπτομερής ανασκόπηση των βασικών μεθόδων που υιοθετούνται συνήθως. Πιο συγκεκριμένα, αρχικά γίνεται μια ανασκόπηση των τεχνικών που χρησιμοποιούνται για την αναγνώριση των υποψήφιων όρων σε ένα κείμενο η οποία ακολουθείται από τις μεθόδους επιλογής των υποψήφιων όρων. Οι μέθοδοι αυτής της παραγράφου στοχεύουν στο να ξεχωρίσουν τους «κατάλληλους» ή «καταλληλότερους» από τους «λιγότερο κατάλληλους» ή «χειρότερους» όρους. Σε αυτή την κατηγορία μεθόδων συναντώνται και στατιστικά μέτρα συχνότητας όπως TF (Term Frequency) ή TF-IDF (Term Frequency Inverse Document Frequency) για την βαθμολόγηση των υποψήφιων όρων. Επιπλέον αναλύονται τεχνικές Μηχανικής Μάθησης που χρησιμοποιούνται συνήθως για την επιλογή των όρων αυτών. Τέλος, πραγματοποιείται μια εκτενής ανασκόπηση στα χαρακτηριστικά που χρησιμοποιούν οι διάφορες προσεγγίσεις με σκοπό να αναγνωρίσουν εάν ένας όρος είναι αντιπροσωπευτικός για το υπό εξέταση κείμενο. Στη συνέχεια, με βάση τη βιβλιογραφική ανασκόπηση που πραγματοποιήθηκε οι προσεγγίσεις Εξαγωγής Πληροφορίας από κείμενα ταξινομούνται σε τέσσερις διαφορετικές κατηγορίες. Σε αυτό το σημείο πρέπει να αναφερθεί πως αυτός είναι ένας αδρός διαχωρισμός και πως μια προσέγγιση μπορεί να ανήκει σε περισσότερες τις μιας κατηγορίες. Το Κεφάλαιο αυτό ολοκληρώνεται με την επισκόπηση των διαφορετικών εφαρμογών όπου η Εξαγωγή Πληροφορίας έχει υιοθετηθεί με επιτυχία, εξετάζονται τα πλεονεκτήματα που προκύπτουν από την ενσωμάτωση τέτοιων μεθόδων αλλά και οι δυνατότητες που προκύπτουν για τις Ψηφιακές Βιβλιοθήκες με την υιοθέτηση τέτοιων τεχνικών.

Στο **Κεφάλαιο 3** επιχειρείται η γεφύρωση του χάσματος μεταξύ της επιλογής χαρακτηριστικών (feature selection) και του μεγέθους του χώρου των χαρακτηριστικών (feature size) για την αποτελεσματική εξαγωγή πληροφορίας, χρησιμοποιώντας Αυτό-Οργανούμενους Χάρτες (SOM), ένα Νευρωνικό Δίκτυο μη εποπτευόμενης Μηχανικής Μάθησης, κάτω από διαφορετικά σενάρια διαμόρφωσης για την ταξινόμηση ενός συνόλου δεδομένων πολλαπλής ετικέτας, συγκεκριμένα του συνόλου δεδομένων Reuters Mod Apte’. Η κατασκευή των διανυσμάτων βασίζεται σε μια απλή, αλλά αποτελεσματική διαδικασία που στοχεύει στη μετατροπή των διανυσμάτων έτσι ώστε να φέρουν μία μόνο ετικέτα αντί για πολλαπλές. Η προτεινόμενη λύση βελτιώνει την αποδοτικότητα ταξινόμησης όχι μόνο από την άποψη της ακρίβειας αλλά και από τους υπολογιστικούς πόρους και το χρόνο για την εκπαίδευση του δικτύου που απαιτούνται. Διεξήχθησαν εκτεταμένα πειράματα, χρησιμοποιώντας διαφορετικές παραμέτρους σχετικά με το μέγεθος του χάρτη και του διανύσματος, και τους κύκλους εκπαίδευσης, χρησιμοποιώντας επίσης το περιεχόμενο γύρω από μια λέξη, για να

εκτιμηθεί ο αντίκτυπός τους στην απόδοση του ταξινομητή. Επιπλέον, προτάθηκε ένας έξυπνος αλγόριθμος για την επιλογή ετικετών, με στόχο να δείξει ότι οι γειτονικοί κόμβοι στον χάρτη επηρεάζουν την επιλογή των ετικετών για έναν συγκεκριμένο κόμβο. Σύμφωνα με τα πειράματα που διεξήχθησαν, η προσέγγισή αυτή επιτυγχάνει αύξηση 10% στις βαθμολογίες Μάκρο-μέσου F1, 30% μείωση στο μέγεθος των διανυσμάτων που χρησιμοποιήθηκαν και περίπου 34% μικρότερους χάρτες σε σύγκριση με το βασικό σενάριο.

Στο **Κεφάλαιο 4** αναλύεται το πρόβλημα της αναζήτησης/σύστασης βιβλίων για την υποστήριξη καθηγητών και φοιτητών στον εντοπισμό βιβλιογραφικών πηγών από μια συλλογή ηλεκτρονικών βιβλίων. Ένα θέμα που έχει μεγάλη σημασία τόσο για τα πανεπιστήμια όσο και για τις Ψηφιακές Βιβλιοθήκες και, ως εκ τούτου, υποκινεί την ανάπτυξη ενός συστήματος συστάσεων. Αυτό το Κεφάλαιο αναλύει τις τεχνικές που εφαρμόστηκαν για την αυτόματη ταξινόμηση ενός πολύπλευρου (multi-disciplinary) συνόλου δεδομένων που δημιουργήθηκε από e-books από τη συλλογή Springer, το οποίο διατίθεται μέσω της συνδρομής των Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, χρησιμοποιώντας ένα μη εποπτευόμενο Νευρωνικό Δίκτυο (NN), συγκεκριμένα το SOM και δύο διαφορετικές αρχιτεκτονικές Νευρωνικών Δικτύων Βαθιάς Μάθησης (DNN), συγκεκριμένα, ένα δίκτυο Μακριάς Βραχυπρόθεσμης Μνήμης (Long-Short Term Memory - LSTM) και ένα Συνελικτικό Νευρωνικό Δίκτυο (CNN) σε συνδυασμό με ένα LSTM (CNN + LSTM) σε διάφορα σενάρια διαμόρφωσης. Για την κατασκευή των διανυσμάτων αξιοποιούνται πληροφορίες που εξήχθησαν από τον πίνακα περιεχομένων (ToC) κάθε βιβλίου χρησιμοποιώντας το σχήμα **Σταθμισμένης Συχνότητας Όρων** (TF-IDF), για την πρώτη περίπτωση και τον **Keras Tokenizer**, για τη δεύτερη. Πραγματοποιήθηκαν εκτενή πειράματα χρησιμοποιώντας διάφορες διαμορφώσεις και βήματα προ επεξεργασίας, διαμορφώσεις των νευρωνικών δικτύων και μεγέθη διανυσμάτων και λεξιλογίου για να εκτιμηθεί η επίδρασή τους στην απόδοση του ταξινομητή. Επιπλέον, προκύπτει ότι η μέθοδος της πλειοψηφίας είναι πιο κατάλληλη για την επιλογή της κυρίαρχης ετικέτας για έναν καθορισμένο κόμβο στην περίπτωση του SOM. Η πειραματική ανάλυση έδειξε τη σκοπιμότητα ανάπτυξης ενός συστήματος συστάσεων για την υποστήριξη καθηγητών και φοιτητών στον εντοπισμό σχετικών πηγών, βάσει μιας λεπτομερούς θεματικής περιγραφής (π.χ. της περίληψης ή του πίνακα περιεχομένων ενός βιβλίου) αντί για μερικές λέξεις-κλειδιά. Στα πειράματα που διεξήχθησαν, το υποσύστημα που χρησιμοποίησε το DNN (LSTM) είχε την καλύτερη απόδοση, με βαθμολογίες F1 67% για τις 26 κατηγορίες και 80% για τις 5 γενικές κατηγορίες, ενώ το SOM πέτυχε βαθμολογίες F1 μικρότερες κατά 5% και στις δύο περιπτώσεις.

Στο **Κεφάλαιο 5** προτείνεται μια μεθοδολογία για τη δημιουργία μιας Πύλης Διασυνδεδεμένων Δεδομένων με χρήση τεχνολογιών Σημασιολογικού Ιστού με στόχο την ενσωμάτωση των μηχανισμών αυτόματης εξαγωγής πληροφορίας ταξινόμησης που αναλύθηκαν στα Κεφάλαια 3 και 4 και των αποτελεσμάτων αυτών, με απώτερο σκοπό τον εμπλουτισμό μεταδεδομένων, έτσι ώστε να υποβοηθηθεί η αποτελεσματικότερη αναζήτηση και ανάκτηση πληροφοριών από τους τελικούς χρήστες. Ο στόχος της μεθοδολογίας αυτής είναι η σημασιολογική επεξεργασία και σύνδεση δεδομένων από διαφορετικές πηγές. Με χρήση των μεθόδων αυτόματης εξαγωγής πληροφορίας ταξινόμησης που αναπτύχθηκαν, η πληροφορία που εξάγεται μπορεί να χρησιμοποιηθεί αποδοτικά κατά την αναζήτηση ή και την πλοήγηση στα πλαίσια της Πύλης. Για να επιτευχθεί όμως η αποτελεσματική αναζήτηση και πλοήγηση πολύ σημαντική κρίνεται η ποιότητα της πληροφορίας που εξάγεται. Για το λόγο αυτό, είναι πολύ σημαντικό να έχουν εξαχθεί ποιοτικοί και ταυτόχρονα ακριβείς περιγραφείς. Με χρήση των μηχανισμών αυτόματης ή ημιαυτόματης Εξαγωγής Πληροφορίας ταξινόμησης μειώνεται η ανάγκη για χειροκίνητη επισημείωση του περιεχομένου από ανθρώπους

ειδικούς στο αντικείμενο μια διαδικασία χρονοβόρα και μη εφικτή όσο το περιεχόμενο της Πύλης αυξάνεται. Η προτεινόμενη μεθοδολογία ενσωματώνει πολλαπλά κανάλια εισόδου για διαφορετικές ομάδες χρηστών και κίνητρα για την επαναχρησιμοποίηση, την αναφορά, τον σχολιασμό και την κοινή χρήση του περιεχομένου. Η υλοποίηση βασίζεται σε ένα Σύστημα Διαχείρισης Περιεχομένου. Σε αυτήν την προσέγγιση χρησιμοποιείται το Drupal, αλλά οποιοδήποτε άλλο CMS μπορεί να χρησιμοποιηθεί εναλλακτικά. Παρόλο που στην περίπτωση αυτή οι πηγές μπορεί προέρχονται κυρίως από δεδομένα ενός ιδρύματος, για παράδειγμα από την Ψηφιακή Βιβλιοθήκη του Εθνικού Μετσόβιου Πολυτεχνείου (ΕΜΠ), η χρήση της προτεινόμενης αρχιτεκτονικής μπορεί να επεκταθεί άμεσα σε επιπλέον πεδία.

Στο **Κεφάλαιο 6**, τέλος, συνοψίζονται τα κυριότερα συμπεράσματα της παρούσας Διατριβής, ενώ σκιαγραφούνται και οι μελλοντικές κατευθύνσεις της εν λόγω ερευνητικής προσπάθειας.

2. Αυτόματη Εξαγωγή Πληροφορίας από κείμενα

Η Εξαγωγή Πληροφορίας από κείμενα, ενώ αποτελεί μια ιδέα που έχει απασχολήσει την ερευνητική κοινότητα από καιρό, τα τελευταία χρόνια έχει διαμορφωθεί κάτω από διαφορετικό πρίσμα, λόγω της έλευσης και μεγάλης ανάπτυξης του Σημασιολογικού Ιστού. Οι μέθοδοι Εξαγωγής Πληροφορίας από κείμενα παρουσιάζουν μεγάλη ποικιλομορφία και εφαρμόζονται σε ένα πλήθος πεδίων. Παρόλο που ο συγκεκριμένος τομέας παρουσιάζει πολύ μεγάλο ερευνητικό ενδιαφέρον αλλά και πληθώρα εφαρμογών σε διαφορετικά πεδία, χαρακτηρίζεται από «φτωχά» αποτελέσματα σε σχέση με άλλες εργασίες Ανάλυσης Φυσικής Γλώσσας (Natural Language Processing - NLP). Ο βασικός λόγος είναι ότι δεν υπάρχει κάποιο «ορθό» σύνολο από λέξεις ή φράσεις για ένα δεδομένο κείμενο το οποίο να χαίρει της ευρείας αποδοχής της ερευνητικής κοινότητας έτσι ώστε να αποτελέσει τη βάση σύγκρισης. Οι φράσεις που ανατίθενται από ειδικούς στο εκάστοτε πεδίο θεωρούνται πως αποτελούν το «χρυσό πρότυπο» (gold standard), αλλά υπάρχουν διαφωνίες για το πιο ακριβώς είναι το πρότυπο αυτό. Ακόμα, οι λέξεις ή οι φράσεις που θα εξαχθούν πρέπει να είναι κατανοητές, γραμματικά ορθές ενώ ταυτόχρονα πρέπει να παρέχουν και την καλύτερη δυνατή θεματική κάλυψη. Η βασική δυσκολία στην αποτελεσματική εξαγωγή πληροφορίας έγκειται στο να καθοριστεί ποιες λέξεις ή φράσεις είναι οι πιο σχετικές και αποδίδουν καλύτερα το νόημα του κειμένου. Υπάρχουν διάφοροι παράγοντες που επηρεάζουν το πρόβλημα αυτό, όπως η έκταση του κειμένου, διαφοροποιήσεις στη διάρθρωση του κειμένου, εναλλαγές στις θεματικές κατηγορίες αλλά και ενδεχόμενη έλλειψη συσχέτισης μεταξύ των θεματικών κατηγοριών ενός δεδομένου κειμένου [2].

Στο παρόν Κεφάλαιο, αρχικά γίνεται μια εισαγωγή στο χώρο της Εξαγωγής Πληροφορίας από κείμενα/πηγές. Στη συνέχεια γίνεται μια ενδελεχής ανασκόπηση των βασικών μεθόδων που ακολουθούνται συνήθως. Γενικά, μπορούμε να πούμε πως οι μέθοδοι αυτές χωρίζονται σε τέσσερις διαφορετικές κατηγορίες: α) **Προσεγγίσεις Εξαγωγής** (Extraction), β) **Προσεγγίσεις Ανάθεσης** (Assignment), γ) **Μεικτές Προσεγγίσεις** (Hybrid) και δ) **Προσεγγίσεις Πρόβλεψης** (Predictive). Προφανώς ο παραπάνω είναι ένας αδρός διαχωρισμός και, όπως αναλύεται και στη συνέχεια του κεφαλαίου, υπάρχουν αρκετά κριτήρια τα οποία μπορούν να χρησιμοποιηθούν προκειμένου να ταξινομηθούν λεπτομερέστερα οι εν λόγω προσεγγίσεις. Ακόμα, παρουσιάζεται ένα πλήθος διαφορετικών εφαρμογών όπου η Εξαγωγή Πληροφορίας έχει υιοθετηθεί με επιτυχία, εξετάζονται τα πλεονεκτήματα που προκύπτουν από την ενσωμάτωση τέτοιων μεθόδων αλλά και οι δυνατότητες που προκύπτουν για τις Ψηφιακές Βιβλιοθήκες με την υιοθέτηση αυτών των τεχνικών.

Η διάρθρωση του κεφαλαίου γίνεται ως εξής: αρχικά γίνεται αναφορά στα κίνητρα και στα οφέλη που προκύπτουν από την μελέτη του πεδίου, ενώ στην επόμενη ενότητα παρουσιάζονται οι προσεγγίσεις ανά κατηγορία, όπως αυτές αναφέρθηκαν παραπάνω. Στη συνέχεια, αναλύονται κάποιες από τις περισσότερο αντιπροσωπευτικές εφαρμογές των τεχνικών Εξαγωγής Πληροφορίας από κείμενα, ενώ γίνεται αναφορά και στα πλεονεκτήματα που οι μέθοδοι αυτές μπορεί να προσφέρουν σε διάφορα πεδία εφαρμογών. Τέλος, περιγράφεται η προτεινόμενη μεθοδολογία με ταυτόχρονη αναφορά των πλεονεκτημάτων που απορρέουν από τη χρήση της συγκεκριμένης προσέγγισης.

2.1. Κίνητρα

Στις μέρες μας, οι Ψηφιακές Βιβλιοθήκες και ειδικότερα αυτές που χρησιμοποιούνται ως αποθετήρια της ερευνητικής παραγωγής ενός Ιδρύματος έχουν να διαχειριστούν έναν κύκλο εργασιών που αποτελείται από πολλές διακριτές φάσεις. Μια από αυτές είναι και η αναγνώριση/ταυτοποίηση του συγγραφέα ενός άρθρου [3]. Η διαδικασία αυτή παρουσιάζει αρκετές δυσκολίες που οφείλονται σε ποικίλους παράγοντες όπως η συνωνυμία (synonymy), η πολυσημία (polysemy) ή ακόμα και η λανθασμένη εισαγωγή του συγγραφέα κατά τη διαδικασία κατάθεσης του τεκμηρίου στο σύστημα διαχείρισης της βιβλιοθήκης. Εξίσου σημαντική εργασία του παραπάνω κύκλου αποτελεί και ο εμπλουτισμός ενός τεκμηρίου με μετά-δεδομένα (metadata enrichment), πέραν του συγγραφέα, διαδικασία που γίνεται κατά κύριο λόγο όταν το τεκμήριο καταχωρίζεται στην Ψηφιακή Βιβλιοθήκη ή στο Ιδρυματικό Αποθετήριο (εφεξής ΙΑ). Η διαδικασία αυτή έχει το μειονέκτημα ότι απαιτεί εξειδικευμένο προσωπικό και είναι ταυτόχρονα χρονοβόρα και με δυσθεώρητο κόστος, όσο ο αριθμός των τεκμηρίων στην Ψηφιακή Βιβλιοθήκη αυξάνεται, καθιστώντας τη διαδικασία αυτή μη κλιμακώσιμη και μη χρηστική για τις ψηφιακές βιβλιοθήκες και τις συλλογές που περιέχονται σε αυτές [4].

Τόσο στον τομέα των Ψηφιακών Βιβλιοθηκών όσο και στο πεδίο της επιστήμης των Πληροφοριών μια λέξη-κλειδί (keyword) ορίζεται ως μια λέξη η οποία περιγράφει ευκρινώς και με ακρίβεια το θέμα ή ένα μέρος του θέματος ενός εγγράφου [5]. Για τον ίδιο λόγο οι βιβλιοθηκονόμοι χρησιμοποιούν θεματικές επικεφαλίδες (subject headings). Οι θεματικές επικεφαλίδες αποτελούν όρους ή φράσεις που περιγράφουν το περιεχόμενο ενός εγγράφου και επιπρόσθετα μπορούν να χρησιμοποιηθούν για να ομαδοποιήσουν και να οργανώσουν τις συλλογές μιας βιβλιοθήκης. Για παράδειγμα, έγγραφα τα οποία σχετίζονται με κάποιο θέμα μπορούν να ομαδοποιηθούν και να υπάρχει σε αυτά γρήγορη πρόσβαση με την επιλογή μιας συγκεκριμένης θεματικής επικεφαλίδας.

Όπως οι θεματικές επικεφαλίδες έτσι και οι λέξεις-κλειδιά και οι σχετικοί με το θέμα όροι μπορούν να επιλεγούν από ένα τυποποιημένο σύνολο περιγραφέων, όπως ένα ελεγχόμενο λεξιλόγιο (controlled vocabulary). Η προσέγγιση αυτή όμως περιορίζει την εφαρμογή σε ένα συγκεκριμένο πεδίο (domain-dependent) που καθορίζεται από το δεσμευμένο λεξιλόγιο που χρησιμοποιείται κάθε φορά και είναι δύσκολο να εφαρμοστεί σε ένα άλλο σύνολο δεδομένων. Ένας άλλος τρόπος για την ανάθεση όρων ή λέξεων-κλειδιά είναι η εξαγωγή τους από το σώμα του κειμένου ή, όπως κάνουν συνήθως οι βιβλιοθηκονόμοι, μόνο από τον τίτλο του κειμένου. Παρόλα αυτά αν οι όροι εξάγονταν με έναν αυτόματο τρόπο και από αλλά μέρη του κειμένου, όπως για παράδειγμα από την περίληψη ή τον πίνακα περιεχομένων ενός βιβλίου, θα μπορούσε κανείς να έχει πιο ακριβείς περιγραφείς που να αποδίδουν καλύτερα το νόημα ενός δεδομένου κειμένου αλλά και να χρησιμοποιηθούν αποδοτικά σε ένα σύστημα αυτόματης ταξινόμησης ή ομαδοποίησης παρόμοιων κειμένων.

Η έννοια της Εξαγωγής Πληροφορίας από κείμενα αναφέρεται στην εξαγωγή μιας συγκεκριμένου μεγέθους λίστας όρων (term list), κάθε αντικείμενο της οποίας μπορεί να αποτελείται από δύο ή περισσότερες λέξεις αντί για μια μεμονωμένη λέξη-κλειδί [6]. Στόχος αυτής της λίστας είναι να αποτυπώσει τα βασικά σημεία ή τις βασικές έννοιες ενός δεδομένου εγγράφου. Σε σχέση με την ευρετηρίαση πλήρους κειμένου (full-text indexing), όπου όλοι οι όροι που εμφανίζονται στο κείμενο αποθηκεύονται σε κάποια μονάδα αποθήκευσης, όπως μια βάση δεδομένων, η ευρετηρίαση με βάση όρους ή λέξεις-κλειδιά απαιτεί λιγότερο αποθηκευτικό χώρο και παρέχει μια συνοπτική επισκόπηση των θεματικών περιοχών μιας συλλογής ή ενός μεμονωμένου κειμένου. Ακόμα, οι όροι αυτοί παρέχουν μια συνοπτική περιγραφή του περιεχομένου ενός κειμένου και είναι επίσης χρήσιμοι

για λειτουργίες όπως: κατηγοριοποίηση (categorization, clustering), ευρετηρίαση (indexing), αναζήτηση (search), και σύνοψη (summarization). Ακόμα, η Εξαγωγή Πληροφορίας μπορεί να συμβάλλει στην ποσοτικοποίηση της σημασιολογικής ομοιότητας με άλλα κείμενα αλλά και στον καθορισμό των εννοιών που σχετίζονται με συγκεκριμένους τομείς γνώσης.

Τέλος, κατά τη δημοσίευση ενός κειμένου σε ένα επιστημονικό περιοδικό ζητείται από τους συγγραφείς, εκτός από το κείμενο της δημοσίευσης και μια λίστα με αντιπροσωπευτικούς όρους. Συνήθως, οι όροι αυτοί παρέχονται είτε απ' ευθείας από τους συγγραφείς, είτε από εξειδικευμένο προσωπικό το οποίο είναι υπεύθυνο για την κατάρτιση της λίστας αυτής. Η διαδικασία αυτή μπορεί να είναι επίπονη και χρονοβόρα και επίσης μειονεκτεί στο ότι είναι χειρωνακτική και άρα υπόκειται σε λάθη. Ακόμα, είναι μια υποκειμενική διαδικασία στην οποία πολύ σημαντικό ρόλο παίζει ο ανθρώπινος παράγοντας, διότι διαφορετικός άνθρωπος μπορεί να καταρτίσει διαφορετική λίστα όρων. Η παραπάνω περίπτωση αποτελεί ένα κλασικό πρόβλημα στο οποίο η εφαρμογή μεθόδων Εξαγωγής Πληροφορίας μπορεί να δώσει λύση.

2.2. Οφέλη

Όλα τα προαναφερθέντα προβλήματα είναι δυνατόν να ξεπεραστούν εάν η παραπάνω διαδικασία αυτοματοποιηθεί με κάποιο τρόπο. Στο σημείο αυτό μπορούμε να πούμε ότι το πρόβλημα μετατρέπεται από «εξαγωγή πληροφορίας από κείμενα με χειροκίνητο τρόπο» σε «εξαγωγή πληροφορίας με αυτόματο ή ημιαυτόματο τρόπο». Όπως είναι φανερό από τα παραπάνω, η χρήση των όρων αυτών μπορεί να επιτελέσει πολλούς και διαφορετικούς σκοπούς, με τη σύνοψη του θέματος ενός επιστημονικού περιοδικού (summarization) να αποτελεί έναν από αυτούς. Σε αυτή την περίπτωση οι όροι που προκύπτουν από την εξαγωγή μπορεί να αποτελέσουν μια «ειδική» μορφή περίληψης [7]. Άμεσο πλεονέκτημα αυτού, είναι να δύναται ο εκάστοτε χρήστης να αναγνωρίσει σε ελάχιστο χρόνο εάν το θέμα ενός δεδομένου κειμένου είναι μέσα στα ενδιαφέροντά του ή όχι.

Ένα ακόμα πλεονέκτημα των όρων και των λέξεων-κλειδιά είναι η πολυ-λειτουργικότητα (multi-functionality) που τα διακρίνει. Οι λέξεις-κλειδιά αλλά και οι λίστες όρων μπορεί να χρησιμοποιηθούν για ποικίλες εφαρμογές NLP όπως κατηγοριοποίηση κειμένου (text clustering) και ταξινόμηση (classification) [8]. Δύο ακόμα περιοχές όπου οι μέθοδοι Εξαγωγής Πληροφορίας μπορούν να εφαρμοστούν με επιτυχία, αποτελούν η ανάκτηση πληροφοριών που βασίζεται στο περιεχόμενο (content-based retrieval) και η θεματική αναζήτηση (topic search). Ακόμα, η πληροφορία που εξάγεται με αυτές τις τεχνικές μπορεί να χρησιμοποιηθεί για λόγους ευρετηρίασης σε ένα ευρετήριο. Σε αυτή την περίπτωση ένας χρήστης είναι δυνατόν να βρει ένα άρθρο που μπορεί να καλύπτει τις ανάγκες του μέσα σε ένα μικρό χρονικό διάστημα. Η πληροφορία αυτή μπορεί να χρησιμοποιηθεί αποδοτικά κατά την αναζήτηση [9] ή και την πλοήγηση είτε στον Παγκόσμιο Ιστό. Οι όροι που έχουν εξαχθεί μπορούν να αποτελέσουν την είσοδο μιας μηχανής αναζήτησης αντί για μεμονωμένες λέξεις-κλειδιά έτσι ώστε η τελευταία να επιστρέψει πιο σχετικά ή/και πιο ακριβή αποτελέσματα. Τα αποτελέσματα αυτά μπορούν να βελτιστοποιηθούν περαιτέρω εάν η λίστα των όρων συνδυαστεί (ή κάποιοι από τους όρους της λίστας αντιστοιχηθούν) με όρους που προέρχονται από κάποιο ελεγχόμενο λεξιλόγιο. Έτσι μπορεί να καταστεί δυνατή η ανάκτηση περισσότερο σχετικών, πιο γενικών ή ακόμα και πιο ειδικών αποτελεσμάτων, λόγω του ότι τα ελεγχόμενα λεξιλόγια ή οι θησαυροί ενσωματώνουν συνήθως ένα είδος ιεραρχίας. Ένας παράγοντας που παίζει πολύ σημαντικό ρόλο στη συγκεκριμένη περίπτωση είναι η ποιότητα των όρων που θα εξαχθούν από

το κείμενο. Για το λόγο αυτό, είναι πολύ σημαντικό να έχουν εξαχθεί ποιοτικοί και ταυτόχρονα ακριβείς όροι.

Απόρροια αυτού θα είναι μια μικρότερη σε μέγεθος και ταυτόχρονα πιο σχετική λίστα αποτελεσμάτων με περισσότερο ποιοτικά αποτελέσματα. Όπως είναι φανερό από τα παραπάνω, παρόλο που οι μέθοδοι Εξαγωγής Πληροφορίας μπορούν να χρησιμοποιηθούν για να επιτελέσουν διαφορετικούς σκοπούς, υπάρχει μια κοινή απαίτηση: **μια μικρή λίστα από σύνθετους όρους ή λέξεις** οι οποίες θα αποτυπώνουν το βασικό νόημα ενός εγγράφου [6].

2.3. Τεχνικές εξαγωγής πληροφορίας

Η Εξαγωγή Πληροφορίας με αυτόματο τρόπο συνήθως αποτελεί μια διαδικασία 2 σταδίων: α) Αναγνώριση Υποψήφιων Όρων (Candidate Identification) και β) Επιλογή Όρων (Candidate Selection). Αρχικά, επιλέγεται ένα σύνολο από λέξεις ή/και σύνθετους όρους που θα μπορούσε να καλύπτει το περιεχόμενο ενός κειμένου ενώ σε επόμενο στάδιο επιλέγονται οι «καλύτερες» ή οι πιο αντιπροσωπευτικές από αυτές για ένα δεδομένο κείμενο (π.χ., εκείνες που συγκεντρώνουν την υψηλότερη βαθμολογία με βάση κάποιο σχήμα στάθμισης) [10].

2.3.1. Αναγνώριση υποψήφιων όρων

Το πρώτο και βασικότερο βήμα της διαδικασίας είναι η αναγνώριση των υποψήφιων όρων. Μια «εξαντλητική» (brute-force) μέθοδος μπορεί να εξετάσει όλες τις λέξεις ή τους σύνθετους όρους σε ένα δεδομένο κείμενο ως υποψήφιους. Το υπολογιστικό κόστος μιας τέτοιας μεθόδου είναι δυσθεώρητο, όσο το μέγεθος του υπό εξέταση κειμένου αυξάνεται, αλλά και λόγω του ότι όλες οι λέξεις ή όροι που βρίσκονται σε ένα κείμενο δεν είναι δυνατόν να αποδίδουν εξίσου το περιεχόμενο του. Για το λόγο αυτό, συνήθως χρησιμοποιούνται ευρετικές μέθοδοι για να καταλήξουν σε ένα υποσύνολο υποψήφιων όρων. Οι πιο συχνά χρησιμοποιούμενες ευρετικές μέθοδοι περιλαμβάνουν: την αφαίρεση των stop-words και των σημείων στίξης [6], τον διαχωρισμό των λέξεων που περιλαμβάνουν συγκεκριμένα τμήματα λόγου (Part of Speech - POS) ή των σύνθετων όρων που περιέχουν συγκεκριμένες αλληλουχίες τμημάτων λόγου (POS patterns) [55]. Τέλος, μπορεί να χρησιμοποιούνται εξωτερικές βάσεις γνώσης όπως είναι το WordNet[1] ή η Wikipedia[2] ως σώματα αναφοράς (reference corpora) για την ταξινόμηση ενός υποψήφιου όρου ως «κατάλληλου» ή μη για το υπό εξέταση κείμενο.

2.3.2. Επιλογή υποψήφιων όρων

Πλήθος ερευνητικών εργασιών στοχεύουν στο να ξεχωρίσουν τους «κατάλληλους» ή «καταλληλότερους» από τους «λιγότερο κατάλληλους» ή «χειρότερους» υποψήφιους όρους. Οι πιο απλές από αυτές τις μεθόδους, όπως θα αναλυθεί και σε επόμενη ενότητα του παρόντος, βασίζονται σε **στατιστικά μέτρα συχνότητας** όπως TF (Term Frequency) ή TF-IDF (Term Frequency Inverse Document Frequency) για την βαθμολόγηση των υποψήφιων όρων, θεωρώντας πως οι όροι που έχουν την τάση να συναντώνται συχνά μέσα σε ένα κείμενο είναι πολύ πιθανό να είναι αντιπροσωπευτικοί για αυτό το κείμενο σε σχέση με κάποιο άλλο εξωτερικό σύνολο κειμένων αναφοράς (external reference corpus). Δυστυχώς, οι επιδόσεις αυτών των μετρικών μπορεί να είναι μέτριες, καθώς υπάρχουν εργασίες που έχουν δείξει ότι οι πιο συχνά εμφανιζόμενοι όροι δεν είναι και απαραίτητα οι καταλληλότεροι για ένα δεδομένο κείμενο. Ακόμα, υπάρχουν προσεγγίσεις που μπορεί να χρησιμοποιούν ένα συνδυασμό στατιστικών ή ευρετικών μεθόδων [11]. Το μειονέκτημα

[1] https://wordnet.princeton.edu
[2] https://www.wikipedia.org

μιας τέτοιας προσέγγισης είναι αφ' ενός ότι έχει αρκετούς περιορισμούς αλλά συνήθως έχει και χειρότερες επιδόσεις αν αλλαχθεί το σύνολο δεδομένων πάνω στο οποίο εφαρμόζεται. Οι περισσότερο προηγμένες μέθοδοι χρησιμοποιούν τεχνικές μηχανικής μάθησης που χωρίζονται σε δύο κατηγορίες: α. Εποπτευόμενη Μηχανική Μάθηση (Supervised Machine Learning) και β. Μη-Εποπτευόμενη Μηχανική Μάθηση (Unsupervised Machine Learning). Στην πρώτη περίπτωση το επιθυμητό αποτέλεσμα χρησιμοποιείται έτσι ώστε να καθοδηγήσει τον αλγόριθμο ως προς το ποιο θα πρέπει να είναι το αποτέλεσμα βάσει μιας δεδομένης εισόδου ενώ στη δεύτερη δεν υπάρχει καθοδήγηση και έτσι ο αλγόριθμος αναγνωρίζει τα μοτίβα που προκύπτουν από τα δεδομένα εισόδου και τα χρησιμοποιεί στη συνέχεια για να καθορίσει το τελικό αποτέλεσμα. Οι αλγόριθμοι που χρησιμοποιούνται παρουσιάζουν ποικιλομορφία και εξαρτώνται από την κατηγορία Μηχανικής Μάθησης (εφεξής ΜΜ) στην οποία ανήκουν. Ένας αδρός διαχωρισμός των τεχνικών δίνεται στον παρακάτω πίνακα (Πίνακας 1).

Πίνακας 1. Τεχνικές μηχανικής μάθησης ανά κατηγορία.

Εποπτευόμενη ΜΜ	Μη-Εποπτευόμενη ΜΜ
Naïve Bayes	Νευρωνικά Δίκτυα
Conditional Random Fields (CRF)	Clustering (k-means ή hierarchical clustering)
Support Vector Machines (SVM)	Expectation–maximization algorithm (EM)
Νευρωνικά Δίκτυα - Supervised	Self-Organizing Maps (SOM)

2.3.2.1. Μη-εποπτευόμενη Μηχανική Μάθηση

Όπως αναφέρθηκε και παραπάνω, στην περίπτωση αυτή ο αλγόριθμος δεν χρειάζεται ένα εκτεταμένο και συνάμα σωστά επισημασμένο σύνολο δεδομένων εκπαίδευσης (training data) για τον καθορισμό του αποτελέσματος. Σε μια εποχή που ο βασικός όγκος δεδομένων, που διακινείται στον Ιστό ή στα πιο στενά πλαίσια μιας βιβλιοθήκης, δεν είναι επισημασμένος, αυτή η ιδιότητα των προσεγγίσεων μη-εποπτευόμενης ΜΜ μπορεί να αποτελέσει ένα ουσιαστικό πλεονέκτημα σε σχέση με τις υπόλοιπες προσεγγίσεις. Βέβαια, προφανώς και αυτές οι προσεγγίσεις πέρα από τα εμφανή πλεονεκτήματα που παρέχουν, παρουσιάζουν και μειονεκτήματα. Για παράδειγμα, μπορεί να καταλήξουν σε συμπεράσματα τα οποία δεν είναι ορθά εάν γενικευτούν σε διαφορετικές θεματικές περιοχές ενώ μέχρι πρότινος η απόδοσή τους ήταν χειρότερη από αυτή των μεθόδων εποπτευόμενης ΜΜ.

Μια λίγο διαφορετική μέθοδος μη-εποπτευόμενης ΜΜ χρησιμοποιεί μια τεχνική κατάταξης που βασίζεται σε γράφο, στην οποία η σημαντικότητα ενός υποψήφιου όρου καθορίζεται από την σχετικότητά του με άλλους υποψήφιους. Σε αυτή την περίπτωση η σχετικότητα μπορεί να ποσοτικοποιηθεί είτε με την μέτρηση της συχνότητας συν-εμφάνισης μεταξύ δύο όρων είτε με την μέτρηση της σημασιολογικής τους ομοιότητας. Η μέθοδος αυτή ακολουθεί τη θεώρηση ότι οι πιο σημαντικοί υποψήφιοι όροι σχετίζονται με έναν μεγαλύτερο αριθμό από υποψήφιους, η πλειονότητα των οποίων θεωρούνται επίσης σημαντικοί υπό την έννοια ότι μπορούν να περιγράψουν επιτυχώς το περιεχόμενο ενός κειμένου. Παρόλα αυτά, αυτή η θεώρηση δεν εξασφαλίζει ότι οι όροι που θα επιλεγούν καλύπτουν όλες τις θεματικές κατηγορίες του κειμένου, αν και έχουν προταθεί παραλλαγές αυτής της προσέγγισης που προσπαθούν να αντισταθμίσουν τη συγκεκριμένη αδυναμία [10]. Ουσιαστικά, ένα έγγραφο αναπαρίσταται ως ένας γράφος οι κόμβοι του οποίου αναπαριστούν υποψήφιους όρους ενώ οι ακμές του συνδέουν όρους που είναι σχετικοί μεταξύ τους. Οι ακμές αυτές σε ορισμένες περιπτώσεις μπορεί να περιέχουν βάρη τα οποία

καταδεικνύουν το βαθμό της σχετικότητας μεταξύ των υποψηφίων όρων. Στη συνέχεια, ένας αλγόριθμος εκτελείται πάνω στο γράφο έτσι ώστε οι όροι με τη μεγαλύτερη βαθμολογία να επιστρέφονται ως οι πλέον αντιπροσωπευτικοί για ένα κείμενο.

2.3.2.2. Εποπτευόμενη Μηχανική Μάθηση

Οι μέθοδοι που ανήκουν σε αυτή την κατηγορία χρησιμοποιούν ένα εκτεταμένο επισημασμένο σύνολο δεδομένων εκπαίδευσης έτσι ώστε να καταλήξουν σε μια συνάρτηση που χαρτογραφεί ένα σύνολο μεταβλητών εισόδου που ονομάζονται χαρακτηριστικά ή ιδιότητες (features) σε κάποια επιθυμητή τιμή εξόδου. Ιδανικά, αυτή η συνάρτηση θα μπορεί να προβλέπει σωστά τις τιμές εξόδου (οι οποίες δεν είναι γνωστές) για νέα παραδείγματα με βάση μόνο αυτά τα χαρακτηριστικά. Για το λόγο αυτό υπάρχουν ερευνητικές εργασίες που έχουν ως στόχο, το σχεδιασμό των κατάλληλων χαρακτηριστικών για ένα δεδομένο σύνολο δεδομένων, μια διαδικασία που ονομάζεται «μηχανική χαρακτηριστικών» (feature design/engineering). Μια ανασκόπηση σχετικά με τα χαρακτηριστικά που επιλέγονται συνήθως αλλά και μια κατηγοριοποίηση αυτών γίνεται στην επόμενη παράγραφο του παρόντος.

Οι προσεγγίσεις αυτής της κατηγορίας συνήθως είχαν καλύτερη απόδοση σε σχέση με τις προσεγγίσεις μη-εποπτευόμενης ΜΜ. Έχουν όμως τον περιορισμό ότι η απόδοσή τους επηρεάζεται πάρα πολύ από την ποιότητα και την πληρότητα των δεδομένων εκπαίδευσης που χρησιμοποιούνται. Συνήθως, είναι αρκετά δύσκολο να βρεθεί ένα καλό και συνάμα πλήρες σύνολο δεδομένων εκπαίδευσης. Τέλος, ενυπάρχει και ο κίνδυνος το μοντέλο εκπαίδευσης που θα κατασκευαστεί να μην μπορεί να λειτουργήσει καλά για παραδείγματα που δεν έχει ξαναδεί. Για το λόγο αυτό, έχουν προταθεί μέθοδοι που αξιολογούν το μοντέλο και τη διαδικασία εκπαίδευσης γενικότερα (π.χ. με χρήση τεχνικών cross-validation). Οι σημαντικότερες προσεγγίσεις και από τις δύο κατηγορίες θα αναλυθούν σε επόμενη παράγραφο.

2.3.3. Χαρακτηριστικά επιλογής υποψήφιων όρων

Στην υπό-ενότητα αυτή γίνεται μια αναφορά στον τρόπο επιλογής υποψήφιων όρων από ένα κείμενο και μια εκτενής ανασκόπηση στα χαρακτηριστικά που χρησιμοποιούν οι διάφορες προσεγγίσεις. Απώτερος σκοπός της ανασκόπησης αυτής είναι η ανάλυση των τρόπων επιλογής όρων για την εξαγωγή πληροφορίας και η ταξινόμηση των προσεγγίσεων σε σχέση με τα χαρακτηριστικά που χρησιμοποιούνται στην εξαγωγή έτσι ώστε στη συνέχεια να επιλεγεί το υποσύνολο αυτών που θα χρησιμοποιηθεί στη συγκεκριμένη Διατριβή. Τα χαρακτηριστικά τα οποία μπορούν να χρησιμοποιηθούν για να υποβοηθήσουν την εξαγωγή πληροφορίας μπορούν να κατηγοριοποιηθούν σε τρεις βασικές κατηγορίες: (1) συνοχής κειμένου (document cohesion features), δηλαδή χαρακτηριστικά που εκφράζουν τη σχέση μεταξύ του κειμένου και των λέξεων αυτού, (2) συνοχής φράσεων (keyphrase cohesion features), δηλαδή χαρακτηριστικά που απεικονίζουν τη σχέση μεταξύ των σύνθετων όρων του κειμένου, και (3) συνοχής όρων (term cohesion features), δηλαδή χαρακτηριστικά που προσδιορίζουν τη σχέση μεταξύ των απλών όρων που απαρτίζουν έναν σύνθετο όρο [12].

Οι προσεγγίσεις που παρουσιάστηκαν στο [13] και στο [14] χρησιμοποιούν τις ευρετηριασμένες λέξεις από ένα κείμενο ως υποψήφιες για εξαγωγή ενώ στο [15] και [16] η εξαγωγή των υποψηφίων όρων γίνεται μέσα από ένα σύνολο χειρωνακτικά δημιουργημένων κανόνων τυπικών εκφράσεων (regular expression rules). Σύμφωνα με το [17] και οι δύο παραπάνω προσεγγίσεις μειονεκτούν στο ότι δεν ασχολούνται ενεργά με την ανάλυση των όρων, καθώς στη συγκεκριμένη εργασία οι συγγραφείς θεωρούν πως μπορεί να υπάρξουν βελτιώσεις στην ποιότητα

της πληροφορίας που εξάγεται εάν πρωτίστως πραγματοποιηθεί μια προσεκτική ανάλυση τις δομής και τις σύνθεσης των υποψήφιων όρων. Η ιδέα αυτή σίγουρα θα μπορούσε να βελτιώσει τους όρους που τελικά θα εξαχθούν, όμως απαιτεί πρότερη γνώση του πεδίου Ανάλυσης Φυσικής Γλώσσας (NLP).

Συντακτικά, οι σύνθετοι όροι σε ένα δεδομένο κείμενο μπορεί να σχηματιστούν είτε από απλά ουσιαστικά (simplex words) ή από φράσεις ουσιαστικών (noun phrases –NPs). Οι φράσεις ουσιαστικών μπορεί να είναι μια αλληλουχία από ουσιαστικά και λέξεις που τα προσδιορίζουν όπως επίθετα (adjectives) ή επιρρήματα (adverbs), παρόλο που τα τελευταία μπορεί να εμφανίζονται με μικρή συχνότητα σε ένα κείμενο. Ακόμα, οι σύνθετοι όροι μπορεί να ενσωματώνουν μια περιφραστική έκφραση (prepositional phrase – PP). Όταν οι όροι παίρνουν τη μορφή μιας φράσης ουσιαστικών που ακολουθείται από μια περιφραστική έκφραση, υπάρχουν συγκεκριμένες προθέσεις που μπορούν να εμφανιστούν, οι οποίες όμως εξαρτώνται κάθε φορά από τη δομή της γλώσσας του κειμένου. Οι ανωτέρω συνδυασμοί συσχετίζονται καλά με τα σχήματα λόγου (POS) που χρησιμοποιούνται στην πλειονότητα των σύγχρονων συστημάτων εξαγωγής πληροφορίας.

Η ανάλυση που πραγματοποιήθηκε στο [17], καταδεικνύει επιπρόσθετα γλωσσολογικά σχήματα (linguistic patterns) και παραλλαγές αυτών που οι προηγούμενες μελέτες είχαν παραβλέψει. Από την προαναφερθείσα μελέτη προκύπτει ότι οι αντιπροσωπευτικοί όροι μπορεί να εμφανιστούν σαν απλοί σύνδεσμοι και σε σπανιότερες περιπτώσεις ως σύνδεσμοι που αποτελούνται από περισσότερο περίπλοκες φράσεις ουσιαστικών. Βέβαια, υπάρχει και η περίπτωση σε ένα κείμενο να περιέχονται όροι που μπορεί να είναι περισσότερο περίπλοκοι σε σχέση με τους προαναφερθέντες. Παρομοίως, συντομογραφίες και τύποι που δείχνουν κτήση θεωρούνται από τους συγγραφείς ως σχήματα που συναντώνται συνήθως.

Όπως είναι φανερό από τα παραπάνω, οι συγκεκριμένες τεχνικές απαιτούν γνώσεις Υπολογιστικής Γλωσσολογίας (Computational Linguistics) και χρειάζεται εκ νέου ανάλυση κάθε φορά που το υποκείμενο σύνολο δεδομένων διαφοροποιείται. Για το λόγο αυτό οι συγκεκριμένες προσεγγίσεις δεν θα αποτελέσουν αντικείμενο μελέτης της παρούσας Διατριβής.

2.3.3.1. Τεχνικές επιλογής των υποψήφιων όρων

Οι τεχνικές που θα χρησιμοποιηθούν για την επιλογή των όρων που θα εξαχθούν από ένα δεδομένο κείμενο είναι φανερό πως αποτελούν απόφαση ύψιστης σημασίας για ένα σύστημα εξαγωγής πληροφορίας, καθώς θα καθορίσουν σε μεγάλο βαθμό τόσο τον αριθμό όσο και την ποιότητα των εξαγόμενων όρων. Το βήμα της επιλογής των υποψηφίων όρων αποτελεί προϋπόθεση για την εξαγωγή των τελευταίων, καθώς στην πλειονότητα των προσεγγίσεων οι “Ν” όροι που συγκεντρώνουν τη μεγαλύτερη βαθμολογία από το σύνολο των όρων που έχουν προκύψει από τη διαδικασία της επιλογής είναι και αυτοί που τελικά χρησιμοποιούνται ως αντιπροσωπευτικοί και συνεπώς μπορούν να αποδώσουν επιτυχώς το νόημα ή τη θεματική κατηγορία του κειμένου. Οι μέθοδοι αυτές μπορεί να βασίζονται είτε σε τεχνικές ευρετηρίασης [12] είτε σε κανόνες τυπικών εκφράσεων [15]. Όπως αναφέρθηκε και στην προηγούμενη παράγραφο στο [17] οι συγγραφείς εστιάζουν περισσότερο στην γλωσσολογική ανάλυση των όρων και στην επιλογή συγκεκριμένων κατηγοριών που τους ενδιαφέρουν. Ακόμα, κάποιες τεχνικές θέτουν κάποιον περιορισμό στο μήκος του όρου που θα εξαχθεί χρησιμοποιώντας ευρετικούς κανόνες (length heuristics).

Μια ακόμα δημοφιλής τεχνική είναι η συχνότητα εμφάνισης ενός όρου μέσα στο κείμενο. Η λογική πίσω από αυτή την προσέγγιση σχετίζεται με το γεγονός πως ένας όρος που έχει υψηλή συχνότητα εμφάνισης μέσα στο κείμενο είναι πολύ πιθανό να είναι αντιπροσωπευτικός για το

δεδομένο. Βέβαια, στον κανόνα αυτό υπάρχουν και εξαιρέσεις καθώς όροι οι οποίοι έχουν εξαχθεί χειροκίνητα από κάποιον συγγραφέα, και άρα να θεωρούνται αντιπροσωπευτικοί για το δεδομένο κείμενο, μπορεί να μην έχουν υψηλή συχνότητα εμφάνισης μέσα στο κείμενο. Ακόμα, κάποιες τεχνικές χρησιμοποιούν κανόνες για αποκοπή προθεμάτων ή καταλήξεων από τις υποψήφιες φράσεις (alternation rules). Τέλος, ανάλογα με την προσέγγιση μπορεί να χρησιμοποιηθούν και κανόνες εξαγωγής (extraction rules) που μπορεί να εξάγουν απλά ουσιαστικά, φράσεις ουσιαστικών, περιφραστικές εκφράσεις ή και σχήματα λόγου.

2.3.3.2.Χαρακτηριστικά συνοχής κειμένου

Τα χαρακτηριστικά που εμπίπτουν σε αυτή την κατηγορία υποδεικνύουν πόσο σημαντικοί είναι οι υποψήφιοι όροι για ένα δεδομένο κείμενο. Το πιο σημαντικό χαρακτηριστικό της κατηγορίας είναι η **Σταθμισμένη Συχνότητα Όρων** (TF-IDF) δεδομένου ότι εμφανίζεται στην πλειονότητα των προσεγγίσεων που έχουμε συναντήσει. Το χαρακτηριστικό αυτό εξετάζει τη συχνότητα των όρων που εμφανίζονται σε ένα κείμενο. Παρόλη την μεγάλη απήχηση της, μειονέκτημα αποτελεί το γεγονός ότι απαιτεί ένα αρκετά μεγάλο σύνολο δεδομένων για να μπορέσει να υπολογιστεί με ακρίβεια.

Επιπροσθέτως, χαρακτηριστικά που ανήκουν σε αυτή την κατηγορία είναι η απόσταση (distance) και οι πληροφορίες τμήματος (section information). Όπως είναι φανερό τα δύο τελευταία χαρακτηριστικά σχετίζονται με την τοπικότητα των όρων μέσα στο κείμενο. Αυτό βασίζεται στην ιδέα ότι αντιπροσωπευτικοί όροι ενός κειμένου είναι πιθανό να εμφανίζονται είτε στην αρχή είτε στο τέλος αυτού. Πιο συγκεκριμένα, αυτό συναντάται συνήθως σε δομημένες αναφορές (structured reports), άρθρα και περιοδικά αλλά και σε δεδομένα ειδήσεων. Όλες αυτές οι πηγές δεδομένων έχουν κοινό χαρακτηριστικό ότι έχουν μια έννοια δομής και έτσι είναι δυνατόν να εκμεταλλευτεί κάποιος την έννοια της τοπικότητας.

Στο [16] οι συγγραφείς χρησιμοποίησαν τις πληροφορίες ενός συγκεκριμένου τμήματος του εγγράφου όπου βρέθηκε κάποιος όρος. Αυτή η ιδιότητα που βασίζεται στην τοπικότητα των όρων μπορεί να χρησιμοποιηθεί για να προσδιορίσει σημαντικά τμήματα μέσα σε ένα κείμενο. Για παράδειγμα, στην παραπάνω εργασία οι συγγραφείς έδωσαν μεγαλύτερη βαρύτητα σε όρους που βρέθηκαν στην περίληψη, στην εισαγωγή, στο επίλογο, στην αρχή του κάθε τμήματος, στον τίτλο αλλά και τις αναφορές.

Αντίστοιχα, στο [17] χρησιμοποιήθηκαν κάποιες επιπρόσθετες πληροφορίες τμήματος εκτός από αυτές που έχουν ήδη αναφερθεί. Αυτές είναι, όροι που μπορεί να εμφανίστηκαν σε άλλα τμήματα ενός δεδομένου κειμένου ή στη βιβλιογραφική ανασκόπηση αυτού αλλά και κάποιες παραλλαγές τους, όπως υποσύνολα λέξεων που εμφανίζονται στους τίτλους των τμημάτων (section headers) και στους τίτλους των βιβλιογραφικών αναφορών. Μια ακόμα παραλλαγή αποτελεί το σύνολο των όρων που εμφανίστηκαν σε περισσότερα του ενός από τα παραπάνω τμήματα, γεγονός που δείχνει πως υπάρχει συσχέτιση μεταξύ των τμημάτων. Έτσι, οι όροι που προέρχονται από συγκεκριμένα τμήματα μπορεί να έχουν συγκεκριμένα βάρη καθώς με αυτό τον τρόπο μπορεί να φανεί η σημαντικότητα του κάθε τμήματος στην εξαγωγή. Αυτό πρακτικά σημαίνει ότι οι όροι που εμφανίζονται στην περίληψη μπορεί να έχουν μεγαλύτερη βαρύτητα σε σχέση με αυτούς που εμφανίζονται στη βιβλιογραφική ανασκόπηση ή στους τίτλους των τμημάτων.

Τέλος, ένα ακόμα χαρακτηριστικό που σχετίζεται με την απόσταση που αναφέρθηκε στα προηγούμενα, είναι η τελευταία εμφάνιση ενός όρου. Η θέση της τελευταίας εμφάνισης ενός όρου μπορεί να καταδεικνύει την σημαντικότητα αυτού, καθώς οι όροι πέρα από την περίληψη και την εισαγωγή τείνουν να εμφανίζονται και στον επίλογο. Παρακάτω, δίνονται τα χαρακτηριστικά συνοχής

κειμένου σε σχέση με τις εργασίες που έχουν χρησιμοποιηθεί (Πίνακας 2). Σε αυτό το σημείο πρέπει να επισημανθεί ότι όλα τα χαρακτηριστικά που αναφέρονται στον συγκεκριμένο πίνακα μπορούν να χρησιμοποιηθούν σε προσεγγίσεις εποπτευόμενης αλλά και μη-εποπτευόμενης ΜΜ.

Πίνακας 2. Χαρακτηριστικά που μετρούν τη συνοχή τις κειμένου σε συνάρτηση με τις εργασίες που χρησιμοποιούνται.

	Χαρακτηριστικά συνοχής κειμένου				
	TF-IDF	Απόσταση / Πρώτη εμφάνιση	Πληροφορίες Τμήματος	Πρόσθετες Πληροφορίες Τμήματος	Απόσταση / Τελευταία εμφάνιση
Frank & συνεργάτες [14]	X	X			
Witten & συνεργάτες [13]	X	X			
Mendelyan & Witten [20]	X	X			
Nguyen & Kan [16]	X		X		
Kim & Kan [17]	X	X	X	X	X
Turney [18]	X	X			
Barker & Corrnachia [7]	X[3]				
VRL 2009 [31]	X				
Hulth & Megyesi [19]	X[4]	X			

[3] Στη συγκεκριμένη μελέτη χρησιμοποιήθηκε μόνο η TF.
[4] Έχουν χρησιμοποιηθεί και ξεχωριστά η TF και IDF

2.3.3.3 Χαρακτηριστικά συνοχής φράσεων

Η λογική πάνω στην οποία βασίζεται η χρήση αυτής της κατηγορίας χαρακτηριστικών για την εξαγωγή των υποψήφιων όρων, είναι ότι οι όροι ενός κειμένου τείνουν να συσχετίζονται μεταξύ τους καθώς είναι σημασιολογικά συσχετισμένοι με το θέμα του υπό εξέταση κειμένου. Βέβαια η λογική αυτή είναι συνεπής μόνο όταν το κείμενο περιγράφει ένα και μόνο θέμα, καθώς σε διαφορετική περίπτωση για να πραγματοποιηθεί αποτελεσματική εξαγωγή πληροφορίας ταξινόμησης θα πρέπει το κείμενο να αποδομηθεί με κάποιο τρόπο στις διαφορετικές θεματικές κατηγορίες.

Τα βασικά χαρακτηριστικά που χρησιμοποιούνται για τη μέτρηση της συνοχής μεταξύ των όρων είναι: η συν-εμφάνιση ενός άλλου υποψήφιου όρου στο ίδιο τμήμα, η επικάλυψη των τίτλων και τέλος η συνοχή των όρων μεταξύ τους. Η ιδέα πίσω από το πρώτο χαρακτηριστικό βασίζεται στο γεγονός πως όροι που συνυπάρχουν σε περισσότερα από ένα βασικά τμήματα του κειμένου είναι πιο πιθανό να αποτελούν αντιπροσωπευτικούς όρους για το υπό εξέταση κείμενο. Στο [17] χρησιμοποιείται ο αριθμός των τμημάτων όπου υπάρχουν συν-εμφανίσεις υποψήφιων όρων για τη μέτρηση της συνοχής τους. Επίσης, οι τίτλοι πολλές φορές μπορεί να χρησιμοποιηθούν για να αναπαραστήσουν το θέμα ενός κειμένου. Έτσι οι τίτλοι από μια συλλογή μπορεί να χρησιμοποιηθούν για την αναγνώριση του θέματος ενός κειμένου πριν από την επιλογή των υποψήφιων όρων.

Πίνακας 3. Χαρακτηριστικά που μετρούν τη συνοχή μεταξύ των όρων ενός κειμένου με αναφορά στις εργασίες που χρησιμοποιούνται. Ο συμβολισμός S - Supervised, U-Unsupervised, αναφέρεται στο εάν το συγκεκριμένο χαρακτηριστικό μπορεί να χρησιμοποιηθεί τόσο σε μεθόδους Εποπτευόμενης και Μη-Εποπτευόμενης ΜΜ.

Χαρακτηριστικά συνοχής όρων			
	Συν-εμφάνιση (S,U)	Επικάλυψη τίτλων (S)	Συνοχή (S,U)
Kim & Kan [17]	X	X	X
Turney [18]			X

Στο [18] ο συγγραφέας ενσωματώνει στο σύστημα του, ένα χαρακτηριστικό συνοχής όρων υπολογίζοντας τη σημασιολογική ομοιότητα (semantic similarity) μεταξύ των “N” υποψήφιων όρων με τη μεγαλύτερη βαθμολογία σε σχέση με τους υπόλοιπους. Για να είναι δυνατή αυτή η σύγκριση στη συγκεκριμένη εργασία χρησιμοποιήθηκε ένα ξεχωριστό σύνολο δεδομένων. Αντίθετα στο [17], λόγω της έλλειψης ενός αντίστοιχου συνόλου δεδομένων για τον υπολογισμό της σημασιολογικής ομοιότητας, χρησιμοποιείται μια παραλλαγή της παραπάνω ιδιότητας που ονομάζεται ομοιότητα με βάση τα συμφραζόμενα (contextual similarity). Για τον υπολογισμό αυτής χρησιμοποιείται ένας 2-διάστατος πίνακας όπου σε κάθε γραμμή εμφανίζεται ένας υποψήφιος όρος ενώ σε κάθε στήλη μια γειτονική του λέξη. Τέλος, χρησιμοποιείται το μέτρο του συνημίτονου για τον υπολογισμό της ομοιότητας μεταξύ των όρων.

2.3.3.4 Άλλα χαρακτηριστικά

Σε αυτή την ενότητα παρουσιάζονται άλλα χαρακτηριστικά που έχουν προταθεί σε διάφορες ερευνητικές προσεγγίσεις και μπορούν να χρησιμοποιηθούν επιπρόσθετα για να βελτιώσουν τη διαδικασία επιλογής υποψήφιων όρων. Ένα από αυτά είναι η χρήση της ανάλυσης των συστατικών των σύνθετων όρων (S,U). Μια υψηλή συσχέτιση μεταξύ των όρων που απαρτίζουν έναν σύνθετο όρο καταδεικνύει ότι ο υπό εξέταση όρος είναι όντως αντιπροσωπευτικός. Στην εργασία των Park και συνεργατών [15] για τον υπολογισμό αυτό χρησιμοποιήθηκε το “Dice coefficient” για τους σύνθετους όρους. Μια παραλλαγή του “Dice coefficient” χρησιμοποιείται και στην εργασία που παρουσιάστηκε στο [17], ενώ στο [16] έχουν χρησιμοποιηθεί τα συνώνυμα (S) ως χαρακτηριστικό. Το συγκεκριμένο

όμως μπορεί να θεωρηθεί πως είναι εξαρτώμενο-πεδίου και άρα επηρεάζεται σε μεγάλο βαθμό από το εκάστοτε σύνολο δεδομένων.

Ένα ακόμα χαρακτηριστικό που έχει χρησιμοποιηθεί στο [19] είναι οι αλληλουχίες σχημάτων λόγου -POS sequences (S). Στη συγκεκριμένη εργασία, οι συγγραφείς δείχνουν πως οι όροι που ενυπάρχουν μέσα σε σχήματα λόγου τείνουν να είναι παρόμοιοι. Παρόλα αυτά και αυτό το χαρακτηριστικό είναι άμεσα εξαρτώμενο από το σύνολο δεδομένων που χρησιμοποιείται. Στην εργασία των Barker & Cornacchia [7] αλλά και στο [20] χρησιμοποιείται το μήκος των φράσεων (Length of Key phrases) (S,U) ενώ στο [16] η αλληλουχία προθέματος (suffix sequence). Το δεύτερο χαρακτηριστικό χρησιμοποιείται για να δείξει την ροπή της Αγγλικής γλώσσας να χρησιμοποιεί μορφολογία που προέρχεται από τα λατινικά για την πλειονότητα των τεχνικών όρων. Όπως είναι φανερό και αυτό είναι εξαρτημένο από το υποκείμενο σύνολο δεδομένων και έτσι χρησιμοποιείται μόνο σε προσεγγίσεις εποπτευόμενης ΜΜ (S).

Στο [18] χρησιμοποιείται η έννοια της συχνότητας του όρου (S). Αυτό σημαίνει πως για έναν όρο P, σε ένα κείμενο D με ένα σύνολο δεδομένων εκπαίδευσης C, η συχνότητα ενός όρου είναι οι φορές που ο συγκεκριμένος όρος έχει ανατεθεί από έναν συγγραφέα σε ένα κείμενο που ανήκει στο C, αν από το σύνολο εκπαίδευσης αφαιρεθεί το D. Αυτό σημαίνει ότι εάν ένας όρος έχει επιλεγεί από πολλούς συγγραφείς, τότε είναι πιο πιθανό να είναι αντιπροσωπευτικός για το κείμενο. Όπως και τα χαρακτηριστικά που αναφέρθηκαν παραπάνω έτσι και το συγκεκριμένο δεν είναι ανεξάρτητο του υποκείμενου συνόλου δεδομένων. Η ίδια προσέγγιση επαυξημένη με τα χαρακτηριστικά που περιεγράφηκαν παραπάνω ακολουθείται και στο [19].

Αντίστοιχα στην εργασία των Sugiyama και συνεργατών [21] χρησιμοποιούνται πέντε διαφορετικά χαρακτηριστικά βάση των οποίων γίνεται η εξαγωγή των όρων από τις προτάσεις. Οι λέξεις που εξάγονται χρησιμοποιούνται στη συνέχεια για την εκπαίδευση ενός ταξινομητή για την αναγνώριση προτάσεων σε μια ερευνητική εργασία που πρέπει να επισημανθούν ως αναφορές [21]. Τα χαρακτηριστικά αυτά είναι 1-grams, 2-grams και ουσιαστικά που γράφονται με κεφαλαία γράμματα (proper nouns). Ακόμα, χρησιμοποιούνται τοπικά χαρακτηριστικά με την έννοια ότι αναλύεται η θέση στην οποία θα βρεθεί η πρόταση μέσα στο κείμενο αλλά και η προηγούμενη και η επόμενη πρόταση σε σχέση με αυτήν. Τέλος, στη συγκεκριμένη εργασία έχει υιοθετηθεί και ένα χαρακτηριστικό που βασίζεται στην ορθογραφία, το οποίο χρησιμοποιείται για τον έλεγχο της πρότασης για συγκεκριμένα χαρακτηριστικά μορφοποίησης όπως: συγκεκριμένος τρόπος γραφής και ορθογραφίας, προτάσεις που περιέχουν αριθμούς ή κεφαλαία γράμματα κλπ.

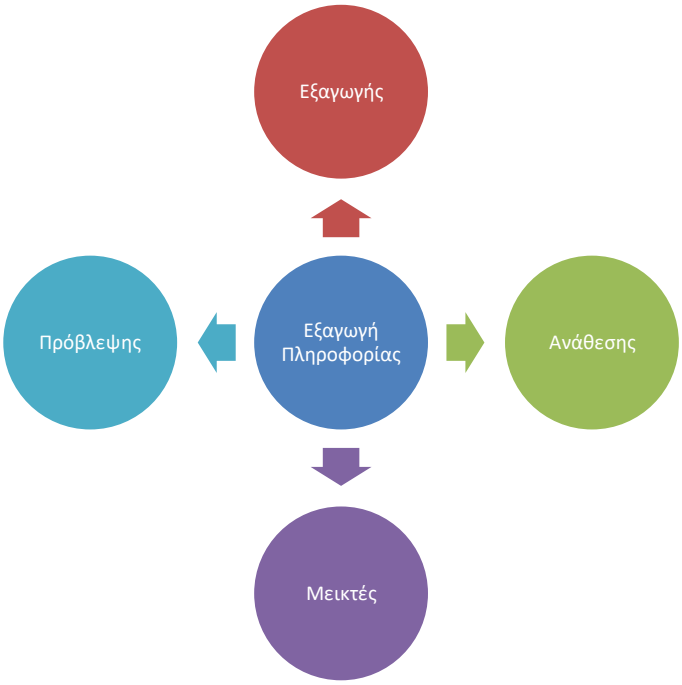
Ένα ενδιαφέρον χαρακτηριστικό που όμως μπορεί να χρησιμοποιηθεί μόνο σε προσεγγίσεις που χρησιμοποιούν κάποιο δεσμευμένο λεξιλόγιο, παρουσιάστηκε στο [20]. Εκεί, πέρα από τα κλασσικά χαρακτηριστικά συνοχής κειμένου όπως παρουσιάστηκαν στον Πίνακα 2, χρησιμοποιείται ακόμα ο βαθμός του κόμβου ενός όρου. Ο βαθμός ενός κόμβου αντιπροσωπεύει τον αριθμό των συνδέσμων ενός θησαυρού ή δεσμευμένου λεξιλογίου που συνδέουν τον όρο με άλλους υποψήφιους όρους. Αυτό έχει άμεση πρακτική εφαρμογή καθώς εάν ένα έγγραφο περιγράφει μια συγκεκριμένη θεματική περιοχή, τότε καλύπτει το μεγαλύτερο μέρος από τους όρους θησαυρού που σχετίζονται με αυτό το θέμα. Ως εκ τούτου, οι υποψήφιοι όροι με υψηλό βαθμό κόμβου είναι πιο πιθανό να είναι σημαντικοί.

Σε αντίθεση με τις προαναφερθείσες προσεγγίσεις που βασίζονταν στη χρήση χαρακτηριστικών που σχετίζονται μόνο με το κείμενο, στο [22] γίνεται προσπάθεια να ενσωματωθούν και να χρησιμοποιηθούν σε αυτή τη διαδικασία περισσότερο πλούσιες

αναπαραστάσεις του εγγράφου, αλλά και τοπολογικά χαρακτηριστικά που έχουν προκύψει από τη χρήση τεχνολογιών όπως η οπτική αναγνώριση χαρακτήρων (OCR), το μέγεθος της γραμματοσειράς και η θέση των όρων μέσα στο κείμενο [23]. Ακόμα, χρησιμοποιούνται χαρακτηριστικά όπως αριθμοί που μπορεί να καταδεικνύουν διαφορετικά τμήματα στο έγγραφο ή μπορεί να χρησιμοποιηθούν για την αναγνώριση των επικεφαλίδων διαφορετικών τμημάτων σε ένα κείμενο. Ένα χαρακτηριστικό που επίσης χρησιμοποιήθηκε στο [22] είναι τα σημεία στίξης, όπως για παράδειγμα σύμβολα που είναι σε θέση να διαχωρίσουν διευθύνσεις ηλεκτρονικού ταχυδρομείου και διευθύνσεις στον Παγκόσμιο Ιστό. Η προσέγγιση αυτή στοχεύει στη αναγνώριση της λογικής δομής και των σημασιολογικών πληροφοριών που περικλείονται σε ένα έγγραφο ή μια ερευνητική εργασία με χρήση των παραπάνω χαρακτηριστικών. Όλα τα παραπάνω χαρακτηριστικά χρησιμοποιούνται στο σύνολο τους και από το σύστημα Enlil [24]. Τέλος στο [25], χρησιμοποιείται ένας συνδυασμός των παραπάνω λεξικολογικών και τοπολογικών χαρακτηριστικών, όπως αριθμοί, σημεία στίξης και άλλες ιδιότητες που βασίζονται στην ορθογραφία.

2.4. Ταξινόμηση Προσεγγίσεων

Στην παράγραφο αυτή γίνεται μια προσπάθεια ταξινόμησης των προσεγγίσεων σε κατηγορίες με βάση τη βιβλιογραφική ανασκόπηση που έχει πραγματοποιηθεί. Ένας αδρός διαχωρισμός των κατηγοριών θα μπορούσε να είναι: α. **οι παραδοσιακές προσεγγίσεις**, β. **οι μεικτές προσεγγίσεις** και γ. **οι προσεγγίσεις πρόβλεψης** (Εικόνα 1). Οι παραδοσιακές προσεγγίσεις εξαγωγής πληροφορίας μπορεί να χωριστούν σε δύο μεγάλες κατηγορίες [26]. Μπορούν να θεωρηθούν είτε ως μια διαδικασία εξαγωγής (extraction) είτε ως μια διαδικασία ανάθεσης (assignment). Στην πρώτη κατηγορία, οι λέξεις ή όροι που εμφανίζονται στο σώμα του κειμένου αναλύονται με σκοπό τελικά να επιλεγούν οι πιο αντιπροσωπευτικοί. Στη δεύτερη κατηγορία, τα κείμενα κατηγοριοποιούνται σε έναν προκαθορισμένο αριθμό ομάδων με βάση κάποιους δεδομένους όρους. Οι μεικτές προσεγγίσεις ουσιαστικά αποτελούν συνδυασμό των δύο προηγούμενων. Πρόσφατα έχουν προταθεί κάποιες προσεγγίσεις που βασίζονται στην πρόβλεψη, οι οποίες στοχεύουν, χρησιμοποιώντας τη γνώση των συμφραζομένων (context) μιας λέξης, να προβλέψουν την ίδια την λέξη ή, αντίθετα, αν γνωρίζουν την λέξη να προβλέψουν τα συμφραζόμενα γύρω από αυτήν. Εκτενέστερη αναφορά στην κατηγορία αυτή θα γίνει σε επόμενη παράγραφο. Πλέον, υπάρχει μια στροφή ερευνητικά προς προσεγγίσεις που χρησιμοποιούν μοντέλα που βασίζονται στην πρόβλεψη (predictive models) παρά στην χρήση ευρετικών (heuristics) ή μοντέλων μέτρησης (count models).



Εικόνα 1. Ταξινόμηση Προσεγγίσεων Εξαγωγής Πληροφορίας.

Πέρα από αυτό το διαχωρισμό υπάρχουν και άλλοι τρόποι ταξινόμησης ενός συστήματος Εξαγωγής Πληροφορίας από κείμενα, οι οποίοι δίνονται παρακάτω (Εικόνα 2).

Εικόνα 2. Ταξινόμηση προσεγγίσεων εξαγωγής πληροφορίας με βάση διαφορετικά κριτήρια.

Οι βασικότεροι από αυτούς είναι:

- Εξαγωγή με χρήση στατιστικών μεθόδων (Statistical Methods)
- Εξαγωγή με χρήση ευρετικών μεθόδων (Heuristic Methods)

- Εξαγωγή με χρήση τεχνικών Εποπτευόμενης ΜΜ (Supervised Machine Learning Methods)
- Εξαγωγή με χρήση τεχνικών Μη-Εποπτευόμενης ΜΜ (Un-Supervised Machine Learning Methods)
- Εξαγωγή με χρήση τεχνικών ανάλυσης φυσικής γλώσσας (NLP)
- Εξαγωγή με αυτόματο τρόπο (Automatic Extraction)
- Εξαγωγή με ημιαυτόματο τρόπο (Semi-automatic Extraction)
- Εξαγωγή ανεξαρτήτως πεδίου (Generic Extraction)
- Εξαγωγή με βάση συγκεκριμένο πεδίο (Domain-Specific Extraction)
- Εξαγωγή με χρήση κάποιου ελεγχόμενου λεξιλογίου (Thesaurus – based Extraction)

Όπως είναι φανερό από αυτόν τον αδρό διαχωρισμό, τα κριτήρια κατηγοριοποίησης παρουσιάζουν μεγάλη ποικιλομορφία και επίσης είναι δυνατόν μια προσέγγιση να ενσωματώνει παραπάνω από ένα. Για το λόγο αυτό επιλέχθηκε στην παρούσα Διατριβή οι προσεγγίσεις να ταξινομηθούν με βάση τις τέσσερις βασικές κατηγορίες που αναφέρθηκαν παραπάνω, δηλαδή κάθε τεχνική θα κατηγοριοποιηθεί με βάση το αν στηρίζεται σε διαδικασία εξαγωγής είτε σε διαδικασία ανάθεσης και το αν αποτελεί μεικτή προσέγγιση ή προσέγγιση πρόβλεψης. Οι προσεγγίσεις πρόβλεψης που προσφάτως έχουν συγκεντρώσει την προσοχή της ερευνητικής κοινότητας, θα αποτελέσουν ξεχωριστή ενότητα του παρόντος.

2.4.1. Προσεγγίσεις εξαγωγής

Στη συγκεκριμένη ενότητα κατηγοριοποιούνται οι προσεγγίσεις εξαγωγής. Σε αυτή την περίπτωση οι όροι δεν περιορίζονται σε αυτούς που εμφανίζονται σε κάποια ιεραρχία ή δεσμευμένο λεξιλόγιο. Αντίθετα, οι όροι μπορεί να είναι οποιοιδήποτε αρκεί να εμφανίζονται μέσα στο υπό εξέταση κείμενο. Έτσι, χρησιμοποιώντας ένα σύνολο κειμένων εκπαίδευσης, οι τεχνικές ΜΜ μπορούν να χρησιμοποιηθούν για να καθορίσουν ποια είναι εκείνα τα χαρακτηριστικά που είναι σε θέση να καθορίσουν ποιοι όροι μπορούν να χρησιμοποιηθούν ως αντιπροσωπευτικοί για να περιγράψουν επιτυχώς το νόημα ενός κειμένου και ποιοι όχι.

Στο [7] παρουσιάζεται μια προσέγγιση στην οποία η εξαγωγή έχει ως στόχο την παραγωγή με αυτόματο τρόπο της σύνοψης του θέματος ενός επιστημονικού περιοδικού (summarization). Στόχος της προσέγγισης αυτής είναι να δημιουργηθεί η σύνοψη του υποκείμενου θέματος χρησιμοποιώντας τεχνικές εξαγωγής πληροφορίας, το αποτέλεσμα των οποίων θα τοποθετηθεί στην αρχική σελίδα του άρθρου ή γενικότερα κάποιου δεδομένου κειμένου. Έτσι, η προσέγγιση αυτή έγκειται στην εξαγωγή προτάσεων που εμφανίζουν υψηλή βαθμολογία σύμφωνα με τον αλγόριθμο που προτείνουν. Ο αλγόριθμος αυτός εξάγει φράσεις που αποτελούνται από ουσιαστικά (noun phrases) και επιλέγονται με βάση το μήκος τους, τη συχνότητα εμφάνισής τους μέσα στο κείμενο αλλά και τη συχνότητα εμφάνισης της ρίζας του ουσιαστικού (base noun). Η συγκεκριμένη προσέγγιση θα μπορούσε να χαρακτηριστεί και ως ευρετική, καθώς κατά κύριο λόγο βασίζεται σε ένα σύνολο ευρετικών κανόνων. Σε αυτή την περίπτωση οι φράσεις αποτελούν μια ειδική μορφή περίληψης. Άμεσο πλεονέκτημα αυτού είναι να δύναται ο εκάστοτε χρήστης να αναγνωρίσει σε ελάχιστο χρόνο εάν το θέμα ενός δεδομένου κειμένου είναι μέσα στα ενδιαφέροντά του ή όχι.

Ο Park και συνεργάτες [15] ταξινομούν τους υποψήφιους όρους με μια δική τους θεώρηση που βασίζεται στο βαθμό συσχέτισης των όρων με ένα δεδομένο πεδίο, με βάση συγκεκριμένους συνδυασμούς λέξεων. Η προσέγγιση αυτή βασίζεται σε ένα σύνολο στατιστικών μεθόδων. Σκοπό της μεθόδου αποτελεί η αυτόματη εξαγωγή ειδικών λεξιλογίων (glossaries) που σχετίζονται με κάποιο πεδίο χρησιμοποιώντας μεγάλα σύνολα κειμένων. Τέτοιου είδους ειδικά λεξιλόγια είναι πολύ χρήσιμα για την αναγνώριση των βασικών εννοιολογικών αντικειμένων μιας συγκεκριμένης θεματικής περιοχής. Πιο συγκεκριμένα, τα λεξιλόγια αυτά παρουσιάζουν έννοιες της περιοχής με τέτοιο τρόπο ώστε στη συνέχεια να μπορούν να χρησιμοποιηθούν από άλλες εφαρμογές.

Μια ενδιαφέρουσα προσέγγιση που παρουσιάστηκε στο [19], στοχεύει στο να δείξει πως όροι ή λέξεις κλειδιά που έχουν εξαχθεί αυτόματα μπορούν να χρησιμοποιηθούν επιτυχώς για την κατηγοριοποίηση κειμένων. Σε αυτή την εργασία, παρουσιάζονται πειράματα στα οποία λέξεις-κλειδιά, που έχουν εξαχθεί αυτόματα, χρησιμοποιούνται ως είσοδος σε έναν αλγόριθμο μάθησης, τόσο αυτόνομα όσο και σε συνδυασμό με το πλήρες κείμενο από το οποίο προέρχονται. Η συγκεκριμένη προσέγγιση χρησιμοποιεί τεχνικές εποπτευόμενης ΜΜ, ενώ το σύνολο των δεδομένων εκπαίδευσης έχει επισημειωθεί χειροκίνητα. Ως ένα πρώτο βήμα για την εξαγωγή των λέξεων-κλειδιών από ένα έγγραφο, οι υποψήφιοι όροι επιλέγονται από το κείμενο με τρεις διαφορετικούς τρόπους. Ο πρώτος αφορά στατιστικές μεθόδους, ενώ οι άλλοι δύο ακολουθούν μια περισσότερο γλωσσολογική προσέγγιση. Ακόμα, χρησιμοποιούνται συγκεκριμένα χαρακτηριστικά για την επιλογή των όρων τα οποία αναλύονται στην προηγούμενη παράγραφο. Η τελική επιλογή των όρων προκύπτει από το συνδυασμό των τριών παραπάνω μοντέλων.

Στο [16] οι Nguyen και Kan επέλεξαν τη διαδικασία της εξαγωγής έναντι της ανάθεσης καθώς δεν είχαν στη διάθεσή τους θεματικές επικεφαλίδες ή κάποια οντολογία για τη μέτρηση της αποτελεσματικότητας της ανάκτησης των όρων από ένα κείμενο. Ένας ακόμα λόγος επιλογής της εξαγωγής έναντι της ανάθεσης σύμφωνα με τους συγγραφείς είναι ότι οι μέθοδοι που βασίζονται στην εξαγωγή και όχι στην ανάθεση παράγουν συνήθως ένα ευρύτερο φάσμα όρων, το οποίο όπως υποστηρίζουν μπορεί να είναι καταλληλότερο για την αξιολόγηση της σχετικότητας της εξαγόμενης πληροφορίας. Η προσέγγιση τους ακολουθεί τεχνικές εποπτευόμενης ΜΜ και εφαρμόζεται σε επιστημονικές δημοσιεύσεις. Τα χαρακτηριστικά που έχουν επιλέξει είναι σε θέση να αναγνωρίσουν τη θέση των όρων μέσα σε ένα κείμενο, λόγω του ότι υπάρχουν λογικά εξαγόμενα τμήματα στο κείμενο μιας επιστημονικής εργασίας. Ακόμα, προτείνεται η χρήση μορφολογικών χαρακτηριστικών που συναντώνται συνήθως στους επιστημονικούς/τεχνικούς όρους. Πιο συγκεκριμένα, τέτοιου είδους μορφολογικά χαρακτηριστικά εξετάζουν εάν ο υπό εξέταση όρος περιέχει κάποιο ακρωνύμιο ή χρησιμοποιεί κάποια συγκεκριμένη (τεχνική) ορολογία. Έτσι, εμπλούτισαν το σύνολο των ιδιοτήτων που χρησιμοποιεί ο αλγόριθμος KEA με τα παραπάνω χαρακτηριστικά. Ο αλγόριθμος KEA αποτελεί έναν από τους πιο γνωστούς αλγορίθμους εξαγωγής και αναλύεται σε επόμενη παράγραφο του παρόντος. Σύμφωνα, με τα αποτελέσματα που παρουσίασαν, σε συγκεκριμένες περιπτώσεις ο αλγόριθμός τους έχει καλύτερη απόδοση από τον KEA [13].

Η προσέγγιση των Wan και Xiao όπως παρουσιάστηκε στο [27] αφορά την αυτόματη εξαγωγή όρων από ένα και μόνο κείμενο χρησιμοποιώντας πληροφορίες που ενυπάρχουν στο ίδιο το κείμενο για την ομαδοποίηση. Η υπόθεση που χρησιμοποιείται σε αυτή την προσέγγιση είναι ότι τα κείμενα τα οποία ανήκουν στο ίδιο ή παρόμοιο πεδίο αλληλοεπιδρούν μεταξύ τους όσο αφορά στην θέση των ίδιων ή παρόμοιων λέξεων μέσα στο κείμενο. Αρχικά, ομαδοποίησαν τα κείμενα και στη συνέχεια χρησιμοποίησαν έναν αλγόριθμο κατάταξης που βασίζεται σε γράφο για την κατάταξη των υποψήφιων όρων ενός κειμένου, λαμβάνοντας υπόψη τις αμοιβαίες αλληλεπιδράσεις άλλων

κειμένων που ανήκουν στην ίδια ομάδα με το υπό εξέταση κείμενο. Επίσης, η συγκεκριμένη προσέγγιση δεν αφορά επιστημονικές δημοσιεύσεις αλλά δεδομένα ειδήσεων. Η επιλογή αυτή έγινε λόγω του ότι τέτοιου είδους δεδομένα αποτελούν πολύ μεγάλο όγκο του Παγκόσμιου Ιστού σήμερα αλλά και διότι συνήθως οι συγγραφείς τους δεν αναθέτουν σε αυτά αντιπροσωπευτικούς όρους.

Μια ακόμα προσέγγιση που ανήκει σε αυτή την κατηγορία παρουσιάζεται στο [21], η οποία επικεντρώνεται στην αυτόματη εξαγωγή και ανάλυση των προτάσεων που χρήζουν αναφοράς σε μια δημοσίευση. Αυτό θα μπορούσε να βοηθήσει στην ταυτοποίηση ενός τμήματος κειμένου σχετικά με το αν χρειάζεται να μπει σε αυτό αναφορά ή όχι. Σε αυτή την προσέγγιση χρησιμοποιείται ένας ταξινομητής (classifier) που υλοποιεί τεχνικές ΜΜ, λαμβάνοντας υπόψη απλά χαρακτηριστικά που υπάρχουν μέσα στο κείμενο όπως ουσιαστικά αλλά και τις λέξεις που υπάρχουν στην αρχή και το τέλος κάθε πρότασης. Σκοπός αυτού του ταξινομητή είναι να εντοπίσει εάν μια πρόταση χρειάζεται αναφορά ή όχι. Η τεχνική αυτή χρησιμοποιεί μεθόδους επεξεργασίας φυσικής γλώσσας και μπορεί να εφαρμοστεί αποδοτικά σε ένα έξυπνο περιβάλλον συγγραφής, ώστε οι συγγραφείς να βοηθιούνται να αναγνωρίσουν τα σημεία που χρειάζονται αναφορά. Η χρήση τεχνικών εξαγωγής πληροφορίας από κείμενα αλλά και εύρεσης σχετικών - με το αντικείμενο του συγγραφέα - άρθρων, θα μπορούσε περαιτέρω να υποβοηθά τη συγγραφή προτείνοντας στον συγγραφέα σχετικά άρθρα που πιθανόν εκείνος θα ήθελε να ενσωματώσει στην εργασία του.

Στο [25] ερευνάται η χρήση της τεχνικής CRF (Conditional Random Fields) για αυτόματη εξαγωγή σημαντικών μετά-δεδομένων από άρθρα στον τομέα της ιατρικής. Σύμφωνα με την προσέγγιση των συγγραφέων τα μετά-δεδομένα σε τέτοιου είδους άρθρα χωρίζονται σε δυο κατηγορίες: α. τα τυποποιημένα μετά-δεδομένα (formulaic metadata) όπως ο τίτλος, ο συγγραφέας, το ίδρυμα και β. τα μετά-δεδομένα που προέρχονται από το πλήρες κείμενο του άρθρου. Σε αυτή τη μελέτη καταδεικνύεται ότι η εξαγωγή των τυποποιημένων μετά-δεδομένων υποστηρίζεται αρκετά καλά με τη χρήση της CRF τεχνικής αλλά για την εξαγωγή των μετά-δεδομένων της δεύτερης κατηγορίας είναι απαραίτητη η χρήση επιπλέον χαρακτηριστικών που ανήκουν στον τομέα της επεξεργασίας φυσικής γλώσσας.

2.4.2. Προσεγγίσεις ανάθεσης

Η παρούσα ενότητα διερευνά τεχνικές ανάθεσης αντιπροσωπευτικών όρων σε ένα κείμενο. Η συγκεκριμένη κατηγορία είναι ευρέως γνωστή και ως κατηγοριοποίηση κειμένου (text categorization), όπου όλοι οι όροι πρέπει να εμφανίζονται σε ένα προκαθορισμένο δεσμευμένο λεξιλόγιο (controlled vocabulary). Αυτό το δεσμευμένο λεξιλόγιο αποτελεί τις κατηγορίες με βάση τις οποίες θα γίνει η ανάθεση των όρων στα κείμενα. Έτσι, το πρόβλημα ανάγεται στο να βρεθεί μια αντιστοιχία μεταξύ κάποιων δεδομένων κειμένων και των κατηγοριών, χρησιμοποιώντας ένα σύνολο κειμένων για την εκπαίδευση του αλγορίθμου. Στη συνέχεια, η αντιστοιχία αυτή μπορεί να επιτευχθεί με την κατάρτιση ενός ταξινομητή (classifier) για κάθε κατηγορία, χρησιμοποιώντας κείμενα που ανήκουν στην κατηγορία ως θετικά δείγματα και τα υπόλοιπα ως αρνητικά. Σκοπός της συγκεκριμένης διαδικασίας είναι η εκπαίδευση του συστήματος έτσι ώστε στη συνέχεια να μπορεί να ταξινομήσει νέα κείμενα χωρίς να χρειαστεί νέα φάση εκπαίδευσης.

Μια τέτοια προσέγγιση παρουσιάστηκε από τους Frank και συνεργάτες, οι οποίοι χρησιμοποιήσαν τη μέθοδο “Naive Bayes” για την εξαγωγή πληροφορίας θεωρώντας πως το πρόβλημα μπορεί να αναχθεί σε ένα πρόβλημα ταξινόμησης (classification) [14]. Έτσι χρησιμοποιήσαν δύο κλάσεις για να επιτύχουν την ταξινόμηση αυτή. Στη μια κλάση ανήκουν τα θετικά δείγματα ενώ στην δεύτερη τα αρνητικά δείγματα για κάθε κατηγορία. Σε αυτό το σημείο θα

πρέπει να αναφερθεί πως ο αλγόριθμος ενθυλακώνει μια φάση προ-επεξεργασίας για την εξαγωγή από το κείμενο των υποψήφιων όρων με χρήση τεχνικών ανάλυσης φυσικής γλώσσας και ευρετικών μεθόδων. Ακόμα, η προσέγγιση αυτή χρησιμοποιεί στατιστικές μεθόδους για να καθορίσει εάν ένας όρος μπορεί να χρησιμοποιηθεί ως αντιπροσωπευτικός σε ένα κείμενο ή όχι. Τα χαρακτηριστικά που χρησιμοποιεί είναι το σχήματα στάθμισης TF-IDF, ένα σύνθετο σχήμα που αποτελείται από το γινόμενο του μέτρου TF (Term Frequency) και IDF (Inverse document Frequency), και η απόσταση από την πρώτη εμφάνιση ενός όρου μέσα στο κείμενο. Η ταξινόμηση γίνεται με χρήση του αλγορίθμου KEA [13]. Τέλος, καταλήγουν πως τα αποτελέσματα του αλγορίθμου μπορούν περαιτέρω να βελτιωθούν αν η συλλογή των κειμένων έχει επιλεγεί προσεκτικά πριν από την εφαρμογή του αλγορίθμου ή εάν ο αλγόριθμος εφαρμοστεί σε ένα συγκεκριμένο πεδίο και όχι γενικά. Πιο συγκεκριμένα, ο αλγόριθμος εμφανίζει βελτιωμένα αποτελέσματα εάν τα κείμενα που έχουν επιλεγεί για την εκπαίδευση ανήκουν στην ίδια κατηγορία όπως και τα υπό εξέταση.

Στην ίδια κατηγορία υπόκειται και η εργασία του Turney [6], όπου το πρόβλημα της αυτόματης εξαγωγής πληροφορίας από ένα κείμενο αντιμετωπίζεται ως ένα πρόβλημα εποπτευόμενης ΜΜ. Στην προσέγγιση αυτή γίνεται η υπόθεση πως όλο το κείμενο μπορεί να χωριστεί σε όρους και εν συνεχεία ο αλγόριθμος πρέπει να μάθει να ταξινομεί κάθε έναν από αυτούς ως θετικό ή αρνητικό παράδειγμα. Σύμφωνα με τον συγγραφέα ένας ιδανικός αλγόριθμος εξαγωγής πληροφορίας θα έπρεπε να είναι σε θέση να αναγνωρίζει τους ίδιους όρους όπως ένας συγγραφέας σε ποσοστό 90%. Ο αλγόριθμος που χρησιμοποιείται στη συγκεκριμένη προσέγγιση είναι ονομάζεται GenEx και αποτελεί υβρίδιο καθώς είναι το αποτέλεσμα του συνδυασμού δύο άλλων αλγορίθμων, των Genitor και Extractor. Ο πρώτος αποτελεί έναν γενετικό αλγόριθμο ενώ ο δεύτερος έναν παραμετροποιημένο αλγόριθμο εξαγωγής πληροφορίας. Η λειτουργία του βασίζεται στην απόδοση ενός αριθμητικού αποτελέσματος στους όρους ενός δεδομένου κειμένου, και την παραγωγή μιας λίστας όρων με τη μεγαλύτερη βαθμολογία. Ο τρόπος ανάθεσης της βαθμολογίας βασίζεται σε ένα σύνολο παραμέτρων. Οι παράμετροι αυτές είναι ουσιαστικά εννέα διαφορετικά συντακτικά χαρακτηριστικά (syntactic features), όπως το μήκος του εκάστοτε όρου και η συχνότητα εμφάνισης της ρίζας μιας λέξης.

Ο Turney [18] προσεγγίζει το πρόβλημα της εξαγωγής πληροφορίας εστιάζοντας στη μείωση της «ασυναρτησίας» των όρων που μπορεί να εξαχθούν από ένα σύστημα εξαγωγής. Αυτό είναι που τον διαφοροποιεί και από τις υπόλοιπες προσεγγίσεις που παρουσιάστηκαν έως τώρα. Θεωρεί πως ενώ η πλειοψηφία των όρων που εξάγονται από ένα κείμενο μπορεί να σχετίζονται μεταξύ τους, υπάρχουν και ορισμένοι που δεν ακολουθούν αυτό τον κανόνα καθώς δεν παρουσιάζουν κάποια σημασιολογική συσχέτιση ούτε με τους υπόλοιπους αλλά ούτε και μεταξύ τους. Για το λόγο αυτό, η εργασία του προτείνει βελτιώσεις στον αλγόριθμο εξαγωγής φράσεων KEA [13], [14] έτσι ώστε να αυξηθεί η συσχέτιση που παρουσιάζουν μεταξύ τους οι εξαγόμενοι από τον αλγόριθμο όροι. Η προσέγγιση του χρησιμοποιεί στατιστικά μέτρα μεταξύ των όρων που έχουν εξαχθεί από τον αλγόριθμο ως ένδειξη ότι μπορεί να είναι σημασιολογικά συσχετισμένοι. Σύμφωνα με τα αποτελέσματα της εργασίας του, αποδεικνύεται ότι οι όροι που εξάγονται με αυτό τον τρόπο είναι περισσότερο ποιοτικοί και μπορούν να εφαρμοστούν γενικότερα και όχι σε ένα συγκεκριμένο πεδίο. Αυτό πρακτικά σημαίνει πως είναι δυνατόν ο αλγόριθμος να εκπαιδευτεί σε ένα πεδίο και να εφαρμοστεί σε κάποιο άλλο. Η παρατήρηση αυτή ουσιαστικά αποτελεί βελτίωση της εργασίας που παρουσιάστηκε στο [13], η οποία είχε τον περιορισμό ότι τα αποτελέσματά της χειροτέρευαν αισθητά εάν εφαρμοζόταν σε ένα διαφορετικό πεδίο σε σχέση με το πεδίο πάνω στο οποίο είχε εκπαιδευθεί ο αλγόριθμος.

Στο [4] περιγράφεται μια μέθοδος αυτόματης εξαγωγής μετά-δεδομένων από κείμενα με χρήση τεχνικών ΜΜ και πιο συγκεκριμένα με χρήση Support Vector Machines (SVM). Στη συγκεκριμένη περίπτωση οι συγγραφείς θεωρούν την εξαγωγή πληροφορίας ως μια διαδικασία ανάθεσης που βασίζεται κατά κύριο λόγο στην ανάλυση των επικεφαλίδων ερευνητικών εργασιών. Η μέθοδος ταξινομεί πρώτα κάθε γραμμή της επικεφαλίδας σε μία ή περισσότερες από 15 προκαθορισμένες κλάσεις. Μια επαναληπτική διαδικασία σύγκλισης στη συνέχεια χρησιμοποιείται για να βελτιωθεί η ταξινόμηση ανά γραμμή χρησιμοποιώντας τις ετικέτες των κλάσεων που έχουν προκύψει από το μοντέλο πρόβλεψης των γειτονικών γραμμών της προηγούμενης επανάληψης.

Ακόμα, μια ενδιαφέρουσα προσέγγιση αποτελεί η [28], όπου παρουσιάζεται μια μέθοδος για την αναγνώριση των αντιπροσωπευτικών όρων σε ένα κείμενο έτσι ώστε να είναι δυνατή η αναγνώριση του θέματος του, με σκοπό να χρησιμοποιηθεί στη συνέχεια σε εφαρμογές κατηγοριοποίησης κειμένων. Η υπόθεση που γίνεται στη συγκεκριμένη προσέγγιση είναι ότι η χρήση όλων των όρων που εμφανίζονται σε ένα κείμενο ως υποψήφιων, μπορεί να οδηγήσει σε μεγάλη ευρύτητα τον χώρο των υποψήφιων. Για το λόγο αυτό οι όροι δεν επιλέγονται από όλο το κείμενο αλλά από συγκεκριμένα πέντε (5) κατηγορίες αυτού: α) τον τίτλο, β) την πρώτη πρόταση του κειμένου, γ) συγκεκριμένους σύνθετους όρους, δ) λέξεις σχετικές με ένα πεδίο και ε) οντότητες ονομάτων (named entities). Οι δύο πρώτες κατηγορίες βασίζονται στην έννοια του χωρισμού του κειμένου σε ζώνες με σκοπό την ανάκτηση των λέξεων αυτών, ενώ οι δύο τελευταίες στη συχνότητα εμφάνισης τους σε κείμενα που σχετίζονται με κάποιο πεδίο. Τέλος, και σε αυτή την περίπτωση, οι όροι εξάγονται με τη βοήθεια τεχνικών ΜΜ όπως αυτές που παρουσιάστηκαν παραπάνω.

Πέρα από τις παραπάνω προσεγγίσεις που στην πλειοψηφία τους χρησιμοποιούν μεθόδους ΜΜ, μια σχετικά διαφορετική τεχνική η οποία βασίζεται στις πιθανότητες παρουσιάστηκε από τους Lyong και συνεργάτες [22]. Στην εργασία τους περιγράφουν το εργαλείο SectLabel, το οποίο αποτελεί προσθήκη ενός λογισμικού ανοικτού κώδικα για την ανίχνευση της λογικής δομής ενός εγγράφου από υπάρχοντα αρχεία σε μορφότυπο PDF, χρησιμοποιώντας τη μέθοδο CRF (Conditional Random Fields). Το συγκεκριμένο σύστημα διαχωρίζει το πρόβλημα της αναγνώρισης της λογικής δομής ενός εγγράφου σε δύο υπό-διαδικασίες: α) ταξινόμηση με βάση τη λογική του δομή και β) γενική ταξινόμηση με βάση συγκεκριμένο τμήμα του. Στην πρώτη περίπτωση, ένα κείμενο θεωρείται ως μια ταξινομημένη συλλογή από γραμμές όπου ο βασικός στόχος είναι η ανάθεση κάθε γραμμής του κειμένου σε μια σημασιολογική κατηγορία, αναπαριστώντας έτσι το λογικό του ρόλο. Στη δεύτερη περίπτωση, οι τίτλοι κάθε τμήματος εξάγονται με σκοπό τον καθορισμό του λογικού σκοπού του τμήματος.

2.4.3. Μεικτές προσεγγίσεις

Σε αυτή την κατηγορία μπορούν να περιληφθούν προσεγγίσεις οι οποίες συνδυάζουν τις δύο παραπάνω, δηλαδή συνδυάζουν τόσο τη διαδικασία της ανάθεσης όσο και της εξαγωγής. Ένας από τους πιο γνωστούς αλγορίθμους εξαγωγής όρων από κείμενα αποτελεί ο αλγόριθμος KEA [13]. Ο KEA στοχεύει στην αναγνώριση των υποψήφιων φράσεων με χρήση λεξικολογικών μεθόδων μέσω του υπολογισμού των τιμών των χαρακτηριστικών για κάθε μια υποψήφια φράση και στη συνέχεια χρησιμοποιεί έναν αλγόριθμο μηχανικής μάθησης για να αποφασίσει ποιες από τις υποψήφιες φράσεις είναι κατάλληλες για το υπό εξέταση κείμενο. Ο αλγόριθμος αρχικά εκπαιδεύεται πάνω σε κείμενα για τα οποία οι υποψήφιες φράσεις είναι γνωστές και στη συνέχεια χρησιμοποιεί το εκπαιδευμένο μοντέλο για να αναγνωρίσει φράσεις σε νέα (άγνωστα) κείμενα.

Στο [14] προτείνεται μια νέα έκδοση του αλγορίθμου KEA, ο αλγόριθμος KEA++ οποίος συνδυάζει τις επιμέρους διαδικασίες της εξαγωγής και της ανάθεσης σε μια συνολική διαδικασία, όπου οι όροι ή φράσεις που εξάγονται από τα κείμενα θα πρέπει να υπάρχουν σε κάποιο ελεγχόμενο λεξιλόγιο. Τέλος, η διαδικασία ευρετηρίασης μπορεί να επεκταθεί με την ανάλυση των σημασιολογικών πληροφοριών που σχετίζονται με τους όρους των εγγράφων όπως η σχέση μεταξύ των όρων που υπάρχουν στο κείμενο αλλά και με άλλους όρους που βρίσκονται στο λεξιλόγιο.

Εξέλιξη αυτού αποτελεί η εργασία της Mendelyan [26], η οποία επικεντρώνεται στον καθορισμό σύνθετων όρων και όχι λέξεων-κλειδιά ως περιγραφέων για ένα κείμενο. Οι σύνθετοι αυτοί όροι μπορεί να έχουν προκύψει από ένα ελεγχόμενο λεξιλόγιο και μπορεί να εμφανίζονται μέσα στο κείμενο αλλά όχι απαραιτήτως. Σύμφωνα με την παραπάνω εργασία, οι όροι αυτοί αποτελούν χρήσιμα μετά-δεδομένα τόσο για φυσικές όσο και για Ψηφιακές Βιβλιοθήκες καθώς υποβοηθούν την οργάνωση των συλλογών που υπάρχουν σε μια βιβλιοθήκη με βάση το περιεχόμενο των εγγράφων.

Ο αλγόριθμος KEA++, πραγματοποιεί ευρετηρίαση εγγράφων με τη βοήθεια ενός θησαυρού όπως παρουσιάστηκε στο [20]. Σε αυτή την προσέγγιση χρησιμοποιούνται τεχνικές ΜΜ και σημασιολογικές πληροφορίες σχετικές με τους όρους που εμπεριέχονται σε ένα ελεγχόμενο λεξιλόγιο. Το βασικότερο πλεονέκτημα της προσέγγισης αυτής έναντι της κλασικής εξαγωγής πληροφορίας, είναι η χρήση ενός ελεγχόμενου λεξιλογίου, το οποίο εξαλείφει την εμφάνιση όρων που δεν έχουν νόημα ή είναι λάθος, και επίσης όπως αναφέρουν οι συγγραφείς παρατήρησαν μια αξιοσημείωτη βελτίωση στην επίδοση του αλγορίθμου KEA. Ένα ακόμα πλεονέκτημα που παρέχει η διαδικασία αυτή έναντι της κλασικής διαδικασίας ανάθεσης όρων, η οποία χρησιμοποιεί ήδη ένα ελεγχόμενο λεξιλόγιο, είναι η απαίτηση για ένα πολύ μικρό πλήθος δεδομένων εκπαίδευσης του υποκείμενου συστήματος. Τέλος, στην εργασία αυτή η διαδικασία αξιολόγησης που εφαρμόστηκε καταδεικνύει ότι η απόδοση του συστήματος είναι ανεξάρτητη από το μέγεθος του ελεγχόμενου λεξιλογίου που χρησιμοποιείται. Παρόλα αυτά, η εργασία αυτή παρουσιάζει τον περιορισμό ότι είναι «εξαρτημένη πεδίου», δεδομένου ότι αν αλλάξει το δεσμευμένο λεξιλόγιο που χρησιμοποιείται τότε ο αλγόριθμος θα πρέπει να περάσει από μια νέα φάση εκπαίδευσης.

Οι δύο παραπάνω προσεγγίσεις βασίζουν ένα τμήμα της υλοποίησής τους, στους αλγορίθμους Porter [29] και Lovins [30], οι οποίοι είναι οι δύο από τους πιο δημοφιλείς αλγορίθμους για την αποκοπή καταλήξεων αγγλικών λέξεων. Και οι δύο χρησιμοποιούν ευρετικούς κανόνες για την αφαίρεση ή τον μετασχηματισμό των καταλήξεων ενός κειμένου. Σε αυτό το σημείο υπάρχει και η εναλλακτική προσέγγιση της χρήσης ενός λεξιλογίου που θα απαριθμεί ρητά το λήμμα για κάθε λέξη που θα μπορούσε να προκύψει σε ένα δεδομένο κείμενο. Παρόλα αυτά, προτιμάται η χρήση των ευρετικών κανόνων έναντι του λεξιλογίου, λόγω του κόπου που απαιτείται για την κατασκευή και τη διατήρηση ενός τέτοιου λεξιλογίου αλλά και της υπολογιστικής πολυπλοκότητας που απαιτείται για τη χρήση του. Ο αλγόριθμος του Lovins είναι περισσότερο επιθετικός από αυτόν του Porter στη διαδικασία αποκοπής. Επίσης, είναι περισσότερο πιθανό να αντιστοιχίσει δύο λέξεις που έχουν την ίδια ρίζα, αλλά έχει και μεγαλύτερη ανοχή σε λάθη [6]. Τέλος, έχει παρατηρηθεί πως η επιθετική αφαίρεση όπως αυτή που εφαρμόζει ο αλγόριθμος του Lovins είναι καλύτερη για την εξαγωγή πληροφορίας από κείμενα σε αντίθεση με τη συντηρητική αφαίρεση του Porter. Πιο πρόσφατα έχει αναπτυχθεί μια βιβλιοθήκη που ενσωματώνει μια νεότερη βελτιωμένη έκδοση του αλγορίθμου του Porter αλλά και ένα σύνολο άλλων αλγορίθμων αποκοπής για διαφορετικές γλώσσες. Η βιβλιοθήκη

ονομάζεται Snowball[5], και έχει επίσης δημιουργηθεί από τον Porter αλλά από το 2014 και έπειτα διατηρείται και αναπτύσσεται περαιτέρω μέσω της κοινότητας.

Μια ακόμα προσέγγιση που ανήκει στην κατηγορία των μεικτών προσεγγίσεων παρουσιάζεται στο [31], όπου χρησιμοποιείται μια μέθοδος μη-εποπτευόμενης ΜΜ για την αυτόματη εξαγωγή αντιπροσωπευτικών όρων από κείμενα με απώτερο στόχο την κατασκευή λεξικών για ένα συγκεκριμένο πεδίο. Τα λεξικά αυτά θα μπορούσαν στη συνέχεια να χρησιμοποιηθούν σε εφαρμογές επεξεργασίας φυσικής γλώσσας. Βασικό χαρακτηριστικό της συγκεκριμένης προσέγγισης αποτελεί η λεγόμενη «σχετικότητα του πεδίου» (domain-specificity). Η έννοια αυτή βασίζεται σε στατιστική επεξεργασία της χρήσης των λέξεων σε ένα κείμενο και υιοθετεί τις ιδιότητες TF και IDF που αναφέρθηκαν σε προηγούμενη παράγραφο του παρόντος, οποίες εφαρμόζονται σε συγκεκριμένο πεδίο, για να είναι δυνατή η εφαρμογή της έννοιας της σχετικότητας του πεδίου που αναφέρθηκε παραπάνω. Όπως είναι φανερό και αυτή η προσέγγιση χρησιμοποιεί τόσο στατιστικές όσο και ευρετικές μεθόδους για τη εξαγωγή των όρων. Ο λόγος για τον οποίο κατατάσσεται στις μεικτές προσεγγίσεις είναι διότι η αξιολόγηση της συγκεκριμένης μεθόδου έγινε για δύο συγκεκριμένες εργασίες: α) κατηγοριοποίηση κειμένου και β) εξαγωγή πληροφορίας. Σύμφωνα με τους συγγραφείς η επιλογή των συγκεκριμένων εργασιών μόνο τυχαία δεν ήταν, αλλά αποσκοπούσε στο να επιδείξει ότι οι φράσεις ή οι όροι που σχετίζονται με ένα συγκεκριμένο πεδίο είναι σε θέση να παρέχουν καλύτερη αναπαράσταση του θέματος του κειμένου σε σχέση με τους περισσότερο γενικούς όρους. Έτσι, η εργασία της κατηγοριοποίησης ενός κειμένου αξιολογήθηκε με την κατασκευή ενός μοντέλου ταξινόμησης με χρήση τριών διαφορετικών εργαλείων: ενός POS tagger, η λειτουργία του οποίου βασίζεται στις πιθανότητες, εύρεση του λήμματος των όρων με χρήση του λογισμικού morpha και τέλος με την κατασκευή ενός ταξινομητή. Η δεύτερη εργασία, αυτή της εξαγωγής πληροφορίας, πραγματοποιήθηκε με χρήση των δύο λογισμικών που χρησιμοποιήθηκαν και στην προηγούμενη περίπτωση και τέλος με την εφαρμογή της μεθόδου επιλογής υποψήφιων όρων από ένα κείμενο που παρουσιάστηκε στο [16] αλλά και τα χαρακτηριστικά: α) TF-IDF και β) την απόσταση ενός όρου σε ένα δεδομένο κείμενο για τη μοντελοποίηση της τοπικότητας των φράσεων, όπως αυτή παρουσιάστηκε στο [13] και στο [14].

Οι τεχνικές που χρησιμοποιούνται με επιτυχία στα [4], [22] έχει γίνει προσπάθεια να συνδυαστούν στο [24]. Εκεί παρουσιάζεται το σύστημα Enlil, το οποίο ενσωματώνει την πιθανοτική μέθοδο CRF. Η CRF χρησιμοποιείται για να εξάγει από το κείμενο τους συγγραφείς και τα ιδρύματα από τα οποία προέρχονται, ενώ η τεχνική SVM χρησιμοποιείται στη συνέχεια για την αντιστοίχιση μεταξύ των συγγραφέων και των ιδρυμάτων στα οποία οι τελευταίοι ανήκουν. Η μέθοδος αυτή ανήκει στην κατηγορία των μεικτών προσεγγίσεων, καθώς το πρώτο βήμα της μοντελοποιείται ως μια διαδικασία επισήμανσης ακολουθιών (sequence labelling), η οποία χρησιμοποιεί ένα μοντέλο εποπτευόμενης ΜΜ (CRF) για την εξαγωγή των συγγραφέων και των ιδρυμάτων. Στη συνέχεια πραγματοποιείται η διαδικασία σχεσιακής ταξινόμησης (relational classification) με χρήση της τεχνικής SVM.

Παρακάτω, γίνεται μια ταξινόμηση των προσεγγίσεων ανάλογα με τα κριτήρια που συγκεντρώνει η κάθε μια, ανεξάρτητα από την κατηγορία στην οποία ανήκει (Πίνακας 4).

[5] https://snowballstem.org/algorithms/

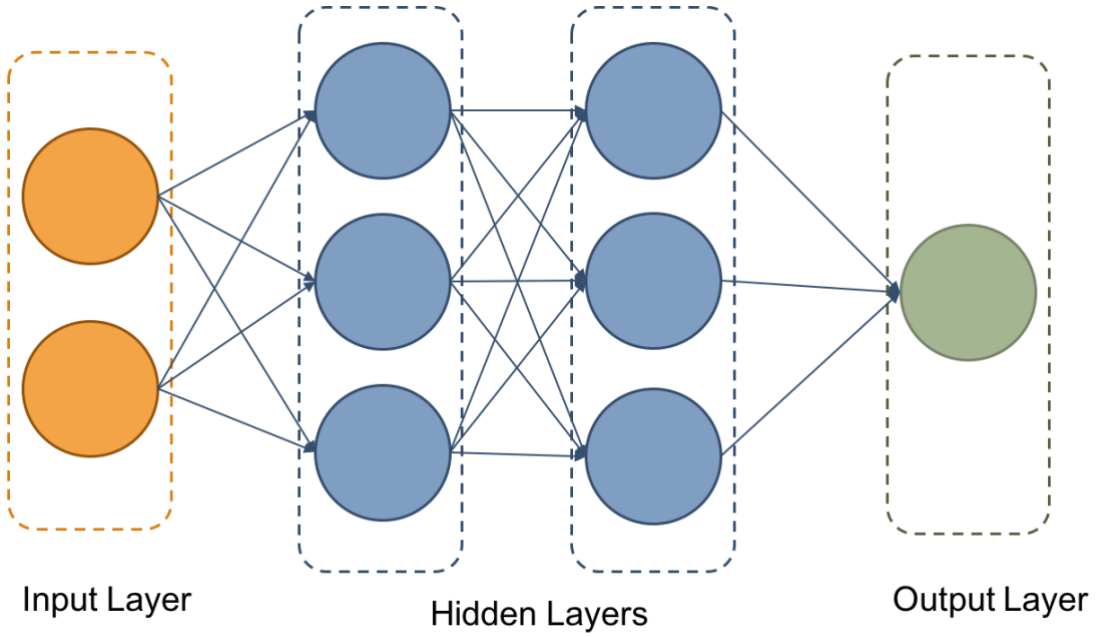
Πίνακας 4. Συγκριτικός πίνακας ταξινόμησης προσεγγίσεων ως προς τα χαρακτηριστικά που χρησιμοποιούν (E = Extraction, As = Assignment, S= Statistical, H= Heuristics, ML-S= Machine Learning – Supervised, A = Automatic, S-A=Semi-Automatic, G=General, D-S=Domain-Specific, T= Thesaurus)

Εργασία	E	As	S	H	ML- S	ML- U	NLP	A	S-A	G	D-S	T
Frank και συνεργάτες (KEA) [14]	Όχι	Ναι	Ναι	Ναι	Ναι	Όχι	Ναι	Ναι	Όχι	Ναι	Ναι	Όχι
Medelyan & Witten (KEA++) [26], [20]	Ναι	Ναι	Ναι	Ναι	Ναι	Ναι	Όχι	Ναι	Όχι	Όχι	Ναι	Ναι
Turney (GenEx) [6]	Όχι	Ναι	Όχι	Ναι	Ναι	Όχι	Ναι	Ναι	Όχι	Ναι[6]	Ναι[7]	Όχι
Turney [18]	Όχι	Ναι	Ναι	Όχι	Ναι	Όχι	Όχι	Ναι	Όχι	Ναι	Όχι	Όχι
Park και συνεργάτες [15]	Ναι	Όχι	Ναι	Ναι	Όχι	Όχι	Ναι	Ναι	Όχι	Όχι	Ναι	Όχι
Wan & Xiao [27]	Ναι	Όχι	Ναι	Όχι	Ναι	Όχι	Όχι	Ναι	Όχι	Ναι	Όχι	Όχι
Barker & Corrnachia [7]	Ναι	Όχι	Ναι	Ναι	Ναι[8]	Όχι	Ναι	Ναι	Ναι	Ναι	Όχι	Ναι[9]
VRL 2009 [31], [28]	Ναι	Ναι	Όχι	Ναι[10]	Όχι	Ναι	Ναι	Ναι	Όχι	Όχι	Ναι	Όχι
Nguyen & Kan [16]	Ναι	Όχι	Ναι	Ναι	Ναι	Ναι	Ναι	Ναι	Όχι	Όχι	Ναι	Όχι
Sugiyama και συνεργάτες [21]	Ναι	Όχι	Όχι	Ναι	Ναι	Όχι	Ναι	Ναι	Όχι	Όχι	Ναι[11]	Όχι
Hulth & Megyesi [19]	Ναι	Όχι	Ναι	Όχι	Ναι	Όχι	Ναι	Όχι	Ναι	Ναι	Όχι	Όχι
Han και συνεργάτες [4]	Όχι	Ναι	Όχι	Ναι	Ναι	Όχι	Ναι	Ναι	Όχι	Ναι	Ναι	Ναι
Luong και συνεργάτες [22]	Όχι	Ναι	Ναι	Όχι	Ναι	Όχι	Ναι	Ναι	Όχι	Ναι	Όχι	Όχι
Do και συνεργάτες [24]	Όχι	Ναι	Ναι[12]	Όχι	Ναι[13]	Όχι	Όχι	Ναι	Όχι	Ναι	Όχι	Όχι
Lin και συνεργάτες [25]	Ναι	Όχι	Ναι	Όχι	Όχι	Όχι	Ναι	Όχι	Ναι	Όχι	Ναι[14]	Όχι

[6] Με χρήση του αλγορίθμου C4.5
[7] Με χρήση του αλγορίθμου GenEx
[8] Χρησιμοποιεί τον Extractor του Turney [6]
[9] Χρησιμοποιεί ένα λεξιλόγιο διαθέσιμο στον Ιστό
[10] Επιλέγονται οι φράσεις με την υψηλότερη βαθμολογία με βάση ένα κατώφλι. Το κατώφλι αυτό υπολογίζεται με ευρετικές μεθόδους.
[11] Σε αυτή τη μέθοδο χρησιμοποιείται συγκεκριμένο σύνολο δεδομένων: ACL Anthology Reference Corpus (ACL-ARC). Αποτελείται από δημοσιεύσεις σε περιοδικά και πρακτικά συνεδρίων στα πεδία της Ανάλυσης Φυσικής Γλώσσας και της Υπολογιστικής Γλωσσολογίας.
[12] Λόγω της χρήσης CRF
[13] Με χρήση SVM
[14] Χρησιμοποιεί άρθρα κλινικών μελετών

2.4.4. Προσεγγίσεις πρόβλεψης

Στη συγκεκριμένη ενότητα θα γίνει μια προσπάθεια κατηγοριοποίησης των προσεγγίσεων που χρησιμοποιούν μοντέλα πρόβλεψης με σκοπό να εκτελέσουν με επιτυχία διάφορες εργασίες ανάλυσης φυσικής γλώσσας. Οι προσεγγίσεις αυτές συνήθως βασίζονται σε τεχνικές Βαθιάς Μάθησης (Deep Learning). Η Βαθιά Μάθηση αποτελεί ένα προηγμένο υποσύνολο της ΜΜ. Από τεχνικής απόψεως είναι ένας ευρέως χρησιμοποιούμενος όρος για ένα σύνολο Νευρωνικών Δικτύων που αποτελούνται από τρία ή περισσότερα επίπεδα. Πιο συγκεκριμένα, έχει τουλάχιστον ένα επίπεδο εισόδου ένα επίπεδο εξόδου και ένα ενδιάμεσο «κρυφό» επίπεδο. Μια αδρή αναπαράσταση τέτοιου τύπου δικτύων δίνεται στην Εικόνα 3. Βασικό χαρακτηριστικό αυτής της κατηγορίας Νευρωνικών Δικτυών είναι ότι μπορούν να αναγνωρίζουν πρότυπα (patterns) και να τα ταξινομούν παίρνοντας ως είσοδο ένα σύνολο από διανύσματα. Μέσα στις βασικές τους εφαρμογές είναι ότι έχουν τη δυνατότητα να αναγνωρίζουν διαφορετικά είδη πληροφοριών όπως φωνή, εικόνες ή πρόσωπα, παρόμοια κείμενα κλπ. Πιο συγκεκριμένα, ένα τέτοιο νευρωνικό δίκτυο αν λάβει ως είσοδο ένα σύνολο από δεδομένα χρονοσειρών, που μπορεί να αναπαριστούν ήχο, εικόνα ή κείμενο, μπορεί να αναλύσει ένα τμήμα της δεδομένης ακολουθίας εισόδου και να συμπληρώσει το τμήμα που υπολείπεται. Για παράδειγμα, εάν λάβει ως είσοδο το περιεχόμενο γύρω από μια λέξη μπορεί να προβλέψει με ακρίβεια την ίδια τη λέξη.



Εικόνα 3. Αναπαράσταση αρχιτεκτονικής Νευρωνικών Δικτύων Βαθιάς Μάθησης.

Ένα από τα βασικά πλεονεκτήματα αυτής της κατηγορίας νευρωνικών δικτύων και ένα χαρακτηριστικό που τα κάνει κατάλληλη επιλογή για την παρούσα Διατριβή είναι ότι μπορούν να χρησιμοποιηθούν αποτελεσματικά στην επεξεργασία μη-δομημένων δεδομένων. Επίσης, δεν χρειάζεται να έχει επισημειωθεί το σύνολο των δεδομένων για να μπορέσει ο αλγόριθμος να ανακαλύψει πρότυπα. Για το λόγο αυτό, αυτού του είδους τα δίκτυα μπορούν να εφαρμοστούν σε ένα πλήθος διαφορετικών εφαρμογών όπως για παράδειγμα, ομιλία που πρέπει να αντιστοιχηθεί με κείμενο, ανάλυση εικόνας και βίντεο, κείμενα που πρέπει να ομαδοποιηθούν με βάση τα συμφραζόμενα (context) ή το συναίσθημα (sentiment). Σύμφωνα με τις μέχρι τώρα μελέτες φαίνεται ότι αποδίδουν καλύτερα σε προβλήματα που αφορούν ομιλία, όραση, γλώσσα κλπ., ενώ είναι μια

κατηγορία Νευρωνικών Δικτυών που μπορεί να προσαρμοστεί σχετικά εύκολα σε νέα πεδία εφαρμογής.

Ακόμα, τα συγκεκριμένα δίκτυα έχουν τη δυνατότητα να επεκτείνουν τα υπάρχοντα ή να «δημιουργούν» νέα χαρακτηριστικά από ένα περιορισμένο σύνολο χαρακτηριστικών που ενυπάρχουν στα δεδομένα εκπαίδευσης, καθώς ο αλγόριθμος τείνει να αναζητεί νέα χαρακτηριστικά που συσχετίζονται με αυτά που είναι ήδη γνωστά. Η συγκεκριμένη δυνατότητα αυτού του τύπου δικτύων τους δίνει το πλεονέκτημα να μπορούν να μειώνουν την ανάγκη για εκτεταμένη μηχανική χαρακτηριστικών, που όπως αναφέρθηκε και προηγουμένως αποτελεί ένα από τα πιο χρονοβόρα βήματα των μεθόδων ΜΜ καθώς απαιτεί ειδικούς στο είδος για την κατασκευή των χαρακτηριστικών. Επίσης, συνήθως τα χαρακτηριστικά αυτά εξαρτώνται σε μεγάλο βαθμό από το υποκείμενο σύνολο δεδομένων κάτι που επιβάλλει μια νέα φάση κατασκευής χαρακτηριστικών εάν το σύνολο των υπό εξέταση δεδομένων μεταβληθεί.

Βέβαια, παρόλα τα εγγενή πλεονεκτήματα που αναφέρθηκαν παραπάνω, ο συγκεκριμένος τύπος Νευρωνικών Δικτύων έχει και κάποια μειονεκτήματα. Αρχικά, δεδομένου ότι δεν υπάρχει κάποια καθοδήγηση στον αλγόριθμο προκειμένου να είναι σε θέση να ανακαλύψει πρότυπα, απαιτεί ένα αρκετά μεγάλο σύνολο δεδομένων πάνω στο οποίο θα εφαρμοστεί. Ακόμα, λόγω της αρχιτεκτονικής του αλλά και του μεγάλου συνόλου δεδομένων που καλείται να επεξεργαστεί η εκπαίδευση του συστήματος είναι υπολογιστικά ακριβή.

Επιπροσθέτως, δεν βασίζεται σε ισχυρή θεωρητική θεμελίωση με αποτέλεσμα να μην είναι δυνατόν να τεκμηριωθεί γιατί το σύστημα κατέληξε σε ένα συγκεκριμένο συμπέρασμα καθώς τα εσωτερικά «κρυφά» επίπεδα δημιουργούν τα δικά τους χαρακτηριστικά επεκτείνοντας τα υπάρχοντα. Αντίθετα, στις παραδοσιακές προσεγγίσεις ΜΜ τα χαρακτηριστικά κατασκευάζονται, από τη σύλληψή τους μέχρι και την υλοποίηση τους, από τους ειδικούς στο εκάστοτε πεδίο οι οποίοι είναι σε θέση να γνωρίζουν πόσο ακριβώς θα συνεισφέρει το κάθε χαρακτηριστικό στην τελική απόφαση του αλγορίθμου. Όπως αναφέρθηκε και παραπάνω, λόγω της μη ισχυρής θεωρητικής θεμελίωσης τους είναι δύσκολο να γίνει κατανοητό το τι ακριβώς έχει μάθει το σύστημα σε αντίθεση με κάποιον άλλο παραδοσιακό ταξινομητή όπως τα δέντρα απόφασης, ή για παράδειγμα τη μέθοδο Naïve Bayes που είναι πολύ πιο εύκολο να γίνει κατανοητός ο τρόπος λειτουργίας τους.

Όπως έγινε φανερό και από τα παραπάνω, τα πλεονεκτήματα της τεχνικής αυτής είναι αδιαμφισβήτητα. Για να είναι δυνατή η εφαρμογή τους σε προβλήματα Ανάλυσης Φυσικής Γλώσσας βασική απαίτηση είναι η αναπαράσταση των λέξεων με χρήση διανυσμάτων, έτσι ώστε να μπορούν να χρησιμοποιηθούν μετρικές ομοιότητας ή απόστασης μεταξύ των λέξεων, όπως του συνημίτονου. Η χρήση των διανυσμάτων λέξεων έχει καθιερωθεί, καθώς αυτού του είδους τα διανύσματα έχουν την ικανότητα να ενσωματώνουν την παραπάνω ιδιότητα εγγενώς. Γενικά, υπάρχουν δύο τρόποι /μοντέλα για τη δημιουργία αυτών των διανυσμάτων: α. Μοντέλα Μέτρησης (Distributional Semantics ή Count Models) και β. Μοντέλα Πρόβλεψης (Word Embeddings ή Predictive Models), τα οποία βασίζονται στην ίδια υπόθεση “The Distributional Hypothesis” [32], [33]. Η υπόθεση αυτή αναφέρει ότι “Παρόμοιες λέξεις έχουν την τάση να εμφανίζονται σε παρόμοια περιεχόμενα”. Πέρα από αυτή την υπόθεση τα δύο μοντέλα έχουν το ίδιο γλωσσολογικό υπόβαθρο, χρησιμοποιούν τα ίδια δεδομένα και έχουν μεταξύ τους μια ισχυρή μαθηματική συσχέτιση [34].

Η πρώτη κατηγορία αναφέρεται στα παραδοσιακά μοντέλα μέτρησης όπως αυτά αναλύθηκαν σε προηγούμενη παράγραφο του παρόντος. Τα μοντέλα αυτά συνήθως χρησιμοποιούν κάποιο είδος πίνακα (matrix) όπως τον word-context PMI. Επειδή αυτοί οι πίνακες είναι «αραιοί» (sparce) και

αρκετά μεγάλων διαστάσεων, για την αποδοτική επεξεργασία τους χρησιμοποιούνται τεχνικές μείωσης των διαστάσεων (dimensionality reduction) όπως Singular Value Decomposition (SVD).

Η δεύτερη κατηγορία που από πολλούς θεωρείται πλέον «τεχνολογία αιχμής», βασίζεται στα Νευρωνικά Δίκτυα Βαθιάς Μάθησης (Deep Learning Neural Networks). Η βασική ιδέα πίσω από αυτή την κατηγορία μοντέλων είναι ότι τα τμήματα των λέξεων ενός κειμένου κωδικοποιούνται σε ένα διάνυσμα που αναπαριστά ένα σημείο σε κάποιον «χώρο» λέξεων (word space). Έτσι, υπάρχει ένας χώρος N-διαστάσεων επαρκής για την κωδικοποίηση όλης της σημασιολογίας μιας γλώσσας. Ο μόνος περιορισμός που υπάρχει σχετικά με τις διαστάσεις αυτού του χώρου είναι ότι πρέπει να είναι μικρότερος από το σύνολο των λέξεων που υπάρχουν σε μια δεδομένη γλώσσα. Κάθε διάσταση αυτού του χώρου μπορεί να κωδικοποιεί κάποιο είδος νοήματος που μεταφέρεται με την ομιλία. Οι διαστάσεις αυτού του χώρου μπορεί να κωδικοποιούν χρόνο (παρελθοντικό ή μελλοντικό), αριθμό (ενικό ή πληθυντικό), γένος (αρσενικό, θηλυκό) κλπ. Ακόμα, πλεονεκτήματα της χρήσης τέτοιων διανυσμάτων για την κωδικοποίηση των λέξεων είναι ότι απαιτούν λιγότερη προ-επεξεργασία και ότι δεν απαιτούν τη χρήση ευρετικών αλγορίθμων για την εξαγωγή πληροφορίας από κείμενα. Ως εκ τούτου για την κατασκευή ενός τέτοιου συστήματος δεν χρειάζεται μεγάλο υπόβαθρο γνώσεων γλωσσολογίας. Τέλος, οι κατανεμημένες αναπαραστάσεις των λέξεων σε ένα χώρο διανυσμάτων βοηθούν τους αλγόριθμους μάθησης να επιτύχουν καλύτερη απόδοση σε εργασίες Ανάλυσης Φυσικής Γλώσσας με την ομαδοποίηση παρόμοιων λέξεων.

Σε αυτό το σημείο θα πρέπει να τονιστεί ότι τα αναφερόμενα στις επόμενες παραγράφους δεν εστιάζουν αμιγώς στην εξαγωγή πληροφορίας αλλά ουσιαστικά θα γίνει μια προσπάθεια να αναφερθούν σημαντικές ερευνητικές εργασίες που χρησιμοποιούν Νευρωνικά Δίκτυα Βαθιάς Μάθησης στον τομέα της Ανάλυσης Φυσικής Γλώσσας (NLP).

Η πρώτη προσέγγιση της κατηγορίας αυτής παρουσιάζεται στο [35] όπου και προτείνεται μια ενοποιημένη αρχιτεκτονική που βασίζεται σε νευρωνικά δίκτυα και ο αντίστοιχος αλγόριθμος μάθησης. Σκοπός της συγκεκριμένης προσπάθειας είναι ο σχεδιασμός μιας ευέλικτης αρχιτεκτονικής που να έχει τη δυνατότητα να μαθαίνει χρήσιμες αναπαραστάσεις ενώ ταυτόχρονα αποφεύγει την υπερβολικά εκτεταμένη μηχανική χαρακτηριστικών. Σύμφωνα με τους συγγραφείς το σύστημα αυτό μπορεί να χρησιμοποιηθεί ως βάση για την κατασκευή ενός πολύ αποδοτικού συστήματος επισημείωσης το οποίο θα απαιτεί τους λιγότερους δυνατούς υπολογιστικούς πόρους. Στη συγκεκριμένη προσέγγιση η μεγαλύτερη βελτίωση στην απόδοση του συστήματος παρατηρείται με τη μεταφορά των μη-εποπτευόμενων εσωτερικών αναπαραστάσεων των διανυσμάτων σε μοντέλα αναφοράς που χρησιμοποιούν τεχνικές εποπτευόμενης ΜΜ. Η παραπάνω τεχνική εφαρμόζεται σε διαφορετικές εργασίες όπως επισημείωση τμημάτων λόγου (POS tagging), τμηματοποίηση κειμένων (chunking), αναγνώριση οντοτήτων με βάση το όνομα (named-entity recognition) και επισημείωση σημασιολογικών ρόλων (semantic role labelling).

Η εργασία που παρουσιάζεται στο [36] αφορά στην πράξη την υλοποίηση του αλγορίθμου Skip-Gram with Negative Sampling (SGNS), ο οποίος αποτελεί σύμφωνα με τους συγγραφείς μια αποδοτική μέθοδο μάθησης η οποία δίνει ως έξοδο υψηλής ποιότητας κατανεμημένες αναπαραστάσεις διανυσμάτων. Οι συγκεκριμένες αναπαραστάσεις έχουν τη δυνατότητα να ενσωματώνουν ένα μεγάλο αριθμό από αναπαραστάσεις λέξεων οι οποίες είναι ακριβείς τόσο συντακτικά όσο και σημασιολογικά. Επίσης, παρουσιάζουν ένα σύνολο από επεκτάσεις που βελτιώνουν την ποιότητα των διανυσμάτων αλλά και την ταχύτητα της φάσης εκπαίδευσης από δύο ως δέκα φορές. Μερικές από αυτές τις βελτιώσεις αφορούν: α) υπό-δειγματοληψία των πιο συχνά

εμφανιζόμενων λέξεων (sub-sampling of frequent words), β) διαφορετική προσεγγιστική συνάρτηση, αντί για "hierarchical softmax" χρησιμοποιείται "negative sampling", κ.α. Ακόμα, παρουσιάζουν μια μέθοδο για την εύρεση σύνθετων όρων μέσα σε ένα κείμενο, όπου προτείνουν τη χρήση διανυσμάτων φράσεων αντί για διανύσματα λέξεων. Θεωρούν πως η χρήση διανυσμάτων φράσεων είναι ο τρόπος για να ξεπεράσουν έναν εγγενή περιορισμό των αναπαραστάσεων λέξεων, καθώς οι τελευταίες δεν μπορούν να απεικονίσουν τη σειρά των λέξεων και άρα δεν έχουν τη δυνατότητα να αναπαραστήσουν ιδιωματικές εκφράσεις. Η προσέγγιση ακόμα ενσωματώνει και μια φάση προ-επεξεργασίας, όπου όλες οι λέξεις που εμφανίζονται λιγότερο από πέντε (5) φορές στα δεδομένα εκπαίδευσης δεν χρησιμοποιούνται. Βασικό κομμάτι τόσο της συγκεκριμένης προσέγγισης αλλά και των υπολοίπων αυτής της κατηγορίας, είναι η επιλογή των υπέρ-παραμέτρων (hyperparameters) του δικτύου, επιλογή η οποία καθορίζεται σε ένα βαθμό από την υποκείμενη εργασία. Τέλος, σύμφωνα με τα αποτελέσματα που παρουσιάστηκαν, υπάρχουν συγκεκριμένες επιλογές που επηρεάζουν την απόδοση ενός τέτοιου συστήματος και αφορούν την επιλογή της αρχιτεκτονικής για την κατασκευή του υποκείμενου μοντέλου, το μέγεθος των διανυσμάτων των λέξεων/ φράσεων, το ρυθμό υπό-δειγματοληψίας αλλά και το μέγεθος του παραθύρου που χρησιμοποιείται για τον καθορισμό των συμφραζομένων (context) κατά τη φάση της εκπαίδευσης.

Αντίστοιχα μοντέλα πρόβλεψης με βάση τα συμφραζόμενα (context-predicting models) αλλά και μια αρκετά εκτεταμένη σύγκριση με τα μοντέλα μέτρησης (context-counting) παρουσιάζονται στο [37]. Τα μοντέλα πρόβλεψης με βάση τα συμφραζόμενα έχουν σχετικά πρόσφατα ενσωματωθεί ερευνητικά στον τομέα της αναδιανεμητικής σημασιολογίας (distributional semantics) τα οποία ονομάζονται νευρωνικά γλωσσικά μοντέλα (neural language models ή embeddings). Σε αυτή την περίπτωση οι πληροφορίες που αντλούνται από τα συμφραζόμενα παρέχουν μια σχετικά καλή προσέγγιση της έννοιας των λέξεων καθώς σημασιολογικά παρόμοιες λέξεις τείνουν να έχουν παρόμοιες κατανομές περιεχομένου [32]. Το σύνολο δεδομένων που χρησιμοποιείται σε αυτή την εργασία αποτελείται από 2,8 εκατομμύρια τμήματα (tokens) που έχουν εξαχθεί από διαφορετικές πηγές (ukWaC, English Wikipedia, British National Corpus). Για την κατασκευή του μοντέλου μέτρησης χρησιμοποιούνται συμμετρικά παράθυρα μήκους 2-5 λέξεων από κάθε πλευρά της λέξης-στόχου (target), με 2 διαφορετικά σχήματα ανάθεσης βαρών το "Pointwise Mutual Information - PMI" και το "Localized Mutual Information - LMI". Αναφορικά με την κατασκευή των μοντέλων πρόβλεψης, και εδώ χρησιμοποιούνται παράθυρα μήκους 2-5 λέξεων από κάθε πλευρά του κεντρικού αντικειμένου (central element), τα οποία στη συνέχεια εκπαιδεύονται χρησιμοποιώντας το σύνολο εργαλείων Word2Vec και πιο συγκεκριμένα τον αλγόριθμο CBOW ο οποίος είναι υπολογιστικά αποδοτικότερος για μοντέλα που αφορούν μεγαλύτερα σύνολα δεδομένων. Όπως και στην προηγούμενη προσέγγιση έτσι και εδώ χρησιμοποιείται ένας ταξινομητής που χρησιμοποιεί την συνάρτηση "hierarchical softmax", ο οποίος είναι ένας υπολογιστικά αποδοτικός τρόπος για να εκτιμηθεί η συνολική κατανομή πιθανότητας χρησιμοποιώντας ένα επίπεδο εξόδου (output layer) που να είναι ανάλογο με τον λογάριθμο της πολυπλοκότητας τη λέξης[15] αντί για την ίδια τη λέξη (W). Στη συνέχεια πραγματοποιήθηκε αξιολόγηση και των δύο μοντέλων σε διαφορετικές εργασίες. Οι εργασίες αυτές αφορούν σημασιολογική ομοιότητα (Semantic Relatedness), αναγνώριση συνώνυμων λέξεων (Synonym detection), κατηγοριοποίηση εννοιών (Concept categorization), εργασίες αναλογιών (Analogy tasks). Σύμφωνα με τα αποτελέσματα που παρουσιάζουν, οι συγγραφείς καταλήγουν ότι τα αποτελέσματα των μοντέλων πρόβλεψης είναι τόσο καλά που υπάρχουν αρκετοί λόγοι να χρησιμοποιήσει κανείς τη νέα αυτή αρχιτεκτονική.

[15] log(unigram: perplexity(W))

2.5. Εφαρμογές

Όπως έγινε φανερό και από την ταξινόμηση των προσεγγίσεων σε κατηγορίες, όπως αυτές παρουσιάστηκαν παραπάνω, οι εφαρμογές τους είναι ανεξάντλητες. Σε αυτή την ενότητα της παρούσας εργασίας θα γίνει μια αναφορά στις βασικές εφαρμογές στις οποίες η εξαγωγή πληροφορίας παίζει κυρίαρχο ρόλο.

Με την επέκταση και ευρύτατη χρήση του Διαδικτύου και των εταιρικών ενδοδικτύων (intranet), η διαχείριση των εγγράφων αποκτά πολύ μεγάλη σημασία. Πολλοί ερευνητές θεωρούν πως η παροχή ποιοτικών μετά-δεδομένων αποτελεί βασικό παράγοντα επιτυχίας στη διαχείριση των εγγράφων. Τα μετά-δεδομένα αποτελούν μετά-πληροφορίες που αφορούν ένα έγγραφο ή ένα σύνολο από έγγραφα. Ειδικότερα στον βιβλιογραφικό τομέα υπάρχουν διάφορα πρότυπα που αφορούν τα μετά-δεδομένα αυτά, όπως το Dublin Core Metadata Element Set, ή η μορφοποίηση MARC (Machine-Readable Cataloguing). Τα πρότυπα αυτά έχουν προδιαγράψει ένα πεδίο για την αποθήκευση όρων που αφορούν το υποκείμενο τεκμήριο. Κάθε έγγραφο που κατατίθεται σε μια Ψηφιακή Βιβλιοθήκη είναι απαραίτητο να συνοδεύεται και από ένα σύνολο μετά-δεδομένων. Επειδή αυτή η διαδικασία βαραίνει συνήθως είτε τον συγγραφέα είτε τον βιβλιοθηκονόμο, θα ήταν πολύ βοηθητικό εάν υπήρχε μια αυτόματη διαδικασία παραγωγής μετά-δεδομένων που αφορούν το τεκμήριο.

Οι βιβλιοθηκονόμοι όταν καλούνται να εντοπίσουν τα βασικά στοιχεία ενός εγγράφου, ψάχνουν μέσα σε αυτό για να εντοπίσουν όρους που να μπορούν να αντιπροσωπεύσουν το περιεχόμενο. Αυτή η διαδικασία επιτυγχάνεται επισημαίνοντας τμήματα κειμένου με χρήση κάποιας ειδικής γραμματοσειράς ή αλλάζοντας το χρώμα υποβάθρου του κειμένου. Υπάρχουν κάποια προγραμματιστικά εργαλεία που υποβοηθούν τη διαδικασία αυτή επισημαίνοντας αυτόματα αυτά τα σημεία του κειμένου. Στη συνέχεια χρησιμοποιώντας τα τμήματα αυτά είναι δυνατόν να παραχθεί μια περίληψη του κειμένου αυτού που να βασίζεται στα τμήματα που έχουν επισημανθεί με την διαδικασία που αναφέρθηκε προηγουμένως. Λόγω του ότι η διαδικασία αυτή εμπεριέχει μεγάλο βαθμό εξάρτησης από τον ανθρώπινο παράγοντα θα ήταν πολύ αποδοτικό εάν υπήρχε ένας αυτόματος τρόπος παραγωγής των όρων αυτών που θα μπορούσε να υποβοηθήσει την επιλογή από τους ειδικούς του εκάστοτε πεδίου.

Ακόμα, στην πλειοψηφία τους οι βιβλιογραφικές βάσεις δεδομένων περιέχουν ένα μεγάλο αριθμό μετά-δεδομένων που προέρχονται από δημοσιεύσεις κάνοντας τες με αυτό τον τρόπο εύκολο να ανακτηθούν στα πλαίσια μιας αναζήτησης. Βασική χρήση αυτών των αποθετηρίων αποτελεί η αξιολόγηση του αντίκτυπου (impact) της έρευνας συγκεκριμένων ερευνητών στην κοινότητα. Το βασικό πρόβλημα σε αυτή την κατεύθυνση δημιουργείται όταν διαφορετικά άτομα μοιράζονται το ίδιο όνομα. Σε αυτή την περίπτωση, δημοσιεύσεις που ανήκουν σε διαφορετικούς ερευνητές μπορεί να μπερδευτούν μεταξύ τους εμποδίζοντας έτσι την επιστημονική συλλογή δεδομένων, την ανάκτηση πληροφοριών αλλά και υποβαθμίσουν την επιρροή ενός ερευνητή [3]. Το σύνολο των παραπομπών ενός ερευνητή χρησιμοποιείται ευρέως ως δείκτης αξιολόγησης της σημαντικότητας και του αντίκτυπου του έργου του. Υπό αυτό το πρίσμα, η αυτόματη εξαγωγή πληροφορίας ταξινόμησης έχει πολλά να προσφέρει εάν προταθεί ένα σύστημα που να μπορεί να πραγματοποιεί αυτή τη διαδικασία με επιτυχία. Το παραπάνω πρόβλημα μπορεί εύκολα να επεκταθεί για να συμπεριλάβει πέρα από τους συγγραφείς μια δημοσίευσης και το/τα ιδρύματα στα οποία ανήκουν όπως παρουσιάστηκε και στην εργασία [24].

Μια ακόμα βασική εφαρμογή των όρων που εξάγονται από ένα κείμενο αποτελεί η χρήση τους για την ευρετηρίαση. Μια αλφαβητική λίστα από φράσεις η οποία έχει προκύψει από μια συλλογή εγγράφων ή από τμήματα ενός μακροσκελούς εγγράφου μπορεί να υποστηρίξει το ρόλο ενός ευρετηρίου. Η λίστα αυτή μπορεί να έχει προκύψει με χρήση ενός αλγορίθμου όπως ο GenEx [6]. Σε αυτή την περίπτωση μπορεί ένας χρήστης να εισάγει μια λέξη ή ένα τμήμα μιας λέξης, ενώ σε επόμενο βήμα το σύστημα μπορεί να παράγει με τη μορφή λίστας, όρους, σύνθετους ή μη, οι οποίοι έχουν παραχθεί με έναν αυτόματο τρόπο, κάθε ένας από τους οποίους θα περιλαμβάνει τη λέξη που εισήγαγε ο χρήστης στην αρχή. Επιλέγοντας τον κατάλληλο όρο το σύστημα επιστρέφει στον χρήστη την λίστα με όλα τα άρθρα που είναι σχετικά με αυτόν.

Πέρα από την ευρετηρίαση, πολύ βασική είναι η χρήση των τεχνικών εξαγωγής πληροφορίας και σε διεπαφές και περιβάλλοντα συγγραφής στον Ιστό (web authoring environments) [21]. Σε τέτοια περιβάλλοντα οι τεχνικές εξαγωγής μπορεί να υποβοηθήσουν ερευνητές/συγγραφείς να αναλύσουν αλλά και να επιλέξουν παραπομπές από διάφορες ερευνητικές εργασίες. Μια ακόμα εφαρμογή θα μπορούσε να είναι η αναγνώριση των σημείων που χρήζουν αναφοράς μέσα σε ένα κείμενο. Τέλος, τέτοιου είδους συστήματα μπορούν να χρησιμοποιηθούν αποδοτικά και για την ανακάλυψη περιπτώσεων λογοκλοπής, επισημαίνοντας τμήματα του εγγράφου που θα έπρεπε να έχουν αναφορά και στα οποία αυτή έχει παραλειφθεί.

Μια ακόμα εφαρμογή των τεχνικών αυτών που σχετίζεται με τα περιβάλλοντα συγγραφής στον Ιστό, αφορά ένα σύστημα παροχής προτάσεων μαθησιακών αντικειμένων με σκοπό την επαναχρησιμοποίηση τους [38]. Οι προτάσεις ενός τέτοιου συστήματος θα μπορούσαν να βασίζονται στην αυτόματη αναγνώριση του υποκείμενου περιεχομένου. Ο στόχος της έρευνας σε αυτό το πεδίο θα ήταν η ανάλυση των δυνατοτήτων επαναχρησιμοποίησης των μαθησιακών αντικειμένων αλλά και η υποστήριξη της ανακάλυψης και της επαναχρησιμοποίησης εντός των διαθέσιμων συγγραφικών εργαλείων. Οι προτάσεις αυτές, για να είναι αξιοποιήσιμες από τους συγγραφείς θα πρέπει να παρέχονται την ίδια στιγμή που το περιεχόμενο επεξεργάζεται από τον συγγραφέα.

Εκτός από τον τομέα των βιβλιοθηκών, η εφαρμογή των τεχνικών εξαγωγής πληροφορίας μπορεί να έχει πάρα πολλά οφέλη και στα πεδία της βιολογίας και της ιατρικής. Η εφαρμογή τους σε αυτά τα πεδία εκτείνεται πέρα από την εξαγωγή των τυποποιημένων μετά-δεδομένων (τίτλος, συγγραφέας κλπ.) καθώς οι τεχνικές αυτές μπορεί να εφαρμοστούν για την εξαγωγή πληροφορίας από το πλήρες κείμενο που αφορούν κρίσιμες παραμέτρους μιας κλινικής μελέτης, όπως διαμήκεις μεταβλητές (longitudinal variables) και ιατρικές παρεμβατικές μεθόδους [25]. Η εξαγωγή τέτοιου είδους πληροφοριών μπορεί να βοηθήσει τόσο τους αναγνώστες να διεξάγουν σημασιολογική αναζήτηση άρθρων σε βάθος όσο και τους φορείς χάραξης πολιτικής (policy makers) και τους κοινωνιολόγους να παρακολουθούν τις τάσεις στον τομέα της έρευνας μακροπρόθεσμα.

Επιπρόσθετα, πολύ μεγάλη μπορεί να είναι και η συνεισφορά των μεθόδων εξαγωγής πληροφορίας στα πλαίσια μιας αναζήτησης. Όπως αναφέρθηκε σε προηγούμενο Κεφάλαιο του παρόντος, η χρήση μιας μηχανής αναζήτησης είναι συχνά μια επαναληπτική διαδικασία. Κατά τη διαδικασία μιας κλασικής αναζήτησης ο εκάστοτε χρήστης χρησιμοποιεί λέξεις-κλειδιά για να πραγματοποιήσει μια αναζήτηση γύρω από το θέμα που τον ενδιαφέρει, ανακτώντας νέες πληροφορίες σε κάθε νέο έγγραφο που ανακαλύπτει, ενώ ταυτόχρονα, συνδυάζει τις πληροφορίες αυτές έτσι ώστε να βελτιστοποιήσει την αναζήτησή του χρησιμοποιώντας νέες λέξεις ή φράσεις. Υπό αυτή την έννοια, η διαδικασία αυτή μπορεί να απλοποιηθεί χρησιμοποιώντας νέες προγραμματιστικές διεπαφές που ενσωματώνουν τεχνολογίες εξαγωγής πληροφορίας από κείμενα

και έχουν δημιουργηθεί με σκοπό να αυτοματοποιήσουν κατά το δυνατό την παραπάνω διαδικασία. Υπάρχουν διεπαφές που έχουν υλοποιηθεί και υποστηρίζουν την επαναληπτική βελτιστοποίηση ερωτημάτων. Η λειτουργία μιας τέτοιας προγραμματιστικής διεπαφής παρουσιάζεται στο άρθρο του Turney [18], όπου ένας χρήστης θέτει ένα ερώτημα στην μηχανή αναζήτησης και του επιστρέφονται άρθρα σχετικά με το ερώτημα που έθεσε αλλά ταυτόχρονα του παρέχει και προτάσεις για να περιορίσει την αναζήτησή του. Οι όροι αναζήτησης που θέτει ο χρήστης συνδυάζονται μεταξύ τους έτσι ώστε η λίστα των αποτελεσμάτων να μικραίνει με κάθε επανάληψη της αναζήτησης.

Μια ακόμα εφαρμογή των μεθόδων αυτόματης εξαγωγής πληροφορίας αποτελεί η ανάλυση των **προτύπων χρήσης** (usage patterns) στα αρχεία ενός διακομιστή Ιστού (web server logs). Αποτελεί μια ιδιαίτερη εφαρμογή που βασίζεται στη ανάγκη που έχουν οι διαχειριστές ενός διακομιστή Ιστού να γνωρίζουν τι αναζητούν συνήθως οι επισκέπτες των ιστοσελίδων τους. Οι πληροφορίες αυτές συνήθως αποθηκεύονται στα λεγόμενα αρχεία καταγραφής Ιστού (log files). Έχουν υλοποιηθεί πλέον και διατίθενται εμπορικά αρκετά προϊόντα που κάνουν αυτή την ανάλυση. Συνήθως τα λογισμικά αυτά παράγουν μια περίληψη των προτύπων κίνησης, ενώ παράγουν και μια λίστα με τα πιο δημοφιλή αρχεία που επισκέπτονται συνήθως οι χρήστες σε έναν διακομιστή Ιστού. Αν συνδυαστεί ένα τέτοιο πρόγραμμα με ένα πρόγραμμα εξαγωγής πληροφορίας, θα ήταν δυνατόν να χρησιμοποιηθούν οι όροι που έχουν εξαχθεί, έτσι ώστε να παραχθεί μια λίστα με τους πιο συχνά χρησιμοποιούμενους. Η λίστα αυτή μπορεί να δώσει στους διαχειριστές της ιστοσελίδας μια ιδέα για το ποια είναι τα πιο δημοφιλή θέματα που αναζητούν οι χρήστες που την επισκέπτονται.

2.6. Σύνοψη και συμπεράσματα

Στο Κεφάλαιο αυτό επιχειρήθηκε μια σύντομη εισαγωγή και επισκόπηση των μεθόδων Εξαγωγής Πληροφορίας από κείμενα. Σε αυτήν την παράγραφο συνοψίζονται τα σημαντικότερα σημεία του κεφαλαίου και εξετάζονται οι προκλήσεις που προκύπτουν για κάθε μια από τις κατηγορίες προσεγγίσεων που αναλύθηκαν παραπάνω. Από την ανάλυση των προσεγγίσεων καταδεικνύεται η αυξανόμενη ενσωμάτωση των μεθόδων αυτών σε ένα πλήθος εφαρμογών, όπως ο εμπλουτισμός μεταδεδομένων των τεκμηρίων, που μπορεί στη συνέχεια να υποβοηθήσουν την πλοήγηση και την αναζήτηση αυτών. Επίσης, στη συγκεκριμένη ενότητα γίνεται αναφορά και στον τρόπο αξιολόγησης των μεθόδων, μιας διαδικασίας αρκετά δύσκολης και σε ορισμένες περιπτώσεις αδύνατης, λόγω της απουσίας μιας κοινής μεθοδολογίας αξιολόγησης και άμεσης σύγκρισης των μεθόδων αυτών.

Κοινό συμπέρασμα το οποίο προκύπτει από την παρατήρηση των υπαρχουσών μεθόδων, όπως αυτές παρουσιάστηκαν στις προηγούμενες παραγράφους, είναι ότι οι περισσότερες αφορούν κυρίως πολύ συγκεκριμένες εφαρμογές, χωρίς την παροχή κάποιου λογισμικού είτε ανοικτού είτε κλειστού κώδικα. Ακόμα, οι περισσότερες αποτελούν επιστημονικά ευρήματα χωρίς άμεση πρακτική εφαρμογή, ενώ μια ακόμα παράμετρος που πρέπει να ληφθεί υπόψη είναι ότι ελάχιστες εξ' αυτών συντηρούνται ενεργά. Επιπρόσθετα, στην πλειοψηφία τους οι εν λόγω μέθοδοι δεν είναι πλήρως αυτοματοποιημένες και είναι απαραίτητη η χειροκίνητη εξαγωγή των όρων από ειδικούς, κατά κύριο λόγο για τα δεδομένα εκπαίδευσης που χρησιμοποιούνται από τους αλγορίθμους ΜΜ. Ακόμα, πολλές μέθοδοι στα πλαίσια της αξιολόγησής τους τείνουν να συγκρίνουν τα αποτελέσματά τους με αυτά των ειδικών του αντίστοιχου πεδίου. Είναι προφανές πως ο ανθρώπινος παράγοντας είναι περισσότερο ικανός να εντοπίσει αντιπροσωπευτικούς όρους εντός ενός κειμένου σε σχέση με οποιαδήποτε αυτοματοποιημένη μέθοδο. Ως εκ τούτου, οι ημι-αυτόματες μέθοδοι τείνουν να δίνουν καλύτερα αποτελέσματα εξαγωγής σε σχέση με τις πλήρως αυτοματοποιημένες, ενώ οι τελευταίες

εμπεριέχουν λιγότερη χειρωνακτική εργασία και για αυτό το λόγο είναι λιγότερο χρονοβόρες και περισσότερο ανθεκτικές στα λάθη, καθώς περιορίζεται αισθητά η παρέμβαση του ανθρώπινου παράγοντα.

Μια από τις βασικότερες τεχνικές που χρησιμοποιούνται στην πλειονότητα των προσεγγίσεων αφορά μεθόδους ΜΜ οι οποίες συνδυάζονται και με άλλες τεχνικές όπως οντολογίες ή ελεγχόμενα λεξιλόγια, στατιστικές αλλά και ευρετικές μεθόδους για να πετύχουν το επιθυμητό αποτέλεσμα. Παρακάτω μνημονεύονται προκλήσεις που ανακύπτουν ξεχωριστά για κάθε κατηγορία ξεκινώντας από τις **προσεγγίσεις εξαγωγής**. Εδώ θα μπορούσαμε να παρατηρήσουμε πως η εξαγωγή πληροφορίας σπάνια αποτελεί το βασικό στόχο ενός συστήματος αλλά τα αποτελέσματα της εξαγωγής και ο τρόπος εφαρμογής τους είναι που προσδίδουν προστιθέμενη αξία στην προσέγγιση, καθώς μπορούν να χρησιμοποιηθούν άμεσα από τους τελικούς χρήστες. Όσες προσεγγίσεις ανήκουν στη συγκεκριμένη κατηγορία και χρησιμοποιούν αποκλειστικά ευρετικές μεθόδους είναι αρκετά χρήσιμες για τη μοντελοποίηση ενός συγκεκριμένου πεδίου αλλά δεν μπορούν να εφαρμοστούν με τον ίδιο τρόπο και σε άλλα πεδία χωρίς τροποποιήσεις. Μια ακόμα πρόκληση που θα μπορούσε να αναφερθεί, είναι ότι μεγάλο ποσοστό των προσεγγίσεων εφαρμόζουν την εξαγωγή είτε στο πλήρες κείμενο μιας επιστημονικής δημοσίευσης είτε σε κάποιο υποσύνολο αυτής ενώ ελάχιστες τείνουν να εφαρμοστούν σε διαφορετικά πρωτογενή δεδομένα όπως άρθρα ειδήσεων ή άλλο περιεχόμενο που δημοσιεύεται στον Ιστό. Πολλές φορές οι προσεγγίσεις εξαγωγής επιλέγονται με γνώμονα το αν υπάρχει διαθέσιμο και μπορεί να χρησιμοποιηθεί κάποιο λεξιλόγιο, οντολογίες ή κάποιο σύνολο θεματικών επικεφαλίδων. Ένα ακόμα σημαντικό κριτήριο, είναι όμως και η ποιότητα και η επαναχρησιμοποίηση των παραγόμενων όρων που είναι και το ζητούμενο σε τέτοιες εφαρμογές.

Σχετικά με τις **προσεγγίσεις ανάθεσης** η πλειονότητα των προσεγγίσεων που παρουσιάστηκαν τείνουν να κατηγοριοποιούν τις φράσεις σε δύο βασικές ομάδες, με βάση το εάν ο υποψήφιος όρος είναι ή όχι αντιπροσωπευτικός για ένα κείμενο. Βασική πρόκληση σε αυτή την κατεύθυνση αποτελεί είτε η διεύρυνση αυτών των ομάδων με σκοπό τη βελτιστοποίηση της διαδικασίας είτε η εξαγωγή των διαφορετικών ομάδων με αυτόματο τρόπο, καθώς στην πλειονότητα των προσεγγίσεων αυτές παρέχονται έπειτα από εμπειρική παρατήρηση. Επίσης, σε αυτή την κατηγορία εντοπίζεται έλλειψη συστημάτων που να μπορούν να χρησιμοποιήσουν ουσιαστικά τις προτεινόμενες προσεγγίσεις, όπως για παράδειγμα μέσα από την παροχή μιας διεπαφής όπου ακόμα και μη ειδικοί χρήστες θα μπορούσαν να χρησιμοποιήσουν με σκοπό την επαναχρησιμοποίηση και ανακάλυψη νέας γνώσης.

Για όλους τους παραπάνω λόγους, κρίνεται αναγκαία η ανάπτυξη εργαλείων αυτόματης εξαγωγής πληροφορίας ταξινόμησης από κείμενα, εργασία που αποτελεί βασικό ερευνητικό πεδίο της παρούσας Διατριβής. Στο επόμενο Κεφάλαιο του παρόντος πραγματοποιείται μια εκτεταμένη πειραματική αξιολόγηση της χρήσης των Αυτό-Οργανούμενων Χαρτών για την αυτόματη ταξινόμηση ενός συνόλου δεδομένων πολλαπλής ετικέτας ενώ στο Κεφάλαιο 4 αναλύεται μια μεθοδολογία για την αυτόματη ταξινόμηση πολλαπλών κλάσεων (multi-class) μιας συλλογής ηλεκτρονικών βιβλίων (e-books), που βασίζεται στην τεχνητή νοημοσύνη χρησιμοποιώντας πληροφορία από τον πίνακα περιεχομένων.

3. Εκτεταμένη πειραματική αξιολόγηση της χρήσης Αυτό-Οργανούμενων Χαρτών για την αυτόματη ταξινόμηση ενός συνόλου δεδομένων πολλαπλής ετικέτας

3.1. Περίληψη

Αυτό το τμήμα της Διατριβής στοχεύει στη γεφύρωση του χάσματος μεταξύ της επιλογής χαρακτηριστικών και του μεγέθους του χώρου χαρακτηριστικών, χρησιμοποιώντας διαφορετικά μεγέθη Αυτό-Οργανούμενων Χαρτών (SOM), κάτω από διαφορετικά σενάρια επιλογών για την ταξινόμηση ενός συνόλου δεδομένων πολλαπλής ετικέτας (multi-label), πιο συγκεκριμένα της συλλογής Reuters Mod' Apte. Η επιλογή ορθογώνιων χαρτών βασίζεται σε μια ευρετική διαδικασία για την εύρεση του κατάλληλου μεγέθους αυτών. Η κατασκευή των διανυσμάτων βασίζεται σε μια απλή, αλλά αποτελεσματική διαδικασία που στοχεύει στον μετασχηματισμό των διανυσμάτων από πολλαπλής ετικέτας σε μονής ετικέτας. Η προτεινόμενη λύση, βελτιώνει την αποτελεσματικότητα της ταξινόμησης, όχι μόνο όσον αφορά την ακρίβεια, αλλά και ως προς τους απαιτούμενους υπολογιστικούς πόρους και το χρόνο που απαιτείται για την εκπαίδευση του δικτύου. Διεξήχθησαν εκτεταμένα πειράματα, με χρήση διαφορετικών επιλογών για τα μεγέθη τόσο χαρτών όσο και διανυσμάτων, κύκλων εκπαίδευσης, καθώς και χρήση των συμφραζόμενων γύρω από μια λέξη, για να εκτιμηθεί η επίδρασή τους στην απόδοση του ταξινομητή. Επιπλέον, προτείνεται ένας ευφυής αλγόριθμος για την επιλογή των ετικετών, με σκοπό να δείξει ότι οι γειτονικοί κόμβοι επάνω στο χάρτη επηρεάζουν την επιλογή ετικετών για έναν συγκεκριμένο κόμβο. Σύμφωνα με τα πειράματα που διεξήχθησαν, η προσέγγιση επιτυγχάνει αύξηση κατά 10% στη βαθμολογία F1 Μάκρο-μέσου όρου, 30 φορές μείωση της διάστασης των διανυσμάτων και περίπου 34 φορές μικρότερους χάρτες σε σχέση με το βασικό σενάριο σύγκρισης.

3.2. Εισαγωγή

Η εξέλιξη του Διαδικτύου έχει δημιουργήσει νέα πρότυπα στον τρόπο με τον οποίο τα ηλεκτρονικά έγγραφα διατίθενται, αναπροσαρμόζονται, αναζητούνται και ανακτώνται. Επιπλέον, ο αριθμός των ηλεκτρονικών εγγράφων που διατίθενται μέσω του Διαδικτύου αυξάνεται συνεχώς, δημιουργώντας έτσι προκλήσεις όσον αφορά τη διαθεσιμότητα, την ανάκτησή τους και τη μακροπρόθεσμη διατήρησή τους. Στο πλαίσιο αυτό, καθίσταται σημαντική η αυτόματη ταξινόμηση των εγγράφων και ειδικότερα η ταξινόμηση πολλαπλής ετικέτας, στην οποία κάθε στοιχείο δεδομένων διαθέτει περισσότερες από μία ετικέτες κάθε φορά. Επιπλέον, η ταξινόμηση πολλαπλής ετικέτας είναι πιο περίπλοκη σε σύγκριση με την ταξινόμηση πολλαπλών κλάσεων, καθώς η υπόθεση ότι οι ετικέτες που ανατίθενται σε κάθε στοιχείο είναι στοχαστικά ανεξάρτητες δεν ισχύει πάντοτε [39].

Μεγάλο κομμάτι έρευνας επικεντρώνεται στην ταξινόμηση πολλαπλών κλάσεων [4], ενώ μερικές από τις πιο πρόσφατες ερευνητικές εργασίες που αφορούν την ταξινόμηση πολλαπλής ετικέτας στοχεύουν στο μετασχηματισμό του προβλήματος σε ένα σύνολο δυαδικών προβλημάτων ταξινόμησης (binary classification) [41]. Στην τελευταία περίπτωση, για κάθε κατηγορία που υπάρχει στο σύνολο δεδομένων, δημιουργείται ένας δυαδικός ταξινομητής και εκπαιδεύεται στο σύνολο δεδομένων, έτσι ώστε καθένας από αυτούς να μπορεί να διακρίνει αν ένα δείγμα ανήκει σε μια συγκεκριμένη κατηγορία ή όχι. Δυστυχώς, η προσέγγιση αυτή επαρκεί μόνο υπό την προϋπόθεση ότι οι κατηγορίες είναι ανεξάρτητες [39]. Στο σύνολο δεδομένων Reuters, που επιλέχθηκε στην

περίπτωσή μας, οι ετικέτες που αντιστοιχούν σε ένα στοιχείο του συνόλου δεδομένων είναι συσχετισμένες, και επομένως δεν είναι δυνατή η μετατροπή του προβλήματος σε ένα σύνολο δυαδικών προβλημάτων ταξινόμησης.

Παραδοσιακά, η αυτόματη ταξινόμηση ενός συνόλου ηλεκτρονικών εγγράφων σε ένα σύνολο διαφορετικών κλάσεων που υποδηλώνεται με τη χρήση διαφορετικών ετικετών, αποτελεί διαδικασία εποπτευόμενης ΜΜ [39]. Σε αυτή την περίπτωση, το σύστημα μαθαίνει ένα σύνολο παραμέτρων με βάση τις ετικέτες που χρησιμοποιούνται κατά τη διάρκεια της εκπαίδευσης και στη συνέχεια χρησιμοποιείται το σύνολο δεδομένων ελέγχου (testing) για την αξιολόγηση της απόδοσης του ταξινομητή. Ωστόσο, αυτή η προσέγγιση απαιτεί ένα μεγάλο, καλά επισημασμένο σύνολο δεδομένων και θεωρείται αναποτελεσματική με σύνολα δεδομένων που έχουν μη ισορροπημένη κατανομή των δειγμάτων μεταξύ των κλάσεων. Αυτός ο περιορισμός επιβάλλει την επιλογή μη εποπτευόμενων προσεγγίσεων ΜΜ έναντι των εποπτευόμενων.

Αντίθετα, οι Αυτό-Οργανούμενοι Χάρτες (Self-Organizing Maps – SOM εφεξής) είναι μια μη-εποπτευόμενη μέθοδος ΜΜ, η οποία μπορεί να συγκεντρώσει παρόμοια έγγραφα μαζί χρησιμοποιώντας μέτρα ομοιότητας διανυσμάτων, χωρίς πρότερη γνώση των κλάσεων. Στην περίπτωση αυτή, οι ετικέτες των δεδομένων εκπαίδευσης θα χρησιμοποιηθούν μόνο για να καθοδηγήσουν την ανάθεση των ετικετών στο σύνολο δεδομένων ελέγχου μετά τη φάση της εκπαίδευσης. Μια άλλη διαφοροποίηση όσον αφορά τις εποπτευόμενες τεχνικές είναι ότι ένας εκπαιδευμένος χάρτης περιέχει αναγνωρίσιμες αλλά μη οριοθετημένες ομάδες (clusters) [39]. Στις ομάδες που δημιουργούνται, παρόμοια έγγραφα τείνουν να μοιράζονται κοινές ετικέτες και συνεπώς το SOM διατηρεί εγγενώς τη συσχέτιση μεταξύ των ετικετών, καθώς οι σχετικές ετικέτες μπορούν να βρεθούν είτε στον ίδιο είτε σε έναν γειτονικό κόμβο στον χάρτη.

Σε αυτό το Κεφάλαιο της παρούσας Διατριβής, εστιάζουμε στη χρήση του SOM για την ταξινόμηση ενός συνόλου δεδομένων, πολλαπλής ετικέτας, κάτω από διαφορετικές επιλογές διαμόρφωσης. Σε αυτό το πλαίσιο, συγκρίνονται αρκετά μεγέθη πλέγματος, χρησιμοποιώντας διαφορετικό σύνολο παραμέτρων, όπως για παράδειγμα το μέγεθος των διανυσμάτων, τους κύκλους εκπαίδευσης, τα συμφραζόμενα μιας λέξης κλπ., για να αποδειχθεί ότι συγκεκριμένες διαμορφώσεις υπερτερούν έναντι άλλων state-of-the-art μεθόδων. Ακόμα, χρησιμοποιούνται και συγκρίνονται τετράγωνοι και ορθογώνιοι χάρτες για να αξιολογηθεί η επίπτωσή τους στην απόδοση του ταξινομητή. Από όσο γνωρίζουμε, αυτή είναι η πρώτη προσπάθεια να πραγματοποιηθεί μια τέτοια σύγκριση.

Η επιλογή των ορθογώνιων χαρτών, βασίστηκε σε μια τεχνική που προτάθηκε στο [40] για την επιλογή του μεγέθους του χάρτη για ένα SOM, η οποία παρουσιάζεται λεπτομερώς στις επόμενες υποενότητες. Επιπλέον, προκειμένου να αποφύγουμε την εκτεταμένη μηχανική χαρακτηριστικών, χρησιμοποιήσαμε ένα απλό σύνολο χαρακτηριστικών που βασίζεται σε στατιστικά λέξεων και μπορεί εύκολα να εφαρμοστεί σε οποιοδήποτε σύνολο δεδομένων χωρίς τροποποιήσεις. Επιπλέον, παρουσιάζεται ένας έξυπνος αλγόριθμος για να επιλεγούν οι ετικέτες που θα ανατεθούν σε ένα στοιχείο με βάση τις ετικέτες που υπάρχουν στους γειτονικούς κόμβους, καθώς και για τον προσδιορισμό των καταλληλότερων ετικετών όταν δεν υπάρχει ετικέτα, κάτι που αποτελεί μια ακόμα συνεισφορά αυτής της Διατριβής.

Αρκετά τμήματα των αλγορίθμων που παρουσιάζονται παρακάτω, θα ενσωματωθούν σε ένα πειραματικό σύστημα ταξινόμησης εγγράφων που θα χρησιμοποιηθεί στο Εθνικό Μετσόβιο Πολυτεχνείο (ΕΜΠ), εξυπηρετώντας δύο διαφορετικούς σκοπούς. Πρώτον, να υποστηριχθεί η

χειρωνακτική επισήμανση (annotation) εγγράφων και μαθησιακού υλικού, μια δαπανηρή και χρονοβόρα διαδικασία που υπόκειται σε λάθη, και δεύτερον, να υποβοηθηθούν οι συγγραφείς κατά την εισαγωγή ηλεκτρονικών εγγράφων στο σύστημα παρέχοντας χρήσιμες πληροφορίες σχετικά με το ποιες είναι οι καταλληλότερες κατηγορίες για ένα δεδομένο έγγραφο.

Οι υποενότητες αυτού του κεφαλαίου οργανώνονται ως εξής. Αρχικά, παρέχεται μια εμπεριστατωμένη επανεξέταση των πρόσφατων προσεγγίσεων σχετικά με την ταξινόμηση εγγράφων πολλαπλής ετικέτας, την επιλογή χαρακτηριστικών και τις προσεγγίσεις ταξινόμησης που χρησιμοποιούν μεθόδους μη εποπτευόμενης ΜΜ και ειδικότερα του SOM. Στη συνέχεια, περιγράφεται το σύνολο δεδομένων που χρησιμοποιείται για τη διεξαγωγή των πειραμάτων, αλλά και η προτεινόμενη λύση λεπτομερώς. Η επόμενη υποενότητα αναφέρεται στην πειραματική αξιολόγηση και πιο συγκεκριμένα στα πειράματα και στα μέτρα αξιολόγησης που χρησιμοποιήθηκαν αλλά και στα αποτελέσματα που προέκυψαν, ακολουθούμενη από μια σύγκριση και συνολική συζήτηση επί των αποτελεσμάτων. Τέλος, προτείνονται ορισμένες βελτιώσεις αλλά και ερευνητικά ζητήματα που πρέπει να αντιμετωπιστούν στο μέλλον.

3.3. Σχετικές ερευνητικές εργασίες

3.3.1. Ταξινόμηση πολλαπλής ετικέτας

Ο κύριος στόχος της ταξινόμησης εγγράφων είναι η κατανομή κάθε στοιχείου σε μία ή περισσότερες κλάσεις, ανάλογα με το αν πρόκειται για εργασία ταξινόμησης πολλαπλών κλάσεων ή πολλαπλής ετικέτας. Με βάση τα δεδομένα εκπαίδευσης, το σύστημα είναι ικανό να ταξινομεί νέα αντικείμενα στην αντίστοιχη κλάση. Σε περιπτώσεις ταξινόμησης πολλαπλών κλάσεων, κάθε στοιχείο ταξινομείται σε μία μόνο κατηγορία, ενώ σε μια εργασία ταξινόμησης πολλαπλής ετικέτας κάθε στοιχείο μπορεί να ανήκει σε περισσότερες από μία κατηγορίες [42]. Ο κύριος στόχος της εργασίας αυτής είναι να προτείνει το καλύτερο σχήμα ταξινόμησης, που αποσκοπεί στη μείωση της διάστασης των χρησιμοποιούμενων διανυσμάτων ενώ αυξάνει την ακρίβεια του ταξινομητή [43]. Σύμφωνα με το [39], η ταξινόμηση πολλαπλής ετικέτας θέτει συγκεκριμένες προκλήσεις καθώς οι ετικέτες που αντιστοιχίζονται σε ένα στοιχείο μπορεί να συσχετίζονται. Κατά συνέπεια, η θεώρηση της ταξινόμησης πολλαπλής ετικέτας ως ένα σύνολο δυαδικών εργασιών ταξινόμησης δεν είναι πάντα ορθή.

Έχουν προταθεί διάφορες προσεγγίσεις για την επίλυση του προβλήματος της ταξινόμησης πολλαπλής ετικέτας. Η εργασία που παρουσιάζεται στο [43] στοχεύει στη σύγκριση δύο δημοφιλών αλγορίθμων, δηλαδή των k-Nearest Neighbour (k-NN) και SVM, χρησιμοποιώντας χαρακτηριστικά, όπως τα μέτρα TF και IDF. Τα ευρήματά τους υποδεικνύουν ότι το SVM είναι υπολογιστικά ακριβό όσο το μέγεθος του συνόλου δεδομένων αυξάνεται, ενώ ο k-NN επηρεάζεται από μη σχετικά χαρακτηριστικά. Τέλος, ένα αυξημένο μέγεθος διανυσμάτων μπορεί να επιφέρει αυξημένες υπολογιστικές απαιτήσεις. Το [44] επιχειρεί να συγκρίνει τον k-NN με τον Naive Bayes αλλά και Term-Graph αλγορίθμους χρησιμοποιώντας το σύνολο δεδομένων Reuters-21578 [45], όπου ο k-NN φαίνεται να υπερτερεί. Η δουλειά που παρουσιάζεται στο Κεφάλαιο αυτό της Διατριβής χρησιμοποιεί σχήμα στάθμισης TF-IDF ως χαρακτηριστικό και επικεντρώνεται στον καθορισμό του κατάλληλου μεγέθους για το SOM, διατηρώντας παράλληλα τη διάσταση των διανυσμάτων που χρησιμοποιούνται σχετικά μικρή ώστε να απορρίπτονται άσχετα χαρακτηριστικά που μπορεί να επιδεινώσουν την απόδοση του ταξινομητή.

Μια προσπάθεια αξιολόγησης δύο διαφορετικών συστημάτων για ταξινόμηση πολλαπλής ετικέτας παρουσιάστηκε στο [46], υποδεικνύοντας ότι η απόδοση της ταξινόμησης εξαρτάται σε μεγάλο βαθμό από το σύνολο δεδομένων που χρησιμοποιείται και από τον στόχο της ταξινόμησης. Το πρώτο σύστημα βασίστηκε σε ένα μοντέλο διαδοχικής απεικόνισης δεδομένων (sequential data representation model), και πιο συγκεκριμένα στο ιεραρχικό SOM, ενώ το δεύτερο χρησιμοποιούσε το Vector Space Model (VSM) για την αντιπροσώπευση εγγράφων και Latent Signal Indexing (LSI) για την απεικόνιση των διανυσμάτων σε ένα μικρότερο χώρο διαστάσεων. Τα αποτελέσματά τους υποδηλώνουν ότι το LSI λειτουργεί πιο αποτελεσματικά σε έγγραφα με μια ετικέτα, ενώ η χρήση του ιεραρχικού SOM ξεπερνά το LSI στην περίπτωση που υπάρχουν πολλαπλές ετικέτες. Στην περίπτωσή μας, ο SOM χρησιμοποιήθηκε επειδή το επιλεγμένο σύνολο δεδομένων στοχεύει στην ταξινόμηση πολλαπλής ετικέτας, αποδεικνύοντας ότι όντως υπάρχει συσχέτιση μεταξύ των ετικετών που αποδίδονται σε ένα στοιχείο, δικαιολογώντας έτσι την επιλογή του. Μια ακόμα προσέγγιση πολλαπλής ετικέτας που ασχολείται με το πρόβλημα της ύπαρξης περισσότερων κλάσεων από τα δείγματα μελετήθηκε στο [47]. Εκεί προτείνεται και αναλύεται μια προσέγγιση δύο σταδίων που περιλαμβάνει έναν ευρετικό αλγόριθμο για την κατασκευή ενός αρχικού ταξινομητή και μια απλή μέθοδο για την επαύξηση αυτού του ταξινομητή αφαιρώντας μη σχετικές ετικέτες από την έξοδό του. Μια εποπτευόμενη γενετική (generative) προσέγγιση για την ταξινόμηση προφορικών εγγράφων πολλαπλής ετικέτας που μπορεί να εφαρμοστεί και για την ταξινόμηση κειμένων, ήτοι το Μοντέλο Μείγματος Συσχέτισης Ετικετών (LCMM), παρουσιάστηκε στο [48]. Το LCMM αποτελείται από ένα μοντέλο συσχέτισης ετικέτας και ένα μοντέλο συνθηκών (conditioned) πολλαπλών ετικετών το οποίο μπορεί να θεωρηθεί ως ένα εποπτευόμενο μοντέλο μίξης ετικετών (mixture model). Οι συγγραφείς υποστηρίζουν επίσης ότι η προσέγγισή τους μπορεί να αντιμετωπίσει αποτελεσματικά το πρόβλημα της ύπαρξης περισσότερων κλάσεων από παραδείγματα αλλά χωρίς να παρέχουν επαρκή στοιχεία για το πώς αυτό επιτυγχάνεται. Και οι δύο προσεγγίσεις είναι παρόμοιες με τη δική μας, καθώς το σύνολο δεδομένων Reuters έχει μια μη ισορροπημένη κατανομή δειγμάτων μεταξύ των κλάσεων, με αποτέλεσμα να υπάρχουν μερικές κλάσεις μόνο με λίγα δείγματα. Στην περίπτωσή μας, εκτός από την ιδιότητα γειτνίασης που χρησιμοποιείται για να επωφεληθούμε από τις ετικέτες που έχουν εκχωρηθεί στους κοντινότερους κόμβους, χρησιμοποιούμε επίσης και μέτρα ακρίβειας Μάκρο-μέσου όρου για να αξιολογήσουμε την απόδοση του ταξινομητή σε κατηγορίες που έχουν μικρό αριθμό δειγμάτων. Επιπλέον, το Μοντέλο Μείγματος Συσχέτισης Ετικετών (LCMM) που παρουσιάστηκε στο [48], παρόλο που ανήκει στην κατηγορία εποπτευόμενης ΜΜ, μπορεί να αποδειχθεί αποτελεσματικό και στην περίπτωσή μας και ως τέτοιο παραμένει ως ανοικτή ερευνητική κατεύθυνση.

Ένας αλγόριθμος ταξινόμησης ημι-εποπτευόμενης ΜΜ που προέκυψε από την προσαρμογή δύο διαφορετικών σημασιολογικών μοντέλων ταξινόμησης παρουσιάστηκε στο [49]. Σε αντίθεση με τις εποπτευόμενες προσεγγίσεις που παρουσιάστηκαν προηγουμένως, από όσο γνωρίζουμε, αυτή η εργασία είναι η πρώτη απόπειρα εφαρμογής αλγορίθμων ταξινόμησης ημι-εποπτευόμενης ΜΜ που βασίζεται σε σημασιολογικά χαρακτηριστικά που σχετίζονται με την εκάστοτε κλάση (class-based semantics). Τα πειράματα που διεξήχθησαν υποδεικνύουν ότι η ενσωμάτωση σημασιολογικών μοντέλων μπορεί να βελτιώσει την ταξινόμηση όσον αφορά στην επίτευξη καλύτερης ακρίβειας σε μικρά σύνολα δεδομένων, ενώ εξαλείφει τον θόρυβο σε μεγάλα σύνολα δεδομένων.

Μια προσέγγιση που χρησιμοποιεί Νευρωνικά Δίκτυα Βαθιάς Μάθησης (Deep Neural Networks – DNN εφεξής) για την ταξινόμηση εγγράφων πολλαπλής ετικέτας ενός συνόλου δεδομένων ειδήσεων στην τσεχική γλώσσα μελετήθηκε στο [50]. Η επιλογή αυτού του τύπου

νευρωνικών δικτύων βασίστηκε στα ευρήματα ότι τα DNNs έχουν το πλεονέκτημα ότι μαθαίνουν από απλά χαρακτηριστικά και έτσι είναι ευκολότερο να γενικεύσουν την προσέγγιση για διαφορετικά σύνολα δεδομένων, δείχνοντας έτσι ότι επιτυγχάνουν καλύτερα αποτελέσματα ταξινόμησης σε σχέση με το βασικό σενάριο σύγκρισης (baseline scenario). Στην παρούσα προσέγγιση επιλέχθηκε ο SOM, που συνδυάστηκε με μια προσεκτικά σχεδιασμένη διαδικασία επιλογής χαρακτηριστικών βασισμένη σε στατιστικά λέξεων, αρκετά γενική ώστε να εφαρμοστεί σε άλλα σύνολα δεδομένων χωρίς σημαντικές τροποποιήσεις.

3.3.2. Επιλογή Χαρακτηριστικών

Μια άλλη σημαντική πτυχή κατά την κατασκευή ενός συστήματος αυτόματης ταξινόμησης κειμένων είναι η κβαντοποίηση διανυσμάτων (vector quantization) και η επιλογή χαρακτηριστικών, καθώς συμβάλλει στη μείωση του χώρου χαρακτηριστικών [51]. Το μέγεθος του συνόλου χαρακτηριστικών που θα χρησιμοποιηθεί θεωρείται πολύ σημαντικό, καθώς έχει αποδειχθεί ότι η αύξηση του χώρου χαρακτηριστικών δεν συνεπάγεται αναγκαστικά καλύτερη ακρίβεια στην ταξινόμηση, καθώς με ένα τεράστιο χώρο χαρακτηριστικών η απόδοση του ταξινομητή μπορεί βαθμιαία να επιδεινωθεί [50], [52]. Ένα άλλο πλεονέκτημα όταν χρησιμοποιείται λογικός αριθμός χαρακτηριστικών είναι η μείωση της επιβάρυνσης στο χρόνο επεξεργασίας, συνοδευόμενη από βελτιωμένη απόδοση ταξινόμησης, καθώς ένας μεγάλος αριθμός μη σχετικών ή περιττών χαρακτηριστικών μπορεί να επηρεάσει σημαντικά την απόδοση του ταξινομητή [52], [53]. Για την επιλογή των χαρακτηριστικών, έχουν προταθεί διαφορετικές προσεγγίσεις οι οποίες μπορούν να χωρισθούν αυθαίρετα σε δύο διαφορετικές κατηγορίες, είτε σε αυτές που βασίζονται σε στατιστικά λέξεων είτε σε αυτές που έχουν έναν περισσότερο γλωσσικό προσανατολισμό. Η παρούσα Διατριβή επικεντρώνεται στην πρώτη κατηγορία, καθώς στοχεύει στη δημιουργία ενός συστήματος που μπορεί εύκολα να εφαρμοστεί σε διαφορετικά σύνολα δεδομένων, ενώ η δεύτερη κατηγορία συνεπάγεται εκτεταμένη μηχανική χαρακτηριστικών, που συνήθως διεξάγεται από εμπειρογνώμονες στον τομέα, και είναι δύσκολο να γενικευθούν όταν επιλεχθεί ένα διαφορετικό σύνολο δεδομένων.

Η πρώτη κατηγορία περιλαμβάνει μεθόδους επιλογής χαρακτηριστικών με βάση τη συχνότητα εγγράφων (Document Frequency - DF) καθώς επίσης και σχετικά μέτρα συχνότητας, Information Gain, Mutual Information και Pointwise Mutual Information, x2-test, Term Strength [54] καθώς και διάφορους συνδυασμούς αυτών των μεθόδων [43]. Αυτές οι προσεγγίσεις συνήθως χρησιμοποιούν είτε απλά χαρακτηριστικά όπως 1- grams ή πιο σύνθετα, όπως 2- ή 3- grams σε συνδυασμό με χαρακτηριστικά τοπολογίας όπως η σχετική θέση της πρώτης και / ή της τελευταίας εμφάνισης ενός όρου [19]. Η χρήση του μέτρου TF σε συνδυασμό με έναν αλγόριθμο εξαγωγής χαρακτηριστικών που βασίζεται στο λήμμα κάθε λέξης (stemmer-based) έχει μελετηθεί στο [41] και έχει αξιολογηθεί με χρήση διαφορετικών ταξινομητών. Εδώ, επιλέξαμε το σχήμα στάθμισης TF-IDF σε συνδυασμό με τον Porter stemmer, καθώς η χρήση του TF από μόνη της δεν λαμβάνει υπόψη την μοναδικότητα ενός όρου σε σχέση με ολόκληρο το σύνολο δεδομένων.

Στο [28], χρησιμοποιήθηκε ένα ασαφές (fuzzy) Bag-of-Words μοντέλο για την αναπαράσταση των εγγράφων, με στόχο την καταγραφή των σημασιολογικών εννοιών (semantic meanings) υψηλού επιπέδου από τα δεδομένα. Αντί της διεξαγωγής ακριβούς αντιστοίχισης λέξεων, χρησιμοποιείται ασαφής συσχέτιση (mapping) βασισμένη στη σημασιολογική συσχέτιση των λέξεων, ποσοτικοποιημένη με τα μέτρα ομοιότητας συνημίτονου. Ωστόσο, η προσέγγιση που παρουσιάστηκε στο [28] δεν μπορούσε να εφαρμοστεί στην περίπτωσή του συνόλου δεδομένων Reuters που χρησιμοποιείται σε αυτό το Κεφάλαιο καθώς είναι μια συλλογή ειδήσεων, που δεν είναι πολύ

εκτεταμένη από την άποψη του μήκους των αντικειμένων και με διαφορετική θεματολογία, και επομένως είναι εξαιρετικά δύσκολο να είναι αποτελεσματική στην εξαγωγή σημασιολογικών σχέσεων. Από την άλλη πλευρά, η προαναφερθείσα προσέγγιση μπορεί να θεωρηθεί αποτελεσματική εάν επιλεχθεί ένα άλλο σύνολο δεδομένων, όπως για παράδειγμα το σύνολο δεδομένων ηλεκτρονικών βιβλίων από το Springer που παρουσιάζεται στο επόμενο Κεφάλαιο.

Μια προσέγγιση εξαγωγής όρων που βασίζεται σε ζώνες (zone-based) και χρησιμοποιεί αντιπροσωπευτικές λέξεις για την ταξινόμηση κειμένου, παρουσιάστηκε στο [28], σε συνδυασμό με μη εποπτευόμενες μεθόδους ΜΜ που βασίζονται στο σχήμα στάθμισης TF-IDF για την εξαγωγή λέξεων-κλειδιών. Τα ευρήματά τους υποδεικνύουν ότι η χρήση λέξεων αντιπροσωπευτικών του θέματος, υπερτερεί σε σχέση με το απλό 1-gram μοντέλο. Το σύνολο δεδομένων που επιλέχθηκε στη δική μας περίπτωση, επιβάλλει τη χρήση χαρακτηριστικών που βασίζονται μόνο σε στατιστικά λέξεων, καθώς η προσέγγιση που παρουσιάζεται στο [28] επικεντρώνεται περισσότερο στην εξαγωγή πληροφορίας παρά στην ταξινόμηση. Ωστόσο, τα χαρακτηριστικά που προτείνονται στο [28] ενδέχεται να είναι πολύ χρήσιμα αν επιλεγεί διαφορετικό σύνολο δεδομένων. Επιπλέον, πειραματιστήκαμε όχι μόνο με 1-grams αλλά και με τα συμφραζόμενα μιας λέξης (2-grams και 3-grams) για την εκτίμηση της επίδρασής τους στην απόδοση του ταξινομητή.

Στο [52] προτάθηκε η μεγιστοποίηση χαρακτηριστικών (feature maximization) ως ένας τρόπος για την επίτευξη αποτελεσματικής επιλογής χαρακτηριστικών, καθώς σύμφωνα με τους συγγραφείς, είναι ένα εξαιρετικά χρήσιμο μέτρο που ευνοεί τις ομάδες με μεγάλο F-measure. Στη μελέτη μας, αντίθετα, πειραματιστήκαμε χρησιμοποιώντας μια σχετικά απλή διαδικασία επιλογής χαρακτηριστικών, δοκιμασμένη κάτω από διαφορετικά σενάρια διαμόρφωσης για να προσδιορίσουμε το κατάλληλο μέγεθος διανύσματος για το υποκείμενο σύνολο δεδομένων. Δεδομένου ότι η κατανομή των δειγμάτων στις κλάσεις είναι άνιση, χρησιμοποιήσαμε μέτρα αξιολόγησης Μίκρο- και Μάκρο-μέσου όρου για να αξιολογήσουμε την απόδοση του ταξινομητή.

Μέθοδοι πιο κοντά στην ανάλυση φυσικής γλώσσας θεωρούνται τα φίλτρα σχημάτων λόγου (POS, part-of-speech), τμήματα φράσεων ουσιαστικών (Noun Phrase-NP) και άλλα σχετικά μοντέλα σχημάτων λόγου [19]. Στο [46], εφαρμόστηκε ένας απλός POS tagger [55] για να επιλεχθούν τα ουσιαστικά ενδιαφέροντος από το έγγραφο μετά από την αφαίρεση των πληροφοριών ετικέτας (tag information) και των δεδομένων που δεν αποτελούν κείμενο. Το φιλτράρισμα των σχημάτων λόγου έχει επίσης προταθεί για την καλύτερη αναπαράσταση εγγράφων [56]. Η χρήση λεξικολογικών και συντακτικών χαρακτηριστικών για τη μείωση του χώρου των χαρακτηριστικών προτάθηκε επίσης στο [57]. Επιπλέον, στο [58] έχει προταθεί η χρήση νέων χαρακτηριστικών χωρίς επίβλεψη (unsupervised) βασισμένων σε έναν Latent Dirichlet Allocation (LDA) stemmer και στους σημασιολογικούς χώρους HAL και COALS. Παρόλο που ο αλγόριθμος LDA χρησιμοποιείται εκτεταμένα για να εξάγει αναπαραστάσεις υψηλού επιπέδου από έγγραφα, μειονεκτεί διότι απαιτεί εκτεταμένη εκλέπτυνση (finetuning) υπερπαραμέτρων που σχετίζονται με τον αριθμό των κλάσεων ή την κατανομή λέξεων και θεματολογίας για την εύρεση των βέλτιστων τιμών αυτών [6]. Για να ξεπεραστεί αυτό το πρόβλημα, οι συγγραφείς προτείνουν μια μέθοδο που βασίζεται σε μια διαδικασία δύο σταδίων. Αρχικά, επεκτείνεται ο χώρος αναπαράστασης χρησιμοποιώντας ένα σύνολο από χώρους θεματολογίας (topic spaces) και στη συνέχεια συμπιέζει τον χώρο, αφαιρώντας τις λιγότερο σχετικές διαστάσεις, χρησιμοποιώντας το σύνολο δεδομένων Reuters για τα πειράματά τους. Στην περίπτωσή μας, χρησιμοποιείται η batch έκδοση του αλγόριθμου SOM, για να ξεπεραστεί η ανάγκη εύρεσης της βέλτιστης τιμής για την υπερπαράμετρο που αφορά το ρυθμό εκμάθησης. Επιπλέον, το SOM είναι

ένα μη εποπτευόμενο νευρωνικό δίκτυο που έχει εγγενώς την ικανότητα να ομαδοποιεί παρόμοια διανύσματα στον ίδιο ή γειτονικό υπο-χώρο.

Μια διαφορετική προσέγγιση παρουσιάστηκε στο [4] όπου διαρθρωτικά στοιχεία (structural elements) από το έγγραφο χρησιμοποιήθηκαν ως χαρακτηριστικά για την αυτόματη εξαγωγή μετά-δεδομένων εγγράφων χρησιμοποιώντας SVMs. Επιπλέον, σε αυτή την εργασία εισήχθη και μια μέθοδος εξόρυξης χαρακτηριστικών βασισμένη σε κανόνες με τη χρήση συμφραζόμενων πληροφοριών. Όπως αναφέρθηκε και προηγουμένως, η παρούσα προσέγγιση χρησιμοποιεί στατιστικά στοιχεία λέξεων καθώς στοχεύουμε στη δημιουργία μιας διαδικασίας επιλογής γενικών χαρακτηριστικών, η οποία θα εφαρμόζεται εύκολα σε διαφορετικά σύνολα δεδομένων.

3.3.3. Self-Organising Maps

Στην πλειονότητα των προσεγγίσεων, για την κατανομή εγγράφων σε μία ή περισσότερες κλάσεις, χρησιμοποιούνται τεχνικές εποπτευόμενης ΜΜ, καθώς θεωρούνται πιο αποτελεσματικές. Ένα μειονέκτημα αυτών των προσεγγίσεων είναι ότι απαιτούν ένα σωστά σχολιασμένο, μεγάλης κλίμακας σύνολο δεδομένων με καλή ισορροπία δειγμάτων μεταξύ των διαφόρων κλάσεων [28], [59]. Το σύνολο δεδομένων Reuters έχει μια άνιση κατανομή δειγμάτων μεταξύ των κλάσεων, δημιουργώντας έτσι προκλήσεις στην ενσωμάτωση εποπτευόμενων τεχνικών.

Αντίθετα, το SOM είναι ένα νευρωνικό δίκτυο μη εποπτευόμενης ΜΜ, ικανό να συγκεντρώνει παρόμοια δεδομένα, ανεξάρτητα από τις ετικέτες που τους έχουν ανατεθεί [60]. Το SOM έχει εγγενώς τη δυνατότητα να χαρτογραφήσει και να εμφανίσει διανύσματα υψηλότερης διάστασης στον δισδιάστατο χώρο [60], [40], [61]. Σε άλλες προσεγγίσεις, τα VSM και LSI έχουν χρησιμοποιηθεί για να επιτευχθεί η μείωση των διαστάσεων [43], [46]. Στο [62], χρησιμοποιήθηκε η μέθοδος Τυχαίας Χαρτογράφησης (Random Mapping) για διανύσματα μεγάλων διαστάσεων, για να επιτευχθεί γρήγορα ο υπολογισμός της ομοιότητας, καθώς άλλες μέθοδοι μείωσης διαστάσεων είναι πολύ δαπανηρές για να χρησιμοποιηθούν. Αν και μειώνεται η διάσταση των αρχικών διανυσμάτων, οι αμοιβαίες ομοιότητες μεταξύ τους διατηρούνται στο μειωμένο χώρο. Ως εκ τούτου, επιδεικνύουν ότι η ακρίβεια που επιτυγχάνεται είναι σχεδόν εξίσου καλή με την αρχική επειδή το εσωτερικό γινόμενο μεταξύ των χαρτογραφημένων διανυσμάτων ακολουθεί πιστά το εσωτερικό γινόμενο των αρχικών διανυσμάτων.

Το SOM έχει εκτεταμένες εφαρμογές σε διαφορετικούς τομείς, όπως στα χρηματοοικονομικά, στις βιοϊατρικές αναλύσεις, στη βιομηχανική και στατιστική ανάλυση αλλά και στην οργάνωση κειμένων. Τρία διαφορετικά παραδείγματα που αφορούν τη διαχείριση συλλογών εγγράφων παρουσιάστηκαν στο [40], μαζί με αρκετές υποδείξεις σχετικά με τον τρόπο επίτευξης της μείωσης των διαστάσεων στην κατασκευή πολύ μεγάλων SOM. Οι συγγραφείς υποστηρίζουν επίσης ότι η χρήση πολύ μεγάλων SOM δεν είναι μια καλή προσέγγιση μοντελοποίησης, καθώς αυξάνει σημαντικά την πολυπλοκότητα, το χρόνο επεξεργασίας για τις εργασίες ενημέρωσης της γειτονιάς κάθε κόμβου (κατά την εκπαίδευση) και αναζήτησης, αλλά επίσης δημιουργεί πολύ μεγάλες ανάγκες σε υπολογιστικούς πόρους. Ωστόσο, αυτή η προσέγγιση ακολουθείται στο [39], όπου συγκεκριμένες επιλογές, όπως η χρήση ιδιαίτερα μεγάλων διανυσμάτων χαρακτηριστικών και η χρήση της βηματικής αναδρομικής έκδοσης (stepwise recursive) αντί της χρήσης της batch έκδοσης του αλγόριθμου SOM, επέβαλαν τη χρήση υπολογιστών μεγάλης κλίμακας (High Performance Computing - HPC) για την εκτέλεση των δαπανηρών υπολογισμών. Στη προσέγγιση που παρουσιάζεται στη συγκεκριμένη Διατριβή, με βάση τα ευρήματα που αναφέρθηκαν στο [40] έγινε εκτεταμένος πειραματισμός με διαφορετικά μεγέθη διανύσματος και πλέγματος για να δείξουμε ότι η χρήση πολύ

μεγάλων διανυσμάτων και χαρτών δεν συνεπάγεται αναγκαστικά καλύτερη ακρίβεια στην ταξινόμηση, ενώ υπαγορεύει σημαντικές απαιτήσεις σχετικά με τους πόρους και το χρόνο εκπαίδευσης.

Μια μεικτή προσέγγιση, που λειτουργεί τόσο με online όσο και με offline τρόπο, παρουσιάστηκε στο [63] για να αντιμετωπίσει την υψηλή διάσταση των διανυσμάτων, ενώ παράλληλα υπήρξε προσπάθεια για την ενίσχυση της απόδοσης του ταξινομητή. Στην παρούσα Διατριβή επιλέχθηκε η batch έκδοση του αλγόριθμου SOM, η οποία είναι σημαντικά ταχύτερη από την online έκδοση, ενώ χρησιμοποιήθηκε το σχήμα στάθμισης TF-IDF που αντιπροσωπεύει καλύτερα το σύνολο δεδομένων, ενώ με κατάλληλη επιλογή του μεγέθους των διανυσμάτων εξασφαλίστηκε ότι δεν θα συμπεριληφθούν μη σχετικά χαρακτηριστικά που μπορεί να μειώσουν την απόδοση του ταξινομητή. Το [59] επικεντρώνεται κυρίως στην ανάλυση της συν-αναφοράς, αλλά η μέθοδος είναι αρκετά γενική ώστε να εφαρμόζεται και σε άλλα σύνολα δεδομένων. Αυτή η εργασία ευθυγραμμίζεται με την προσέγγισή μας καθώς επιλέξαμε το SOM για να μπορέσουμε να διαχειριστούμε μικρότερα σύνολα δεδομένων που μπορεί να μην είναι καλά σχολιασμένα ή η κατανομή των δειγμάτων στις διαφορετικές κλάσεις να μην είναι ισορροπημένη.

Η ενσωμάτωση τεχνικών εξόρυξης κειμένου για την υπέρβαση των ελλείψεων του αρχικού αλγόριθμου SOM με επέκταση του χάρτη τόσο ιεραρχικά όσο και πλευρικά εξετάστηκε στο [64]. Ομοίως, στο [65] χρησιμοποιούνται επίσης τεχνικές εξόρυξης κειμένων που στοχεύουν στην ανάπτυξη ενός ανεξάρτητου από τη γλώσσα μοντέλου με την ενσωμάτωση δύο χαρτών SOM, δηλαδή ενός χάρτη σε επίπεδο λέξεων και ενός χάρτη σε επίπεδο εγγράφων. Στην περίπτωσή μας, αντί να αλλάξουμε τον βασικό αλγόριθμο SOM, αναπτύξαμε έναν ευφυή αλγόριθμο για την επιλογή των ετικετών.

Οι ConSOM (conceptual SOM), Fast SOM και η κυκλική παραλλαγή (cyclic variant) V-SOM [159] παρουσιάστηκαν στο [66]. Ο ConSOM στοχεύει στην αναγνώριση των μορφολογικών χαρακτηριστικών κάθε γλώσσας με βάση τις υποκείμενες οντολογικές γνώσεις [158], ενώ ο Fast SOM χρησιμοποίησε μια πρωτότυπη μέθοδο υπολογισμού των αντιπροσωπευτικών ιδιοτήτων και της ομοιότητας μεταξύ των εγγράφων για την επιτάχυνση της ταξινόμησης (clustering) μεγάλων συλλογών κειμένων. Σε αυτή τη Διατριβή χρησιμοποιήσαμε, όπως προαναφέραμε, την batch έκδοση του SOM, ακολουθώντας μια διαφορετική προσέγγιση για την κατασκευή των διανυσμάτων, με στόχο τη δημιουργία διανυσμάτων που θα μπορούσαν να διαχειριστούν από το SOM, καθώς ο αλγόριθμος δεν μπορεί να αντιμετωπίσει εγγενώς πολλαπλές ετικέτες που έχουν εκχωρηθεί στο ίδιο διάνυσμα.

Μια υβριδική προσέγγιση που χρησιμοποιεί τον αλγόριθμο Naive Bayes για τη δημιουργία των διανυσμάτων και τον SOM ως ένα πολυδιάστατο ταξινομητή μη εποπτευόμενης ΜΜ, έχει προταθεί στο [67], με βάση την ιδέα ότι ο συνδυασμός των δύο αλγορίθμων θα ενισχύσει την αποτελεσματικότητα της ταξινόμησης ενώ θα υπερνικήσει τις ελλείψεις του αλγορίθμου Naïve Bayes. Στη δική μας περίπτωση, χρησιμοποιήσαμε το SOM τόσο για τη μείωση των διαστάσεων των διανυσμάτων όσο και για την εκπαίδευση, ενώ το σχήμα στάθμισης TF-IDF χρησιμοποιήθηκε για την επιλογή των χαρακτηριστικών κατά τη διαδικασία κατασκευής των διανυσμάτων.

Εν κατακλείδι, η προτεινόμενη λύση στοχεύει στην αξιοποίηση των πλεονεκτημάτων των σύγχρονων προσεγγίσεων για την ταξινόμηση ενός συνόλου δεδομένων πολλαπλής ετικέτας, όσον αφορά στις διαδικασίες επιλογής χαρακτηριστικών αλλά και εκπαίδευσης του ταξινομητή, καθώς και στην απόδοση των ετικετών στα δεδομένα ελέγχου, με στόχο την ενίσχυση της απόδοσης του

ταξινομητή όσον αφορά την ακρίβεια, ελαχιστοποιώντας ταυτόχρονα το χρόνο επεξεργασίας και τους απαιτούμενους υπολογιστικούς πόρους. Συγκεκριμένα, η προσέγγισή μας ενσωματώνει μια συγκεκριμένη διαδικασία για την κατασκευή διανυσμάτων για την αντιμετώπιση των πολλαπλών ετικετών σε κάθε διάνυσμα, αποφεύγοντας την μετατροπή του προβλήματος σε ένα σύνολο δυαδικών προβλημάτων ταξινόμησης. Επιπλέον, η διαδικασία επιλογής χαρακτηριστικών έχει σχεδιαστεί ώστε να είναι σχετικά γενική, με βάση στατιστικά λέξεων και επομένως εφαρμόζεται εύκολα σε διαφορετικά σενάρια χρήσης. Χρησιμοποιούμε την batch έκδοση του αλγορίθμου SOM για την εκπαίδευση του ταξινομητή, ο οποίος θεωρείται πιο σταθερός και σημαντικά ταχύτερος από την αντίστοιχη online έκδοση και επιλέγουμε το μέγεθος του χάρτη με βάση μια ευρετική διαδικασία, όπως θα παρουσιαστεί λεπτομερώς παρακάτω. Τέλος, προτείνεται ένας έξυπνος αλγόριθμος για την απόδοση των ετικετών, ο οποίος λαμβάνει υπόψη τις ετικέτες που έχουν ανατεθεί στους γειτονικούς κόμβους για περαιτέρω βελτίωση της απόδοσης του ταξινομητή.

3.4. Περιγραφή του συνόλου δεδομένων

Το σύνολο δεδομένων Reuters [45] αποτελεί ένα από τα συνηθέστερα χρησιμοποιούμενα σύνολα δεδομένων για τη συγκριτική αξιολόγηση των προσεγγίσεων ταξινόμησης. Το αρχικό σύνολο δεδομένων (Reuters-22173) είναι μια συλλογή από 22173 άρθρα ειδήσεων από το Reuters που δημοσιεύτηκαν το έτος 1987. Το γεγονός ότι περιέχει έγγραφα χωρίς κλάσεις και τυπογραφικά λάθη οδήγησε στη δημιουργία του Reuters-21578. Επιπλέον, έχουν δημιουργηθεί υποσύνολα με διαφορετικά χαρακτηριστικά για τη διευκόλυνση της σύγκρισης των αποτελεσμάτων μεταξύ των διαφορετικών προσεγγίσεων.

Το υποσύνολο Mod Apte (ή RV1) [68], το οποία σύμφωνα με το [42] είναι το συνηθέστερα χρησιμοποιούμενο υποσύνολο του σώματος δεδομένων Reuters, αποτελείται από 12.902 έγγραφα για 90 κλάσεις, με 9.603 δείγματα για εκπαίδευση και 3.299 δείγματα για έλεγχο. Αυτά τα δείγματα θα μπορούσαν να μειωθούν περαιτέρω σε 7.769 δείγματα εκπαίδευσης και 3.019 δείγματα ελέγχου επιλέγοντας τις κλάσεις που περιέχουν τουλάχιστον ένα δείγμα εκπαίδευσης και ένα ελέγχου. Το υποσύνολο Apte 115’ περιέχει τουλάχιστον ένα δείγμα εκπαίδευσης για καθεμία από τις 115 κλάσεις, αλλά αυτό δεν ισχύει και για τα δείγματα ελέγχου. Επιπλέον, το υποσύνολο Apte 10’ έχει επίσης χρησιμοποιηθεί εκτενώς καθώς περιέχει τις δέκα κλάσεις που έχουν τον μεγαλύτερο αριθμό δειγμάτων από το Mod Apte. Σε αυτή την εργασία, χρησιμοποιήθηκε το υποσύνολο που περιέχει τουλάχιστον ένα δείγμα εκπαίδευσης και ελέγχου, που είναι διαθέσιμο μέσω του πακέτου Python Natural Language Toolkit (NLTK) [69]. Από αυτό το σύνολο δεδομένων, εφαρμόζοντας ορισμένα βήματα προεπεξεργασίας, όπως θα αναλυθεί στην επόμενη ενότητα, ανακτήθηκε επίσης το υποσύνολο Apte '10.

Στον Πίνακα 5 παρουσιάζεται η κατανομή μεταξύ των δειγμάτων εκπαίδευσης και ελέγχου για τις δέκα κλάσεις με τα περισσότερα δείγματα. Τα συνολικά δείγματα εκπαίδευσης για τις 10 πρώτες κλάσεις είναι 7192 ενώ τα αντίστοιχα δείγματα ελέγχου είναι 2787 και αντιπροσωπεύουν το 92% του συνολικού αριθμού των δειγμάτων εκπαίδευσης και ελέγχου.

Πίνακας 5. Κατανομή μεταξύ των δειγμάτων εκπαίδευσης και ελέγχου για τις 10 κλάσεις με τα περισσότερα δείγματα του συνόλου δεδομένων Reuters.

Κλάση	# δειγμάτων εκπαίδευσης	# δειγμάτων ελέγχου
EARN	2877	1087
ACQ	1650	719
MONEY-FX	538	179

GRAIN	433	149
CRUDE	389	189
TRADE	368	117
INTEREST	347	131
WHEAT	212	71
SHIP	197	89
CORN	181	56
Total	**7192**	**2787**

Πίνακας 6. Κατανομή μεταξύ του πλήθους των κλάσεων που ανατίθενται σε κάποιο έγγραφο και του αριθμού των εγγράφων.

# εγγράφων	# πλήθος κλάσεων
9160	1
1173	2
255	3
91	4
52	5
27	6
9	7
7	8
5	9
3	10
2	11
1	12
2	14
1	15

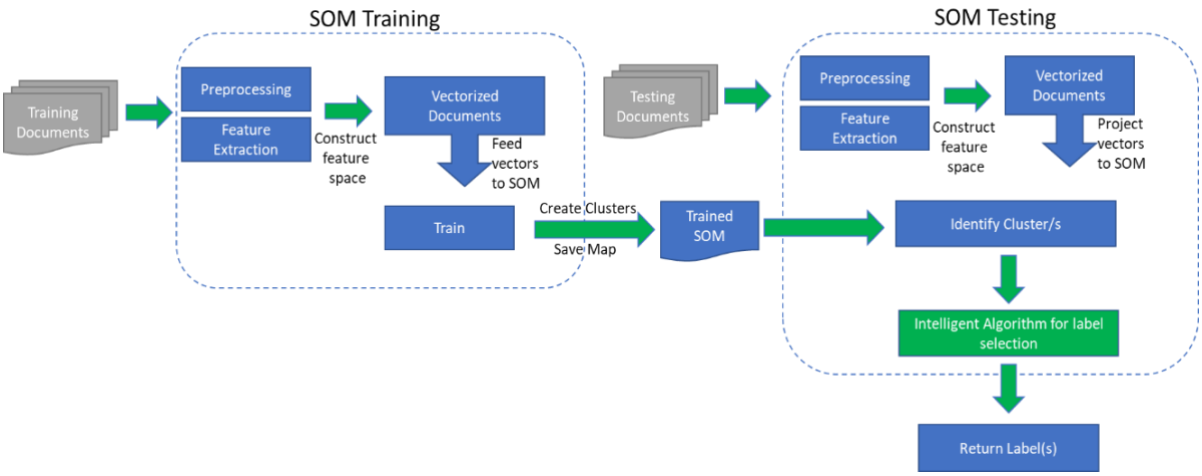
Επιπλέον, από τον Πίνακα 6 μπορεί να εξαχθεί ότι η πλειονότητα των εγγράφων ανήκει είτε σε μία είτε σε δύο κλάσεις. Ως εκ τούτου, μπορούμε με ασφάλεια να υποθέσουμε ότι οι δέκα κλάσεις αντιπροσωπεύουν την πλειονότητα του σώματος και συνεπώς η κατανομή των δειγμάτων στις κλάσεις είναι άνιση. Το γεγονός αυτό δημιουργεί προκλήσεις όσον αφορά στα μέτρα ποιότητας που θα χρησιμοποιηθούν για τη μέτρηση της απόδοσης του ταξινομητή, καθώς η ακρίβεια στις μικρότερες κλάσεις ενδέχεται να παρασυρθεί από την ακρίβεια στις μεγαλύτερες κλάσεις. Έτσι, τα μέτρα Μίκρο- και Μάκρο-μέσου όρου θα χρησιμοποιηθούν για τη μέτρηση της απόδοσης του ταξινομητή, όπως θα αναλυθεί στην επόμενη ενότητα.

3.5. Ταξινόμηση με χρήση Αυτό-Οργανούμενων χαρτών

Σε αυτή την ενότητα αναλύεται η προσέγγιση που σχεδιάσαμε και εφαρμόσαμε για την ταξινόμηση του σώματος δεδομένων Reuters χρησιμοποιώντας τον παραδοσιακό αλγόριθμο μη εποπτευόμενης ΜΜ, SOM. Όπως αναφέρθηκε προηγουμένως, αυτό είναι ένα πρόβλημα ταξινόμησης πολλαπλής ετικέτας. Ωστόσο, ο SOM δεν μπορεί να χειριστεί πολλαπλές ετικέτες για το ίδιο διάνυσμα εγγενώς. Έτσι, στην παρούσα Διατριβή προτείνουμε μια νέα προσέγγιση για το χειρισμό των πολλαπλών ετικετών που αντιστοιχούν σε κάθε διάνυσμα. Επιπλέον, η προτεινόμενη προσέγγιση στοχεύει στη χρήση ενός μικρότερου συνόλου χαρακτηριστικών για την κατασκευή των διανυσμάτων, χρησιμοποιώντας τόσο τετράγωνους όσο και ορθογώνιους χάρτες. Επιπλέον, παρουσιάζεται ένας έξυπνος αλγόριθμος για τη λήψη απόφασης σχετικά με τις ετικέτες που θα εκχωρηθούν σε ένα στοιχείο δεδομένων καθώς και για τον προσδιορισμό των καταλληλότερων

ετικετών για ένα στοιχείο όταν δεν υπάρχει καθόλου ετικέτα στον κόμβο. Η παρακάτω εικόνα απεικονίζει το διάγραμμα ροής της προτεινόμενης μεθόδου.

Η λεπτομερής επεξήγηση της λύσης περιγράφεται στις επόμενες υποενότητες. Γενικά, το αριστερό μέρος του διαγράμματος ροής απεικονίζει τη διαδικασία εκπαίδευσης ενώ το δεξί τη διαδικασία ελέγχου. Συγκεκριμένα, το σύνολο δεδομένων εκπαίδευσης δίδεται ως είσοδος στο σύστημα, περνά μια διαδικασία προεπεξεργασίας προκειμένου να εξαχθούν τα πιο σημαντικά χαρακτηριστικά που αποτελούν το χώρο των χαρακτηριστικών. Στη συνέχεια, τα διανύσματα που δημιουργούνται από την προηγούμενη διαδικασία τροφοδοτούνται στο SOM για να ξεκινήσει η διαδικασία εκπαίδευσης. Η έξοδος αυτής της διαδικασίας είναι ένας εκπαιδευμένος χάρτης, που δίδεται ως είσοδος μαζί με το σύνολο δεδομένων ελέγχου για να επιστραφούν οι κατάλληλες ετικέτες για το σύνολο δεδομένων ελέγχου. Θα πρέπει να σημειωθεί ότι τόσο τα διανύσματα του συνόλου εκπαίδευσης όσο και αυτά του συνόλου ελέγχου κατασκευάζονται χρησιμοποιώντας την ίδια διαδικασία που εξηγείται παρακάτω λεπτομερώς. Τέλος, τα διανύσματα των δεδομένων ελέγχου προβάλλονται επάνω στον εκπαιδευμένο SOM για να επιλεχθούν οι καταλληλότερες ετικέτες για κάθε ένα από αυτά, με βάση έναν έξυπνο αλγόριθμο, μια ακόμα συνεισφορά αυτής της Διατριβής.



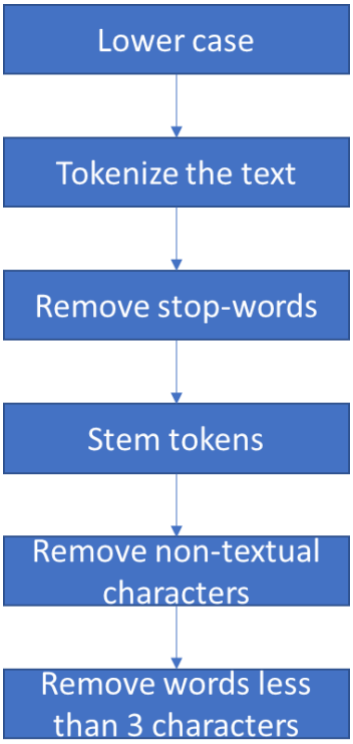
Εικόνα 4. Διάγραμμα ροής της προτεινόμενης μεθόδου. Το αριστερό μέρος αναπαριστά τη διαδικασία εκπαίδευσης ενώ το δεξί τη διαδικασία ελέγχου. Το αποτέλεσμα της διαδικασίας εκπαίδευσης είναι ένας εκπαιδευμένος SOM χάρτης, που δίδεται ως είσοδος μαζί με τα δεδομένα ελέγχου για να επιστραφούν τελικά οι κατάλληλες ετικέτες για αυτά με χρήση ενός έξυπνου αλγορίθμου για την επιλογή των ετικετών.

3.5.1. Προεπεξεργασία

Κατά την προεπεξεργασία, τα δεδομένα εισόδου μετατρέπονται σε μορφή που μπορεί να επεξεργαστεί το νευρωνικό δίκτυο SOM. Αυτό το βήμα είναι κοινό για την πλειονότητα των προσεγγίσεων που παρουσιάστηκαν προηγουμένως και συνήθως περιλαμβάνει βήματα όπως: την κατάτμηση του κειμένου, την απομάκρυνση stop-words και την εύρεση του λήμματος για κάθε λέξη (stemming). Στην προσέγγισή μας, αυτή η διαδικασία περιλαμβάνει έξι διαδοχικά βήματα όπως συνοψίζονται στην παρακάτω εικόνα.

Πιο συγκεκριμένα, το κείμενο μετατρέπεται αρχικά σε πεζά και στη συνέχεια γίνεται τμηματοποίηση έτσι ώστε να ανακτηθεί μια λίστα λέξεων για κάθε στοιχείο δεδομένων. Ως επόμενο βήμα, αφαιρούνται τα stop-words χρησιμοποιώντας την αντίστοιχη λειτουργία που είναι διαθέσιμη μέσω του πακέτου NLTK της Python. Η εύρεση του λήμματος (stemming) συνήθως αναφέρεται σε μια ευρετική διαδικασία που στοχεύει στο να μετασχηματίσει παραγωγικά σχετικές μορφές μιας λέξης σε μια κοινή μορφή βάσης, με την απογύμνωση λέξεων από τις καταλήξεις τους [70]. Για την

απομάκρυνση των καταλήξεων, χρησιμοποιήθηκε ο αλγόριθμος του Porter [29], καθώς έχει αποδειχθεί εμπειρικά πολύ αποτελεσματικός, και επίσης εκτελεί λιγότερο επιθετική αποκοπή από τον παλαιότερο αλγόριθμο του Lovins [30]. Τα τελευταία δύο βήματα περιλαμβάνουν την αφαίρεση οποιουδήποτε χαρακτήρα που δεν αποτελεί κείμενο, όπως αριθμητικών χαρακτήρων και σημείων στίξης αλλά και αφαίρεση "πολύ μικρών" λέξεων (με λιγότερους από τρεις χαρακτήρες). Με δεδομένο ότι στοχεύουμε στην παροχή μιας γενικής διαδικασίας, αυτή η παράμετρος μπορεί να τροποποιηθεί κατά την αρχική διαμόρφωση. Σημειώνεται ότι το τελευταίο βήμα θα πρέπει να πραγματοποιηθεί μετά την απομάκρυνση των καταλήξεων. Η έξοδος αυτής της διαδικασίας είναι ένα λεξιλόγιο όρων για κάθε στοιχείο δεδομένων. Στην επόμενη υποενότητα παρουσιάζεται το σχήμα στάθμισης που χρησιμοποιήθηκε για την επιλογή χαρακτηριστικών καθώς και η κατασκευή των διανυσμάτων που αντιπροσωπεύουν κάθε στοιχείο δεδομένων.



Εικόνα 5. Τα βήματα της διαδικασίας προεπεξεργασίας των εγγράφων του συνόλου δεδομένων Reuters.

3.5.2. Επιλογή Χαρακτηριστικών – Κατασκευή Διανυσμάτων

Αυτό το βήμα στοχεύει στην επιλογή των πιο χρήσιμων χαρακτηριστικών για την αναπαράσταση ενός εγγράφου με τη μορφή διανύσματος που θα τροφοδοτηθεί στη συνέχεια στο SOM. Το μέγεθος των διανυσμάτων εξαρτάται από τον αριθμό των επιλεγμένων χαρακτηριστικών που αποτελούν το χώρο των χαρακτηριστικών. Είναι προφανές ότι όσο υψηλότερη είναι η διάσταση των διανυσμάτων, τόσο μεγαλύτερες είναι οι υπολογιστικές απαιτήσεις της εργασίας ταξινόμησης.

Έστω $S = \{d_1, d_2, \dots, d_K\}$, η συλλογή των εγγράφων του συνόλου δεδομένων Reuters. Κάθε έγγραφο που ανήκει στη συλλογή S αναπαρίσταται από ένα n-διάστατο διάνυσμα της ακόλουθης μορφής: $d_i = \{w_{i1}, w_{i2}, \dots, w_{in}\}$, όπου w_{in} είναι το βάρος του εγγράφου d_i για το χαρακτηριστικό n. Τα βάρη υπολογίζονται με βάση στατιστικά λέξεων. Πιο συγκεκριμένα το σχήμα στάθμισης $TF - IDF$ χρησιμοποιήθηκε για την κατασκευή των διανυσμάτων. Το $TF - IDF$ είναι ένα σύνθετο σχήμα στάθμισης που αποτελείται από τα μέτρα Συχνότητας Όρων (Term Frequency - TF) και Αντίστροφης

Συχνότητας Εγγράφου (Inverse Document Frequency - IDF). Το TF ενός όρου t για ένα έγγραφο d, είναι απλά το πλήθος των εμφανίσεων του όρου t στο d. Το μέτρο IDF καθορίζει πόσο σπάνια είναι μια λέξη ανάμεσα σε όλα τα έγγραφα μιας συλλογής. Ο ορισμός του μέτρου IDF δίνεται παρακάτω:

$$IDF(t, D) = log \frac{N}{|\{d \in D : t \in d\}|} \qquad (1)$$

Στην ανωτέρω εξίσωση, το N υποδηλώνει τον συνολικό αριθμό των εγγράφων της συλλογής, ενώ ο παρονομαστής δηλώνει τον αριθμό των εγγράφων όπου εμφανίζεται ο όρος t. Το $TF - IDF$ είναι το προϊόν των μέτρων TF και IDF όπως ορίζεται παρακάτω.

$$TF - IDF(t, d) = TF(t, d) * IDF(t, D) \qquad (2)$$

Τα διανύσματα κατασκευάστηκαν χρησιμοποιώντας τον TF-IDF Vectorizer διαθέσιμο μέσω του πακέτου Python scikit-learn [71]. Μετά τον υπολογισμό των τιμών $TF - IDF$ για όλες τις λέξεις στο σύνολο δεδομένων, οι x λέξεις με τις υψηλότερες βαθμολογίες επιλέγονται για την κατασκευή των διανυσμάτων. Όπως θα αναλυθεί λεπτομερώς σε επόμενη ενότητα το x κυμαίνεται από 250-3000 λέξεις. Ένα παράδειγμα των χαρακτηριστικών που επιλέγονται μέσα από αυτή τη διαδικασία δίνεται στην Εικόνα 6 όπου επιλέχθηκαν μόνο μεμονωμένες λέξεις (1-grams), ενώ στην Εικόνα 7 επιλέχθηκαν και συμφραζόμενα των λέξεων (n-grams). Πειράματα διεξήχθησαν χρησιμοποιώντας και τις δύο ρυθμίσεις για να εκτιμηθεί η επίπτωση των συμφραζόμενων λέξεων στην ταξινόμηση. Μετά την κατασκευή του, κάθε διάνυσμα κανονικοποιήθηκε χρησιμοποιώντας το πρότυπο "l2 norm" ή αλλιώς την Ευκλείδεια απόσταση. Στο τελευταίο βήμα, ετικέτες που ανήκουν στο αντίστοιχο δείγμα δεδομένων ανατέθηκαν σε κάθε διάνυσμα.

Το διάνυσμα που φέρει πολλαπλές ετικέτες απεικονίζεται στον Πίνακα 7. Κάθε διάνυσμα αποτελείται από ένα σύνολο χαρακτηριστικών $F = \{f_1, f_2, \dots, f_n\}$ και από ένα σύνολο ετικετών $L = \{l_1, l_2, \dots, l_k\}$. Όπως αναφέρθηκε προηγουμένως, τα προβλήματα ταξινόμησης πολλαπλής ετικέτας μετατρέπονται συνήθως σε ένα σύνολο δυαδικών προβλημάτων ταξινόμησης. Αντί της κατασκευής πολλαπλών δυαδικών ταξινομητών, προτείνουμε τη μετατροπή του προβλήματος σε ένα πρόβλημα πολλαπλών κλάσεων και έναν ευφυή αλγόριθμο για την ανάθεση των ετικετών. Κάθε διάνυσμα που έχει περισσότερες από μία ετικέτες χωρίζεται σε περισσότερα διανύσματα καθένα από τα οποία έχει μία ετικέτα από το αρχικό σύνολο ετικετών L, δηλαδή ένα διάνυσμα με 3 διαφορετικές ετικέτες θα γίνει 3 διανύσματα που αντιπροσωπεύονται από το ίδιο σύνολο χαρακτηριστικών F αλλά έχουν μία ετικέτα. Αυτός ο μετασχηματισμός απεικονίζεται στον Πίνακα 8, λαμβάνοντας υπόψη το διάνυσμα που παρουσιάζεται στον Πίνακα 7.

```
'crude',
'csr',
'cubic',
'cumul',
'curb',
'currenc',
'current',
'custom',
'cut',
'cyacq',
'cyclop',
'daili',
'dairi',
'damag',
'dart',
'data',
'date',
'dauster',
'david',
'day',
'deal',
'dealer',
```

Εικόνα 6. Έξοδος του συστήματος για την επιλογή χαρακτηριστικών με χρήση 1-grams.

```
'crude',
'crude oil',
'crude oil price',
'csr',
'cubic',
'cubic feet',
'cumul',
'curb',
'currenc',
'current',
'current account',
'current account deficit',
'current level',
'current year',
'custom',
'custom repurchas',
'custom repurchas agreement',
'cut',
'cyacq',
'cyclop',
'daili',
'dairi',
'damag',
'dart',
'data',
'date',
```

Εικόνα 7. Έξοδος του συστήματος για την επιλογή χαρακτηριστικών με χρήση n-grams.

Πίνακας 7. Το διάνυσμα με πολλαπλές ετικέτες

Sample1	$f_1, f_2, \dots, f_n$	l_1, l_2, l_k
Sample2	$f_1, f_2, \dots, f_n$	l_1, l_3, l_{k-1}

Πίνακας 8. Το διάνυσμα που προκύπτει μετά το διαχωρισμό των ετικετών

Sample1	$f_1, f_2, \dots, f_n$	l_1
Sample1	$f_1, f_2, \dots, f_n$	l_2
Sample1	$f_1, f_2, \dots, f_n$	l_k
Sample2	$f_1, f_2, \dots, f_n$	l_1
Sample2	$f_1, f_2, \dots, f_n$	l_3
Sample2	$f_1, f_2, \dots, f_n$	l_{k-1}

3.5.3. Ο αλγόριθμος SOM

Ο SOM είναι ένας πίνακας νευρώνων δύο διαστάσεων όπου κάθε νευρώνας αναπαρίσταται από ένα ειδικό διάνυσμα (codebook vector) [40]. Στο SOM μοντέλο, ένα συγκεκριμένο χαρακτηριστικό του διανύσματος εισόδου προβάλλεται σε μια συγκεκριμένη χωρική θέση ενός νευρώνα εξόδου στον χάρτη. Το κύριο πλεονέκτημα για την επιλογή του SOM για εργασίες ταξινόμησης ή ομαδοποίησης είναι ότι το σύστημα έχει τη δυνατότητα να μάθει από τα διανύσματα που διέρχονται από τους νευρώνες εισόδου με έναν Αυτό-οργανούμενο τρόπο. Κάθε νευρώνας έχει ένα βάρος που αλλάζει σε κάθε φάση εκπαίδευσης ανάλογα με το διάνυσμα που δόθηκε ως είσοδος, με βάση ένα μέτρο ομοιότητας (συνήθως την Ευκλείδεια απόσταση). Σε ένα σενάριο βέλτιστου διαχωρισμού, ελαχιστοποιείται η απόσταση από την καλύτερη μονάδα αντιστοίχισης (Best Matching Unit - BMU εφεξής) για κάθε στοιχείο δεδομένων εισόδου, δηλ. ελαχιστοποιείται το Μέσο Σφάλμα Ποσοτικοποίησης (Mean Quantization Error). Ας υποθέσουμε ότι τα στοιχεία δεδομένων εισόδου συνιστούν ένα σύνολο n-διάστατων διανυσμάτων x. Τα codebook διανύσματα αναπαρίστανται ως

m_i, όπου i είναι ο δείκτης του κάθε codebook διανύσματος. Ο αριθμός αυτού του δείκτη εξαρτάται από το επιλεγμένο μέγεθος πλέγματος. Ο δείκτης του εκάστοτε BMU, $c(x)$, μπορεί να βρεθεί με την ακόλουθη εξίσωση:

$$c(x) = arg\min_i\{\|x - m_i\|\} \qquad (3)$$

Στη συνέχεια, το Μέσο Σφάλμα Ποσοτικοποίησης (QE εφεξής) ορίζεται από την ακόλουθη εξίσωση, όπου $p(x)$ είναι η συνάρτηση πυκνότητας πιθανότητας του x. Το QE υπολογίζεται προσδιορίζοντας τη μέση απόσταση των διανυσμάτων από τα codebook διανύσματα από τα οποία αντιπροσωπεύονται.

$$QE = \int_v \|x - m_c\|^2 \, p(x) dV \qquad (4)$$

Ένα άλλο, ποιοτικό μέτρο για την αξιολόγηση της προβολής των διανυσμάτων στον χάρτη είναι το Τοπογραφικό Σφάλμα (Topographic Error - TE εφεξής), που υποδηλώνεται από τον Αλγόριθμο 1, το οποίο μετρά τη διατήρηση της τοπολογίας για ένα δεδομένο σύνολο διανυσμάτων εισόδου. Ο υπολογισμός του είναι αρκετά απλός καθώς για κάθε διάνυσμα εισόδου υπολογίζεται η 1η και η 2η BMU. Το επόμενο βήμα είναι να ελεγχθεί εάν γειτνιάζουν επάνω στο χάρτη. Τέλος, υπολογίζεται το ποσοστό των δεδομένων εισόδου που οι BMU τους δεν γειτνιάζουν σε σύγκριση με τον συνολικό αριθμό των δεδομένων εισόδου και το τελικό αποτέλεσμα κανονικοποιείται στο διάστημα 0-1. Όσο πλησιέστερα είναι το TE στο μηδέν τόσο καλύτερο είναι το μοντέλο, καθώς υποδηλώνει άψογη διατήρηση της τοπολογίας. Σύμφωνα με το [60], υπάρχει ένα σημείο καμπής μεταξύ αυτών των δύο μέτρων καθώς η αύξηση της ποιότητας προβολής συνήθως μειώνει τις ιδιότητες προβολής. Μια άλλη υπόθεση που βασίζεται σε αυτή την εργασία είναι ότι τα μεγαλύτερα μεγέθη πλέγματος τείνουν να μειώνουν μονότονα το QE.

Αλγόριθμος 1: Υπολογισμός TE

Βήμα 1:	**Για κάθε** διάνυσμα εισόδου βρες την 1η και τη 2η BMU
Βήμα 2:	Έλεγξε εάν οι BMU είναι γειτονικές επάνω στο χάρτη.
Βήμα 3:	Υπολόγισε το ποσοστό των διανυσμάτων που τα BMU τους δεν είναι γειτονικά.
Βήμα 4:	Κανονικοποίησε την τιμή στην περιοχή 0-1.

Σε αυτή την εργασία χρησιμοποιείται η βιβλιοθήκη SOMPY που είναι γραμμένη σε Python [72] και βασίζεται στην αρχική βιβλιοθήκη SOMToolbox του λογισμικού Matlab. Η εργασία που παρουσιάστηκε στο [40] έδωσε την έμπνευση και την αιτιολόγηση για αρκετές επιλογές μοντελοποίησης που έχουμε κάνει. Δεδομένου ότι τα στοιχεία δεδομένων εισόδου στην περίπτωσή μας συνιστούν ένα πεπερασμένο σύνολο, επιλέχθηκε η batch έκδοση του SOM. Σε αυτή την έκδοση, όλα τα στοιχεία δεδομένων εισόδου δίδονται στον αλγόριθμο ως μια ενιαία παρτίδα (batch) και όλα τα μοντέλα ενημερώνονται σε μία ενιαία ταυτόχρονη λειτουργία. Μια ευθεία επίδραση της επιλογής αυτού του υπολογισμού, αντί για την επαναληπτική αναδρομική διαδικασία, είναι ότι δεν υπάρχει ανάγκη για την υπέρ-παράμετρο ρυθμού μάθησης (learning-rate) καθιστώντας έτσι τη σύγκλιση ταχύτερη.

Μια άλλη απόφαση μοντελοποίησης που έχει πρωταρχική σημασία ήταν να αποφασιστεί πώς θα αρχικοποιηθούν τα μοντέλα. Σύμφωνα με το [73], υπάρχουν δύο διαφορετικοί τύποι

αρχικοποίησης, δηλαδή ο τυχαίος (random) και ο γραμμικός (linear). Ο γραμμικός είναι προτιμότερος καθώς δείχνει ταχύτερη σύγκλιση. Στην πρώτη περίπτωση, το m_i, επιλέγεται τυχαία, έτσι ώστε να μην είναι γνωστό τίποτα για τα δεδομένα εισόδου, ενώ στην τελευταία περίπτωση οι αρχικές τιμές επιλέγονται ως μια δισδιάστατη αλληλουχία διανυσμάτων που λαμβάνονται κατά μήκος μιας 'ευθείας' (hyperplane) που καλύπτεται από τις δύο μεγαλύτερες κύριες συνιστώσες του x. Σε αυτή την περίπτωση, τα διανύσματα codebook αρχικοποιούνται ώστε να βρίσκονται στον ίδιο χώρο που καλύπτεται από δύο ιδιοδιανύσματα που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές των δεδομένων εισόδου. Αυτό έχει ως αποτέλεσμα τον προσανατολισμό του SOM στην ίδια κατεύθυνση με τα δεδομένα που παίζουν τον σημαντικότερο ρόλο. Χρησιμοποιώντας τα διανύσματα που κατασκευάστηκαν στην προηγούμενη φάση, το μοντέλο SOM αρχικοποιείται χρησιμοποιώντας συγκεκριμένες παραμέτρους που δικαιολογούνται από τις επιλογές μοντελοποίησης που αναλύθηκαν παραπάνω. Συγκεκριμένα, **χρησιμοποιούμε τη μέθοδο υπολογισμού batch με γραμμική αρχικοποίηση και κανονικοποίηση διακύμανσης για τα διανύσματα εισόδου**. Επιπλέον, πριν από τη φάση της εκπαίδευσης, οι ετικέτες για κάθε δείγμα που θα τροφοδοτηθεί στο SOM είναι εκείνες που κατασκευάστηκαν στην προηγούμενη φάση. Τα βήματα που ακολουθήσαμε για τη διαδικασία εκπαίδευσης υποδηλώνονται στον Αλγόριθμο 2, ενώ η διαδικασία εκπαίδευσης απεικονίζεται στην επόμενη εικόνα. Αρχικά, κάθε κόμβος στο χάρτη συσχετίζεται με ένα codebook διάνυσμα. Στη συνέχεια, για κάθε διάνυσμα εκπαίδευσης, βρίσκεται η BMU το βάρος της οποίας είναι πιο κοντά στο διάνυσμα με βάση την Ευκλείδεια απόσταση. Τέλος, ενημερώνονται οι τιμές των codebook διανυσμάτων που βρίσκονται στη γειτονιά της BMU. Αυτό απεικονίζεται στην Εικόνα 8 καθώς οι γειτονικοί κόμβοι έρχονται πιο κοντά στη BMU (σημειωμένο με κόκκινο χρώμα). Αυτή η διαδικασία επαναλαμβάνεται για όλες τις εποχές εκπαίδευσης.

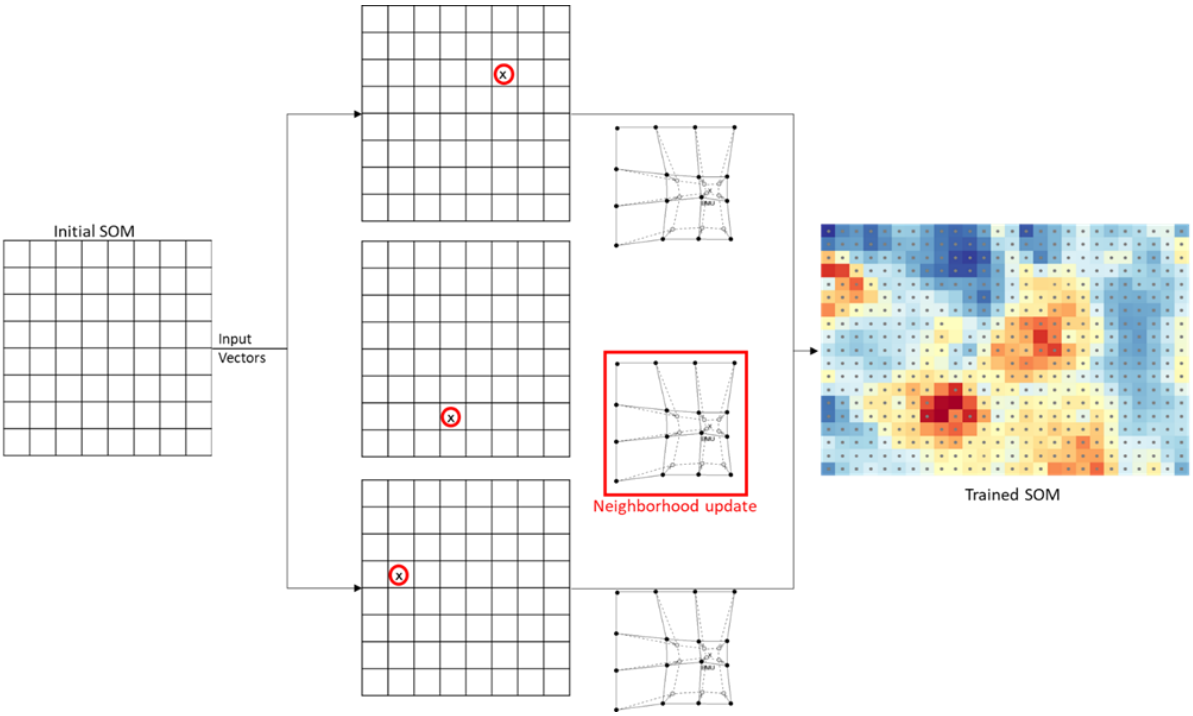
Με τη διαδικασία που αναφέρεται παρακάτω, το νευρωνικό δίκτυο SOM διαμορφώνει ένα ελαστικό δίκτυο που διπλώνει με βάση το σχήμα των δεδομένων εισόδου κατά τη διάρκεια της διαδικασίας εκπαίδευσης. Επιπλέον, τα διανύσματα codebook τείνουν να μετακινούνται εκεί όπου τα δεδομένα είναι πυκνά, ενώ μόνο λίγα από αυτά βρίσκονται εκεί όπου τα δεδομένα είναι αραιά. Με αυτό τον τρόπο, το δίκτυο τείνει να προσεγγίζει τη συνάρτηση πυκνότητας πιθανότητας των δεδομένων εισόδου. Μετά την ολοκλήρωση της φάσης εκπαίδευσης, κάθε BMU συνδέεται με ένα σύνολο ετικετών που αντιστοιχούν στα δείγματα εκπαίδευσης που εκχωρούνται σε κάθε κόμβο στον χάρτη. Σε κάθε BMU στο SOM μπορεί να έχει αντιστοιχιστεί μία ή περισσότερες ετικέτες. Το επόμενο βήμα είναι η προβολή των δειγμάτων ελέγχου στο χάρτη και ο υπολογισμός της BMU για καθένα από αυτά. Μετά την προβολή, κάθε διάνυσμα από το υποσύνολο ελέγχου συσχετίζεται με ένα σύνολο ετικετών που βασίζονται στην BMU που ανατέθηκε επάνω στο χάρτη.

Αλγόριθμος 2: Η διαδικασία εκπαίδευσης (batch) του SOM

Βήμα 1:	Αρχικοποίησε τον αριθμό εποχής εκπαίδευσης στο 0.
Βήμα 2:	Σύνδεσε κάθε κόμβο i του χάρτη χρησιμοποιώντας ένα codebook διάνυσμα m_i, αρχικοποιημένο χρησιμοποιώντας τη γραμμική μέθοδο.
Βήμα 3:	Όρισε μια λίστα για κάθε κόμβο στο χάρτη που θα περιέχει συγκεκριμένα διανύσματα εισόδου $x(t)$ που σχετίζονται με κάθε κόμβο.
Βήμα 4:	Επίλεξε ένα διάνυσμα εκπαίδευσης από το σύνολο $x(t)$.
Βήμα 5:	Για το επιλεγμένο διάνυσμα βρες τη BMU k της οποίας το βάρος είναι πιο κοντά στο επιλεγμένο διάνυσμα με χρήση της Ευκλείδειας απόστασης (Εξίσωση (3)).
Βήμα 6:	Δημιούργησε ένα αντίγραφο του $x(t)$ στην υπο-λίστα που σχετίζεται με τον επιλεγμένο νευρώνα.
Βήμα 7:	Επανέλαβε τα βήματα 2-6 μέχρι να επιλεχθούν όλα τα διανύσματα εκπαίδευσης.

Βήμα 8: Ενημέρωσε των τιμών των codebook διανυσμάτων στην γειτονιά του νευρώνα που κερδίζει.

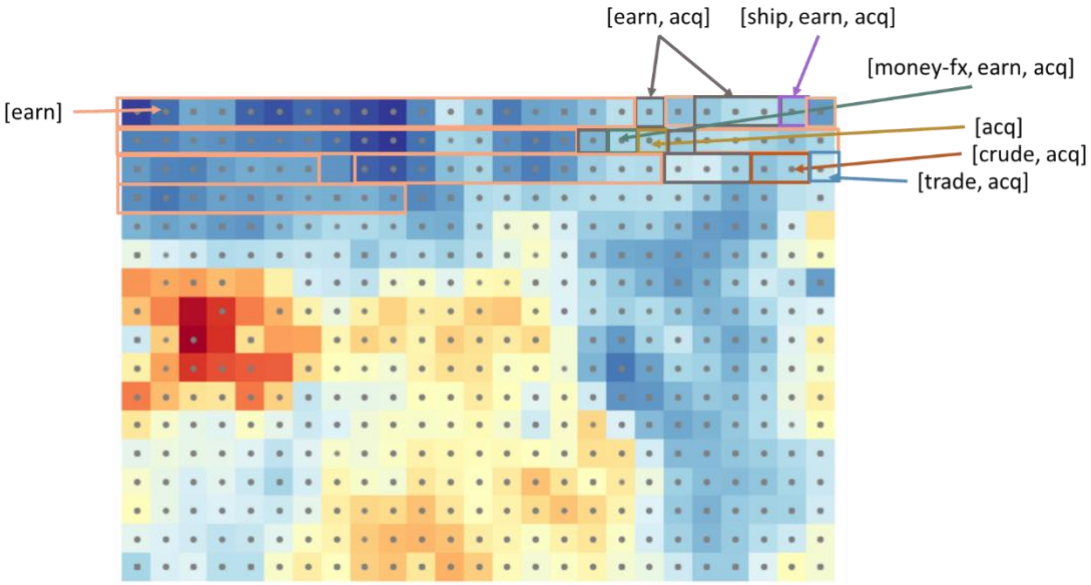
Βήμα 9: Αύξησε του αριθμού εποχής εκπαίδευσης. Εάν η εποχή εκπαίδευσης έχει φτάσει το μέγιστο προκαθορισμένο αριθμό εποχών εκπαίδευσης σταμάτα τη διαδικασία εκπαίδευσης, αλλιώς πήγαινε στο Βήμα 4.



Εικόνα 8. Η διαδικασία εκπαίδευσης SOM και ο χάρτης που προκύπτει. Για κάθε ένα από τα διανύσματα εισόδου βρίσκεται η BMU (επισημειωμένη με x). Σε κάθε επανάληψη, στη διάρκεια της εκπαίδευσης, οι γειτονικοί κόμβοι μετακινούνται πιο κοντά στη BMU. Ο εκπαιδευμένος SOM απεικονίζεται ως U-Matrix για να επισημανθούν οι αποστάσεις μεταξύ των γειτονικών κόμβων και να προαχθεί η παρατήρηση της δομής των συστάδων (clusters): υψηλές τιμές (σκουρότερα χρώματα) καταδεικνύουν τα όρια κάθε συστάδας ενώ ομοιόμορφες περιοχές παρόμοια χρωματισμένες καταδεικνύουν τις ίδιες τις συστάδες.

Στην παρακάτω εικόνα παρουσιάζεται ένα παράδειγμα όπου εμφανίζονται οι ετικέτες για κάποιους κόμβους μετά την ολοκλήρωση της φάσης εκπαίδευσης. Όπως είναι φανερό από την παρακάτω εικόνα οι ετικέτες δημιουργούν συστάδες από κόμβους όπου έγγραφα που ανήκουν στις ίδιες κλάσεις ομαδοποιούνται. Για παράδειγμα, αν ένα έγγραφο ελέγχου τοποθετηθεί στην πάνω αριστερή γωνία του χάρτη τότε με βάση την ανάθεση των ετικετών που βλέπουμε στην εικόνα, το έγγραφο αυτό πιθανότατα θα ανήκει στην κλάση ‘earn’.

Έχουν προταθεί διαφορετικές προσεγγίσεις σχετικά με τον τρόπο επιλογής των ετικετών για τη ανάθεση στα δείγματα ελέγχου. Στην πλειονότητα των προσεγγίσεων είτε αναθέτουν στο δείγμα ελέγχου όλες τις ετικέτες που υπάρχουν στην BMU είτε ελέγχουν την πλειοψηφία για να επιλέξουν τις πιο συχνές. Ωστόσο, οι προσεγγίσεις αυτές υστερούν, καθώς δεν λαμβάνουν υπόψη τους γειτονικούς κόμβους του χάρτη. Εδώ προτείνουμε έναν ευφυή αλγόριθμο για την επιλογή των ετικετών με βάση μια ιδιότητα γειτνίασης, όπως θα αναλυθεί στην επόμενη υποενότητα.



Εικόνα 9. Παράδειγμα εκπαιδευμένου χάρτη και των ετικετών που έχουν ανατεθεί σε κάθε κόμβο.

3.5.4. INTERSECT - INTelligEnt algorithm for label SelECTion

Όπως αναφέρεται στην παραπάνω ενότητα, η διαδικασία επιλογής ετικετών για τα δείγματα ελέγχου συνήθως εκτείνεται σε δύο διαφορετικές προσεγγίσεις που παρουσιάζουν συγκεκριμένες ελλείψεις. Ως εκ τούτου, προτείνουμε μια νέα ευρετική προσέγγιση για την επιλογή ετικετών, προκειμένου να ξεπεραστούν οι περιορισμοί των υφιστάμενων προσεγγίσεων. Ο προτεινόμενος αλγόριθμος έχει σχεδιαστεί ειδικά για σύνολα δεδομένων πολλαπλής ετικέτας, καθώς παίρνει τις κοντινότερες BMU με βάση την επιλεγμένη επάνω στο χάρτη, χρησιμοποιώντας την ευκλείδεια απόσταση και αγνοεί τις ετικέτες που δεν είναι κοινές σε όλους τους γειτονικούς κόμβους. Η λογική πίσω από αυτόν τον ευρετικό αλγόριθμο βασίζεται στην παρατήρηση ότι οι BMU που ανήκουν στο ίδιο σύμπλεγμα τείνουν να μοιράζονται κοινές ετικέτες. Ως εκ τούτου, αντί να χρησιμοποιήσουμε μόνο τις ετικέτες του ίδιου του κόμβου για τη διαδικασία επισήμανσης, ο αλγόριθμός μας επεκτείνεται στους γειτονικούς κόμβους του χάρτη για να αποφασίσει τις ετικέτες, εκμεταλλευόμενος έτσι τις ιδιότητες γειτνίασης των κόμβων για να μην λάβει υπόψη τις ετικέτες που μπορεί να μην είναι σχετικές με το επιλεγμένο δείγμα ελέγχου. Θα πρέπει να σημειωθεί ότι ο ίδιος αλγόριθμος χρησιμοποιείται επίσης για την ανάθεση ετικετών σε κόμβους χωρίς πρόβλεψη, δηλ. κόμβους που δεν έχει ανατεθεί κάποιο σχετικό διάνυσμα εκπαίδευσης. Τα βήματα που εκτελούνται για την επιλογή των ετικετών για κάθε δείγμα ελέγχου περιγράφονται στον Αλγόριθμο 3. Η απόδοση του προτεινόμενου ευρετικού αλγορίθμου αποδεικνύεται στα πειράματα που παρουσιάζονται λεπτομερώς στην επόμενη ενότητα.

Αλγόριθμος 3: INTERSECT για την επιλογή ετικέτας

Βήμα 1:	Επίλεξε τυχαία ένα δείγμα από το υποσύνολο ελέγχου, βρες τη BMU και αποθήκευσε τις ετικέτες
Βήμα 2:	Βρες τους n-κοντινότερους κόμβους και αποθήκευσε τις ετικέτες
Βήμα 3:	Βρες την τομή των δύο συνόλων
Βήμα 4:	Ανάθεσε τις ετικέτες του νέου συνόλου στο δείγμα ελέγχου
Βήμα 5:	**Εάν** δεν υπάρχουν άλλα αντικείμενα στο υποσύνολο ελέγχου σταμάτα τη διαδικασία ανάθεσης ετικετών **Αλλιώς** πήγαινε στο Βήμα 1

3.6. Πειραματική Αξιολόγηση

Ο στόχος αυτής της διαδικασίας είναι να επικυρωθεί η ακρίβεια ταξινόμησης της προτεινόμενης προσέγγισης για το επιλεγμένο σύνολο δεδομένων. Ως εκ τούτου, πραγματοποιήθηκαν εκτεταμένα πειράματα με διαφορετικές επιλογές διαμόρφωσης, ενώ χρησιμοποιήθηκαν τόσο τετράγωνοι όσο και ορθογώνιοι χάρτες για να εκτιμηθεί η επίπτωση της τοπολογίας του χάρτη στην απόδοση του ταξινομητή. Τα επιλεγμένα μέτρα αξιολόγησης, ο αλγόριθμος για τον υπολογισμό του μεγέθους του SOM, ο τρόπος διεξαγωγής των πειραμάτων, καθώς και τα αποτελέσματα που προέκυψαν από αυτά, περιγράφονται λεπτομερώς στις επόμενες υποενότητες.

3.6.1. Μέτρα Αξιολόγησης Απόδοσης

Τα πιο συχνά χρησιμοποιούμενα μέτρα για την αξιολόγηση της απόδοσης ενός ταξινομητή είναι η Ακρίβεια ($Precision$), η Ανάκτηση ($Recall$) και η βαθμολογία F1 ($F1\_score$). Καθώς εστιάζουμε σε μια ταξινόμηση πολλαπλής ετικέτας, τα μέτρα ποιότητας που θα χρησιμοποιηθούν πρέπει να είναι είτε ανά κλάση είτε συνολικά, καθώς η απόδοση των ταξινομητών σε ολόκληρο το σύνολο δεδομένων μπορεί να είναι παραπλανητική, θεωρώντας την απόδοση των λιγότερο συχνών κλάσεων ίδια με αυτή των μεγαλύτερων. Για να αποφευχθεί αυτό, χρησιμοποιήθηκαν τα μέτρα ακριβείας Μίκρο- και Μάκρο-μέσου όρου, με τη χρήση του πακέτου scikit-learn της Python [74], [75]. Ο τυπικός ορισμός των μέτρων ακρίβειας δίνεται στις παρακάτω εξισώσεις.

Η Ακρίβεια ($Precision$) υπολογίζεται διαιρώντας τον αριθμό των πραγματικών θετικών (True Positives - TP) με το άθροισμα του TP και τον αριθμό των ψευδών θετικών (False Positives - FP).

$$Precision = \frac{TP}{TP + FP} \qquad (5)$$

Η Ανάκληση ($Recall$) υπολογίζεται διαιρώντας τον αριθμό των σωστά ταξινομημένων στοιχείων δεδομένων TP με το άθροισμα $TP + FN$, όπου FN τα στοιχεία δεδομένων που ταξινομήθηκαν λανθασμένα ότι δεν ανήκουν στην κλάση.

$$Recall = \frac{TP}{TP + FN} \qquad (6)$$

Η βαθμολογία F1 ($F1\_score$) βασίζεται στα μέτρα ακρίβειας και ανάκλησης όπως αυτά ορίζονται παραπάνω. Πρέπει να σημειωθεί ότι από πλευρά ακρίβειας είτε Μίκρο- είτε Μάκρο- μπορεί να χρησιμοποιηθεί για τη μέτρηση της βαθμολογίας F1. Όσο υψηλότερη είναι η βαθμολογία, τόσο καλύτερη είναι η απόδοση ταξινόμησης. Η βαθμολογία F1 υπολογίζεται με χρήση της παρακάτω εξίσωσης.

$$F1\_score = 2\frac{Precision * Recall}{Precision + Recall} \qquad (7)$$

Ακρίβεια Μίκρο-μέσου όρου: Κάθε πίνακας περιεχομένων στο σύνολο δεδομένων είναι εξίσου σημαντικός. Οι πιο κοινές κλάσεις έχουν μεγαλύτερη επιρροή στη συνολική ποιότητα από τις μικρότερες [99]. Σε αυτήν την περίπτωση, η ακρίβεια είναι ίση με την ανάκληση και είναι επίσης ίση με την αρμονική μέση τιμή τους (harmonic mean). Η ακρίβεια Μίκρο-μέσου όρου αντιπροσωπεύει επίσης τη συνολική ακρίβεια του ταξινομητή και υπολογίζεται από τον αριθμό των σωστά

ταξινομημένων στοιχείων δεδομένων (TP) και τον συνολικό αριθμό στοιχείων δεδομένων n, όπως εκφράζεται στην Εξίσωση 8.

$$Micro\ Average\ Precision = \frac{TP}{n} \qquad (8)$$

Ακρίβεια Μάκρο-μέσου όρου: Σε αυτή την περίπτωση όλες οι κλάσεις είναι εξίσου σημαντικές. Η ποιότητα για κάθε κλάση υπολογίζεται ανεξάρτητα και στη συνέχεια αναφέρεται ο μέσος όρος όλων των κλάσεων. Ο μάκρο-μέσος όρος είναι ένα ικανοποιητικό μέτρο ποιότητας για μικρότερες κατηγορίες [86], [99]. Στην εξίσωση (6), ο όρος TP_j δηλώνει τα σωστά ταξινομημένα στοιχεία δεδομένων που ανήκουν στην κλάση j και το n_j δηλώνει τον συνολικό αριθμό εγγράφων που ανήκουν σε αυτήν την κλάση.

$$Macro\ Average\ Precision = \frac{\sum \frac{TP_j}{n_j}}{|Classes|} \qquad (9)$$

3.6.2. Υπολογισμός μεγέθους SOM

Η επιλογή των ορθογώνιων χαρτών βασίστηκε στην παραδοχή που αναφέρθηκε στο [40], ότι για να βρεθεί ή να προσεγγιστεί το βέλτιστο μέγεθος χάρτη ανάλογα με τα δεδομένα εισόδου πρέπει να επιλεγούν οι δύο μεγαλύτερες κύριες συνιστώσες, δηλαδή αυτές με τις μεγαλύτερες ιδιοτιμές. Σύμφωνα με τους συγγραφείς, το μέγεθος που λαμβάνεται από αυτή τη διαδικασία είναι απλώς μια εικασία, καθώς το ακριβές μέγεθος του χάρτη δεν μπορεί να εκτιμηθεί εκ των προτέρων. Αντίθετα, πρέπει να προσδιοριστεί μέσω δοκιμών, αλλά αυτή η υπόθεση δίνει μια ιδέα για το αρχικό μέγεθος χάρτη.

Σε αυτή τη Διατριβή ακολουθούμε την λύση που παρουσιάζεται στο [75] για τον υπολογισμό του μεγέθους χάρτη που βασίζεται στην αντίστοιχη μέθοδο από το SOM Toolbox στο λογισμικό Matlab. Χρησιμοποιεί μια ευρετική μέθοδο για τον υπολογισμό των κόμβων του χάρτη όπως υποδηλώνεται στην ακόλουθη εξίσωση. Αφού καθοριστεί ο αριθμός των κόμβων, προσδιορίζεται και το μέγεθος χάρτη. Βασικά, οι δύο μεγαλύτερες ιδιοτιμές των δεδομένων εκπαίδευσης υπολογίζονται και με βάση τις τιμές τους, υπολογίζεται ο λόγος μεταξύ των πλευρών του χάρτη όπως υποδεικνύεται από τον ακόλουθο αλγόριθμο. Τα πραγματικά μήκη πλευρών ρυθμίζονται έτσι ώστε το γινόμενό τους να είναι όσο το δυνατόν πιο κοντά στον επιθυμητό αριθμό κόμβων. Με βάση αυτόν τον υπολογισμό, ο ακόλουθος πίνακας απεικονίζει το μέγεθος για 10 και τις 90 κλάσεις του συνόλου δεδομένων Reuters λαμβάνοντας υπόψη 1-grams, 2-grams και 3-grams και διαφορετικά μεγέθη διανυσμάτων, συγκεκριμένα 1000, 2000 και 3000 διαστάσεις ανά διάνυσμα (Πίνακας 9). Η αιτιολόγηση των επιλεγμένων μεγεθών ενισχύεται περαιτέρω από τα αποτελέσματα που παρουσιάζονται στην επόμενη υποενότητα, καθώς τείνουν να παράγουν καλύτερες F1 βαθμολογίες.

$$\boldsymbol{map\ units} = \mathbf{5} * \boldsymbol{vector\ length}^{1/2} \qquad (\mathbf{9})$$

Αλγόριθμος 4: Υπολογισμός μεγέθους SOM

Βήμα 1:	Υπολόγισε τον αριθμό των κόμβων στο χάρτη με χρήση της εξίσωσης (9).
Βήμα 2:	Βρες τις 2 μεγαλύτερες ιδιοτιμές
Βήμα 3:	**Εάν** $eigenvalue\_1 = 0$ **ή** $eigenvalue\_2 * mapunits < eigenvalue\_1$ **τότε** θέσε την μεταβλητή $ratio = 1$
Βήμα 4:	**Αλλιώς** υπολόγισε την μεταβλητή $ratio = \sqrt{eigenvalue\_1/eigenvalue\_2}$
Βήμα 5:	Υπολόγισε $size1 = \min(map\_units, round(\sqrt{map\_units/ratio}))$
Βήμα 6:	Υπολόγισε $size2 = round(map\_units/size1)$
Βήμα 7:	**Επίστρεψε** $size1, size2$

Πίνακας 9. Επιλεγμένο μέγεθος χάρτη με βάση τον αλγόριθμο 4 για διαφορετικά μεγέθη διανυσμάτων.

Μέγεθος διανύσματος		***# n-grams***	***Επιλεγμένο Μέγεθος***	***# κόμβων***
90 κλάσεις	***1000***	*1*	19*26	494
		2	20*24	480
		3	20*24	480
	2000	*1*	18*27	486
		2	19*26	494
		3	20*24	480
	3000	*1*	17*25	425
		2	18*23	414
		3	19*22	418
10 κλάσεις	***1000, 2000, 3000***	*1*	17*25	425
		2	18*23	414
		3	19*22	418

Τα αποτελέσματα που παρουσιάζονται στον πίνακα υποδεικνύουν ότι τα μεγέθη διανύσματος που επιλέγονται "ταιριάζουν" καλά με ένα συγκεκριμένο σύνολο κόμβων. Για παράδειγμα, κατά την εξέταση των αποτελεσμάτων για τις 90 κλάσεις μπορούμε να παρατηρήσουμε ότι οι διαφορές μεταξύ 1000 και 2000 διαστάσεων ανά διάνυσμα είναι αμελητέες. Επιπλέον, για τις 3000 διαστάσεις ανά διάνυσμα παρατηρήσαμε ότι απαιτείται ακόμη μικρότερος αριθμός κόμβων. Αυτό είναι ενδιαφέρον καθώς όσο μικρότερος είναι ο αριθμός των κόμβων, τόσο λιγότερος χρόνος χρειάζεται για την εκπαίδευση. Τα αποτελέσματα που αφορούν τις 10 κλάσεις ακολουθούν το ίδιο μοτίβο, αλλά το μέγεθος του διανύσματος (τουλάχιστον αυτά με τα οποία έχουμε πειραματιστεί) φαίνεται να μην επηρεάζει τον υπολογισμό του μεγέθους του χάρτη για το υποκείμενο σύνολο δεδομένων. Αντί αυτού, η χρήση των συμφραζόμενων φαίνεται να μειώνει τον αριθμό των κόμβων στο χάρτη κατά 2,5% για τα 2-grams και 1,6% για τα 3-grams.

3.6.3. Πειραματική Ανάλυση

Η πειραματική αξιολόγηση διεξήχθη στο υποσύνολο Mod Apte του σώματος δεδομένων Reuters. Τα διανύσματα χαρακτηριστικών κατασκευάστηκαν ακολουθώντας την προσέγγιση που περιεγράφηκε προηγουμένως, ενώ το σύνολο δεδομένων εκπαιδεύτηκε χρησιμοποιώντας τον batch αλγόριθμο SOM. Πραγματοποιήθηκαν διαφορετικοί κύκλοι πειραμάτων χρησιμοποιώντας διαφορετικά μεγέθη χαρτών και επιλογές διαμόρφωσης, τόσο για τις 90 όσο και για τις 10 κλάσεις.

Ο πρώτος γύρος πειραμάτων αποσκοπούσε στην εκτίμηση του ρόλου που παίζει το μέγεθος του χάρτη στην απόδοση της ταξινόμησης. Ως εκ τούτου, χρησιμοποιήθηκαν οι ίδιες επιλογές διαμόρφωσης τόσο για τις 90 κλάσεις όσο και για τις 10 μεγαλύτερες του συνόλου δεδομένων. Πιο

συγκεκριμένα, οι παράμετροι που αφορούν τις επαναλήψεις κατά τη φάση εκπαίδευσης (rough, finetune) ήταν 15 και 7 αντίστοιχα. Δοκιμάστηκαν διαφορετικά μεγέθη διανυσμάτων, δηλαδή διανύσματα χαρακτηριστικών διαστάσεων 250, 500, 1000, 2000, 3000, για να εκτιμηθεί η επίπτωση του μεγέθους λεξιλογίου στην απόδοση του ταξινομητή. Σε αυτόν τον πρώτο γύρο πειραμάτων, χρησιμοποιήθηκαν μόνο μεμονωμένες λέξεις (1-grams) και τετράγωνοι χάρτες. Τα μεγέθη χαρτών που χρησιμοποιήθηκαν ήταν 8*8, 16*16, 32*32, 40*40, 45*45, 50*50, 55*55, 60*60, 64*64, 75*75 και 128*128.

Ο δεύτερος γύρος πειραμάτων αποσκοπούσε στην αξιολόγηση της χρήσης των συμφραζόμενων λέξεων στην απόδοση ταξινόμησης. Ως εκ τούτου, σε αυτόν τον γύρο επαναλάβαμε τα ίδια πειράματα όπως στον προηγούμενο γύρο, χρησιμοποιώντας επίσης 2-grams και 3-grams αναφέροντας τις επιτευχθείσες $F1$ βαθμολογίες.

Ο τρίτος γύρος πειραμάτων είχε στόχο την αξιολόγηση της χρήσης ορθογώνιων χαρτών και διαφορετικών εποχών εκπαίδευσης στα αποτελέσματα ταξινόμησης. Πρέπει να σημειωθεί ότι σε αυτό το γύρο χρησιμοποιήθηκαν μόνο 1-grams (μεμονωμένες λέξεις). Τα χρησιμοποιούμενα μεγέθη διανύσματος ήταν 1000, 2000 και 3000.

Ο τελευταίος γύρος πειραμάτων αποσκοπούσε στην αξιολόγηση της χρήσης των συμφραζόμενων λέξεων στα αποτελέσματα ταξινόμησης. Ως εκ τούτου, επαναλάβαμε τα προηγούμενα πειράματα, χρησιμοποιώντας επίσης 2- grams και 3- grams και μεγέθη χάρτη που παρουσιάζονται στον επόμενο πίνακα. Τα αποτελέσματα των τεσσάρων γύρων πειραμάτων παρουσιάζονται λεπτομερώς στην επόμενη υποενότητα. Επιπλέον, στο τελευταίο κομμάτι της υποενότητας προσπαθούμε να συγκρίνουμε την εργασία μας με αυτήν που παρουσιάζεται στο [39] καταλήγοντας σε κάποιες ενδιαφέρουσες παρατηρήσεις.

3.6.4. Αποτελέσματα

Όπως περιεγράφηκε στην προηγούμενη υποενότητα, πραγματοποιήσαμε πειράματα χρησιμοποιώντας τόσο τις 10 κλάσεις με τα περισσότερα δείγματα όσο και το σύνολο (90) των κλάσεων του συνόλου δεδομένων Reuters. Για να παρατηρήσουμε την απόδοση του SOM, πρώτα πειραματιστήκαμε με διαφορετικά μεγέθη χαρτών και μεγέθη διανυσμάτων χρησιμοποιώντας μόνο τετράγωνους χάρτες και 1-grams.

Όσον αφορά τις 90 κλάσεις παρατηρήσαμε ότι τα αποτελέσματα ακολουθούν το ίδιο μοτίβο, καθώς για κάθε διαφορετικό μέγεθος χάρτη που εξετάστηκε, παρατηρήθηκαν μικρές διαφορές μεταξύ των διαφορετικών μεγεθών διανυσμάτων. Ο ταξινομητής πέτυχε περίπου 0,50 ή 50% χρησιμοποιώντας ένα χάρτη 32*32, φτάνοντας το 0,63 για τα μεγέθη 60*60 και 64*64. Για το ίδιο μέγεθος χαρτών οι διαφορές μεταξύ διαφόρων μεγεθών διανυσμάτων κυμαίνονται από 1-5%. Για τις $F1$ βαθμολογίες Μάκρο-μέσου όρου, φαίνεται ότι η απόδοση του ταξινομητή αυξάνεται με τη χρήση μεγαλύτερων διανυσμάτων καθώς οι υψηλότερες βαθμολογίες επιτεύχθηκαν όταν επιλέχθηκαν διανύσματα με 2000 και 3000 διαστάσεις ανά διάνυσμα ενώ ο ταξινομητής πέτυχε γύρω στο 28% για 1000, 2000 και 3000 διαστάσεις ανά διάνυσμα για χάρτη 64*64. Ο ταξινομητής πέτυχε 28% -30% για μεγέθη χαρτών 40*40, 60*60, 64*64 και 75*75, αλλά εδώ η απόκλιση μεταξύ μεγεθών των διανυσμάτων είναι πιο σημαντική, κυμαίνεται από 0-10%.

Επαναλάβαμε τα πειράματα χρησιμοποιώντας μόνο τις 10 κλάσεις για να αξιολογήσουμε αν μπορούμε να εξάγουμε τα ίδια ή παρόμοια συμπεράσματα. Για τις βαθμολογίες Μίκρο-μέσου Όρου, η τάση παραμένει η ίδια, υπάρχει γραμμική αύξηση στις βαθμολογίες για μεγέθη χαρτών 8*8 έως

32*32 αλλά για μεγέθη χάρτη μεγαλύτερα από 50*50 η απόδοση του ταξινομητή επιδεινώνεται σταδιακά.

Αυτό συμβαίνει λόγω των υψηλών τιμών TE που αναφέρονται στην περίπτωση αυτή, υποδηλώνοντας ότι η τοπολογία δεν διατηρείται για τα συγκεκριμένα δεδομένα εισόδου. Μια άλλη ενδιαφέρουσα παρατήρηση είναι ότι διανύσματα με μικρότερες διαστάσεις παράγουν καλύτερα αποτελέσματα σε σύγκριση με διανύσματα 2000 ή 3000 διαστάσεων. Κατά τη σύγκριση των βαθμολογιών Μίκρο- και Μάκρο-μέσου όρου, οι τάσεις είναι παρόμοιες, με τις υψηλότερες βαθμολογίες $F1$ να επιτυγχάνονται όταν επιλέγονται διανύσματα με 250 και 500 διαστάσεις.

Τα καλύτερα αποτελέσματα που ελήφθησαν ανεξάρτητα από το μέγεθος των διανυσμάτων που χρησιμοποιήθηκαν, για διαφορετικά μεγέθη χαρτών (90, 10 κλάσεις) απεικονίζονται στους Πίνακες 10 και 11 αντίστοιχα, ενώ αναφέρονται και οι τιμές QE και TE. Αυτές οι τιμές θεωρούνται απαραίτητες για την επιλογή του καλύτερου μοντέλου, δεδομένου ότι υπάρχουν συγκεκριμένα μεγέθη χάρτη που παράγουν τα ίδια αποτελέσματα, αλλά οι τιμές QE και TE παρέχουν χρήσιμες πληροφορίες για το μέγεθος του χάρτη που θα επιλεγεί.

Σύμφωνα με τον Πίνακα 10, η υψηλότερη βαθμολογία $F1$ επιτυγχάνεται για μεγέθη χαρτών 50*50 και 55*55 (62%). Και τα δύο μεγέθη δίνουν περίπου το ίδιο TE αλλά επιλέγουμε τον χάρτη 55*55 καθώς παράγει χαμηλότερο QE. Θα πρέπει να σημειωθεί ότι για τον ίδιο λόγο δεν επιλέγουμε τα δίκτυα 60*60 και 64*64 ως καλύτερα, παρόλο που επιτυγχάνουν βαθμολογία $F1$ 63%. Σε αυτή την περίπτωση το TE είναι κοντά στο 1 αν και το QE είναι μικρότερο από τα προηγούμενα εξεταζόμενα πλέγματα. Παρόλα αυτά τα πλέγματα αυτά απορρίπτονται λόγω του TE, καθώς τιμές κοντά στο 1 δείχνουν ότι η τοπολογία δεν διατηρείται. Όπως έχει ήδη αναφερθεί προηγουμένως, στόχος είναι η παραγωγή διαφορετικών σεναρίων διαμόρφωσης και στη συνέχεια να επιλεγούν τα σενάρια εκείνα που δίνουν TE κοντά στο μηδέν, καθώς αυτό σημαίνει ότι διατηρείται η τοπολογία του χάρτη και συνεπώς το επιλεγμένο μοντέλο είναι καλό για τη δεδομένη είσοδο. Τα αποτελέσματα για τη βαθμολογία $F1$ Μάκρο-μέσου όρου τείνουν να σταθεροποιούνται γύρω στο 28-30% για χάρτες μεταξύ 40*40 και 75*75. Έτσι, οι υψηλότερες βαθμολογίες επιτυγχάνονται όταν χρησιμοποιείται ένα πλέγμα 60*60 με μηδενικό TE. Αυτή η επιλογή ενισχύεται περαιτέρω καθώς άλλα δίκτυα που επιτυγχάνουν την ίδια βαθμολογία παράγουν υψηλότερο TE. Όταν παρατηρούμε το πλέγμα 128*128 βλέπουμε ότι η βαθμολογία $F1$ επιδεινώνεται σημαντικά. Αυτό οφείλεται στο ότι το TE είναι κοντά στο 1 και το QE αυξάνεται επίσης σε σύγκριση με τον χάρτη 75*75. Αν παρατηρήσουμε ολόκληρο το σύνολο των αποτελεσμάτων για το QE, για ένα πλέγμα 8*8 το QE είναι 41.74 και μειώνεται σταδιακά καθώς το μέγεθος χάρτη αυξάνεται, ενώ στην τελευταία περίπτωση (128*128 χάρτης) αρχίζει και πάλι να αυξάνεται, πράγμα που σημαίνει ότι υπάρχει συμβιβασμός μεταξύ του μεγέθους SOM, του QE και του TE. Αυτό θα μπορούσε να συμβεί και για άλλα μεγέθη πλέγματος μεταξύ αυτών των δύο, αλλά η συγκεκριμένη Διατριβή δεν επικεντρώνεται στο να βρεθεί το σημείο όπου συμβαίνει αυτό.

Συνεπώς, ο Πίνακας 11 απεικονίζει τα καλύτερα αποτελέσματα για τις 10 κλάσεις του συνόλου δεδομένων Reuters. Για τη βαθμολογία $F1$ Μίκρο-μέσου όρου, επιτυγχάνουμε 79%, για τα πλέγματα 40*40, 45*45 και 50*50, ενώ τα πλέγματα 55*55, 60*60 και 64*64 παράγουν αποτελέσματα με απόκλιση 1-2%. Επιλέγουμε το 40*40 καθώς έχει το χαμηλότερο TE, ενώ μικρές διαφορές μπορούν να παρατηρηθούν για το QE. Αν και η βαθμολογία για αυτό το μέγεθος χαρτών είναι υψηλή, το μοντέλο δεν λειτουργεί καλά καθώς οι τιμές TE είναι υψηλές για την πλειοψηφία των πλεγμάτων που δοκιμάσαμε.

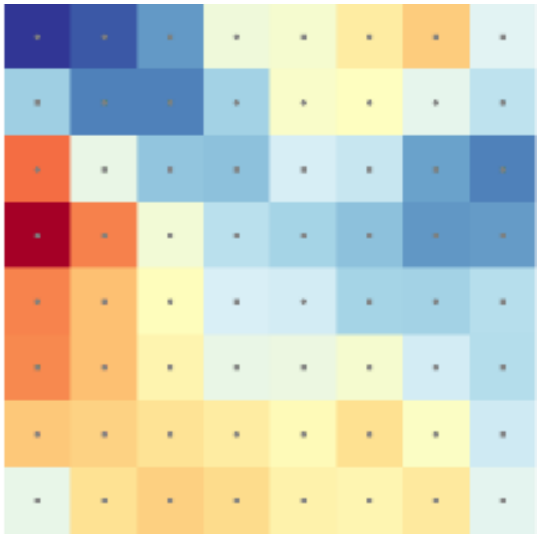
Πίνακας 10. $F1$ βαθμολογίες Μίκρο- και Μάκρο μέσου όρου, TE και QE με χρήση τετράγωνων χαρτών διαφορετικών μεγεθών για τις 90 κλάσεις του συνόλου δεδομένων Reuters (1-grams).

Μέγεθος Χάρτη		8*8	16*16	32*32	40*40	45*45	50*50	55*55	60*60	64*64	75*75	128*128
90 κλάσεις	*Micro*	0.18	0.34	0.54	0.58	0.61	**0.62**	**0.62**	0.63	0.63	0.62	0.47
	TE	0.66	0.09	0.00	0.00	1.00	**0.00**	**0.02**	0.99	0.98	1.00	1.00
	QE	41.74	49.89	38.38	37.38	25.54	**25.17**	**24.69**	24.21	23.98	23.42	29.70
	Macro	0.07	0.14	0.26	0.29	0.28	0.28	0.28	**0.30**	0.28	0.30	0.22
	TE	0.66	0.09	0.00	0.00	0.00	0.00	0.15	**0.00**	0.05	0.98	0.96
	QE	41.74	49.89	47.40	46.30	36.96	45.22	35.74	**43.89**	34.83	42.18	36.70

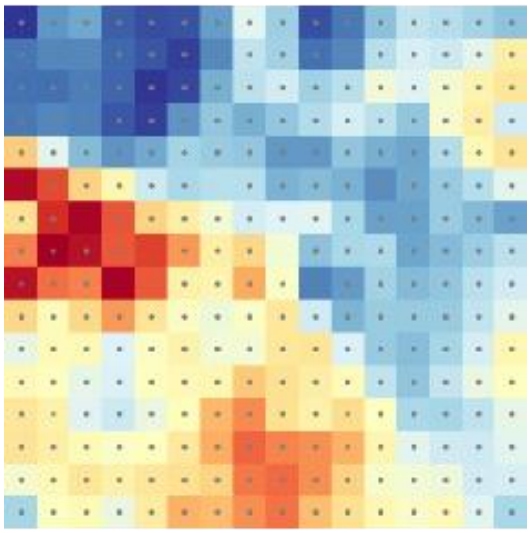
Πίνακας 11. $F1$ βαθμολογίες Μίκρο- και Μάκρο-μέσου όρου, TE και QE με χρήση τετραγωνικών χαρτών διαφορετικών μεγεθών για τις 10 κλάσεις του συνόλου δεδομένων Reuters (1-grams).

Μέγεθος Χάρτη		8*8	16*16	32*32	40*40	45*45	50*50	55*55	60*60	64*64	75*75	128*128
10 κλάσεις	*Micro*	0.47	0.63	0.76	**0.79**	**0.79**	**0.79**	0.78	0.77	0.77	0.72	0.51
	TE	0.00	0.15	0.98	**0.92**	**0.99**	**1.00**	1.00	1.00	0.93	0.98	1.00
	QE	41.01	39.95	17.99	**17.50**	**11.54**	**11.38**	16.77	16.59	10.86	10.62	37.58
	Macro	0.37	0.50	0.62	0.65	0.67	**0.68**	0.66	0.65	0.66	0.62	0.37
	TE	0.65	0.98	0.98	0.99	0.99	**1.00**	1.00	0.43	0.93	0.98	0.00
	QE	20.40	19.46	17.99	11.70	11.54	**11.38**	16.77	11.03	10.86	10.62	14.15

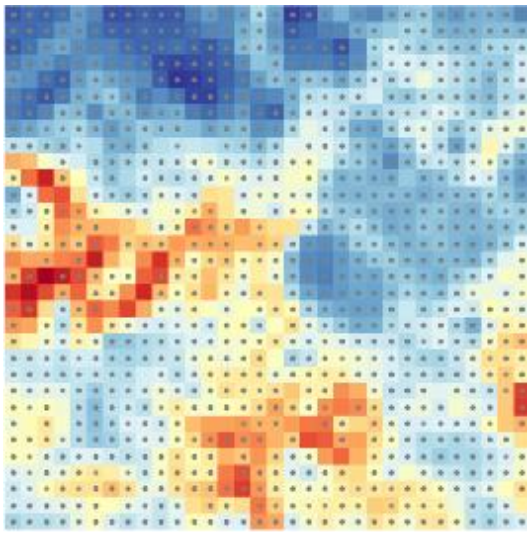
Πίνακας 12. Απεικόνιση διαφορετικών μεγεθών SOM ως U-matrices με 1000 διαστάσεις ανά διάνυσμα, 90 κλάσεις, τετράγωνοι χάρτες.



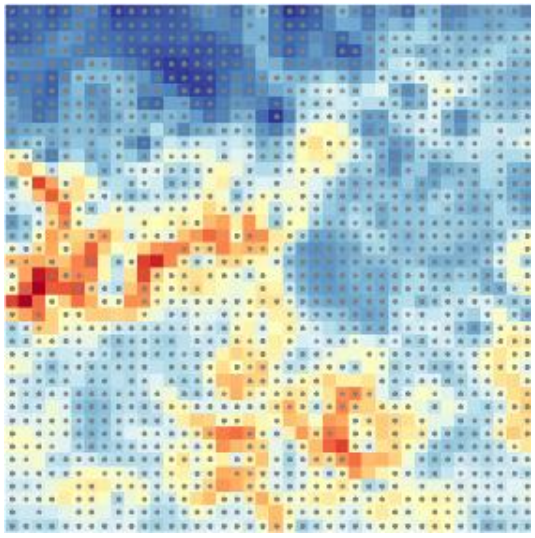
8*8



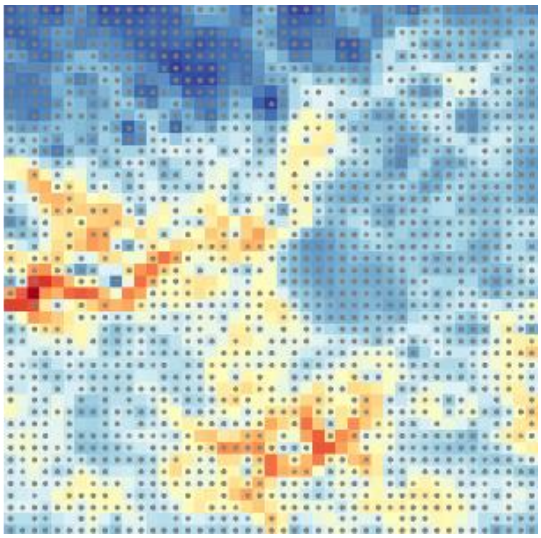
16*16



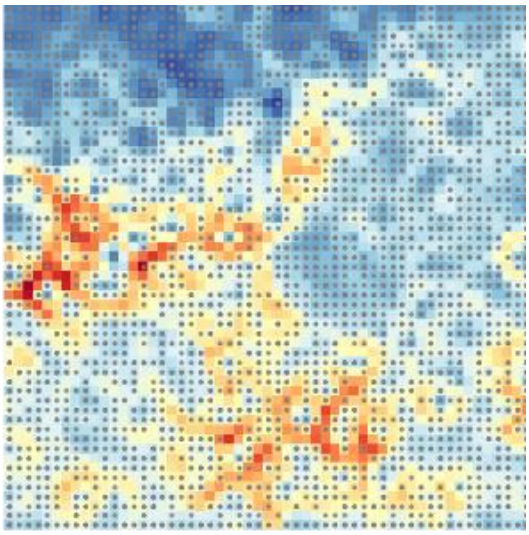
32*32



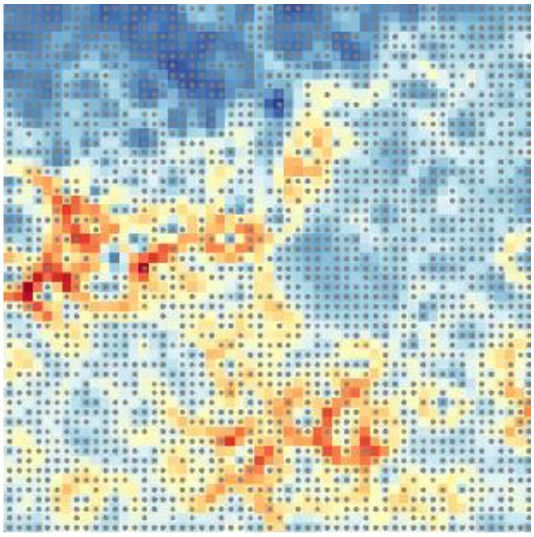
40*40



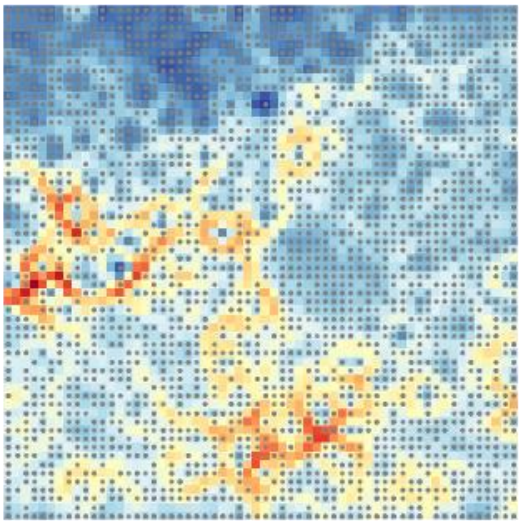
45*45



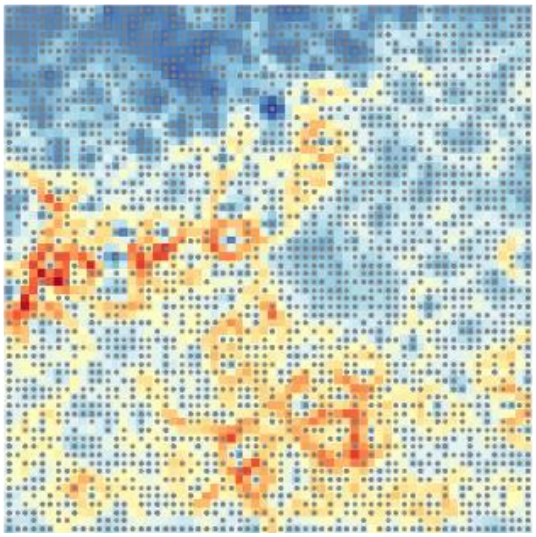
50*50



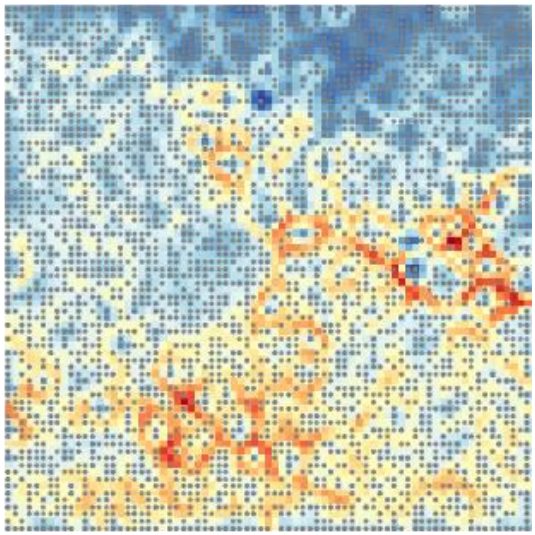
55*55



60*60



64*64



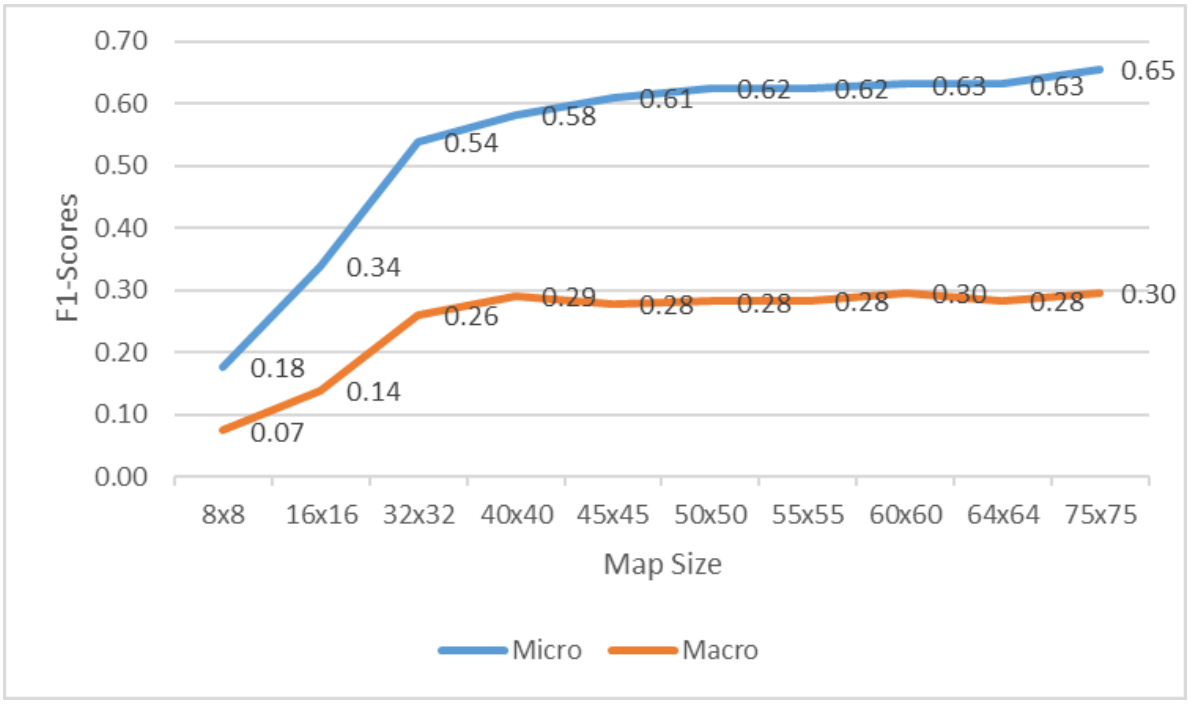
75*75

Ο Πίνακας 12 παρουσιάζει SOMs διαφορετικών μεγεθών για τις 90 κλάσεις ως Unified distance matrices (U-matrices). Κάθε κόμβος στον χάρτη αντιπροσωπεύεται από ένα codebook διάνυσμα, το οποίο έχει τις ίδιες διαστάσεις με τα διανύσματα που χρησιμοποιήθηκαν για την απεικόνιση του συνόλου δεδομένων. Σε αυτό το παράδειγμα, η διάσταση των codebook διανυσμάτων είναι 1000. Οι αντίστοιχοι χάρτες που αφορούν τις 10 κλάσεις δεν παρουσιάζονται εδώ για λόγους συντομίας. Η U-Matrix απεικονίζει τις αποστάσεις μεταξύ γειτονικών κόμβων χαρτών και προάγει την παρατήρηση της δομής των συστάδων (clusters) επάνω στον χάρτη. Υψηλές τιμές της U-Matrix υποδεικνύουν ένα σύνορο μεταξύ των συστάδων που έχουν δημιουργηθεί. Ειδικότερα, ένας σκούρος χρωματισμός αντιστοιχεί σε μεγάλη απόσταση και κατά συνέπεια ένα κενό μεταξύ των τιμών του codebook διανύσματος στον χώρο εισόδου. Οι ομοιόμορφες περιοχές παρόμοιων χρωματισμών υποδεικνύουν τις ίδιες τις συστάδες, που υποδηλώνουν επίσης ομοιότητα μεταξύ των διανυσμάτων που βρίσκονται στην ίδια περιοχή. Είναι προφανές ότι οι χάρτες μικρότερου μεγέθους έχουν περισσότερες ετικέτες συνδεδεμένες σε κάθε κόμβο, ενώ οι μεγαλύτεροι χάρτες έχουν λιγότερες ή και καθόλου. Οι κόμβοι που έχουν προσαρτημένες ετικέτες σημειώνονται χρησιμοποιώντας ένα γκρι σημείο που υποδεικνύει ότι υπάρχει μια ή περισσότερες ετικέτες στο συγκεκριμένο κόμβο. Επιπλέον, μπορούμε να παρατηρήσουμε ότι σε μεγαλύτερα μεγέθη χάρτη διαμορφώνονται σχετικά οριοθετημένες συστάδες που δεν διαφοροποιούνται καθώς αυξάνεται το μέγεθος του χάρτη.

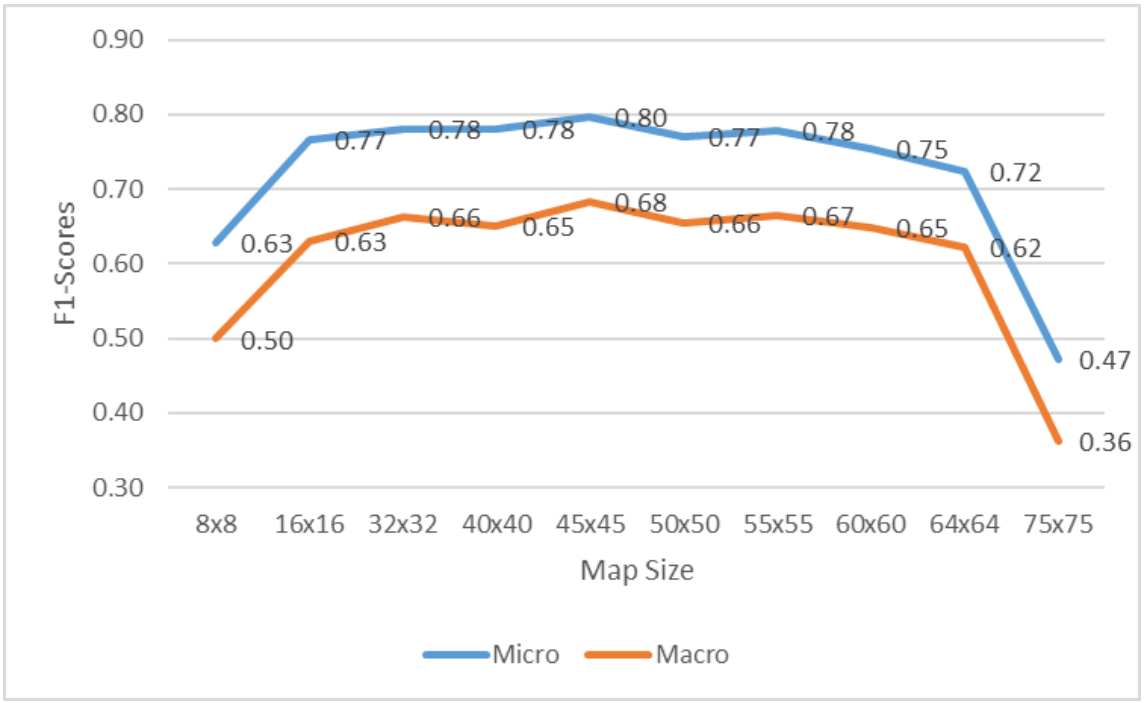
Ο δεύτερος κύκλος των πειραμάτων αποσκοπούσε στην διερεύνηση του τρόπου με τον οποίο τα συμφραζόμενα (context words) επηρεάζουν την ταξινόμηση. Σε αυτόν τον γύρο χρησιμοποιήθηκαν οι ίδιες επιλογές διαμόρφωσης όσον αφορά στο μέγεθος του χάρτη και των διανυσμάτων, αλλά χρησιμοποιήθηκαν επίσης 2-grams και 3-grams. Το επίκεντρο αυτού του γύρου ήταν να παρουσιαστούν οι διαφορές στις βαθμολογίες που αναφέρθηκαν κατά τη χρήση των συμφραζόμενων σε σύγκριση με τα αποτελέσματα που προέκυψαν στον προηγούμενο γύρο. Ένα ενδιαφέρον συμπέρασμα που προέκυψε κατά την εξέταση των αποτελεσμάτων για τις βαθμολογίες $F1$ Μίκρο-μέσου όρου για τις 90 κλάσεις, είναι ότι η χρήση n-grams παράγει καλύτερα αποτελέσματα σε μικρότερα πλέγματα. Για παράδειγμα, σε ένα πλέγμα 8*8, τα 1-grams παράγουν 17% $F1$, ενώ για τα 2- grams και τα 3- grams η αντίστοιχη βαθμολογία είναι 30% και 32%. Τα αντίστοιχα αποτελέσματα για τις βαθμολογίες Μάκρο-μέσου όρου υποδηλώνουν ότι η χρήση των συμφραζόμενων δεν αυξάνει την απόδοση του ταξινομητή για κατηγορίες με μικρό αριθμό δειγμάτων, καθώς η χρήση 1-grams είναι ανώτερη. Όταν παρατηρήσουμε τις βαθμολογίες Μίκρο-μέσου όρου και Μάκρο-μέσου όρου για τις 10 κλάσεις, μπορούμε να υποθέσουμε με ασφάλεια ότι η χρήση των συμφραζόμενων δεν επηρεάζει σημαντικά την απόδοση του ταξινομητή καθώς η απόκλιση στην απόδοση κυμαίνεται γύρω στο 1-2% για το σύνολο των πειραμάτων.

Στην παρακάτω υποενότητα παρουσιάζονται τα αποτελέσματα με χρήση των συμφραζόμενων για διαφορετικά μεγέθη χαρτών. Οι επιλεγμένες βαθμολογίες $F1$ είναι οι υψηλότερες ανεξάρτητα από το μέγεθος των χρησιμοποιούμενων διανυσμάτων, που κυμαίνονται από 250 έως 3000 διαστάσεις ανά διάνυσμα. Οι $F1$ βαθμολογίες Μίκρο-μέσου όρου για τις 90 κλάσεις υποδηλώνουν ότι για κάθε μέγεθος χάρτη παρατηρούνται μικρές διαφορές μεταξύ διαφόρων μεγεθών διανύσματος, καθώς οι βαθμολογίες κυμαίνονται γύρω στο 55% (με απόκλιση 5% μεταξύ των διαφόρων μεγεθών διανύσματος) για χάρτη 32*32, όπου η χρήση μεγαλύτερων διανυσμάτων βελτιώνει την απόδοση. Για τις αντίστοιχες βαθμολογίες Μάκρο-μέσου όρου, μπορούμε να υποθέσουμε με ασφάλεια ότι η απόδοση του ταξινομητή επιδεινώνεται όταν χρησιμοποιούνται διανύσματα με μικρότερες διαστάσεις για πλέγματα μεγαλύτερα από 32*32. Μπορεί επίσης να γίνει αντιληπτό ότι οι υψηλότερες $F1$ βαθμολογίες επιτεύχθηκαν για διανύσματα που έχουν 2000 και 3000 διαστάσεις με την απόκλιση να κυμαίνεται μεταξύ 0% και 13%. Για τις 10 κλάσεις και τις βαθμολογίες Μίκρο-μέσου όρου η τάση παραμένει η ίδια, καθώς υπάρχει γραμμική αύξηση των βαθμολογιών για τα μεγέθη χαρτών 8*8 έως 32*32, ενώ τα αποτελέσματα τείνουν να σταθεροποιούνται για το πλέγμα 40*40, αλλά η απόδοση του ταξινομητή επιδεινώνεται σταδιακά όσο το μέγεθος του χάρτη αυξάνεται λόγω των υψηλών τιμών ΤΕ. Οι υψηλότερες βαθμολογίες $F1$ Μάκρο-μέσου όρου επιτυγχάνονται όταν επιλέγονται διανύσματα με 250 διαστάσεις (για πλέγματα μεγαλύτερα από 40*40), που κυμαίνονται από 59-65% ανάλογα με το μέγεθος του διανύσματος, με την καλύτερη επίδοση να παρατηρείται για έναν χάρτη 50*50.

Οι παρακάτω εικόνες (εικόνα 10 και 11) δείχνουν τα καλύτερα αποτελέσματα που επιτεύχθηκαν για τις $F1$ βαθμολογίες Μίκρο- και Μάκρο μέσου όρου, ανεξάρτητα από το μέγεθος των διανυσμάτων αλλά και τα n-grams που χρησιμοποιήθηκαν τόσο για τις 90 όσο και για τις 10 κλάσεις αντίστοιχα. Για τις 90 κλάσεις, τα καλύτερα αποτελέσματα που επιτυγχάνονται είναι 65% σε ένα πλέγμα 75*75 και 30% σε πλέγματα 60*60 και 75*75 για τις βαθμολογίες Μίκρο- και Μάκρο-μέσου όρου αντίστοιχα. Οι αντίστοιχες βαθμολογίες για τις 10 κλάσεις είναι 80% και 68% για ένα χάρτη 45*45.



Εικόνα 10. Οι καλύτερες $F1$ βαθμολογίες Μίκρο και Μάκρο-μέσου όρου που επιτεύχθηκαν με χρήση τετράγωνων χαρτών για τις 90 κλάσεις του συνόλου δεδομένων Reuters.



Εικόνα 11. Οι καλύτερες $F1$ βαθμολογίες Μίκρο και Μάκρο-μέσου όρου που επιτεύχθηκαν με χρήση τετράγωνων χαρτών για τις 10 κλάσεις του συνόλου δεδομένων Reuters (1,2,3-grams).

Ο τρίτος γύρος πειραμάτων αποσκοπεί στην αξιολόγηση της χρήσης ορθογώνιων χαρτών και διαφορετικού αριθμού κύκλων εκπαίδευσης στα αποτελέσματα ταξινόμησης. Σε αυτόν τον γύρο χρησιμοποιήθηκαν μόνο 1-grams ενώ η διάσταση των διανυσμάτων κυμάνθηκε από 1000 έως 3000. Τα αποτελέσματα που προέκυψαν από αυτό τον γύρο, παρουσιάζονται στον Πίνακα 13 για τις 90 κλάσεις και τον Πίνακα 14 για τις 10 κλάσεις. Αυτοί οι πίνακες παρουσιάζουν τις τιμές με χρήση

τεσσάρων δεκαδικών ψηφίων, έτσι ώστε να επισημαίνονται ευκολότερα οι διαφορές στα αποτελέσματα.

Πίνακας 13. Αποτελέσματα ταξινόμησης με χρήση ορθογώνιων χαρτών για διαφορετικά μεγέθη διανυσμάτων και κύκλων εκπαίδευσης για τις 90 κλάσεις του συνόλου δεδομένων Reuters.

Κύκλοι Εκπαίδευσης				15	20	30	40
90 κλάσεις	Μέγεθος Διανύσματος 1000 Χάρτης 19*26	Micro	Precision	0.5793	**0.5785**	0.5726	0.5724
			Recall	0.8550	**0.8590**	0.8568	0.8568
			F1	0.6907	**0.6914**	0.6864	0.6863
		Macro	Precision	0.2804	0.3157	**0.3195**	0.2874
			Recall	0.4455	0.4909	**0.5083**	0.4646
			F1	0.3442	0.3843	**0.3924**	0.3552
	Μέγεθος Διανύσματος 2000 Χάρτης 18*27	Micro	Precision	0.5778	0.5776	0.5805	**0.5799**
			Recall	0.8558	0.8619	0.8616	**0.8635**
			F1	0.6898	0.6917	0.6937	**0.6938**
		Macro	Precision	**0.3129**	0.2996	0.3027	0.3092
			Recall	**0.4755**	0.4766	0.4615	0.4821
			F1	**0.3774**	0.3679	0.3656	0.3768
	Μέγεθος Διανύσματος 3000 Χάρτης 18*27	Micro	Precision	0.5724	**0.5714**	0.5703	0.5695
			Recall	0.8478	**0.8576**	0.8459	0.8472
			F1	0.6834	**0.6859**	0.6813	0.6811
		Macro	Precision	**0.3319**	0.3278	0.3210	0.3175
			Recall	**0.5237**	0.5206	0.5046	0.4933
			F1	**0.4063**	0.4023	0.3924	0.3864

Πίνακας 14. Αποτελέσματα ταξινόμησης με χρήση ορθογώνιων χαρτών για διαφορετικά μεγέθη διανυσμάτων και κύκλων εκπαίδευσης για τις 10 κλάσεις του συνόλου δεδομένων Reuters.

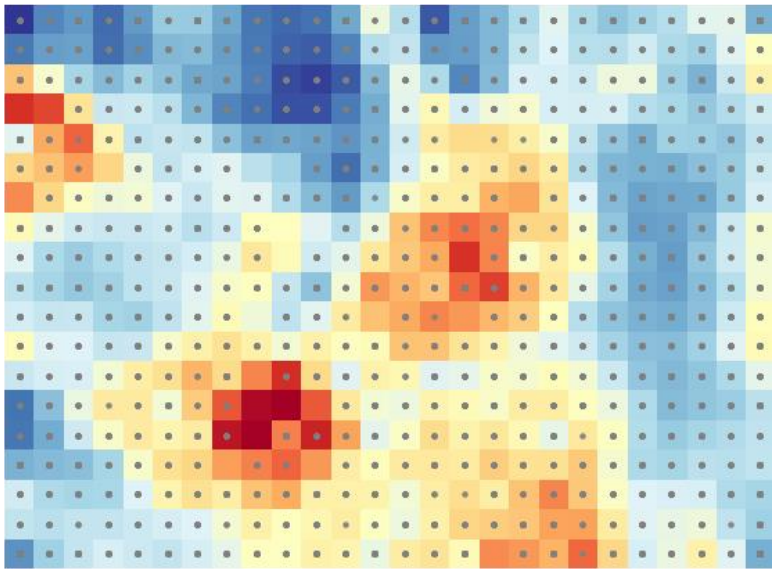
Κύκλοι Εκπαίδευσης				15	20	30	40
10 κλάσεις	Μέγεθος Διανύσματος 1000 Χάρτης 17*25	Micro	Precision	0.7507	0.7626	0.7618	**0.7649**
			Recall	0.9673	0.9637	0.9640	**0.9626**
			F1	0.8453	0.8514	0.8511	**0.8524**
		Macro	Precision	0.6608	**0.6717**	0.6641	0.6686
			Recall	0.9188	**0.9180**	0.9073	0.9090
			F1	0.7687	**0.7757**	0.7669	0.7705
	Μέγεθος Διανύσματος 2000 Χάρτης 17*25	Micro	Precision	0.7578	0.7554	**0.7624**	0.7573
			Recall	0.9563	0.9537	**0.9545**	0.9548
			F1	0.8455	0.8431	**0.8477**	0.8447
		Macro	Precision	0.6668	0.6685	**0.6781**	0.6681
			Recall	0.9233	0.9211	**0.9180**	0.9106
			F1	0.7744	0.7747	**0.7800**	0.7707
	Μέγεθος Διανύσματος 3000 Χάρτης 17*25	Micro	Precision	0.7571	0.7541	**0.7575**	0.7529
			Recall	0.9431	0.9486	**0.9526**	0.9508
			F1	0.8399	0.8403	**0.8440**	0.8403
		Macro	Precision	0.6617	0.6614	**0.6761**	0.6693
			Recall	0.8978	0.8960	**0.9183**	0.9228
			F1	0.7619	0.7610	**0.7788**	0.7759

Σύμφωνα με τον Πίνακα 13, η καλύτερη βαθμολογία $F1$ Μίκρο-μέσου όρου (69,38%) επιτυγχάνεται για διάσταση χάρτη 18*27 και 40 κύκλους εκπαίδευσης με 2000 διαστάσεις ανά διάνυσμα. Δεν είναι ξεκάθαρο πόσοι κύκλοι εκπαίδευσης θα πρέπει να επιλεγούν, καθώς για διανύσματα διαστάσεων 1000 και 3000, οι υψηλότερες βαθμολογίες επιτυγχάνονται για 20 κύκλους εκπαίδευσης. Επιπλέον, η απόκλιση στα αποτελέσματα μεταξύ των διαφορετικών κύκλων εκπαίδευσης για το ίδιο μέγεθος διανύσματος κυμαίνεται από 0,5% (1000 και 3000 μεγέθη διανύσματος) έως 0,4% για διανύσματα 2000 διαστάσεων. Για τη βαθμολογία $F1$ Μίκρο-μέσου όρου, οι υψηλότερες βαθμολογίες επιτυγχάνονται όταν χρησιμοποιούνται 3000 διαστάσεις ανά διάνυσμα και χάρτη 18*27, με την βαθμολογία $F1$ να φτάνει το 40,63% για 15 κύκλους εκπαίδευσης.

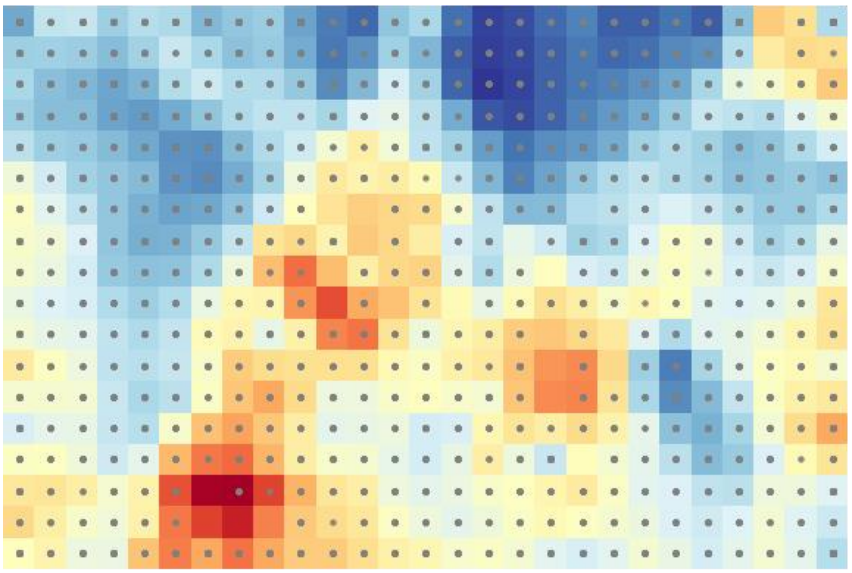
Αντίστοιχα, ο Πίνακας 14 απεικονίζει τα καλύτερα αποτελέσματα για τις 10 κλάσεις. Η υψηλότερη $F1$ βαθμολογία Μίκρο-μέσου όρου (85,24%), δίνεται για 1000 διαστάσεις ανά διάνυσμα και 40 κύκλους εκπαίδευσης, ενώ 84,77% και 84,40% επιτυγχάνεται σε 20 κύκλους εκπαίδευσης για 2000 και 3000 διαστάσεις ανά διάνυσμα αντίστοιχα. Για διανύσματα 2000 και 3000 διαστάσεων, τα αποτελέσματα καταδεικνύουν ότι 30 κύκλοι εκπαίδευσης παράγουν τα καλύτερα αποτελέσματα τόσο για τις βαθμολογίες $F1$ Μίκρο- και Μάκρο-μέσου όρου, παρόλο που η απόκλιση μεταξύ των άλλων κύκλων εκπαίδευσης είναι μικρή (0,5-1%). Για τη βαθμολογία $F1$ Μάκρο-μέσου όρου, η βαθμολογία που επιτυγχάνεται είναι 78% με μέγεθος διανύσματος 2000 στους 30 κύκλους εκπαίδευσης.

Οι Πίνακες 15 και 16 αντιπροσωπεύουν τα SOMs που προκύπτουν για τις 90 και 10 κλάσεις ως U-matrices. Τα μεγέθη του χάρτη που χρησιμοποιήθηκαν είναι παρόμοια, αλλά μπορούμε να παρατηρήσουμε ότι υπάρχει διαφοροποίηση στις συστάδες που δημιουργούνται. Μπορούμε επίσης να παρατηρήσουμε ότι για αυτά τα μεγέθη χάρτη ο αριθμός κόμβων χωρίς ετικέτες μειώνεται σημαντικά σε σύγκριση με τους τετράγωνους χάρτες που παρουσιάστηκαν στον Πίνακα 12.

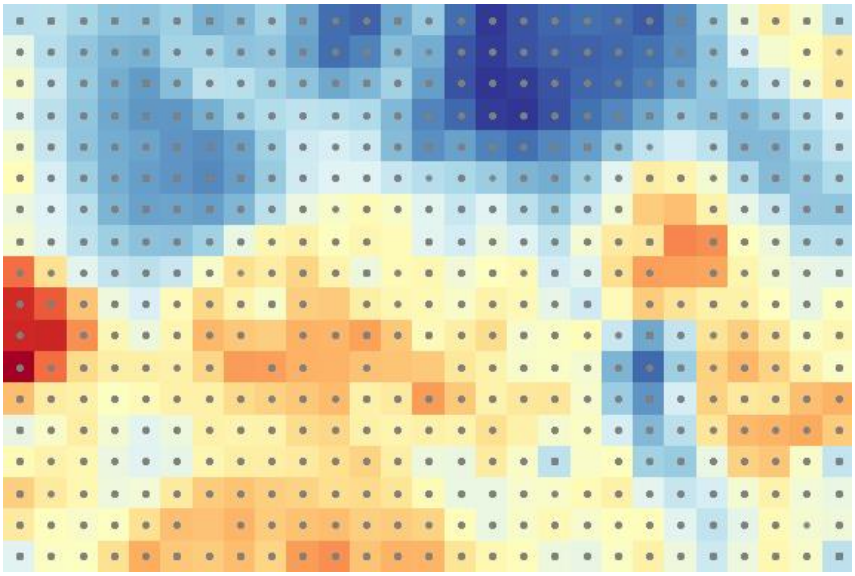
Πίνακας 15. Απεικόνιση διαφορετικών SOM ως U-matrices, 90 κλάσεις, ορθογώνιοι χάρτες.



(a) 19*26, 1000 διαστάσεις

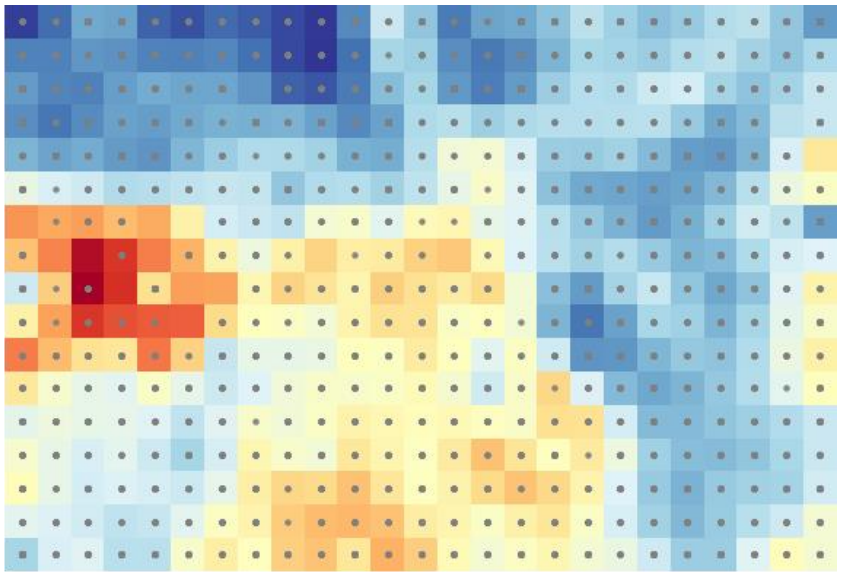


(b) 18*27, 2000 διαστάσεις

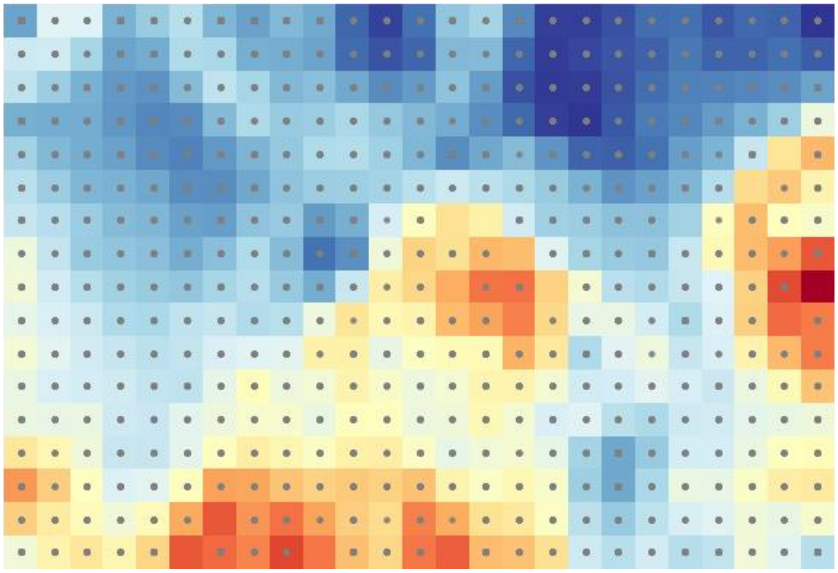


(c) 18*27, 3000 διαστάσεις

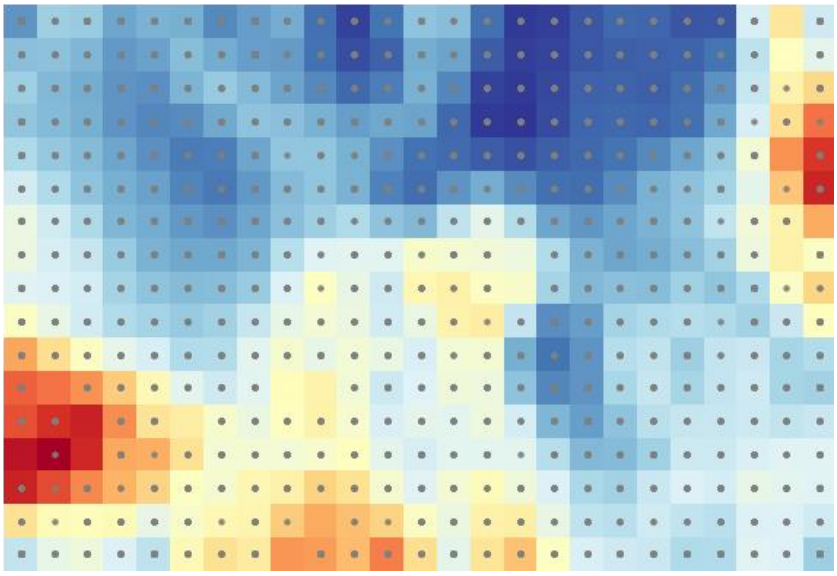
Πίνακας 16. Απεικόνιση διαφορετικών SOM ως U-matrices, 10 κλάσεις, ορθογώνιοι χάρτες



(d) 17*25, 1000 διαστάσεις



(e) 17*25, 2000 διαστάσεις



(f) 17*25, 3000 διαστάσεις

Ο τελευταίος κύκλος πειραμάτων αποσκοπεί στην αξιολόγηση της χρήσης των συμφραζόμενων στα αποτελέσματα ταξινόμησης. Σε αυτό το σύνολο πειραμάτων τα μεγέθη των διανυσμάτων με τα οποία πειραματιστήκαμε είναι επίσης εντός της κλίμακας 1000 - 3000. Οι παρακάτω πίνακες παρουσιάζουν τις τιμές TE και QE προκειμένου να έχουμε μια σαφέστερη εικόνα για το μοντέλο που ταιριάζει καλύτερα στη δεδομένη είσοδο.

Σύμφωνα με τον Πίνακα 17, οι υψηλότερες βαθμολογίες $F1$ Μίκρο-μέσου όρου είναι 0,69 και Μάκρο-μέσου όρου είναι 0,40 αντίστοιχα. Παρόλο που για τη βαθμολογία Μίκρο-μέσου όρου υπάρχουν και άλλοι συνδυασμοί που παράγουν τα ίδια ή παρόμοια αποτελέσματα, η επιλεγμένη διαμόρφωση έχει μηδενικό TE και μικρότερο QE σε σύγκριση με τις υπόλοιπες. Οι αποκλίσεις στα αποτελέσματα ανάμεσα στις διαμορφώσεις κυμαίνονται περίπου στο 1% για τις βαθμολογίες Μίκρο-μέσου όρου και από 1-4% για τις βαθμολογίες Μάκρο-μέσου όρου.

Ο Πίνακας 18 παρουσιάζει τα πειράματα που διεξήχθησαν για τις 10 κλάσεις, όπου οι υψηλότερες βαθμολογίες Μίκρο-μέσου όρου είναι 0,86 και Μάκρο-μέσου όρου είναι 0,78, με 3000

διαστάσεις ανά διάνυσμα χρησιμοποιώντας επίσης 2-grams. Θα πρέπει να σημειωθεί ότι αυτή η διαμόρφωση έχει επίσης μηδενικό TE, παρόλο που το QE δεν είναι το χαμηλότερο σε αυτές τις σειρές πειραμάτων. Επιπλέον, είναι αξιοσημείωτο ότι οι άλλες διαμορφώσεις τείνουν να παράγουν αποτελέσματα με απόκλιση 1% για τις βαθμολογίες Μίκρο-μέσου όρου και 1-2% για τις βαθμολογίες Μάκρο-μέσου όρου. Ως εκ τούτου, δεν μπορεί να συναχθεί εάν μια συγκεκριμένη διαμόρφωση αυξάνει την απόδοση του ταξινομητή.

Πίνακας 17. $F1$ βαθμολογίες Μίκρο- και Μάκρο-μέσου όρου, TE και QE με χρήση ορθογώνιων χαρτών για διαφορετικά μεγέθη διανύσματος για τις 90 κλάσεις του συνόλου δεδομένων Reuters.

90 κλάσεις	Μέγεθος Χάρτη	Micro-Average	Macro-Average	TE	QE
Μέγεθος Διανύσματος 1000	*19*26 (1-grams)*	0.69	0.39	8.64%	27.75
	*20*24 (2-grams)*	0.68	0.36	7.52%	27.50
	*20*24 (3-grams)*	0.68	0.35	19.81%	27.26
Μέγεθος Διανύσματος 2000	*18*27 (1-grams)*	0.69	0.38	0%	39.85
	*20*24 (2-grams)*	**0.69**	**0.40**	**0%**	**39.90**
	*20*24 (3-grams)*	0.69	0.36	0%	39.70
Μέγεθος Διανύσματος 3000	*18*27 (1-grams)*	0.69	0.40	2.51%	49
	*20*24 (2-grams)*	0.69	0.38	0%	49.55
	*20*24 (3-grams)*	0.69	0.37	0%	49.37

Πίνακας 18. $F1$ βαθμολογίες Μίκρο- και Μάκρο-μέσου όρου, TE και QE με χρήση ορθογώνιων χαρτών για διαφορετικά μεγέθη διανύσματος για τις 10 κλάσεις του συνόλου δεδομένων Reuters.

10 κλάσεις	Μέγεθος Χάρτη	Micro-Average	Macro-Average	TE	QE
Μέγεθος Διανύσματος 1000	*17*25 (1-grams)*	0.85	0.78	14.70%	27.51
	*18*23 (2-grams)*	0.85	0.77	98.20%	27.34
	*19*22 (3-grams)*	0.84	0.77	3.29%	27.13
Μέγεθος Διανύσματος 2000	*17*25 (1-grams)*	0.85	0.77	0.16%	39.38
	*18*23 (2-grams)*	0.85	0.77	1.32%	39.65
	*19*22 (3-grams)*	0.84	0.78	14.60%	39.53
Μέγεθος Διανύσματος 3000	*17*25 (1-grams)*	0.84	0.78	0%	48.48
	*18*23 (2-grams)*	**0.86**	**0.78**	**0%**	**49.27**

	*19*22 (3-grams)*	0.84	0.76	15.56%	49.20

Το τελευταίο μέρος αυτής της ενότητας είναι αφιερωμένο στη σύγκριση των πειραματικών μας αποτελεσμάτων με αυτά που παρουσιάζονται στο [39]. Ο Πίνακας 19 συνοψίζει τα αποτελέσματα των δύο προσεγγίσεων για τις 90 κατηγορίες και τετράγωνους χάρτες, ενώ ο Πίνακας 20 συγκρίνει τα αποτελέσματά μας με αυτά που παρουσιάζονται στο [39] χρησιμοποιώντας επίσης ορθογώνιους χάρτες και για τις 90 και για τις 10 κλάσεις του συνόλου δεδομένων Reuters.

Πρέπει να επισημανθεί ότι τα αποτελέσματα που παρουσιάζονται στον Πίνακα 20 σχετικά με τις 10 κλάσεις δεν μπορούν να συγκριθούν άμεσα, δεδομένου ότι στο [39] επιλέχθηκαν μόνο δείγματα που ανήκουν σε μία μόνο κλάση, μετατρέποντας το πρόβλημα σε ταξινόμηση μονής ετικέτας. Παρόλο που σε αυτή την περίπτωση δεν προτείνεται άμεση σύγκριση, η απόκλιση των βαθμολογιών $F1$ μεταξύ των δύο προσεγγίσεων είναι μικρή, 3% για τις βαθμολογίες Μίκρο-μέσου όρου και μόνο 4% για τις βαθμολογίες Μάκρο-μέσου όρου. Για τις 90 κλάσεις, επιτυγχάνουμε 0,78 χρησιμοποιώντας ένα πλέγμα 32*32, ενώ επιτυγχάνουμε 0,69 χρησιμοποιώντας ένα πλέγμα 19*26 για κατηγορίες με μεγάλο αριθμό δειγμάτων. Για τις βαθμολογίες Μάκρο-μέσου όρου επιτυγχάνουμε 0,40 χρησιμοποιώντας ένα πλέγμα 20*24 ενώ επιτυγχάνουμε 0,30 χρησιμοποιώντας ένα πλέγμα 128*128. Για τις 10 κλάσεις επιτυγχάνουμε 0,89 (βαθμολογία Μίκρο-μέσου όρου) και 0,82 (βαθμολογία Μάκρο-μέσου όρου) χρησιμοποιώντας ένα πλέγμα 32*32, ενώ πέτυχαμε 0,86 και 0,78 αντίστοιχα χρησιμοποιώντας ένα πλέγμα 18*23. Όπως αναφέρθηκε στην προηγούμενη ενότητα, η προσέγγισή μας δεν επικεντρώνεται μόνο στην εύρεση της υψηλότερης βαθμολογίας $F1$ αλλά και στην εξέταση των κατάλληλων επιλογών διαμόρφωσης σχετικά με το πόσο καλά τα μοντέλα ταιριάζουν με τη δεδομένη είσοδο εξετάζοντας επίσης μέτρα όπως TE και QE που δεν εξετάζονται στην άλλη προσέγγιση.

Πίνακας 19. Σύγκριση $F1$ βαθμολογιών Μάκρο-μέσου όρου με την προσέγγιση που παρουσιάστηκε στο [39] για τις 90 κλάσεις με χρήση τετράγωνων χαρτών.

Μέγεθος Χάρτη	16*16	32*32	64*64	128*128
Προσέγγιση	**0.14**	**0.26**	**0.30**	**0.30**
Προσέγγιση που παρουσιάστηκε στο [39]	0.13	0.25	0.25	0.30
Χρόνος Εκπαίδευσης (λεπτά)	0.18	0.19	2.23	24.17
Διεργασίες	4	4	4	4
% Διαφορά F1 βαθμολογίας	**+1**	**+1**	**+5**	0

Πίνακας 20. Σύγκριση $F1$ βαθμολογιών για τις 90 και 10 κλάσεις με την προσέγγιση που παρουσιάστηκε στο [39].

		Μέγεθος Χάρτη	Micro	Μέγεθος Χάρτη	Macro
90 κλάσεις	***Προσέγγιση***	19*26	0.69	20*24	**0.40**
	Προσέγγιση που παρουσιάστηκε στο [39]	32*32	0.78	128*128	0.30
10 κλάσεις	***Προσέγγιση***	18*23	0.86	18*23	0.78
	Προσέγγιση που παρουσιάστηκε στο [39]	32*32	0.89	32*32	0.82

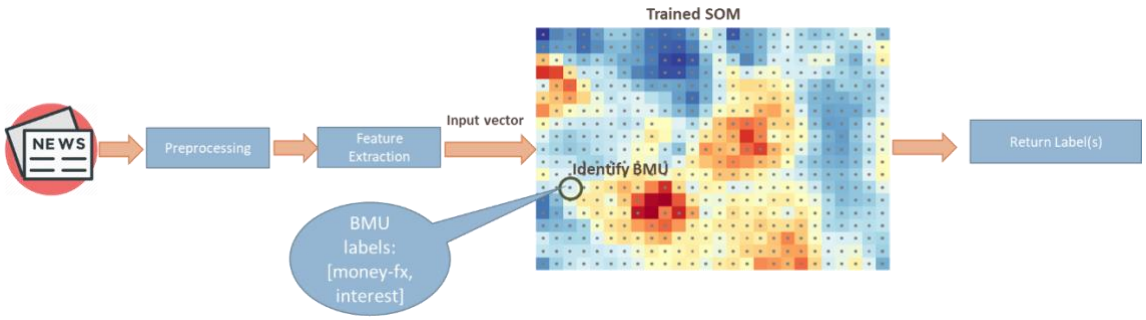
3.7. Συζήτηση

Χρησιμοποιώντας τον παραδοσιακό αλγόριθμο SOM, μπορέσαμε να παρατηρήσουμε τις σχέσεις μεταξύ των κλάσεων, καθώς τα έγγραφα που ανήκουν στην ίδια ή παρόμοια (από άποψη περιεχομένου) κατηγορία τοποθετήθηκαν σε γειτονικούς κόμβους του χάρτη. Αυτό δικαιολογεί την επιλογή του SOM έναντι άλλων παραδοσιακών προσεγγίσεων, καθώς το SOM είναι ικανό να απεικονίσει τέτοιες σχέσεις εγγενώς. Επομένως, με τη χρήση της προτεινόμενης προσέγγισης δεν ήταν απαραίτητο να μειωθεί/μετατραπεί το πρόβλημα σε ένα σύνολο ανεξάρτητων δυαδικών ταξινομητών, που δεν θα λάμβαναν υπόψη τη συσχέτιση μεταξύ των ετικετών. Επίσης, μετατρέψαμε το πρόβλημα ταξινόμησης πολλαπλής ετικέτας σε πρόβλημα πολλαπλών κλάσεων, χρησιμοποιώντας μια προσέγγιση διαφορετικών σταδίων για την κατασκευή των διανυσμάτων. Μετά την κατασκευή των διανυσμάτων, κάθε αντικείμενο του συνόλου δεδομένων που είχε πολλαπλές ετικέτες χωρίστηκε σε περισσότερα αντικείμενα, καθένα από τα οποία έχει μία ετικέτα από το αρχικό σύνολο ετικετών, αλλά διατηρώντας το τμήμα του διανύσματος άθικτο.

Επιπλέον, χρησιμοποιώντας την ευρετική διαδικασία που παρουσιάστηκε στην εξίσωση 9 και στον Αλγόριθμο 4 καταφέραμε να υπολογίσουμε το μέγεθος για το SOM, που υποδεικνύει τη χρήση ορθογώνιων χαρτών. Η επιλογή αυτή υποστηρίζεται περαιτέρω από τα αποτελέσματα που παρουσιάστηκαν προηγουμένως (Πίνακες 19 και 20) που υποδηλώνει ότι η επιλογή του μεγέθους του πλέγματος με βάση αυτή την προσέγγιση έχει καλύτερη απόδοση σε σύγκριση με τα τετράγωνα πλέγματα. Επιπλέον, κατά την αντιστοίχιση ετικετών στα δείγματα ελέγχου, σχεδιάσαμε και υλοποιήσαμε τον INTERSECT, έναν απλό αλλά αποτελεσματικό αλγόριθμο, σύμφωνα με τα αποτελέσματα της ταξινόμησης, ειδικά για τις $F1$ βαθμολογίες Μάκρο-μέσου όρου για το σύνολο δεδομένων Reuters.

Εν κατακλείδι, η προτεινόμενη προσέγγιση θα μπορούσε να χρησιμοποιηθεί ως ένας αποτελεσματικός ταξινομητής μη εποπτευόμενης ΜΜ για δεδομένα ειδήσεων. Σε αυτή την περίπτωση νέα στοιχεία δεδομένων θα μπορούσαν να τοποθετηθούν επάνω στον εκπαιδευμένο

SOM και έτσι να υπολογιστούν αυτόματα οι κατηγορίες στις οποίες ανήκουν χωρίς να χρειαστεί νέος γύρος εκπαίδευσης όπως φαίνεται και στην παρακάτω εικόνα.



Εικόνα 12. Χρήση της προσέγγισης ως ταξινομητή για νέα δεδομένα ειδήσεων.

Ακόμη, η προτεινόμενη προσέγγιση μπορεί να εφαρμοστεί σε ένα διαφορετικό σύνολο δεδομένων με ελάχιστες τροποποιήσεις. Σε αυτή την περίπτωση το σύστημα θα πρέπει να περάσει από μια νέα φάση εκπαίδευσης, όμως θεωρούμε πως σε συνδυασμό με χρήση του Αλγορίθμου 4 για την εύρεση του κατάλληλου μεγέθους για το SOM αυτό θα οδηγήσει στον ταχύτερο καθορισμό του κατάλληλου μεγέθους για το υπό εξέταση σύνολο δεδομένων. Στη συνέχεια θα ακολουθηθεί η αντίστοιχη διαδικασία προεπεξεργασίας και κατασκευής διανυσμάτων αλλά και εκπαίδευσης του δικτύου. Τέλος, τα δείγματα ελέγχου θα χρησιμοποιηθούν για τον καθορισμό της απόδοσης του δικτύου.

Συνοπτικά, η συνολική προσέγγιση που ακολουθήθηκε σε αυτή την εργασία βασίστηκε σε αρκετές πολλά υποσχόμενες επιλογές σχεδιασμού και αλγοριθμικές διαδικασίες, η αξιοπιστία και η αποτελεσματικότητα των οποίων τελικά δικαιολογήθηκαν από τα πειραματικά αποτελέσματα. Αυτές οι επιλογές μπορούν να συνοψιστούν ως εξής: α. στη χρήση του SOM ως μη-εποπτευόμενου ταξινομητή, β. στη μετατροπή των διανυσμάτων από πολλαπλής ετικέτας σε πολλαπλών κλάσεων, λαμβάνοντας επιπλέον μέριμνα για τη διατήρηση των σχέσεων μεταξύ των ετικετών, γ. στη διαδικασία υπολογισμού του μεγέθους για το SOM, δ. στη χρήση της batch SOM έκδοσης που συγκλίνει ταχύτερα και δεν απαιτεί την υπερπαράμετρο του ρυθμού εκμάθησης και ε. στον αλγόριθμο INTERSECT που προτείνουμε για την επιλογή της ετικέτας, λαμβάνοντας υπόψη τις ετικέτες που έχουν δοθεί και στους γειτονικούς κόμβους.

Κατά τη σύγκριση των αποτελεσμάτων μας με την προσέγγιση που παρουσιάστηκε στο [39] μπορέσαμε να καταλήξουμε σε ένα σύνολο από ενδιαφέροντα συμπεράσματα: και οι δύο προσεγγίσεις (τόσο η δική μας προσέγγιση όσο και εκείνη που παρουσιάστηκε στο [39]), διεξήγαγαν πειράματα χρησιμοποιώντας τις 90 και 10 κλάσεις του υποσυνόλου Mod Apte του Reuters, χρησιμοποιώντας τα ίδια μέτρα αξιολόγησης, καθιστώντας έτσι τις προσεγγίσεις συγκρίσιμες. Για να βρεθεί η BMU για κάθε δείγμα ελέγχου, η ευκλείδεια απόσταση επιλέχθηκε και στις δύο προσεγγίσεις.

Επιπλέον, το σχήμα στάθμισης που χρησιμοποιήθηκε για τη διαδικασία επιλογής χαρακτηριστικών βασίστηκε σε στατιστικά λέξεων και ειδικότερα στο σχήμα στάθμισης TF-IDF. Αντί να παίρνουμε όλες τις λέξεις από τα συμφραζόμενα μεταξύ 1-7 λέξεων όπως στο [39], τα πειράματα διεξήχθησαν χρησιμοποιώντας είτε καμία είτε 1 έως 2 λέξεις. Δεδομένου ότι το Reuters είναι ένα σύνολο δεδομένων ειδήσεων, η χρήση ενός εκτεταμένου παραθύρου γύρω από μια λέξη για την εύρεση των συμφραζόμενων δεν δικαιολογείται, καθώς τα ειδησεογραφικά στοιχεία έχουν

περιορισμένο μήκος και τα συμφραζόμενα διαφέρουν σημαντικά από στοιχείο σε στοιχείο για να επωφεληθεί η προσέγγιση από ένα τέτοιο παράθυρο. Τα αποτελέσματα που παρουσιάστηκαν υποστηρίζουν περαιτέρω αυτήν την παρατήρηση, καθώς η χρήση των συμφραζόμενων λέξεων δεν επηρεάζει σημαντικά την απόδοση της ταξινόμησης. Αξίζει να τονιστεί ότι η απόδοση ταξινόμησης βελτιώνεται όταν χρησιμοποιούνται 1-grams για κατηγορίες που έχουν μικρό αριθμό δειγμάτων.

Εκτός αυτού, χρησιμοποιήσαμε περίπου 34 φορές μικρότερο λεξιλόγιο για την κατασκευή των διανυσμάτων σε σχέση με το [39], που κυμαίνεται από 250 έως 3000 διαστάσεις ανά διάνυσμα αντί για τη χρήση 33120 διαστάσεων ανά διάνυσμα (που χρησιμοποιήθηκε στο [39]), γεγονός που θα προκαλούσε σημαντικά αυξημένες απαιτήσεις στον χρόνο υπολογισμού και επεξεργασίας. Πετύχαμε συγκρίσιμα αποτελέσματα με το [39] όταν χρησιμοποιήσαμε όλες τις (90) κλάσεις για τις βαθμολογίες Μίκρο-μέσου όρου και καλύτερα αποτελέσματα στις βαθμολογίες Μάκρο-μέσου όρου.

Επιπλέον, αντί να χρησιμοποιήσουμε την online έκδοση του SOM, χρησιμοποιήθηκε η batch έκδοση. Σύμφωνα με το [40], αυτή η έκδοση είναι προτιμότερη καθώς είναι πιο σταθερή. Μια άλλη διαφορά μεταξύ των δύο προσεγγίσεων έγκειται στη διαδικασία επιλογής της ετικέτας, καθώς στο [39] επιλέχθηκε η μέθοδος της πλειοψηφίας, η οποία αναθέτει στο δείγμα την πιο συχνή ετικέτα, για την ανάθεση των ετικετών στα δείγματα και η εξέταση των γειτονικών BMU γινόταν μόνο εάν δεν υπήρχε ετικέτα στη BMU. Συνεπώς, σε αυτή την περίπτωση εκχωρείται και πάλι μόνο μία ετικέτα σε κάθε κόμβο. Από την άλλη πλευρά, εμείς σχεδιάσαμε και υλοποιήσαμε τον δικό μας αλγόριθμο για την επιλογή ετικετών με βάση τις ετικέτες που ανατίθενται όχι μόνο στον κόμβο αλλά και στους γειτονικούς κόμβους.

Συνολικά, η προσέγγισή μας υποδεικνύει τη χρήση 1-grams και όχι χρήση των συμφραζόμενων γύρω από μια λέξη για αυτό το συγκεκριμένο σύνολο δεδομένων, καθώς η χρήση τους φαίνεται να βελτιώνει τα αποτελέσματα ταξινόμησης κατά 10% για τις $F1$ βαθμολογίες Μάκρο-μέσου όρου. Επιπρόσθετα, πετύχαμε 30 φορές μείωση της διάστασης των διανυσμάτων σε σχέση με την συγκρινόμενη προσέγγιση [39], ενώ η επιλογή του μεγέθους για το SOM με βάση τον αλγόριθμο που παρουσιάστηκε σε αυτή την εργασία οδήγησε στη χρήση χαρτών 34 φορές μικρότερων. Συνολικά, τα αποτελέσματά μας είναι συγκρίσιμα ή και καλύτερα σε ορισμένες περιπτώσεις (κατηγορίες με μικρό αριθμό δειγμάτων), ενώ χρησιμοποιούν λιγότερους πόρους και χρόνο για την εκπαίδευση, λόγω της χρήση 30 φορές μικρότερων διανυσμάτων και 34 φορές μικρότερου μεγέθους χάρτη, καθιστώντας έτσι την προσέγγιση κατάλληλη για καθημερινές εφαρμογές που δεν μπορούν να υποστηρίξουν μια χρονοβόρα διαδικασία εκπαίδευσης. Τέλος, καθορίσαμε και μια μεθοδολογία για την ταξινόμηση νέων δεδομένων ειδήσεων αλλά και για την προσαρμογή της προσέγγισης σε ένα νέο σύνολο δεδομένων.

3.8. Συμπεράσματα και μελλοντικές επεκτάσεις

Σε αυτή την εργασία, ερευνήσαμε τη χρήση τετράγωνων και ορθογώνιων SOMs για την αξιολόγηση της ικανότητας ταξινόμησης της προσέγγισης σε ένα πρόβλημα πολλαπλής ετικέτας. Η διαδικασία επιλογής χαρακτηριστικών είναι όσο το δυνατόν ελαφρύτερη και γενικότερη, προκειμένου να εξασφαλιστεί η δυνατότητα εφαρμογής της σε άλλα σύνολα δεδομένων με μικρές τροποποιήσεις. Η ακρίβεια ταξινόμησης έχει αυξηθεί κατά 10% για τις κατηγορίες με μικρό αριθμό δειγμάτων σε σχέση με άλλες προσεγγίσεις, ενώ υπερέχει από άποψη χρόνου και πόρων στη διαδικασία εκπαίδευσης. Επιπρόσθετα, αν έχουμε εξοικονόμηση πόρων μετά, κατά τη χρήση του μοντέλου, αλλά και με τη δυνατότητα ταξινόμησης νέων δεδομένων ειδήσεων χωρίς επανεκπαίδευση του μοντέλου.

Έχουμε δείξει ότι με τη χρήση της προτεινόμενης διαδικασίας επιλογής χαρακτηριστικών μπορούμε να καταγράψουμε τα πιο σημαντικά χαρακτηριστικά του συνόλου δεδομένων διατηρώντας παράλληλα το μέγεθος του διανύσματος σχετικά μικρό, βελτιώνοντας έτσι την αποδοτικότητα της ταξινόμησης. Επιπλέον, ο αλγόριθμος INTERSECT για την επιλογή της ετικέτας προτάθηκε με βάση την τομή των ετικετών που είχαν ανατεθεί μεταξύ της BMU και των γειτονικών κόμβων αυτής. Η προτεινόμενη προσέγγιση δοκιμάστηκε εκτενώς υπό διαφορετικές ρυθμίσεις διαμόρφωσης χρησιμοποιώντας το συχνά χρησιμοποιούμενο σύνολο δεδομένων συγκριτικής αξιολόγησης για προσεγγίσεις ταξινόμησης πολλαπλής ετικέτας, το υποσύνολο Mod Apte του συνόλου δεδομένων Reuters.

Έχουμε επίσης εξετάσει διάφορες μελλοντικές ερευνητικές κατευθύνσεις για την επέκταση και βελτίωση αυτής της προσέγγισης. Ένα βήμα είναι να πειραματιστούμε με τη χρήση άλλων συνόλων δεδομένων για να δείξουμε τη γενίκευση της προτεινόμενης προσέγγισης στο σύνολό της. Αυτό ισχύει και για την ενσωμάτωση άλλων αλγορίθμων Μηχανικής Μάθησης χωρίς επίβλεψη ή Βαθιάς Μάθησης προκειμένου να εκτιμηθεί η απόδοσή τους σε σύγκριση με το SOM υπό τις ίδιες πειραματικές συνθήκες. Τα παραπάνω αναλύονται στο επόμενο Κεφάλαιο της παρούσας Διατριβής.

4. Αυτόματη ταξινόμηση πολλαπλών κλάσεων (multi-class) μιας συλλογής ηλεκτρονικών βιβλίων (e-books) με βάση τον πίνακα περιεχομένων

4.1. Περίληψη

Η πρόταση/σύσταση συγγραμμάτων με σκοπό την υποστήριξη καθηγητών και φοιτητών στον εντοπισμό σχετικών πηγών είναι υψίστης σημασίας τόσο για τα πανεπιστήμια όσο και για τις ψηφιακές βιβλιοθήκες, και ως εκ τούτου παρέχει κίνητρα για την ανάπτυξη ενός συστήματος συστάσεων (recommendation system). Για το σκοπό αυτό, στο συγκεκριμένο Κεφάλαιο της παρούσας Διατριβής, αναλύεται μια μεθοδολογία για την αυτόματη ταξινόμηση πολλαπλών κλάσεων μιας συλλογής e-books από τη συλλογή Springer, η οποία διατίθεται μέσω της συνδρομής του Συνδέσμου Ελληνικών Ακαδημαϊκών Βιβλιοθηκών (ΣΕΑΒ). Η προσέγγιση που παρουσιάζεται εδώ χρησιμοποιεί δύο διαφορετικά νευρωνικά δίκτυα, ένα νευρωνικό δίκτυο μη εποπτευόμενης Μηχανικής Μάθησης – τους λεγόμενους Αυτό-Οργανούμενους Χάρτες, Self-Organizing Maps (SOM) οι οποίοι μελετήθηκαν στο προηγούμενο Κεφάλαιο - και δύο αρχιτεκτονικές νευρωνικών δικτύων Βαθιάς Μάθησης (Deep Neural Network - DNN), κάτω από διαφορετικά σενάρια διαμόρφωσης. Συγκεκριμένα χρησιμοποιείται ένα δίκτυο Μακράς Βραχείας Μνήμης (Long-Short Term Memory - LSTM) και ένα Συνελικτικό Νευρωνικό Δίκτυο (Convolutional Neural Network - CNN) σε συνδυασμό με ένα LSTM (CNN + LSTM). Οι δύο αυτές αρχιτεκτονικές επιλέχθηκαν έναντι άλλων αρχιτεκτονικών νευρωνικών δικτύων Βαθιάς Μάθησης καθώς με βάση την έρευνα που πραγματοποιήθηκε διαφαίνεται πως είναι οι καταλληλότεροι λόγω της φύσης των δεδομένων που χρησιμοποιούνται. Η αλληλουχία των λέξεων στις επικεφαλίδες των πινάκων περιεχομένων έχει σημασία και η αναγνώριση αυτών των συσχετίσεων μεταξύ διαδοχικών λέξεων μπορούν να βελτιώσουν την ακρίβεια σε ένα σύστημα αυτόματης ταξινόμησης.

Για την κατασκευή των διανυσμάτων αξιοποιούνται πληροφορίες που εξάγονται από τον πίνακα περιεχομένων (ToC) κάθε βιβλίου χρησιμοποιώντας το σχήμα στάθμισης (weighting scheme) TF-IDF (για την πρώτη περίπτωση) και τον Keras Tokenizer (για τη δεύτερη). Σε αυτό το σημείο θα πρέπει να αναφερθεί πως κατά τη διάρκεια των πειραμάτων δοκιμάσαμε και το σχήμα στάθμισης TF-IDF ως τρόπο αναπαράστασης των πινάκων περιεχομένων για τις δύο αρχιτεκτονικές DNN αλλά παρατηρήσαμε ότι η απόδοσή του ήταν χειρότερη και για αυτό το λόγο προτιμήθηκε ο Keras Tokenizer. Πραγματοποιήθηκαν εκτεταμένα πειράματα χρησιμοποιώντας διάφορες διαμορφώσεις και βήματα προεπεξεργασίας, διαμορφώσεις των νευρωνικών δικτύων και μεγέθη διανυσμάτων και λεξιλογίου για να εκτιμηθεί η επίδρασή τους στην απόδοση του ταξινομητή.

Επιπλέον στο Κεφάλαιο αυτό καταλήγουμε στο συμπέρασμα ότι η τεχνική “Majority Voting” είναι πιο κατάλληλη για την επιλογή της επικρατούσας ετικέτας για κάποιο συγκεκριμένο κόμβο στην περίπτωση του SOM. Η ανάλυση των πειραμάτων επιβεβαίωσε τη σκοπιμότητα ανάπτυξης ενός συστήματος συστάσεων για την υποστήριξη καθηγητών και φοιτητών στον εντοπισμό σχετικών πηγών, βάσει μιας λεπτομερούς θεματικής περιγραφής (π.χ. της περίληψης ή του πίνακα περιεχομένων ενός βιβλίου) αντί για μερικές λέξεις-κλειδιά. Στα πειράματα που διεξήχθησαν, το υποσύστημα που χρησιμοποίησε το DNN (LSTM) είχε την καλύτερη απόδοση, με βαθμολογίες $F1$ 67% για 26 κατηγορίες και 80% για 5 γενικές κατηγορίες, ενώ το SOM πέτυχε βαθμολογίες $F1$ μικρότερες κατά 5% και στις δύο περιπτώσεις. Οι 26 κατηγορίες μπορεί να είναι μια από τις

'Anthropology', 'Art', 'Computer Science', 'Linguistics', 'Literature', 'Management', 'Culture', 'Economics', 'Education', 'Engineering', 'Environment', 'Food', 'History', 'Humanities', 'Law', 'Life Sciences', 'Mathematics', 'Medicine', 'Music', 'Organization', 'Physical Sciences', 'Popular works', 'Religion', 'Social Sciences', 'Science', 'Transportation' ενώ οι 5 γενικές κατηγορίες είναι: 'Computer Science', 'Engineering', 'Mathematics', 'Medicine', 'Physics'.

4.2. Εισαγωγή

Κατά την αναζήτηση πηγών, καθηγητές και φοιτητές χρησιμοποιούν συχνά την ψηφιακή βιβλιοθήκη του Ιδρύματός τους ή και άλλων βιβλιοθηκών, εισάγοντας, ως όρους αναζήτησης, λέξεις-κλειδιά, που πιστεύουν ότι είναι σχετικές με το αναζητούμενο τεκμήριο, ως όρους αναζήτησης. Ωστόσο, τα αποτελέσματα που επιστρέφονται από τα συστήματα των ψηφιακών βιβλιοθηκών είναι συχνά μη σχετικά, καθώς μπορεί να ανήκουν σε ένα εντελώς διαφορετικό θέμα ή, αντίστροφα σχετικά βιβλία μπορεί να εξαιρούνται από τη λίστα αποτελεσμάτων, καθώς δεν περιέχουν τους ακριβείς όρους αναζήτησης που δόθηκαν από το χρήστη.

Το πρόβλημα αυτό επιτείνεται καθώς το περιεχόμενο μιας ψηφιακής βιβλιοθήκης αυξάνεται. Τα αποτελέσματα που επιστρέφονται από τα συστήματα της εκάστοτε βιβλιοθήκης ταξινομούνται με βάση την ομοιότητα του όρου αναζήτησης με τους όρους που περιλαμβάνονται στο τεκμήριο (βιβλίο, άρθρο σε περιοδικό, κλπ.). Ωστόσο, η χρήση μιας πιο λεπτομερούς περιγραφής του θέματος που αναζητά ο χρήστης (π.χ. η περιγραφή ενός μαθήματος ή ο πίνακας περιεχομένων ενός γνωστού βιβλίου) μπορεί να οδηγήσει σε πιο ακριβή αποτελέσματα, εάν υπάρχει κατάλληλο σύστημα ταξινόμησης και ανάκτησης το οποίο βασίζεται σε μέτρα ομοιότητας τέτοιων στοιχείων κειμένου. Σε αυτό το Κεφάλαιο διαμορφώνουμε το πρόβλημα της εύρεσης παρόμοιων βιβλίων ως πρόβλημα ταξινόμησης πολλαπλών κλάσεων, χρησιμοποιώντας πληροφορίες από τον Πίνακα Περιεχομένων (ToC) για αποτελεσματική αναζήτηση βιβλίων ή σχετικών πηγών.

Η εργασία της αυτόματης ταξινόμησης εγγράφων, και πιο συγκεκριμένα η ταξινόμηση πολλαπλών κλάσεων (multi-class classification), στην οποία κάθε έγγραφο μπορεί να αντιστοιχιστεί μόνο σε μία κλάση, είναι πολύ σημαντική σε διάφορους τομείς, συμπεριλαμβανομένων των ψηφιακών βιβλιοθηκών. Έχουν υπάρξει διάφορες ερευνητικές μελέτες που εστιάζουν στην ταξινόμηση πολλαπλών κλάσεων [76]-[82], [84], [85]. Η πλειονότητα των προσεγγίσεων στοχεύει στο μετασχηματισμό του προβλήματος ταξινόμησης πολλαπλών κλάσεων σε ένα σύνολο προβλημάτων δυαδικής ταξινόμησης. Σε αυτόν τον μετασχηματισμό, για κάθε κατηγορία που υπάρχει στο σύνολο δεδομένων, δημιουργείται και εκπαιδεύεται ένας δυαδικός ταξινομητής, έτσι ώστε κάθε ταξινομητής να μπορεί να διακρίνει εάν ένα δείγμα ελέγχου ανήκει σε μια συγκεκριμένη κατηγορία ή όχι. Δυστυχώς, αυτή η προσέγγιση θα αρκούσε μόνο υπό την προϋπόθεση ότι το σύνολο δεδομένων είναι πρακτικά αμετάβλητο, καθώς για κάθε νέα κατηγορία που προστίθεται, θα πρέπει να αναπτυχθεί ένας νέος ταξινομητής. Στο σύνολο δεδομένων που εξετάζουμε σε αυτό το Κεφάλαιο το οποίο προέρχεται από τη συλλογή Springer, διαθέσιμη μέσω συνδρομής του Συνδέσμου Ελληνικών Ακαδημαϊκών Βιβλιοθηκών (ΣΕΑΒ), ηλεκτρονικά βιβλία μπορούν να προστεθούν ή να αφαιρεθούν, ενώ νέες κατηγορίες μπορούν επίσης να εισαχθούν σε κάποιο σημείο, και ως εκ τούτου η μετατροπή σε ένα σύνολο δυαδικών προβλημάτων ταξινόμησης δημιουργεί σοβαρές δυσκολίες.

Παραδοσιακά, η αυτόματη ταξινόμηση ενός συνόλου ηλεκτρονικών εγγράφων σε ένα σύνολο διαφορετικών κλάσεων, που υποδηλώνεται από διαφορετικές ετικέτες, είναι μια εργασία ταξινόμησης εποπτευόμενης Μηχανικής Μάθησης [85]. Σε αυτήν την περίπτωση, ο ταξινομητής μαθαίνει με βάση τα επισημασμένα δείγματα από το σύνολο δεδομένων που χρησιμοποιήθηκαν

κατά τη φάση της εκπαίδευσης, και μετά το υποσύνολο δειγμάτων ελέγχου χρησιμοποιείται για την αξιολόγηση της απόδοσής του. Ωστόσο, αυτή η προσέγγιση απαιτεί ένα μεγάλο, καλά επισημασμένο σύνολο δεδομένων, ενώ θεωρείται αναποτελεσματική η εφαρμογή της σε σύνολα δεδομένων που έχουν μη ισορροπημένη κατανομή δειγμάτων μεταξύ των διαφορετικών κλάσεων. Αυτός ο περιορισμός επιβάλλει την επιλογή προσεγγίσεων μη εποπτευόμενης ΜΜ και Βαθιάς Μάθησης έναντι των τεχνικών εποπτευόμενης ΜΜ που παραδοσιακά επιλέγονται. Πρόσφατα, προεκπαιδευμένα γλωσσικά μοντέλα όπως τα Transformers, BERT, XLNet και GPT-2 έχουν προσελκύσει την προσοχή της ερευνητικής κοινότητας. Τα μοντέλα αυτά έχουν κριθεί αποτελεσματικά σε σύγκριση με αντίστοιχα μοντέλα που κατασκευάζονται από το μηδέν. Επιπλέον, η μεταφορά μάθησης, από δεδομένα που χρησιμοποιούνται σε μεθόδους εποπτευόμενης ΜΜ, μπορεί να εφαρμοστεί ως αντίδοτο στο πρόβλημα της ανισορροπίας των δειγμάτων μεταξύ των κλάσεων όπως θα αναλυθεί σε επόμενη ενότητα του παρόντος κεφαλαίου.

Σε αυτό το Κεφάλαιο, περιγράφεται η μεθοδολογία και αναλύονται τα αποτελέσματα μιας έρευνας που συνδυάζει δύο διαφορετικές προσεγγίσεις, συγκεκριμένα το SOM, και δύο διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων Βαθιάς Μάθησης, ένα LSTM και ένα CNN+LSTM. Το SOM, όπως αναλύθηκε και στο Κεφάλαιο 3, είναι μια μέθοδος μη εποπτευόμενης ΜΜ που μπορεί να συγκεντρώσει παρόμοια έγγραφα μαζί, με χρήση μέτρων ομοιότητας διανυσμάτων, χωρίς προηγούμενη γνώση των κλάσεων. Σε αυτήν την περίπτωση, οι ετικέτες των δειγμάτων εκπαίδευσης θα χρησιμοποιηθούν μόνο για να καθοδηγήσουν την εκχώρηση ετικετών στο σύνολο δειγμάτων ελέγχου μετά τη φάση της εκπαίδευσης. Μια ακόμα διαφοροποίηση σε σχέση με τις εποπτευόμενες τεχνικές ΜΜ, είναι ότι ένας εκπαιδευμένος SOM περιέχει αναγνωρίσιμες αλλά μη οριοθετημένες συστάδες (clusters) [40].

Με βάση την ανάλυση που πραγματοποιήθηκε μπορούμε να παρατηρήσουμε ότι στις συστάδες που δημιουργούνται, παρόμοια έγγραφα τείνουν να μοιράζονται κοινές ετικέτες και επομένως ο SOM διατηρεί εγγενώς τη συσχέτιση μεταξύ των ετικετών, καθώς οι σχετικές ετικέτες μπορούν να βρεθούν είτε στον ίδιο είτε σε έναν γειτονικό κόμβο στον χάρτη. Παρομοίως, τα DNN χρησιμοποιούν ενσωματώσεις διανυσμάτων (vector embeddings) για να συγκεντρώσουν παρόμοια έγγραφα σε έναν χώρο διανυσμάτων χαμηλής διάστασης. Σε αντίθεση με το SOM, ωστόσο, η επιλεγμένη προσέγγιση DNN είναι εποπτευόμενη, καθώς οι ετικέτες χρησιμοποιούνται για να υποδείξουν την κλάση κάθε δείγματος κατά τη φάση εκπαίδευσης. Η προσέγγιση DNN επιλέχθηκε καθώς υπάρχουν αναφορές στη βιβλιογραφία ότι παρουσιάζει υψηλή ακρίβεια σε εργασίες ανάλυσης φυσικής γλώσσας, ενώ δεν απαιτεί εκτεταμένη κατασκευή χαρακτηριστικών (feature engineering).

Εδώ, εστιάζουμε στη σύγκριση των δύο προαναφερθέντων NN για την ταξινόμηση πολλαπλών κατηγοριών ενός συνόλου δεδομένων χρησιμοποιώντας διαφορετικές επιλογές διαμόρφωσης. Αυτές αφορούν διάφορα μεγέθη πλέγματος, χρησιμοποιώντας διαφορετικά σύνολα παραμέτρων, όπως το μέγεθος χάρτη και διανύσματος, το σχήμα στάθμισης, καθώς και μέγιστο μήκος του Πίνακα Περιεχομένων (ToC) και το μέγεθος του λεξιλογίου που χρησιμοποιείται έχοντας ως στόχο να δείξουμε ότι τέτοιες επιλογές μπορούν να επηρεάσουν σημαντικά την απόδοση του ταξινομητή.

Με βάση αυτόν τον πειραματισμό και τις παρατηρήσεις που εξάγονται, προτείνουμε μια μεθοδολογία για αυτόματη ταξινόμηση πολλαπλών κλάσεων που βασίζεται αποκλειστικά σε πληροφορίες από τους Πίνακες Περιεχομένων των βιβλίων. Από όσο γνωρίζουμε, αυτή είναι η πρώτη προσπάθεια να πραγματοποιηθεί μια τέτοια σύγκριση. Επιπλέον, για την αποφυγή εκτεταμένης

κατασκευής χαρακτηριστικών, χρησιμοποιήσαμε ένα απλό σύνολο χαρακτηριστικών που βασίζονται σε στατιστικά στοιχεία λέξεων και έτσι μπορούν εύκολα να εφαρμοστούν σε οποιοδήποτε σύνολο δεδομένων με μικρές τροποποιήσεις. Ακόμα, δύο διαφορετικοί αλγόριθμοι χρησιμοποιήθηκαν για την επιλογή της ετικέτας που θα ανατεθεί σε κάθε κόμβο στο χάρτη μετά τη φάση της εκπαίδευσης, συγκεκριμένα ο αλγόριθμος Majority Voting και ο αλγόριθμος INterSECT που παρουσιάστηκε στο προηγούμενο Κεφάλαιο, ο οποίος επιλέγει τις ετικέτες που θα εκχωρηθούν σε έναν κόμβο με βάση τις ετικέτες που υπάρχουν στους γειτονικούς. Καθώς ο INterSECT είναι προσαρμοσμένος σε σενάρια ταξινόμησης πολλαπλής ετικέτας (multi-label), ήταν αναμενόμενο ότι ο αλγόριθμος Majority Voting θα ήταν πιο κατάλληλος για το πρόβλημα πολλαπλών κλάσεων που αναλύεται σε αυτό το Κεφάλαιο. Η συνεισφορά αυτού του κεφαλαίου αναλύεται παρακάτω:

- Δημιουργήσαμε ένα σύνολο δεδομένων που αποτελείται από περίπου 56 χιλιάδες ηλεκτρονικά βιβλία από τον εκδοτικό οίκο Springer, το οποίο διαθέτουμε ελεύθερα στην κοινότητα.
- Προτείνουμε μια μεθοδολογία για την αυτόματη ταξινόμηση ενός συνόλου δεδομένων πολλαπλών κλάσεων χρησιμοποιώντας πληροφορίες από τους πίνακες περιεχομένων των βιβλίων.
- Αποδεικνύουμε ότι η επιλεγμένη διαδικασία επιλογής χαρακτηριστικών μειώνει την ανάγκη για εκτεταμένη κατασκευή χαρακτηριστικών και δεν επηρεάζεται αν το υποκείμενο σύνολο δεδομένων αλλάξει.
- Αξιολογούμε δύο διαφορετικές μεθόδους για την επιλογή ετικέτας για κάθε κόμβο στο σενάριο του SOM. Αποδεικνύουμε ότι ο αλγόριθμος Majority Voting είναι πιο αποτελεσματικός από τον INterSECT.
- Αξιολογούμε τη συνολική προσέγγιση μέσω εκτεταμένων πειραμάτων που διεξήγαμε με χρήση ενός νευρωνικού δικτύου μη εποπτευόμενης ΜΜ (SOM) και δύο διαφορετικών DNN αρχιτεκτονικών (LSTM, CNN+LSTM). Τα αποτελέσματα δείχνουν ότι το LSTM αποδίδει καλύτερα από το SOM σε ποσοστό 5%.

Το υπόλοιπο αυτού του κεφαλαίου οργανώνεται ως εξής. Στην Ενότητα 4.3, παρέχεται μια ανασκόπηση των πρόσφατων προσεγγίσεων σχετικά με την ταξινόμηση εγγράφων πολλαπλών κλάσεων, την επιλογή χαρακτηριστικών καθώς και των προσεγγίσεων ταξινόμησης που χρησιμοποιούν SOM και DNN. Η Ενότητα 4.4 περιγράφει το σύνολο δεδομένων που χρησιμοποιήθηκε για την εκτέλεση των πειραμάτων/προσομοιώσεων, ενώ η Ενότητα 4.5 παρουσιάζει λεπτομερώς την προτεινόμενη μεθοδολογία. Στην Ενότητα 4.6 γίνεται ανάλυση της πειραματικής αξιολόγησης: συγκεκριμένα αναλύονται τα πειράματα που πραγματοποιήθηκαν, τα μέτρα αξιολόγησης που χρησιμοποιήθηκαν και τα αποτελέσματα που ελήφθησαν, ακολουθούμενα από μια συνολική συζήτηση επί αυτών. Τέλος, η Ενότητα 4.7 ολοκληρώνει το Κεφάλαιο αυτό και παρέχει μερικές σκέψεις σχετικά με βελτιώσεις και ερευνητικά ζητήματα που θα μπορούσαν να εξεταστούν στο μέλλον.

4.3. Σχετικές ερευνητικές εργασίες

4.3.1. Ταξινόμηση εγγράφων πολλαπλών κλάσεων

Ο κύριος στόχος μιας εργασίας ταξινόμησης εγγράφων είναι η κατάταξη κάθε στοιχείου σε μία ή περισσότερες κατηγορίες, ανάλογα με το αν πρόκειται για ταξινόμηση πολλαπλών κατηγοριών ή

πολλαπλής ετικέτας [77]. Με βάση τα δεδομένα εκπαίδευσης, το σύστημα μπορεί να ταξινομήσει νέα στοιχεία στην κλάση που αντιστοιχούν. Αυτό μπορεί να γίνει αυτόματα με τη βοήθεια ενός συστήματος ταξινόμησης, το οποίο έχει εκπαιδευτεί σχετικά, επί ενός συνόλου δεδομένων εκπαίδευσης.

Στην πλειονότητα των προσεγγίσεων, για την κατανομή προηγουμένως άγνωστων εγγράφων σε μία ή περισσότερες κλάσεις, χρησιμοποιούνται τεχνικές εποπτευόμενης ΜΜ, καθώς θεωρούνται πιο αποτελεσματικές. Ένα μειονέκτημα αυτών των προσεγγίσεων είναι ότι απαιτούν ένα καλά σχολιασμένο σύνολο δεδομένων μεγάλης κλίμακας που να παρουσιάζει ισορροπία στον αριθμό των δειγμάτων μεταξύ των διαφορετικών κλάσεων [28], [85] και για το λόγο αυτό παρουσιάζουν χαμηλότερη απόδοση στις περιπτώσεις που τα δείγματα είναι δεν ισομερή μεταξύ των κλάσεων. Το σύνολο ηλεκτρονικών βιβλίων του Springer, που χρησιμοποιείται στην περίπτωσή μας, είναι ένα σύνολο δεδομένων πολλαπλών κλάσεων με μια μη ισορροπημένη κατανομή δειγμάτων μεταξύ των κλάσεων, θέτοντας έτσι προκλήσεις στην ενσωμάτωση εποπτευόμενων τεχνικών.

Έχουν προταθεί αρκετές προσεγγίσεις για την επίλυση του προβλήματος της ταξινόμησης πολλαπλών κλάσεων. Η εργασία που προτείνεται στο [76], στοχεύει στη μοντελοποίηση του προβλήματος της εύρεσης της σχέσης μεταξύ δύο εγγράφων ως μια εργασία ταξινόμησης πολλαπλών τάξεων με τη χρήση της Wikipedia και του Wikidata για την εξαγωγή σημασιολογικών σχέσεων, ενώ ο αλγόριθμος SCVD-MS παρουσιάστηκε στο [81], ο οποίος χρησιμοποιεί ενσωματώσεις διανυσμάτων πολλαπλής λογικής (multi-sense embeddings) για τη βελτίωση της ταξινόμησης πολλαπλών κλάσεων στο σύνολο δεδομένων 20NewsGroup[16] ενώ ταυτόχρονα στοχεύει στην αναπαράσταση των διανυσμάτων σε ένα χώρο μικρότερων διαστάσεων σε σχέση με τον προκάτοχό του SCVD. Το ίδιο σύνολο δεδομένων είχε χρησιμοποιηθεί και στο [77], όπου προτάθηκε μια επέκταση των αλγορίθμων Word2Vec και FastText [87]. Σε αυτή την εργασία οι ενσωματώσεις διανυσμάτων είχαν επαυξηθεί με σημασιολογική πληροφορία με την ανάθεση ετικετών με χρήση σχημάτων λόγου (Part-of-Speech - POS) σε κάθε λέξη με τελικό στόχο την αξιολόγηση της απόδοσης του επαυξημένου μοντέλου στην ταξινόμηση πολλαπλών κλάσεων.

Μια άλλη προσέγγιση για την ταξινόμηση πολλαπλών κλάσεων παρουσιάστηκε στο [78], στην οποία εφαρμόστηκε μια βελτιωμένη τεχνική για τη δημιουργία διανυσμάτων με βάση τον αλγόριθμο Naive Bayes, για να επιτύχει τη μείωση του χώρου των διανυσμάτων, σε συνδυασμό με έναν ταξινομητή SVM. Ομοίως, στο [79] παρουσιάστηκε ένα μοντέλο ταξινόμησης πολλαπλών κλάσεων που βασίστηκε στη θεωρία της κβαντικής ανίχνευσης (quantum detection). Η εργασία που παρουσιάστηκε στο [80] διερεύνησε το πρόβλημα της πολύγλωσσης ταξινόμησης συναισθημάτων πολλαπλών κλάσεων χρησιμοποιώντας συνελικτικά νευρωνικά δίκτυα τα οποία εφάρμοσε επάνω σε τρία (3) διαφορετικά σύνολα δεδομένων σε τρείς (3) διαφορετικές γλώσσες. Σύμφωνα με τους συγγραφείς, ο παράγοντας που διαφοροποιεί τη μελέτη τους από άλλες συναφείς μελέτες είναι ότι το παραγόμενο μοντέλο δεν βασίζεται σε συγκεκριμένα εργαλεία, όπως οντολογίες ή λεξικά και δεν απαιτεί μορφολογική ή συντακτική προεπεξεργασία. Χρησιμοποιούν επίσης την τεχνική της υπερδειγματοληψίας για την αντιμετώπιση του προβλήματος της ανισορροπίας των δειγμάτων μεταξύ των διαφορετικών κλάσεων.

Το [82] αντιμετώπισε την ταξινόμηση πολλαπλών κλάσεων από διαφορετική οπτική γωνία. Το HDLTex προτάθηκε σε αυτήν τη μελέτη ως μια προσέγγιση ιεραρχικής ταξινόμησης εφαρμόζοντας

[16] http://qwone.com/~jason/20Newsgroups/

διαφορετικές στοίβες αρχιτεκτονικών νευρωνικών δικτύων βαθιάς μάθησης σε κάθε επίπεδο της ιεραρχίας εγγράφων. Η συν-εκπαίδευση (co-training), μια προσέγγιση ημι-εποπτευόμενης ΜΜ παρουσιάστηκε στο [84] για την αποτελεσματική ταξινόμηση εγγράφων. Χρησιμοποιήθηκαν τρεις διαφορετικές μέθοδοι αναπαράστασης εγγράφων, συγκεκριμένα το σχήμα στάθμισης TF-IDF, η μέθοδος Latent Dirichlet Allocation (LDA) και ο αλγόριθμος Document to Vector (Doc2Vec) ενώ στη συνέχεια η μέθοδος αυτή αξιολογήθηκε σε πέντε (5) διαφορετικά σύνολα δεδομένων πολλαπλών κλάσεων. Σε αυτή την εργασία, οι συγγραφείς προσπαθούν να λύσουν το πρόβλημα των ανεπαρκών πληροφοριών για τις ετικέτες κάθε δείγματος δεδομένων χρησιμοποιώντας μια ημι-εποπτευόμενη προσέγγιση με την ενσωμάτωση διαφορετικών μεθόδων αναπαράστασης εγγράφων για την επίλυση του προβλήματος των αδόμητων εγγράφων που περιέχουν λίγη πληροφορία.

Στην παρούσα εργασία, χρησιμοποιείται ένα μη εποπτευόμενο Νευρωνικό Δίκτυο (SOM) και δύο διαφορετικές αρχιτεκτονικές Βαθιάς Μάθησης (DNN), συγκεκριμένα ένα LSTM και ένα CNN+LSTM. Όπως θα αναλυθεί και σε επόμενη παράγραφο, για την κατασκευή των διανυσμάτων αναπαράστασης των βιβλίων χρησιμοποιείται η πληροφορία από τους πίνακες περιεχομένων, με το σχήμα στάθμισης TF-IDF να εφαρμόζεται στην περίπτωση του SOM, ενώ αξιοποιείται ο Keras Tokenizer στα Δίκτυα DNN. Εναλλακτικές μέθοδοι για την αναπαράσταση των πινάκων περιεχομένων θα μπορούσαν να είναι ο LDA και ο αλγόριθμος Doc2Vec. Όπως αναφέρθηκε και παραπάνω κατά τη διάρκεια των πειραμάτων δοκιμάσαμε και το σχήμα στάθμισης TF-IDF ως τρόπο αναπαράστασης των πινάκων περιεχομένων για τις δύο αρχιτεκτονικές DNN αλλά παρατηρήσαμε ότι η απόδοσή του ήταν χειρότερη και για αυτό το λόγο προτιμήθηκε ο Keras Tokenizer, τα αποτελέσματα του οποίου παρατίθενται στην παράγραφο 4.6.4.

4.3.2. Προεκπαιδευμένα γλωσσικά μοντέλα

Πρόσφατα, πολλές ερευνητικές προσπάθειες βασίστηκαν σε προεκπαιδευμένα γλωσσικά μοντέλα (pre-trained language models). Οι προεκπαιδευμένες ενσωματώσεις διανυσμάτων έχουν κριθεί πιο αποτελεσματικές για αρκετές εργασίες επεξεργασίας φυσικής γλώσσας συγκριτικά με τα μοντέλα που μαθαίνουν από το μηδέν [91]. Η χρήση προεκπαιδευμένων μοντέλων έχει επίσης το πλεονέκτημα της μεταφοράς μάθησης από εποπτευόμενα δεδομένα, τα οποία μπορούν να αποτελέσουν λύση στο πρόβλημα ανισορροπίας των δειγμάτων εκπαίδευσης μεταξύ των διαφορετικών κλάσεων, όπως συζητήθηκε παραπάνω.

Η προσέγγιση που προτείνεται στο [76] στοχεύει στη μοντελοποίηση του προβλήματος του εντοπισμού της σχέσης μεταξύ δύο εγγράφων ως μια εργασία ταξινόμησης εγγράφων πολλαπλών κλάσεων, χρησιμοποιώντας τη Wikipedia και τα Wikidata για την εξαγωγή των σημασιολογικών σχέσεων μεταξύ των ζευγών εγγράφων. Ο προσδιορισμός των σημασιολογικών σχέσεων μεταξύ εγγράφων αξιοποιεί κάποια από τις παρακάτω πέντε μεθόδους: GloVe [90], Paragraph-Vectors [89], BERT [91], XLNet [92] και GPT-2 [93]. Επιπλέον, σε αυτή τη μελέτη, έχουν χρησιμοποιηθεί δύο διαφορετικές αρχιτεκτονικές μετασχηματιστών (Transformers), ένας Vanilla και ένας Siamese. Στην τελευταία περίπτωση, τα προ εκπαιδευμένα γλωσσικά μοντέλα BERT και XLNet έχουν συνδυαστεί σε μια Siamese αρχιτεκτονική.

Σε αντίθεση με τον αλγόριθμο GloVe και τη μέθοδο Paragraph-Vectors, τα οποία είναι ενσωματώσεις διανυσμάτων που δεν λαμβάνουν υπόψη τα συμφραζόμενα, το BERT και το XLNet έχουν τη δυνατότητα να χρησιμοποιούν και τα συμφραζόμενα. Έτσι, ο συνδυασμός ενσωματώσεων που χρησιμοποιούν και τα συμφραζόμενα με την αρχιτεκτονική των μετασχηματιστών οδήγησε στην μη εποπτευόμενη προ εκπαίδευση γλωσσικών μοντέλων βαθιάς μάθησης, η οποία φαίνεται ότι

παράγει αξιοσημείωτα αποτελέσματα σε διάφορες εργασίες NLP. Ωστόσο, μέχρι τώρα, οι μετασχηματιστές δεν έχουν χρησιμοποιηθεί ευρέως σε άλλους τομείς όπως τα συστήματα προτάσεων περιεχομένου. Μόνο στο [95] έχει χρησιμοποιηθεί το BERT για σύσταση επιστημονικών εργασιών.

Σύμφωνα με το [94], ένας μετασχηματιστής είναι το πρώτο μοντέλο μεταγωγής (transduction model) που βασίζεται αποκλειστικά στην αυτό-προσοχή (self-attention) για τον υπολογισμό των αναπαραστάσεων της εισόδου και της εξόδου του, αντικαθιστώντας τα επαναλαμβανόμενα στρώματα που χρησιμοποιούνται συχνότερα σε αρχιτεκτονικές κωδικοποιητή-αποκωδικοποιητή με την αυτό-προσοχή πολλαπλών επιπέδων (multiheaded). Οι συγγραφείς υποστηρίζουν ότι χρησιμοποιώντας την αυτό-προσοχή, η υπολογιστική πολυπλοκότητα για κάθε στρώμα είναι μικρότερη σε σύγκριση με εκείνα των επαναλαμβανόμενων ή συνελικτικών νευρωνικών δικτύων. Επιπλέον, μπορεί να επιτευχθεί μεγαλύτερος βαθμός παραλληλοποίησης όσον αφορά τον υπολογισμό, καθώς, στην περίπτωση αυτή, απαιτούνται σημαντικά λιγότερες διαδοχικές λειτουργίες.

Η μέθοδος BERT στοχεύει στην προ εκπαίδευση βαθιών αμφίδρομων αναπαραστάσεων από μη επισημασμένο κείμενο λαμβάνοντας υπόψη τόσο τα συμφραζόμενα από τη δεξιά πλευρά κάθε λέξης όσο και από τα αριστερά για όλα τα επίπεδα [91]. Η αρχιτεκτονική του BERT μπορεί να γίνει αντιληπτή ως ένας πολυστρωματικός κωδικοποιητής διπλής κατεύθυνσης που βασίζεται στην αρχιτεκτονικών των μετασχηματιστών. Το κύριο πλεονέκτημα αυτής της τεχνικής είναι ότι μπορεί εύκολα να προσαρμοστεί σε διάφορες εργασίες NLP μόνο με την προσθήκη ενός επιπρόσθετου επιπέδου εξόδου στο βασικό προ εκπαιδευμένο μοντέλο. Έχει αποδειχθεί ότι οι προ εκπαιδευμένες αναπαραστάσεις μειώνουν την ανάγκη για αρχιτεκτονικές ειδικά σχεδιασμένες για συγκεκριμένες εργασίες.

Το XLNet είναι μια γενικευμένη μέθοδος αυτόματης προ εκπαίδευσης που επιτρέπει την εκμάθηση αμφίδρομων πλαισίων γύρω από μια λέξη, καθώς το πλαίσιο για κάθε λέξη μπορεί να αποτελείται από τμήματα κειμένου τόσο από τα αριστερά όσο και από τα δεξιά αυτής [92]. Οι συγγραφείς προτείνουν ότι εμπειρικά, κάτω από συγκρίσιμες ρυθμίσεις, η απόδοση του XLNet ξεπερνά αυτή του BERT σε 20 διαφορετικές εργασίες λόγω της αυτό-παλινδρομικής διαμόρφωσής (autoregressive formulation) του. Υποστηρίζουν επίσης ότι το XLNet εγγενώς εξαλείφει την υπόθεση που κάνει ο BERT ότι τα προβλεπόμενα τμήματα κειμένου είναι ανεξάρτητα το ένα από το άλλο.

Το GPT-2 είναι ένα μεγάλο γλωσσικό μοντέλο με 1,5 δισεκατομμύριο παραμέτρους που έχει εκπαιδευτεί σε ένα σύνολο δεδομένων 8 εκατομμυρίων ιστοσελίδων και αξιοποιεί την αρχιτεκτονική των μετασχηματιστών (Transformers) [93]. Ο κύριος στόχος του είναι να προβλέψει την επόμενη λέξη στο σενάριο ότι όλες οι προηγούμενες λέξεις είναι γνωστές. Παρόμοια με τα BERT και XLNet, το GPT-2 είναι μια ενσωμάτωση διανυσμάτων που λαμβάνει υπόψη τα συμφραζόμενα και είναι δέκα φορές μεγαλύτερο (όσον αφορά τις παραμέτρους και τον όγκο των δεδομένων εκπαίδευσης) σε σχέση με τον προκάτοχό του GPT. Έχει αξιολογηθεί σε διάφορες εργασίες, όπως η απάντηση ερωτήσεων (question answering), η κατανόηση της ανάγνωσης (reading comprehension), η σύνοψη (summarization) και η μετάφραση, η δε αξιοσημείωτη απόδοση του GPT-2 πραγματοποιείται χωρίς τη χρήση δεδομένων εκπαίδευσης για συγκεκριμένες εργασίες. Το κύριο συμπέρασμα αυτής της μελέτης είναι ότι αυτές οι εργασίες μπορούν να επωφεληθούν από μη εποπτευόμενες τεχνικές ΜΜ εάν τα δεδομένα και η υπολογιστική ισχύς είναι επαρκή.

Ωστόσο, οι μετασχηματιστές, BERT, XLNet και GPT-2 θεωρούνται αποτελεσματικοί κυρίως σε εργασίες όπως η απάντηση ερωτήσεων, η συλλογιστική φυσικής γλώσσας, η ανάλυση συναισθημάτων, η κατάταξη εγγράφων και η μετάφραση πλήρους κειμένου. Στην περίπτωσή μας, η χρήση των πινάκων περιεχομένων και όχι του πλήρους κειμένου (καθώς αυτό δεν ήταν διαθέσιμο) δημιουργεί προκλήσεις για την ενσωμάτωση τέτοιων τεχνικών, αλλά μπορεί να παρέχει ενδιαφέρουσες πληροφορίες για την εφαρμογή τους σε άλλους τομείς, όπως για παράδειγμα σε συστήματα προτάσεων περιεχομένου και στην ταξινόμηση πολλαπλών κλάσεων χρησιμοποιώντας πληροφορίες από τους πίνακες περιεχομένων των βιβλίων. Το τελευταίο αφήνεται ως μελλοντική επέκταση της παρούσας Διατριβής.

4.3.3. Επιλογή Χαρακτηριστικών

Μια σημαντική πτυχή κατά τη δημιουργία ενός συστήματος αυτόματης ταξινόμησης εγγράφων, είναι η αποτελεσματική αναπαράσταση του κειμένου [97]. Οι περισσότερες από τις προσεγγίσεις που συναντώνται στη βιβλιογραφία κωδικοποιούν τα κείμενα με τη βοήθεια αριθμητικών διανυσμάτων, στα οποία μπορούν να εφαρμοστούν μέθοδοι κβαντοποίησης (vector quantization), με στόχο τη μείωση της διάστασης του χώρου των χαρακτηριστικών (feature space) [51], [101]. Το μέγεθος του χώρου των χαρακτηριστικών παίζει σημαντικό ρόλο στην απόδοση του ταξινομητή [98]. Κατανεμημένες ενσωματώσεις διανυσμάτων όπως τα word2vec [88], Doc2vec [89] και GloVe [90] έχουν αποδειχθεί αποτελεσματικές για την κωδικοποίηση της σημασιολογικής πληροφορίας από κείμενα, ενώ τα αναπαριστούν σε έναν χώρο χαμηλότερων διαστάσεων. Οι μέθοδοι για την αποτελεσματική αναπαράσταση κειμένων θεωρούνται ύψιστης σημασίας, καθώς αν ενσωματωθούν πολλά άσχετα ή περιττά χαρακτηριστικά στα διανύσματα που θα χρησιμοποιηθούν, αυτά μπορεί να επηρεάσουν σημαντικά την απόδοση του ταξινομητή [96], [53].

Διαφορετικές προσεγγίσεις έχουν προταθεί για την αναπαράσταση εγγράφων, όπως το απλό Bag-of-words μοντέλο [32], η Συχνότητα Όρων (Term Frequency -TF) και η Αντίστροφη Συχνότητα Εγγράφου (Inverse Document Frequency - IDF), ο αλγόριθμος LDA [84] οι μετρικές Information Gain [86], [100], Mutual Information και Pointwise Mutual Information [101], το x2-test [99] και η Βαρύτητα ενός όρου (Term Strength) καθώς και συνδυασμοί αυτών, ενώ η TF συνδυάστηκε στο [41] και με έναν αλγόριθμο αποκοπής (stemming. Τροποποιημένα σχήματα στάθμισης που βασίζονται στην συχνότητα των όρων μέσα σε ένα κείμενο προτάθηκαν στο [99], έτσι ώστε να λαμβάνουν υπόψη και τον αριθμό των όρων που λείπουν όταν υπολογίζουν τα βάρη για κάθε όρο μέσα στο κείμενο.

Ο απλός μέσος όρος των διανυσμάτων λέξεων με ένα μη σταθμισμένο σχήμα παρουσιάστηκε στο [36]. Επιπλέον, μελετήθηκε ένας απλός μέσος όρος του σχήματος στάθμισης TF-IDF των διανυσμάτων λέξεων για την παραγωγή διανυσμάτων εγγράφων στο [102]. Η μέθοδος Sparce Composite Document Vector (SCDV) που προτάθηκε στο [103], επέκτεινε τη σταθμισμένη μέση τιμή των διανυσμάτων λέξεων από το επίπεδο προτάσεων στο επίπεδο των εγγράφων χρησιμοποιώντας «ελαφριά» ομαδοποίηση (soft clustering) σε διανύσματα λέξεων ενώ στο [104] η προσέγγιση επεκτάθηκε για να καλύψει επίσης την πολυσημία των λέξεων και να αντιμετωπίσει το πρόβλημα του υψηλού χώρου διαστάσεων. Αυτό επιτεύχθηκε με τη χρήση ενσωματώσεων διανυσμάτων πολλαπλής λογικής (multi-sense embeddings) σε ένα χώρο μικρότερων διαστάσεων για τη μάθηση.

Τα DNNs, όπως τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNN) [104], [105], τα Αναδρομικά Νευρωνικά Δίκτυα (Recursive Neural Networks - RNN) [81] και τα δίκτυα Long Short Term Memory (LSTM) [106], έχουν επίσης χρησιμοποιηθεί για την αναπαράσταση των συντακτικών ιδιομορφιών κατά την αναπαράσταση εγγράφων.

Η χρήση νέων μη εποπτευόμενων μεθόδων μηχανικής μάθησης που βασίζονται σε έναν αλγόριθμο αποκοπής με βάση τη μέθοδο LDA και σημασιολογικούς χώρους έχει προταθεί στο [58]. Ο LDA έχει χρησιμοποιηθεί για την εξαγωγή αναπαραστάσεων υψηλού επιπέδου από έγγραφα, αλλά απαιτεί εκτεταμένη ρύθμιση υπερπαραμέτρων, όπως για παράδειγμα του αριθμού των κλάσεων ή την κατανομή των λέξεων και των θεματικών ενοτήτων [107]. Ο υπολογισμός του μέσου όρου των διανυσμάτων λέξεων με βάση μεθόδους που βασίζονται σε hard clustering προτάθηκε στο [108], ενώ η ασαφής ομαδοποίηση (fuzzy clustering) σε συνδυασμό με το σχήμα στάθμισης TF-IDF στο [103] με σκοπό τη δημιουργία διανυσμάτων εγγράφων.

Σε αυτό το Κεφάλαιο της παρούσας Διατριβής, εστιάζουμε στην επιλογή αντιπροσωπευτικών χαρακτηριστικών κάθε βιβλίου χρησιμοποιώντας τον Keras Tokenizer για τις δύο αρχιτεκτονικές DNN (LSTM και το CNN+LSTM), και το σχήμα στάθμισης TF-IDF σε συνδυασμό με τον αλγόριθμο αποκοπής Snowball [109] για το SOM, για να δημιουργήσουμε τα διανύσματα για κάθε έγγραφο. Όπως αναφέρθηκε και παραπάνω στα πειράματα που διεξάγαμε για τα DNN χρησιμοποιήθηκε και το σχήματα στάθμισης TF-IDF αλλά παρουσίασε χειρότερη επίδοση σε σχέση με τον Keras Tokenizer. Η μεθοδολογία που εφαρμόζεται στην περίπτωσή μας είναι αρκετά γενική και μπορεί εύκολα να εφαρμοστεί σε διαφορετικά σύνολα δεδομένων, αποφεύγοντας έτσι την εκτεταμένη κατασκευή χαρακτηριστικών, η οποία μπορεί να αποφέρει καλύτερα αποτελέσματα για ένα συγκεκριμένο σύνολο δεδομένων, αλλά δεν μπορεί να γενικευθεί όταν το υποκείμενο σύνολο δεδομένων διαφοροποιηθεί.

4.3.4. Αυτό-Οργανούμενοι χάρτες και Νευρωνικά Δίκτυα Βαθιάς Μάθησης (DNN) για την ταξινόμηση πολλαπλών κλάσεων

Στο Κεφάλαιο 3 της παρούσας Διατριβής αναλύθηκε η λειτουργία του μη εποπτευόμενου Νευρωνικού Δικτύου SOM , που είναι ικανό να συγκεντρώσει παρόμοια δεδομένα μαζί ανεξάρτητα από τις ετικέτες που τους έχουν ανατεθεί [110]. Ακόμα μπορεί να αντιστοιχίσει και να αναπαραστήσει διανύσματα υψηλότερης διάστασης στον δισδιάστατο χώρο εγγενώς [40], [110], [61]. Το SOM έχει εφαρμοστεί εκτενώς σε διαφορετικούς τομείς εφαρμογών, συμπεριλαμβανομένης της ταξινόμησης εγγράφων. Το πρόβλημα της διαχείρισης μεγάλων συλλογών εγγράφων με χρήση SOM ακολουθούμενο από ένα σύνολο προτάσεων για την επίτευξη μείωσης του χώρου διανυσμάτων (dimensionality reduction) για μεγάλα SOM μελετήθηκε στο [40]. Σε αυτή την εργασία, οι συγγραφείς αποδεικνύουν ότι η χρήση πολύ μεγάλων SOM αυξάνει την πολυπλοκότητα και το χρόνο επεξεργασίας και συνεπώς δεν αποτελεί καλή προσέγγιση μοντελοποίησης. Αντίθετα, η χρήση πολύ μεγάλων διανυσμάτων χαρακτηριστικών και η χρήση της αναδρομικής εκδοχής (step-wise recursive version) του SOM, απαιτούσαν τη χρήση υπολογιστών υψηλής απόδοσης (High Performance Computing - HPC) μεγάλης κλίμακας για την εκτέλεση των δαπανηρών υπολογισμών [39]. Προκειμένου να αντιμετωπιστεί αποτελεσματικά η ταξινόμηση πολλαπλών κλάσεων, στο [111] μελετήθηκε μια αρχιτεκτονική Cascaded SOM όπου οι κλάσεις αντιμετωπίστηκαν με ιεραρχικό τρόπο. Οι συγγραφείς θεωρούν ότι η προτεινόμενη προσέγγιση μπορεί να εφαρμοστεί σε οποιοδήποτε πρόβλημα ταξινόμησης πολλαπλών κλάσεων με μικρό αριθμό επισημασμένων δειγμάτων.

Σε αντίθεση με το SOM, το οποίο είναι ένας μη εποπτευόμενος ταξινομητής, η Βαθιά Μάθηση είναι μια οικογένεια αποτελεσματικών Νευρωνικών Δικτύων [112] που μπορούν να εκτελέσουν τόσο μη εποπτευόμενη, όσο και εποπτευόμενη, και ημι-εποπτευόμενη μάθηση [113]. Μια μέθοδος ιεραρχικής Βαθιάς Μάθησης για την ταξινόμηση κειμένου έχει προταθεί στο [82] ως αντίδοτο στην υποβάθμιση της απόδοσης των εποπτευόμενων ταξινομητών κατά την αύξηση του αριθμού των

εγγράφων σε σενάρια ταξινόμησης πολλαπλών κλάσεων. Αυτή η μελέτη χρησιμοποιεί αρχιτεκτονικές Βαθιάς Μάθησης όχι μόνο για την ταξινόμηση ενός εγγράφου σε έναν εξειδικευμένο τομέα αλλά και για την οργάνωσή του στον κατάλληλο υπό-τομέα, δημιουργώντας έτσι μια ιεραρχική προσέγγιση ταξινόμησης. Η εργασία που παρουσιάστηκε στο [114] στοχεύει στην επίλυση του προβλήματος της διαδοχικής ταξινόμησης σύντομου κειμένου χρησιμοποιώντας τόσο επαναλαμβανόμενα (RNNs) [116] όσο και συνελικτικά νευρωνικά δίκτυα (CNNs), ενώ η αξιολόγηση πραγματοποιήθηκε σε τρία διαφορετικά σύνολα δεδομένων πολλαπλών κλάσεων που σχετίζονται με την πρόβλεψη διαλόγου (dialog act). Ομοίως, οι συγγραφείς στο [115] χρησιμοποίησαν CNN για την ταξινόμηση κειμένου πολλαπλών κλάσεων χρησιμοποιώντας χαρακτηριστικά σε επίπεδο χαρακτήρων, ενώ RNN χρησιμοποιήθηκαν και στο [82] έτσι ώστε να αναπαρασταθούν χρονικά εξαρτημένες δομές στο κείμενο.

Συνολικά, η λύση που προτείνεται στο Κεφάλαιο αυτό της παρούσας Διατριβής στοχεύει στην αξιοποίηση των πλεονεκτημάτων των προηγμένων προσεγγίσεων για την ταξινόμηση ενός συνόλου δεδομένων πολλαπλών κλάσεων αναφορικά με την επιλογή χαρακτηριστικών και τις διαδικασίες εκπαίδευσης, καθώς και την ανάθεση ετικετών στα δείγματα ελέγχου, με στόχο την ενίσχυση της απόδοσης του ταξινομητή ως προς την ακρίβεια, ελαχιστοποιώντας ταυτόχρονα τον χρόνο επεξεργασίας και τους υπολογιστικούς πόρους που απαιτούνται. Πιο συγκεκριμένα, η προσέγγισή μας ενσωματώνει μια συγκεκριμένη διαδικασία για την κατασκευή διανυσμάτων και κατ' επέκταση για τη διαχείριση των απαιτήσεων της ταξινόμησης πολλαπλών κλάσεων, αποφεύγοντας τη μετατροπή του προβλήματος σε ένα σύνολο προβλημάτων δυαδικής ταξινόμησης. Επιπλέον, η διαδικασία επιλογής χαρακτηριστικών έχει σχεδιαστεί ώστε να είναι σχετικά γενική, με δεδομένο ότι βασίζεται σε στατιστικά λέξεων, έτσι ώστε να μπορεί να εφαρμοστεί εύκολα σε ένα διαφορετικό σενάριο. Χρησιμοποιούμε την batch εκδοχή του αλγορίθμου SOM για την εκπαίδευση, που θεωρείται πιο σταθερός και πολύ πιο γρήγορος από την αναδρομική έκδοση και δύο διαφορετικές αρχιτεκτονικές DNN, συγκεκριμένα ένα LSTM και ένα CNN + LSTM που φαίνεται να είναι αποτελεσματικά σε εργασίες ανάλυσης φυσικής γλώσσας (NLP).

Πιο συγκεκριμένα η LSTM αρχιτεκτονική επιλέχθηκε καθώς είναι αποτελεσματική για την μάθηση ακολουθιών στα δεδομένα, δεν παρουσιάζει το πρόβλημα των «vanishing/exploding gradients» που παρουσιάζουν άλλες αρχιτεκτονικές όπως τα vanilla RNNs κ.α. Τέλος, φαίνεται από τη βιβλιογραφία πως παρουσιάζουν μεγαλύτερη ακρίβεια σε σχέση με άλλες αντίστοιχες αρχιτεκτονικές όταν εφαρμόζονται για την εκμάθηση πιο μεγάλων ακολουθιών στα δεδομένα. Ακόμα, τα CNN έχουν χρησιμοποιηθεί επιτυχώς για σε εργασίες ανάλυσης φυσικής γλώσσας όπως η ταξινόμηση τύπων ερωτήσεων, κατηγοριοποίηση κειμένων μικρού μήκους κ.α.

4.4. Περιγραφή του συνόλου δεδομένων

Το σύνολο δεδομένων που χρησιμοποιείται σε αυτό το Κεφάλαιο δημιουργήθηκε από μια συλλογή από 56403 πολύ-επιστημονικούς τίτλους ηλεκτρονικών βιβλίων από τον Springer, που διατίθενται μέσα από τη συνδρομή του Συνδέσμου Ελληνικών Ακαδημαϊκών Βιβλιοθηκών[17]. Αυτή η συλλογή χρησιμεύει ως αναφορά για τους καθηγητές όταν προτείνουν πηγές για περαιτέρω ανάγνωση, μαζί με το κύριο μαθησιακό υλικό που παρέχουν στους φοιτητές τους. Για τη δημιουργία αυτού του συνόλου δεδομένων, δημιουργήθηκε ένα πρόγραμμα ανάλυσης κειμένου για την εξαγωγή σχετικών πληροφοριών, όπως ο τίτλος, ο υπότιτλος και ο πίνακας περιεχομένων (ToC), από κάθε

[17] https://www.heal-link.gr/en/home-2/

βιβλίο. Οι εξαχθείσες πληροφορίες αποθηκεύτηκαν σε μια βάση δεδομένων για περαιτέρω επεξεργασία. Συγκεκριμένα, για κάθε τίτλο βιβλίου στη βάση δεδομένων περιλαμβάνεται: ο τίτλος, ο υπότιτλος, ο/οι συγγραφείς, το έτος έκδοσης, ο εκδότης, ο πίνακας περιεχομένων, η περίληψη και το θέμα. Ως επόμενο βήμα, μια ομάδα βιβλιοθηκονόμων που εργάζονται στην Ψηφιακή Βιβλιοθήκη του ΕΜΠ επεξεργάστηκε και καταχώρισε τις πληροφορίες του θέματος για κάθε βιβλίο της συλλογής. Το αποτέλεσμα αυτού του βήματος ήταν μια λίστα θεμάτων για κάθε τίτλο βιβλίου. Σε αυτήν τη μελέτη, χρησιμοποιούμε το πρωτεύον πεδίο θέματος ως ετικέτα κάθε βιβλίου. Για να αποκτήσουμε την κύρια ετικέτα θέματος, έχουμε εφαρμόσει μια διαδικασία χαρτογράφησης, η οποία θα αναλυθεί παρακάτω (Εικόνα 13). Ένα παράδειγμα των δεδομένων εισόδου μετά την επιλογή του πρωτεύοντος θέματος μέσω της διαδικασίας χαρτογράφησης παρουσιάζεται στην Εικόνα 14.

Θα πρέπει να σημειωθεί ότι δεν ήταν διαθέσιμα όλα τα παραπάνω πεδία για το σύνολο των ηλεκτρονικών βιβλίων της συλλογής, γεγονός που καθιστούσε την επιλογή των πληροφοριών που χρησιμεύουν ως είσοδος για το σύστημα ταξινόμησής μας πιο σύνθετη. Έτσι, επιλέξαμε να μοντελοποιήσουμε το πρόβλημα της εύρεσης παρόμοιων βιβλίων ως πρόβλημα ταξινόμησης πολλαπλών κλάσεων, χρησιμοποιώντας πληροφορίες από τους Πίνακες Περιεχομένων. Η επιλογή αυτή δεν ήταν σε καμία περίπτωση τυχαία, καθώς ένας σημαντικός αριθμός περιλήψεων έλειπε από τη συλλογή. Επιπλέον, πιστεύουμε ότι η χρήση πληροφοριών από τους πίνακες περιεχομένων των βιβλίων μπορεί να βοηθήσει στην αναπαράσταση των διαφορετικών θεμάτων σε κάθε βιβλίο, καθιστώντας έτσι ευκολότερη την εύρεση παρόμοιων βιβλίων.

Καθώς αυτή η συλλογή είναι ένα σύνολο δεδομένων που χρησιμοποιείται στην πράξη και δεν έχει κατασκευαστεί για τις ανάγκες του συστήματος ταξινόμησης, δεν υπάρχουν διαχωρισμένα δείγματα εκπαίδευσης και ελέγχου. Έπρεπε, συνεπώς, να χωριστεί τυχαία το σύνολο δεδομένων σε υποσύνολα εκπαίδευσης και ελέγχου κατά τη φάση της προ-επεξεργασίας όπως θα αναλυθεί παρακάτω, χρησιμοποιώντας το 80% των δειγμάτων για εκπαίδευση και το 20% για έλεγχο. Αρχικά υπήρχαν 252 διαφορετικές κατηγορίες στο σύνολο δεδομένων, αλλά καθώς ορισμένες από αυτές μπορούσαν να ομαδοποιηθούν, τις αντιστοιχίσαμε σε πιο γενικές και έτσι καταλήξαμε με 26 γενικές κατηγορίες. Η αντιστοίχιση αυτή πραγματοποιήθηκε με τη βοήθεια των βιβλιοθηκονόμων της Ψηφιακής Βιβλιοθήκης του ΕΜΠ. Ένα απόσπασμα αυτής της διαδικασίας αντιστοίχισης απεικονίζεται στην Εικόνα 13. Πραγματοποιήσαμε πειράματα χρησιμοποιώντας αφενός τις 26 προκύπτουσες κατηγορίες, αφ΄ ετέρου τις πέντε (5) από αυτές με τα περισσότερα δείγματα (συγκεκριμένα τις: Medicine, Mathematics, Engineering, Computer Science και Physics). Στον Πίνακα 21, αναφέρεται η κατανομή των δειγμάτων μεταξύ των 26 διαφορετικών κατηγοριών. Τα συνολικά έγγραφα για τις 26 κατηγορίες είναι 56403, ενώ τα συνολικά έγγραφα για τις 5 πολυπληθείς κατηγορίες είναι 35271 που αντιπροσωπεύουν το 63% του συνολικού αριθμού εγγράφων στο σύνολο δεδομένων. Η κατανομή των δειγμάτων για τις 5 κατηγορίες αναφέρεται στον Πίνακα 22.

Πίνακας 21. Κατανομή Βιβλίων για τις 26 κλάσεις του συνόλου δεδομένων SPRINGER.

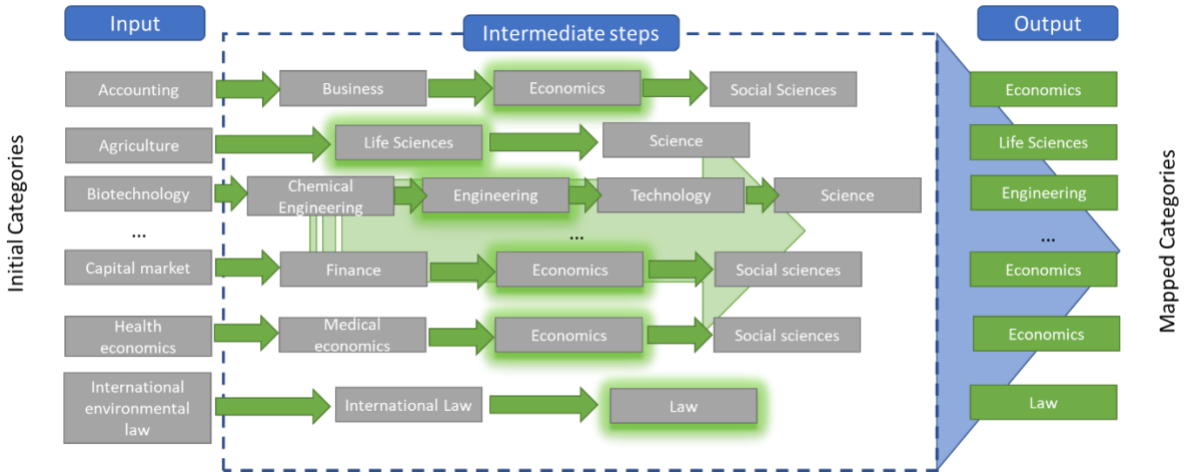
Κλάση	Αριθμός δειγμάτων	Κλάση	Αριθμός δειγμάτων
Anthropology	997	Linguistics	172
Art	63	Literature	5
Computer Science	11220	Management	221
Culture	24	Mathematics	4852
Economics	3767	Medicine	8471
Education	1842	Music	3
Engineering	8162	Organization	11

Environment	1394	Physics	8788
Food	4	Popular works	322
History	105	Religion	38
Humanities	815	Social Sciences	1361
Law	599	Science	34
Life Sciences	3125	Transportation	10
Σύνολο	**56403**		

Επιπλέον, από τον Πίνακα 21 μπορεί να εξαχθεί ότι τα περισσότερα έγγραφα ανήκουν στις κλάσεις Computer Science, Physics, Medicine και Engineering και ότι η κατανομή των δειγμάτων μεταξύ των κλάσεων δεν είναι ισομερής. Αυτό το γεγονός θέτει προκλήσεις σχετικά με τα ποιοτικά μέτρα που θα χρησιμοποιηθούν για τη μέτρηση της απόδοσης του ταξινομητή, καθώς η ακρίβεια στις μικρότερες κατηγορίες μπορεί να θεωρηθεί ψευδώς ότι είναι ίδια με αυτή των μεγαλυτέρων. Επιπλέον, υπάρχει σημαντική επικάλυψη θεμάτων μεταξύ ορισμένων κατηγοριών, π.χ. μεταξύ των κλάσεων Engineering και Computer Science. Ως εκ τούτου είναι απαραίτητο να χρησιμοποιηθούν μέτρα Μίκρο- και Μάκρο-μέσου όρου, για τη μέτρηση της απόδοσης του ταξινομητή όπως θα αναλυθεί σε επόμενη ενότητα.

Πίνακας 22. Κατανομή βιβλίων για τις 5 πολυπληθείς κλάσεις του συνόλου δεδομένων SPRINGER.

Κλάση	Συνολικός Αριθμός δειγμάτων	Αριθμός δειγμάτων εκπαίδευσης	Αριθμός δειγμάτων ελέγχου
Computer Science	11183	8946	2237
Engineering	8047	6437	1610
Mathematics	4206	3364	842
Medicine	8315	6652	1663
Physics	3520	2816	704
Σύνολο	**35271**	**28215**	**7056**



Εικόνα 13. Ένα απόσπασμα της διαδικασίας αντιστοίχισης των 252 αρχικών θεμάτων των ηλεκτρονικών βιβλίων στις 26 γενικές κλάσεις.

	idbook	title	toc	Label
0	99500000	Bioequivalence Requirements in Various Global ...	Brazil / Canada / China / The European Union /...	Medicine
1	99500001	Translating Molecules into Medicines	Pharmaceutical Industry Performance / Scope- p...	Medicine
2	99500003	School Inspectors	1. School Inspectors as Policy Implementers- I...	Education
3	99500004	Corporate Risk Management for International Bu...	Chapter 1- Introduction / Chapter 2- Managemen...	Economics
4	99500005	Equity Valuation and Negative Earnings	Chapter 1- Introduction / Chapter 2-The Models...	Economics

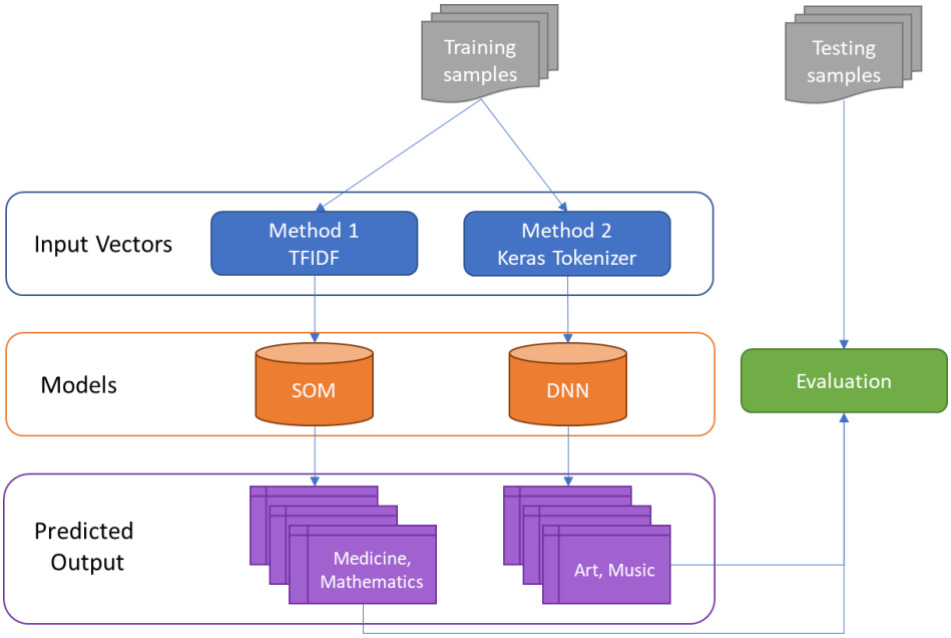
Εικόνα 14. Παράδειγμα των δεδομένων εισόδου μετά την επιλογή της πρωτεύουσας ετικέτας.

4.5. Μεθοδολογία

Σε αυτήν την ενότητα προτείνουμε, σχεδιάζουμε και εφαρμόζουμε μια νέα προσέγγιση για την ταξινόμηση ενός συνόλου δεδομένων πολλαπλών κλάσεων όπως αναλύθηκε και στην προηγούμενη ενότητα, αξιοποιώντας τις πληροφορίες που εξάγονται από τους πίνακες περιεχομένων των βιβλίων. Όπως αναφέρθηκε, έπρεπε να διαχωρίσουμε τυχαία τα δείγματα εκπαίδευσης από τα δείγματα ελέγχου, ενώ οι ετικέτες που χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων προέρχονται από το πεδίο του θέματος. Η Εικόνα 15 απεικονίζει σε υψηλό επίπεδο την προτεινόμενη μεθοδολογία. Η λεπτομερής ανάλυση της λύσης περιγράφεται στις ακόλουθες ενότητες. Σε γενικές γραμμές, η προτεινόμενη προσέγγιση στοχεύει στην εξαγωγή των πιο αντιπροσωπευτικών χαρακτηριστικών από τους πίνακες περιεχομένων των ηλεκτρονικών βιβλίων χρησιμοποιώντας δύο διαφορετικές μεθόδους για την αναπαράστασή τους, συγκεκριμένα το σχήμα στάθμισης TF-IDF στην περίπτωση του SOM και τον Keras Tokenizer[18] στην περίπτωση του LSTM και CNN+LSTM.

Σε αυτό το σημείο πρέπει να τονιστεί ότι τόσο τα διανύσματα του υποσυνόλου εκπαίδευσης όσο και αυτά του υποσυνόλου ελέγχου κατασκευάζονται με την ίδια διαδικασία, η οποία εξηγείται παρακάτω λεπτομερώς. Ως επόμενο βήμα, τα παραγόμενα διανύσματα που προκύπτουν από τους πίνακες περιεχομένων τροφοδοτούνται σε δύο διαφορετικά μοντέλα, δηλαδή στο SOM και στο DNN (LSTM και CNN+LSTM), τα οποία είναι ικανά να προβλέψουν την καταλληλότερη ετικέτα για κάθε δείγμα ελέγχου μόλις ολοκληρωθεί η φάση εκπαίδευσης. Τέλος, η προβλεπόμενη έξοδος των δύο μοντέλων αξιολογείται έναντι των πραγματικών ετικετών από τα δείγματα ελέγχου χρησιμοποιώντας τα μέτρα Μίκρο και Μάκρο-μέσου όρου όπως θα συζητηθεί στην ενότητα Πειραματική Αξιολόγηση.

[18] https://www.tensorflow.org/api_docs/Python/tf/keras/preprocessing/text/Tokenizer



Εικόνα 15. Απεικόνιση της προτεινόμενης μεθοδολογίας.

4.5.1. Προεπεξεργασία

Κατά τη φάση προεπεξεργασίας τα δείγματα εκπαίδευσης μετατρέπονται σε αριθμητικά διανύσματα που θα δοθούν ως είσοδος στον ταξινομητή. Αυτό είναι ένα κοινό βήμα για την πλειονότητα των προσεγγίσεων ταξινόμησης που αναφέρονται στη βιβλιογραφία, και συνήθως περιλαμβάνει ορισμένα από τα ακόλουθα βήματα: α. τμηματοποίηση λέξεων (word tokenization), β. αφαίρεση λέξεων διακοπής (stop-words), γ. λημματοποίηση (lemmatization) και δ. αποκοπή καταλήξεων (stemming). Στην προσέγγισή μας, αυτή η διαδικασία εξαρτάται από τη μέθοδο αναπαράστασης του πίνακα περιεχομένων του βιβλίου που θα επιλεγεί όπως συνοψίζεται στην Εικόνα 16.

Πιο συγκεκριμένα, ακολουθήσαμε ένα διαφορετικό σύνολο βημάτων προεπεξεργασίας όταν χρησιμοποιούσαμε το σχήμα στάθμισης TF-IDF σε σύγκριση με τον Keras Tokenizer. Στην πρώτη περίπτωση, καταργήσαμε τυχόν html-χαρακτήρες και σημεία στίξης, και στη συνέχεια αφαιρέσαμε όλους τους αριθμητικούς χαρακτήρες. Έπρεπε επίσης να προσέξουμε ιδιαίτερα τους λατινικούς αριθμούς που συνήθως βρίσκονταν στους τίτλους των κεφαλαίων. Στη συνέχεια, το κείμενο τμηματοποιείται, για να ανακτηθεί μια λίστα λέξεων από κάθε πίνακα περιεχομένων. Για κάθε λίστα λέξεων που δημιουργήθηκε στο προηγούμενο βήμα, αφαιρέσαμε άρθρα, συνδέσμους και άλλες λέξεις διακοπής (λέξεις όπως “the”, “a”, “an”, “in”, “be”,“can”, κλπ.) χρησιμοποιώντας την αντίστοιχη λειτουργία από το πακέτο NLTK[19] της Python.

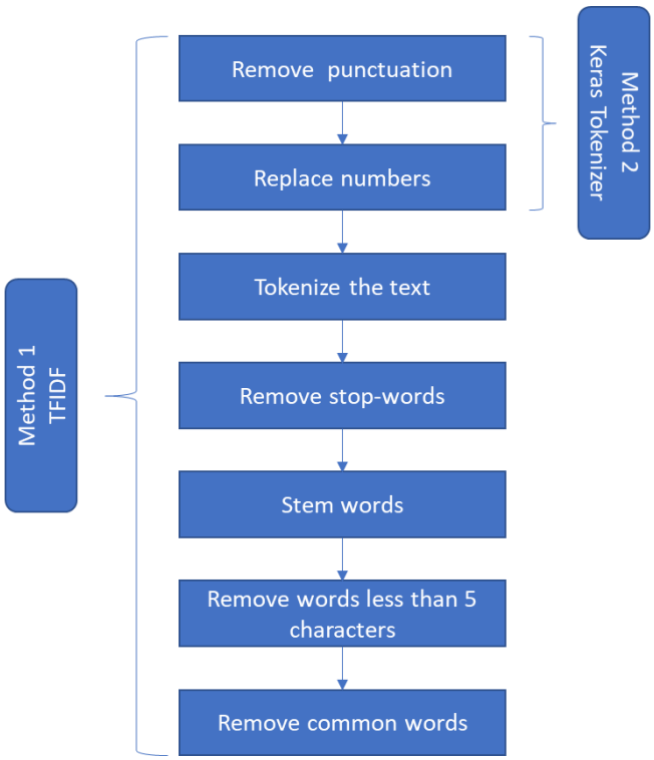
Προκειμένου να μετατρέψουμε κάθε λέξη στη βασική της μορφή, αφαιρώντας τις καταλήξεις, πειραματιστήκαμε με τρείς διαφορετικούς αλγορίθμους αποκοπής, συγκεκριμένα τους Porter [20], Snowball [21] και Lancaster [22]. Το επόμενο βήμα αφορά την αφαίρεση «πολύ μικρών» λέξεων, λέξεων με λιγότερους από πέντε χαρακτήρες στην περίπτωσή μας, μια παράμετρο που μπορεί να τροποποιηθεί κάθε φορά στην αρχική διαμόρφωση. Θα πρέπει να επισημανθεί ότι αυτό το βήμα

[19] http://www.nltk.org/
[20] https://www.nltk.org/_modules/nltk/stem/porter.html
[21] https://www.nltk.org/_modules/nltk/stem/snowball.html
[22] https://www.nltk.org/_modules/nltk/stem/lancaster.html

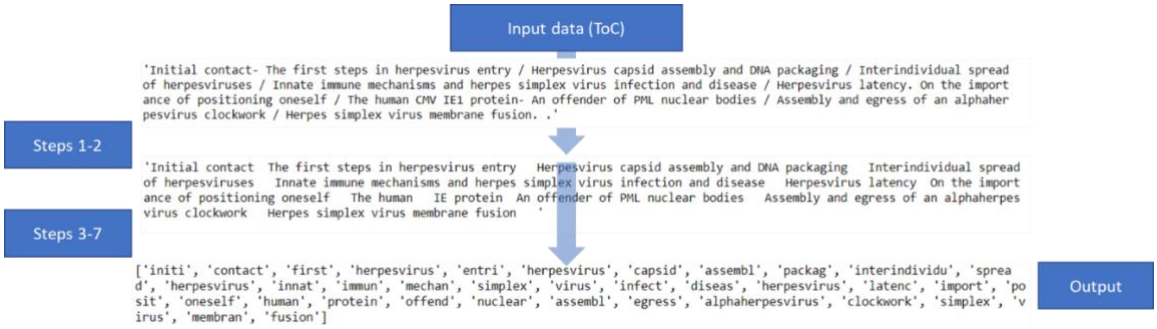
πρέπει να εκτελεστεί μετά την διαδικασία αποκοπής καταλήξεων. Τέλος, η φάση της προ επεξεργασίας ολοκληρώνεται με την αφαίρεση κάποιων συνηθισμένων λέξεων (δηλαδή, λέξεων που συνήθως υπάρχουν στον πίνακα περιεχομένων), όπως 'πρόλογος', 'εισαγωγή', 'κεφάλαιο', 'αναφορές', 'παράρτημα' κ.λπ.

Αντίθετα, μια αρκετά απλή διαδικασία προεπεξεργασίας έχει ακολουθηθεί κατά τη χρήση του Keras Tokenizer για την αναπαράσταση των πινάκων περιεχομένων των ηλεκτρονικών βιβλίων. Στην περίπτωση αυτή, εφαρμόστηκαν μόνο τα πρώτα δύο βήματα, όπως φαίνεται και στην Εικόνα 16. Αυτή η διαδικασία επιλέχθηκε καθώς ο στόχος ήταν να διαφοροποιηθεί από το bag-of-words (BoW) μοντέλο και να χρησιμοποιήσουμε και ακολουθίες λέξεων, καθώς σε αυτή την περίπτωση ακόμη και μικρές λέξεις (συζεύξεις, αντωνυμίες, κλπ.) μπορεί να επηρεάσουν την ταξινόμηση ενός βιβλίου. Το αποτέλεσμα αυτής της διαδικασίας είναι ένα λεξιλόγιο λέξεων-κλειδιών για κάθε πίνακα περιεχομένων. Ένα παράδειγμα της διαδικασίας αυτής δίνεται στην Εικόνα 17.

Στην επόμενη υποενότητα, παρουσιάζονται οι μέθοδοι που επιλέχθηκαν για την αναπαράσταση διανυσμάτων για κάθε πίνακα περιεχομένων, καθώς και τα βήματα που ακολουθήθηκαν για την κατασκευή των διανυσμάτων που αντιπροσωπεύουν κάθε βιβλίο του υπό εξέταση συνόλου δεδομένων.



Εικόνα 16. Τα βήματα της διαδικασίας προεπεξεργασίας που χρησιμοποιήθηκαν για κάθε μέθοδο αναπαράστασης κειμένου.



Εικόνα 17. Παράδειγμα της διαδικασίας προεπεξεργασίας για τη μέθοδο αναπαράστασης TF-IDF. Το αποτέλεσμα είναι ένα λεξιλόγιο λέξεων-κλειδιών για κάθε πίνακα περιεχομένων.

4.5.2. Αναπαράσταση διανυσμάτων των πινάκων περιεχομένων

Αυτό το βήμα στοχεύει στην επιλογή των πιο αντιπροσωπευτικών χαρακτηριστικών για την αναπαράσταση ενός εγγράφου ως διάνυσμα που θα τροφοδοτείται στο SOM και στις δύο αρχιτεκτονικές DNN που έχουν επιλεγεί για την αυτόματη ταξινόμηση ενός συνόδου δεδομένων πολλαπλών κλάσεων. Η διάσταση του διανύσματος εξαρτάται από τον αριθμό των επιλεγμένων χαρακτηριστικών που αποτελούν το χώρο των χαρακτηριστικών. Είναι προφανές ότι όσο υψηλότερη είναι η διάσταση των διανυσμάτων, τόσο πιο υπολογιστικά απαιτητική γίνεται η εργασία ταξινόμησης.

Έστω ότι $S = \{ToC_1, ToC_2, \dots, ToC_K\}$, είναι η συλλογή των πινάκων περιεχομένων του συνόλου δεδομένων ηλεκτρονικών βιβλίων του Springer. Ο πίνακας περιεχομένων (ToC) κάθε βιβλίου που ανήκει στη συλλογή S αναπαρίσταται από ένα διάνυσμα m-διαστάσεων της μορφής: $ToC_i = \{w_{i1}, w_{i2}, \dots, w_{im}\}$, όπου το w_{im} είναι το βάρος του πίνακα περιεχομένων ToC_i για το χαρακτηριστικό m. Τα βάρη υπολογίζονται με χρήση της σχήματος στάθμισης TF-IDF, ένα σύνθετο σχήμα στάθμισης που αποτελείται από το γινόμενο των μέτρων TF και IDF. Το μέτρο TF ενός όρου t σε ένα έγγραφο d, είναι το άθροισμα των φορών που ο όρος t εμφανίζεται στο d ενώ το μέτρο IDF δείχνει πόσο σπάνια είναι μια λέξη σε όλα τα έγγραφα της συλλογής. Η πλήρης μαθηματική διατύπωση των μέτρων IDF και TF-IDF δίνεται στο [83] και έτσι παραλείπεται εδώ για συντομία. Τα διανύσματα κατασκευάστηκαν χρησιμοποιώντας το TF-IDFVectorizer που διατίθεται μέσω του πακέτου scikit-learn[23] της Python. Πρώτον, υπολογίζονται οι τιμές TF-IDF για όλες τις λέξεις στο σύνολο δεδομένων και στη συνέχεια, επιλέγονται οι x λέξεις που έχουν τις υψηλότερες βαθμολογίες TF-IDF για την κατασκευή των διανυσμάτων. Όπως θα αναλυθεί λεπτομερώς σε επόμενη ενότητα, το x κυμαίνεται από 2000-3000 λέξεις. Θα πρέπει να σημειωθεί ότι όταν χρησιμοποιείται ο TF-IDFVectorizer υπάρχει η επιλογή να χρησιμοποιηθούν μόνο μεμονωμένες λέξεις (1-grams) ή n-grams λέξεων ως χαρακτηριστικά. Σε αυτή την περίπτωση, χρησιμοποιούμε μόνο 1-grams λόγω της φύσης του συνόλου δεδομένων. Μετά την κατασκευή των διανυσμάτων, κάθε διάνυσμα κανονικοποιείται χρησιμοποιώντας το "l_2 norm" ή την Ευκλείδεια απόσταση. Στο τελευταίο βήμα, οι ετικέτες που ανήκουν στο αντίστοιχο δείγμα δεδομένων εκχωρήθηκαν σε κάθε διάνυσμα.

Η τελική μορφή του διανύσματος παρουσιάζεται στον Πίνακα 23. Κάθε δείγμα αποτελείται από ένα σύνολο χαρακτηριστικών $F = \{f_1, f_2, \dots, f_n\}$ και την ετικέτα που ανατέθηκε στο προηγούμενο βήμα από το σύνολο των ετικετών $L = \{l_1, l_2, \dots, l_k\}$. Όπως αναφέρθηκε και προηγουμένως, τα προβλήματα ταξινόμησης πολλαπλών κλάσεων συνήθως μετατρέπονται σε ένα

[23] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TF-IDFVectorizer.html

σύνολο δυαδικών προβλημάτων ταξινόμησης. Σε αυτή την περίπτωση αντί να κατασκευάσουμε πολλαπλούς δυαδικούς ταξινομητές (έναν για κάθε ετικέτα που υπάρχει στο σύνολο δεδομένων), προτείνουμε τη χρήση δύο διαφορετικών νευρωνικών δικτύων, ενός δικτύου μη εποπτευόμενης ΜΜ και δύο διαφορετικών αρχιτεκτονικών Βαθιάς Μάθησης όπως αναφέρθηκε και παραπάνω, που παίρνουν ως είσοδο τα διανύσματα που κατασκευάστηκαν στο προηγούμενο βήμα. Η αρχιτεκτονική των νευρωνικών δικτύων αυτών θα παρουσιαστεί λεπτομερώς σε επόμενη υποενότητα του παρόντος.

Πίνακας 23. Τα διανύσματα που κατασκευάστηκαν και οι ετικέτες που ανατέθηκαν σε κάθε ένα.

Δείγμα 1	$f_1, f_2, \dots, f_n$	l_1
Δείγμα 2	$f_1, f_2, \dots, f_n$	l_3
…	…	…
Δείγμα N	$f_1, f_2, \dots, f_n$	l_{k-1}

Η δεύτερη μέθοδος που χρησιμοποιείται σε αυτό το Κεφάλαιο για την κατασκευή των διανυσμάτων είναι ο Keras Tokenizer, συγκεκριμένα το Tokenizer API[24]. Κατασκευάσαμε αρχικά τον Tokenizer τον οποίο στη συνέχεια εφαρμόσαμε στην λίστα των λέξεων που έχει προκύψει από τη φάση της προ επεξεργασίας για κάθε έναν από του πίνακες περιεχομένων στη συλλογή. Για την κατασκευή του Tokenizer χρησιμοποιήσαμε την ακόλουθη διαμόρφωση: α. ορίσαμε τον μέγιστο αριθμό λέξεων που πρέπει να διατηρηθούν (20000 λέξεις σε αυτήν την περίπτωση), με βάση τη συχνότητα των λέξεων, β. να γίνουν πεζά όλες οι λέξεις και γ. χρησιμοποιήθηκε το κενό διάστημα ως διαχωριστικό λέξεων. Όταν ο Tokenizer εφαρμόστηκε στη λίστα των πινάκων περιεχομένων, δημιουργήθηκε ένα ευρετήριο με βάση τη συχνότητα των λέξεων.

Σε αυτήν την περίπτωση, ένας μικρότερος ακέραιος σημαίνει συχνότερη λέξη, ενώ οι πρώτες λέξεις είναι συνήθως λέξεις διακοπής (stop-words) καθώς τείνουν να εμφανίζονται συχνά. Στη συνέχεια, κάθε πίνακας περιεχομένων μετατρέπεται σε μια ακολουθία ακεραίων χρησιμοποιώντας το ευρετήριο λέξεων που κατασκευάστηκε στο προηγούμενο βήμα. Τέλος, κατασκευάζουμε διανύσματα σταθερού μήκους συμπληρώνοντας την ακολουθία ακέραιων με μηδενικά χρησιμοποιώντας την παράμετρο max_length (κυμαίνεται από 800-3000 στην περίπτωσή μας). Θα πρέπει να επισημανθεί ότι χρησιμοποιείται ακριβώς το ίδιο ευρετήριο λέξεων για τη μετατροπή του κειμένου σε ακολουθίες και για τα δείγματα ελέγχου που στη συνέχεια θα τροφοδοτήσουν τις δύο αρχιτεκτονικές νευρωνικών δικτύων Βαθιάς Μάθησης. Το τελικό διάνυσμα σε αυτή την περίπτωση έχει την ίδια μορφή με το διάνυσμα που χρησιμοποιήθηκε στην περίπτωση TF-IDF και παρουσιάζεται στον Πίνακα 23.

4.5.3. Αυτο-Οργανούμενοι Χάρτες - SOM

Το θεωρητικό υπόβαθρο του νευρωνικού δικτύου SOM παρουσιάζεται στο Κεφάλαιο 3. Για την υλοποίηση του νευρωνικού δικτύου SOM στην παρούσα Διατριβή χρησιμοποιήθηκε η βιβλιοθήκη SOMPY σε Python[25], η οποία βασίζεται στη βιβλιοθήκη SOMToolbox του Matlab, μαζί με την ακόλουθη διαμόρφωση. Επιλέξαμε τη batch έκδοση του SOM με τη γραμμική αρχικοποίηση (linear initialization) και την κανονικοποίηση διακύμανσης (variance normalization) για τα διανύσματα εισόδου, καθώς συγκλίνει γρηγορότερα και δεν υπάρχει ανάγκη για την παράμετρο ρυθμού εκμάθησης (learning rate parameter). Η γραμμική αρχικοποίηση είναι προτιμότερη από την τυχαία (random initialization), καθώς τα αρχικά διανύσματα που ανατίθενται στις BMU βρίσκονται στον ίδιο

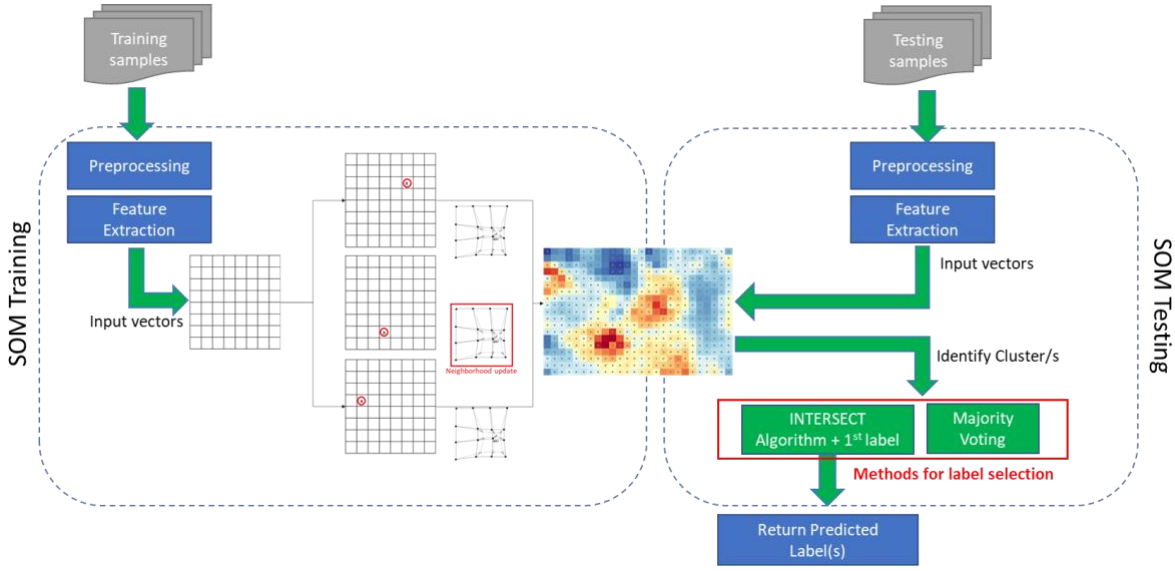
[24]

[25] https://github.com/sevamoo/SOMPY

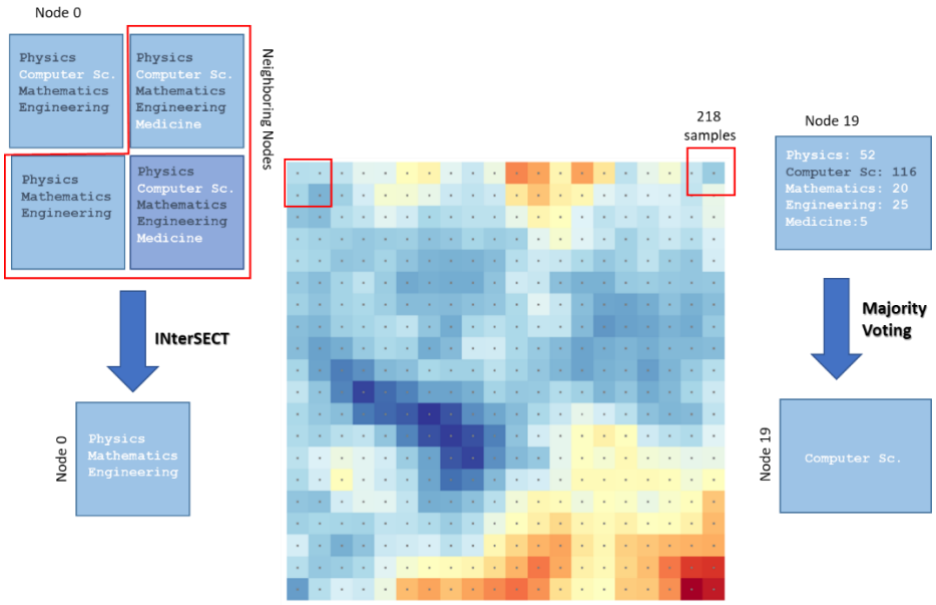
χώρο εισόδου που καλύπτεται από δύο ιδιοδιανύσματα που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές των δεδομένων εισόδου και έτσι το SOM εκτείνεται στον ίδιο προσανατολισμό με τα δεδομένα που παίζουν τον πιο σημαντικό ρόλο.

Η διαδικασία εκπαίδευσης και ελέγχου του SOM δίνεται στην Εικόνα 18. Πρώτον, κάθε κόμβος στο χάρτη σχετίζεται με ένα αρχικό διάνυσμα. Στη συνέχεια, για κάθε διάνυσμα εκπαίδευσης, υπολογίζεται η BMU της οποίας το βάρος είναι πιο κοντά στο διάνυσμα με βάση την ευκλείδεια απόσταση. Τέλος, οι τιμές των αρχικών διανυσμάτων στη γειτονιά της BMU ενημερώνονται προκειμένου να πλησιάσουν τη BMU (επισημαίνεται με κόκκινο - Εικόνα 18). Αυτή η διαδικασία επαναλαμβάνεται για όλες τις εποχές εκπαίδευσης. Με αυτόν τον τρόπο, το SOM σχηματίζει ένα ελαστικό δίκτυο που διπλώνει στο σχήμα που σχηματίζεται από τα δεδομένα εισόδου κατά τη διάρκεια της διαδικασίας εκπαίδευσης. Αφού ολοκληρωθεί η φάση της εκπαίδευσης, κάθε BMU συνδέεται έμμεσα με ένα σύνολο ετικετών που αντιστοιχούν στα δείγματα εκπαίδευσης που έχουν εκχωρηθεί σε κάθε κόμβο του χάρτη. Κάθε BMU στο SOM μπορεί να έχει μια ή περισσότερες ετικέτες.

Το επόμενο βήμα είναι να προβληθούν τα διανύσματα ελέγχου στο χάρτη και να υπολογιστεί η BMU για καθένα από αυτά. Μετά την προβολή, κάθε διάνυσμα από το υποσύνολο ελέγχου, συσχετίζεται με ένα σύνολο ετικετών με βάση τη BMU στην οποία εκχωρήθηκε. Έχουν προταθεί διαφορετικές προσεγγίσεις σχετικά με τον τρόπο επιλογής των ετικετών για κάθε διάνυσμα ελέγχου. Στην ταξινόμηση πολλαπλής ετικέτας μπορεί κανείς είτε να εκχωρήσει όλες τις ετικέτες που υπάρχουν στη BMU, είτε να εφαρμόσει τον αλγόριθμο “Majority Voting” για να επιλέξει την πιο συχνή. Ωστόσο, η πρώτη επιλογή δεν μπορεί να εφαρμοστεί στην περίπτωσή μας, δεδομένου ότι είναι το πρόβλημα που καλούμαστε να αντιμετωπίσουμε είναι ταξινόμηση πολλαπλών κλάσεων και όχι πολλαπλής ετικέτας. Στην ενότητα 3.5.4 προτείναμε έναν έξυπνο αλγόριθμο κατάλληλο ειδικά για προβλήματα ταξινόμησης πολλαπλής ετικέτας, ο οποίος βασίζεται σε πληροφορίες από τους γειτονικούς κόμβους για να αναθέσει την κατάλληλη ετικέτα στον υπό εξέταση κόμβο. Σε αυτή την περίπτωση δοκιμάσαμε να εφαρμόσουμε τον αλγόριθμο INterSECT σε ένα πρόβλημα ταξινόμησης πολλαπλών κλάσεων. Επίσης, χρησιμοποιήσαμε και τον αλγόριθμο Majority Voting με απώτερο στόχο να συγκρίνουμε τα αποτελέσματα. Ένα παράδειγμα της διαδικασίας επιλογής ετικετών με βάση τους αλγόριθμους INterSECT και Majority Voting παρουσιάζεται στην Εικόνα 18. Η απόδοση των προαναφερθεισών μεθόδων για την επιλογή ετικέτας παρουσιάζεται στα πειράματα που διεξήχθησαν, που περιγράφονται λεπτομερώς στην επόμενη ενότητα.



Εικόνα 18. Διάγραμμα ροής της προτεινόμενης μεθόδου με χρήση του SOM. Το αριστερό τμήμα αναπαριστά τη φάση της εκπαίδευσης ενώ το δεξί τη φάση ελέγχου. Το αποτέλεσμα της διαδικασίας εκπαίδευσης είναι ένας SOM ο οποίος δίνεται ως είσοδος μαζί με το υποσύνολο ελέγχου για να βρεθεί η κατάλληλη ετικέτα για κάθε βιβλίο στο υποσύνολο ελέγχου.



Εικόνα 19. Σύγκριση των αλγορίθμων INTERSECT και Majority Voting για την επιλογή ετικετών.

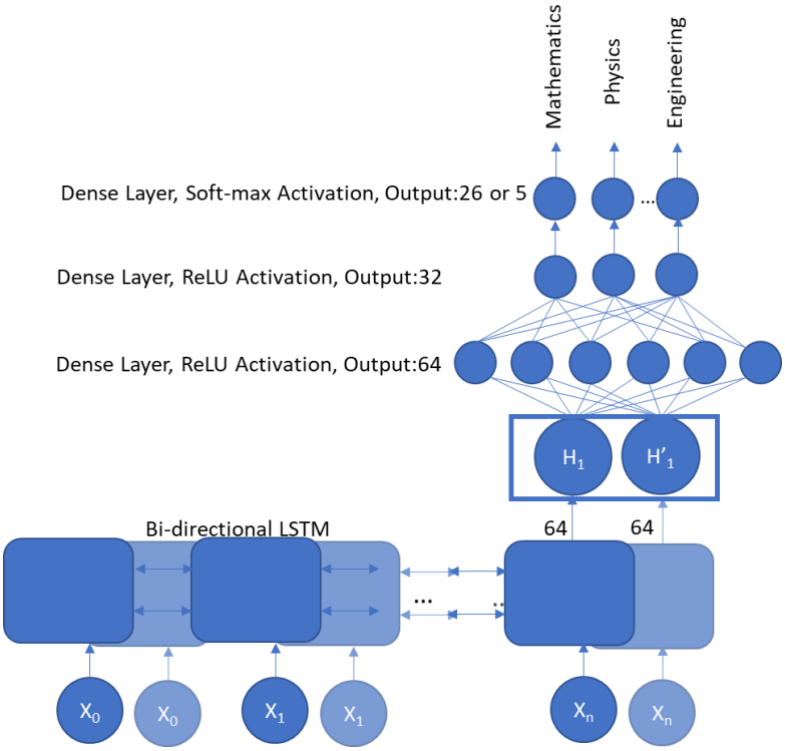
4.5.4. Δίκτυα Βαθιάς Μάθησης

Τα δίκτυα Βαθιάς Μάθησης (DNN) μπορούν να χρησιμοποιηθούν για τη δημιουργία αποδοτικών υπολογιστικών μοντέλων τόσο σε εποπτευόμενα, όσο και σε ημι-εποπτευόμενα και μη εποπτευόμενα σενάρια. Στη βασική του μορφή, το DNN είναι ένα πλήρως συνδεδεμένο σύνολο μονάδων επεξεργασίας (κόμβων) οργανωμένων σε επίπεδα. Το πρώτο επίπεδο είναι το επίπεδο εισόδου, το τελευταίο είναι το επίπεδο εξόδου και όλα τα ενδιάμεσα επίπεδα ονομάζονται κρυμμένα (hidden layers). Τα τροφοδοτούμενα νευρωνικά δίκτυα (feedforward) αποτελούν **προσεγγιστές γενικής λειτουργίας** (universal function approximators), και έτσι είναι δυνατόν να εντοπίσουν μια μη γραμμική αντιστοίχιση από μια είσοδο μεγάλων διαστάσεων σε μια έξοδο μικρότερης διάστασης. Σε αυτό το Κεφάλαιο της παρούσας Διατριβής, πειραματιστήκαμε με διαφορετικές αρχιτεκτονικές

βαθιάς μάθησης και συνδυασμούς υπερπαραμέτρων, προκειμένου να είμαστε σε θέση να ταξινομήσουμε με επιτυχία ένα σύνολο δεδομένων πολλαπλών κλάσεων αξιοποιώντας πληροφορίες από τους πίνακες περιεχομένων αυτών. Σε γενικές γραμμές, έχουμε χρησιμοποιήσει ένα στρώμα ενσωμάτωσης για να μειώσουμε την αραιότητα των δεδομένων εισόδου, ενώ λαμβάνουμε ενσωματώσεις λέξεων (word embeddings) που παρέχουν μια πυκνή αναπαράσταση των λέξεων και της σημασίας τους. Αυτές οι ενσωματώσεις λέξεων μπορούν να χρησιμοποιηθούν και σε άλλα παρόμοια σύνολα δεδομένων αποτελεσματικά.

Σε αυτό το Κεφάλαιο, χρησιμοποιήσαμε ένα μονού επιπέδου LSTM διπλής κατεύθυνσης και μια πολυεπίπεδη αρχιτεκτονική CNN + LSTM με διαφορετικές ρυθμίσεις υπερπαραμέτρων, όπως θα αναλυθεί στις παρακάτω υπό ενότητες. Η LSTM αρχιτεκτονική παρουσιάζεται στην Εικόνα 20 ενώ η αρχιτεκτονική CNN+LSTM στην Εικόνα 21. Η είσοδος στο LSTM είναι μια 64-διάστατη ενσωμάτωση κάθε λέξης. Αυτή η ρύθμιση στοχεύει στη αναπαράσταση των χρονικών εξαρτήσεων μεταξύ των λέξεων σε κάθε πίνακα περιεχομένων. Στην έξοδο έχουμε μια πυκνή αναπαράσταση κάθε πίνακα περιεχομένων, με 64 χαρακτηριστικά που χρησιμοποιούνται για την ταξινόμηση. Το LSTM διπλής κατεύθυνσης αποτελείται από 128 χαρακτηριστικά, που μεταβιβάζονται ως είσοδος στο επόμενο επίπεδο. Χρησιμοποιήθηκαν δύο κρυφά στρώματα με τη ReLU ως συνάρτηση ενεργοποίησης, δίνοντας ως έξοδο 64 και 32 χαρακτηριστικά αντίστοιχα, ακολουθούμενα από το επίπεδο εξόδου όπου χρησιμοποιείται η softmax συνάρτηση ενεργοποίησης. Το επίπεδο εξόδου έχει έναν κόμβο για κάθε ετικέτα ταξινόμησης, 26 ή 5 κατηγορίες σε αυτήν την περίπτωση. Σύμφωνα με το [76], ένα LSTM μπορεί να μεροληπτήσει όταν οι τελευταίες λέξεις έχουν μεγαλύτερη επιρροή από τις προηγούμενες. Προκειμένου να αντιμετωπίσουμε αυτό το πρόβλημα, έχουμε συμπεριλάβει ένα max pooling επίπεδο στο συνελικτικό νευρωνικό δίκτυο (CNN) που θα συζητηθεί παρακάτω, προκειμένου να προσδιοριστούν οι περισσότερο αντιπροσωπευτικές λέξεις ή φράσεις στον πίνακα περιεχομένων κάθε βιβλίου.

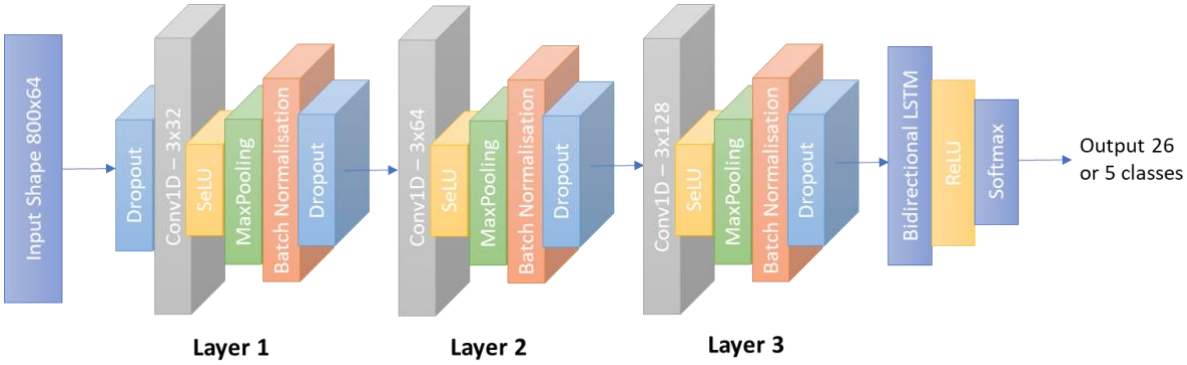
Το βασικό χαρακτηριστικό ενός CNN είναι ότι ενσωματώνει την έννοια των συνελικτικών επιπέδων. Κάθε συνελικτικό επίπεδο συνδέεται με ένα μικρό υποσύνολο των εισόδων συνήθως έναν πυρήνα μεγέθους 3. Όταν τα συνελικτικά επίπεδα στοιβάζονται, το υποσύνολο των εισόδων από τα προηγούμενα στρώματα συνδέεται μόνο με το υποσύνολο εισόδων του επόμενου επιπέδου. Αυτά τα συνελικτικά επίπεδα ονομάζονται χάρτες χαρακτηριστικών και όταν στοιβάζονται μπορούν να λειτουργούν ως πολλαπλά φίλτρα στην είσοδο. Όπως αναφέρθηκε παραπάνω, ένα από τα κύρια πλεονεκτήματα στη χρήση των CNN είναι ότι κάνουν την επιλογή χαρακτηριστικών πιο ισχυρή, αποφεύγοντας έτσι την πιθανή μεροληψία όταν οι τελευταίες λέξεις έχουν μεγαλύτερη επιρροή από τις προηγούμενες. Επιπλέον, με τη χρήση δειγματοληψίας (down-sampling) είναι δυνατή η μείωση του αριθμού των παραμέτρων στο μοντέλο μειώνοντας έτσι την υπολογιστική πολυπλοκότητα αυτού.



Εικόνα 20. Η μονού επιπέδου διπλής κατεύθυνσης LSTM αρχιτεκτονική. Η έξοδος θα είναι ή 26 ή 5 κλάσεις.

Έχουν προταθεί διαφορετικές τεχνικές συγκέντρωσης (pooling techniques), ώστε να μειωθεί το μέγεθος της εξόδου από μια στοίβα επιπέδων στην επόμενη στο δίκτυο [118], ο κύριος στόχος της οποίας είναι να καταγράψει τα πιο αντιπροσωπευτικά χαρακτηριστικά της εισόδου. Η πιο συνηθισμένη μέθοδος συγκέντρωσης και αυτή που χρησιμοποιείται σε αυτό το Κεφάλαιο είναι η μέγιστη συγκέντρωση (max-pooling), όπου η συνάρτηση max () εφαρμόζεται πάνω στα περιεχόμενα του συρόμενου παραθύρου ψηφοφορίας (polling / sliding window). Μια άλλη προσέγγιση είναι να επιλεχθεί ο στατιστικός μέσος όρος (statistical mean) των περιεχομένων. Τέλος, προκειμένου να τροφοδοτηθεί η συγκεντρωμένη έξοδος στο επόμενο επίπεδο, οι χάρτες χαρακτηριστικών πεπλατύνονται (flattened) σε μία στήλη και στη συνέχεια τροφοδοτούνται σε ένα πλήρως συνδεδεμένο σύνολο επιπέδων.

Σε αυτό το Κεφάλαιο χρησιμοποιούμε μια ελαφρώς διαφορετική αρχιτεκτονική σε σχέση με την κλασσική αρχιτεκτονική CNN, η οποία αποτελείται από ένα CNN με 3 επίπεδα, συνδυασμένο με ένα LSTM διπλής κατεύθυνσης. Η CNN+LSTM αρχιτεκτονική παρουσιάζεται στην Εικόνα 21. Κάθε επίπεδο αποτελείται από μίας διάστασης (1D) συνελικτικό επίπεδο με μια Scaled Exponential Linear Unit (SELU) συνάρτηση ενεργοποίησης [119] ακολουθούμενη από ένα max-pooling επίπεδο. Το επίπεδο αποκοπής (dropout layer) χρησιμοποιείται για να επιβραδύνει τη διαδικασία μάθησης, καθώς τα CNN τείνουν να μαθαίνουν γρήγορα με στόχο να παραχθεί ένα καλύτερο μοντέλο στο τέλος. Για αυτό το μοντέλο, χρησιμοποιούμε μια τυπική διαμόρφωση παράλληλων χαρτών μεγέθους 32, 64 και 128 και έναν πυρήνα μεγέθους 3. Το μέγεθος των χαρτών χαρακτηριστικών υποδηλώνει πόσες φορές ερμηνεύεται η είσοδος, ενώ το μέγεθος του πυρήνα υποδεικνύει πόσα χρονικά βήματα λαμβάνονται υπόψη καθώς η ακολουθία εισόδου διαβάζεται ή επεξεργάζεται στους χάρτες χαρακτηριστικών. Τέλος, η έξοδος των τριών επιπέδων τροφοδοτείται στο διπλής κατεύθυνσης LSTM που περιλαμβάνει ένα κρυφό επίπεδο, με συνάρτηση ενεργοποίησης ReLU, ακολουθούμενο από το επίπεδο εξόδου όπου χρησιμοποιείται η συνάρτηση Softmax. Το τελικό αποτέλεσμα είναι, όπως στην προηγούμενη περίπτωση, είτε οι 26 είτε οι 5 κατηγορίες.



Εικόνα 21. Απεικόνιση της αρχιτεκτονικής του Νευρωνικού δικτύου Βαθιάς Μάθησης CNN+LSTM.

4.6. Πειραματική αξιολόγηση

Σε αυτή την υποενότητα θα παρουσιαστεί η πειραματική αξιολόγηση της προτεινόμενης προσέγγισης. Ο κύριος στόχος αυτής της διαδικασίας είναι να επικυρώσει την ακρίβεια ταξινόμησης για το επιλεγμένο σύνολο δεδομένων. Πρώτον, θα παρουσιαστούν τα μέτρα που επιλέχθηκαν και χρησιμοποιούνται για την αξιολόγηση της απόδοσης των ταξινομητών, ακολουθούμενα από τις συναρτήσεις βελτιστοποίησης (optimization functions) που επιλέχθηκαν για τα μοντέλα Βαθιάς Μάθησης, την πειραματική ρύθμιση, καθώς και τα αποτελέσματα που προέκυψαν από τα πειράματα που πραγματοποιήθηκαν.

4.6.1. Μέτρα Απόδοσης

Τα πιο συχνά χρησιμοποιούμενα μέτρα απόδοσης για την αξιολόγηση ενός συστήματος αυτόματης ταξινόμησης κειμένου, είναι η ακρίβεια (Precision), η ανάκληση (Recall) και η βαθμολογία F1 (F1-score) [84]-[86], [99]. Τα μέτρα ακρίβειας μίκρο-και μάκρο-μέσου όρου χρησιμοποιούνται συνήθως για την αξιολόγηση της ακρίβειας των ταξινομητών σε προβλήματα πολλαπλών κλάσεων [86], [99]. Ωστόσο, κατά τη σύγκριση των αποτελεσμάτων μίκρο- και μάκρο μέσου όρου, το αποτέλεσμα μπορεί να είναι πολύ διαφορετικό. Αυτό πηγάζει από τη διαφορά κατά τον υπολογισμό των δύο μέτρων. Στην περίπτωση του μάκρο-μέσου όρου σε όλες τις κατηγορίες δίνονται πανομοιότυπα βάρη, ενώ στην περίπτωση του μίκρο-μέσου όρου τα ίδια βάρη δίδονται σε κάθε απόφαση ταξινόμησης ανά έγγραφο [99]. Εδώ, υπολογίζουμε τόσο την ακρίβεια μίκρο-μέσου όρου όσο και την ακρίβεια μάκρο-μέσου όρου για να λάβουμε μια συνολική εικόνα του τρόπου απόδοσης των ταξινομητών τόσο στις λιγότερο συχνές όσο και στις μεγαλύτερες κλάσεις. Στη συνέχεια, χρησιμοποιούμε τα μέτρα αυτά για να υπολογίσουμε τις βαθμολογίες F1 μίκρο και μάκρο-μέσου όρου για να εκτιμήσουμε την απόδοση των ταξινομητών. Τα μέτρα ακρίβειας μίκρο- και μάκρο-μέσου όρου που διατίθενται μέσω της βιβλιοθήκης scikit-learn [26],[27] της Python χρησιμοποιούνται σε αυτό το Κεφάλαιο, ο μαθηματικός ορισμός των οποίων δίνεται στην ενότητα 3.6.1.

4.6.2. Συναρτήσεις βελτιστοποίησης

Μια κλίση (gradient) είναι η παράγωγος μιας συνάρτησης που έχει περισσότερες της μίας παραμέτρους εισόδου. Καταγράφει την τοπική κλίση μιας συνάρτησης, επιτρέποντάς μας να προβλέψουμε το αποτέλεσμα ενός μικρού βήματος από ένα σημείο προς οποιαδήποτε κατεύθυνση. Η *μείωση της κλίσης* (gradient descent) είναι ένας αλγόριθμος βελτιστοποίησης που χρησιμοποιείται για την ελαχιστοποίηση κάποιας συνάρτησης υπολογίζοντας επαναληπτικά την κατεύθυνση της πιο

[26] http://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html
[27] http://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html

απότομης καθόδου όπως ορίζεται από το αρνητικό της κλίσης. Οι αλγόριθμοι βελτιστοποίησης χρησιμοποιούνται στην ΜΜ για την επικαιροποίηση των παραμέτρων του μοντέλου, δηλαδή τα βάρη των Νευρωνικών Δικτύων.

Απότομη κλίση (steepest gradient) από ένα συγκεκριμένο σημείο είναι η κατεύθυνση σύμφωνα η οποία μειώνει τη συνάρτηση κόστους (cost function) όσο το δυνατόν πιο γρήγορα. Ο *ρυθμός εκμάθησης* (learning rate) αποτυπώνει το μέγεθος των βημάτων που χρειάζεται να ακολουθήσουμε για να βρούμε το τοπικό ελάχιστο. *Συνάρτηση κόστους* (cost function) καθορίζει πόσο καλό είναι το μοντέλο στο να προβλέπει για ένα συγκεκριμένο σύνολο παραμέτρων . Η συνάρτηση κόστους έχει τη δική της καμπύλη και τις αντίστοιχες κλίσεις. Η κλίση της καμπύλης της συνάρτησης κόστους μας λέει πως να επικαιροποιήσουμε τις παραμέτρους του μοντέλου μας έτσι ώστε το τελευταίο να είναι πιο ακριβές στις προβλέψεις. Όσο μεγαλύτερη είναι η κλίση (gradient) τόσο πιο απότομη είναι η καμπύλη και τόσο πιο γρήγορα μπορεί να μάθει το μοντέλο.

Υπάρχουν 3 μέθοδοι μείωσης κλίσης: α. batch gradient descent, β. stochastic gradient descent και γ. mini-batch gradient descent. Στην πρώτη μέθοδο υπολογίζεται το σφάλμα για κάθε παράδειγμα στο σύνολο δεδομένων εκπαίδευσης, αλλά το μοντέλο ενημερώνεται μόνο αφού αξιολογηθούν όλα τα δεδομένα εκπαίδευσης (αυτό συνιστά μια εποχή εκπαίδευσης - training epoch). Μερικά πλεονεκτήματα της μεθόδου αυτής είναι η υπολογιστική της αποτελεσματικότητα, καθώς παράγει ένα σταθερό σφάλμα κλίσης (gradient error) και σταθερή σύγκλιση. Στα μειονεκτήματα συγκαταλέγεται ότι λόγω του σταθερού σφάλματος κλίσης που παρουσιάζει, αυτό μπορεί μερικές φορές να οδηγήσει σε μια κατάσταση σύγκλισης που δεν είναι η καλύτερη που μπορεί να επιτύχει το μοντέλο. Επίσης, απαιτεί ολόκληρο το σύνολο δεδομένων εκπαίδευσης να είναι στη μνήμη και να είναι διαθέσιμο στον αλγόριθμο.

Αντίθετα, η στοχαστική μείωση κλίσης (SGD) το κάνει αυτό για κάθε δείγμα εκπαίδευσης στο σύνολο δεδομένων, πράγμα που σημαίνει ότι ενημερώνει τις παραμέτρους για κάθε δείγμα εκπαίδευσης ένα προς ένα. Ανάλογα με το πρόβλημα, αυτό μπορεί να κάνει αυτή τη μέθοδο γρηγορότερη σε σχέση με την προηγούμενη. Ένα πλεονέκτημα είναι ότι οι συχνές ενημερώσεις μας επιτρέπουν να έχουμε έναν αρκετά λεπτομερή ρυθμό βελτίωσης. Οι συχνές ενημερώσεις, ωστόσο, είναι πιο υπολογιστικά ακριβές σε σχέση με την προσέγγιση του batch gradient descent. Επιπλέον, η συχνότητα αυτών των ενημερώσεων μπορεί να οδηγήσει σε κλίσεις που περιέχουν θόρυβο, οι οποίες μπορεί να προκαλέσουν σημαντικές διαφοροποιήσεις στο ποσοστό σφάλματος που παρατηρείται το οποίο σε κάθε άλλη περίπτωση θα μειωνόταν αργά.

Τέλος, η μέθοδος mini-batch gradient descent αποτελεί έναν συνδυασμό των δύο προηγούμενων. Στην πράξη χωρίζει το σύνολο δεδομένων εκπαίδευσης σε μικρές παρτίδες και εκτελεί μια ενημέρωση για καθεμία από αυτές τις παρτίδες. Αυτό δημιουργεί μια ισορροπία μεταξύ της ευρωστίας της στοχαστικής μείωσης κλίσης (SGD) και της αποτελεσματικότητας της μείωσης κλίσης παρτίδας (batch gradient descent). Τα κοινά μεγέθη παρτίδας (mini-batch) που χρησιμοποιούνται κυμαίνονται μεταξύ 50 και 256, αλλά όπως και για οποιαδήποτε άλλη τεχνική ΜΜ, δεν υπάρχει σαφής κανόνας, επειδή μπορεί να διαφέρει ανάλογα με την εφαρμογή. Αυτή η μέθοδος είναι η πιο κοινή μέθοδος μείωσης κλίσης για την εκπαίδευση ενός νευρωνικού δικτύου και είναι ο πιο συνηθισμένος τύπος μείωσης κλίσης στα Δίκτυα Βαθιάς Μάθησης.

Σε αυτό το Κεφάλαιο χρησιμοποιούνται δύο συναρτήσεις *βελτιστοποίησης που ανήκουν στην κατηγορία της στοχαστικής* μείωσης *κλίσης* (stochastic gradient descent) για τις δύο αρχιτεκτονικές του δικτύου Βαθιάς Μάθησης που αναλύθηκαν παραπάνω, συγκεκριμένα η Root Mean Square

Propagation[28] (RMSProp) και η Adaptive Moment Optimization (Adam) [120] όπως περιγράφεται παρακάτω.

Συνάρτηση βελτιστοποίησης RMSProp: Ο κύριος στόχος της συνάρτησης βελτιστοποίησης RMSProp είναι να διατηρηθεί ένας κινούμενος μέσος όρος (moving average) του τετραγώνου των κλίσεων (square of gradients) και να διαιρεθεί η κλίση (gradient) με τη ρίζα αυτού του μέσου όρου. Η εφαρμογή της συνάρτησης RMSprop που χρησιμοποιείται σε αυτό το Κεφάλαιο χρησιμοποιεί plain momentum και όχι Nesterov momentum.

Συνάρτηση βελτιστοποίησης Adam: Είναι μια μέθοδος μείωσης της στοχαστικής κλίσης (stochastic gradient descent) με βάση την προσαρμοστική εκτίμηση των στιγμών πρώτης τάξης (first-order moments - των μέσων) και δεύτερης τάξης (second-order moments – της μη κεντρικής διακύμανσης). Σύμφωνα με το [120] η συνάρτηση Adam είναι υπολογιστικά αποδοτική και έχει χαμηλές απαιτήσεις όσον αφορά τη μνήμη. Επιπλέον, θεωρείται κατάλληλη για προβλήματα που έχουν πολύ μεγάλο αριθμό παραμέτρων. Σε αυτό το Κεφάλαιο χρησιμοποιούμε τις συναρτήσεις βελτιστοποίησης RMSProp και Adam από το Keras API[29] όπως διατίθενται στο Tensorflow.

4.6.3. Πειραματική Ανάλυση

Η πειραματική αξιολόγηση πραγματοποιείται στη συλλογή ηλεκτρονικών βιβλίων (ebooks) Springer που περιγράφεται στην ενότητα 4.4. Τα διανύσματα χαρακτηριστικών κατασκευάστηκαν σύμφωνα με την προσέγγιση που περιεγράφηκε προηγουμένως, ενώ το σύνολο δεδομένων εκπαιδεύτηκε χρησιμοποιώντας τόσο τον αλγόριθμο SOM όσο και δύο διαφορετικές αρχιτεκτονικές DNN, συγκεκριμένα ένα LSTM και ένα CNN + LSTM. Πραγματοποιήθηκαν δύο κύκλοι πειραμάτων χρησιμοποιώντας διαφορετικά μεγέθη χάρτη και επιλογές διαμόρφωσης (SOM) και διαφορετικό μήκος, μέγεθος λεξιλογίου (DNNs) χρησιμοποιώντας τόσο τις 26 όσο και τις 5 κατηγορίες του συνόλου δεδομένων.

Ο πρώτος γύρος πειραμάτων στόχευε στην αξιολόγηση της απόδοσης του ταξινομητή κατά τη χρήση του σχήματος στάθμισης TF-IDF με το νευρωνικό δίκτυο SOM. Ως εκ τούτου, χρησιμοποιήθηκαν οι ίδιες επιλογές διαμόρφωσης τόσο για ολόκληρο το σύνολο δεδομένων Springer (26 κλάσεις) όσο και για το υποσύνολο των πέντε (5) κλάσεων. Πιο συγκεκριμένα, οι παράμετροι που αφορούσαν τις “rough” και “finetune” εποχές εκπαίδευσης καθορίστηκαν σε 15 και 7 αντίστοιχα. Χρησιμοποιήθηκαν διαφορετικά μεγέθη διανυσμάτων, συγκεκριμένα διανύσματα διαστάσεων 2000, 2500 και 3000, για να εκτιμηθεί η επίδραση του μεγέθους του λεξιλογίου στην απόδοση του ταξινομητή. Σε αυτή την περίπτωση τα μεγέθη χάρτη που χρησιμοποιήθηκαν ήταν 10x10, 16x16, 20x20, 30x30, 40x40, 50x50, 60x60 και 70x70. Σε αυτό το σημείο πρέπει να σημειωθεί ότι τα πειράματα διεξήχθησαν τόσο με τον αλγόριθμο INterSECT όσο και με τον αλγόριθμο Majority Voting για την επιλογή ετικέτας.

Ο δεύτερος γύρος πειραμάτων στόχευε στην αξιολόγηση της χρήσης των DNNs, αντί του SOM, (πιο συγκεκριμένα ενός LSTM και ενός CNN + LSTM, όπως προαναφέρθηκε), στην ταξινόμηση ενός συνόλου δεδομένων πολλαπλών κλάσεων. Σε αυτό τον γύρο, χρησιμοποιήθηκε ο Keras Tokenizer, ενώ πειραματιστήκαμε με διαφορετικά μεγέθη μήκους για τους πίνακες περιεχομένων των βιβλίων, συγκεκριμένα, με 800, 1000, 2000, 2500 και 3000 λέξεις, ενώ το μήκος του συνολικού λεξιλογίου ήταν

[28] https://www.tensorflow.org/api_docs/Python/tf/keras/optimizers/RMSprop
[29] https://keras.io/

20000 λέξεις. Πραγματοποιήθηκαν επίσης πειράματα για να εκτιμηθεί πόσο η παράμετρος του μέγιστου μήκους και η έλλειψη προεπεξεργασίας επηρεάζουν την ακρίβεια των ταξινομητών.

Επιπλέον, οι διάφορες υπερπαράμετροι και οι συναρτήσεις βελτιστοποίησης που παρουσιάζονται στην υπό-ενότητα 0 ρυθμίστηκαν όπως φαίνεται παρακάτω.

LSTM: embedding dimension: 64, max features: 20000, input length: 800, dropout: 0.2-0.25, loss: categorical cross entropy, train batch size: 128, epochs: 10.

CNN: embedding dimension: 64, max features: 20000, input length: 800, dropout: 0.25-0.5, loss: categorical cross entropy, train batch size: 128, epochs: 15, learning rate: 3e-4.

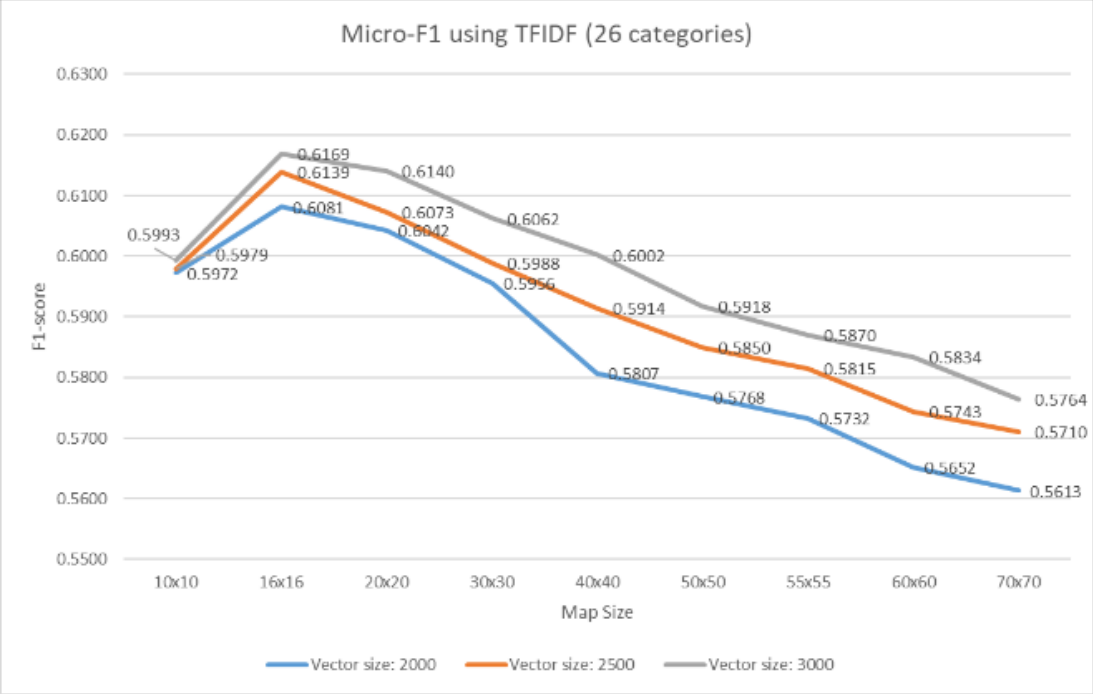
4.6.4. Αποτελέσματα

1ος γύρος: SOM+ TF-IDF

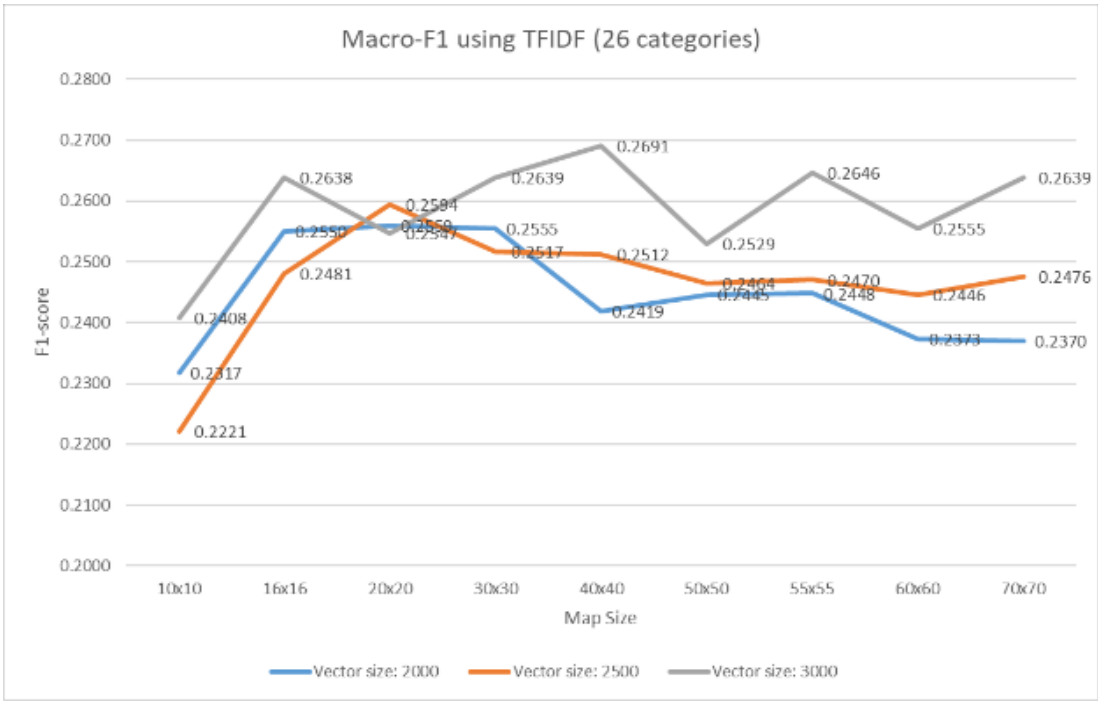
Όπως περιγράφεται στην προηγούμενη υπό-ενότητα, πραγματοποιήσαμε πειράματα χρησιμοποιώντας τόσο τις πέντε (5) πολυπληθείς όσο και τις συνολικά εικοσιέξι (26) κλάσεις του συνόλου δεδομένων Springer. Για να παρατηρήσουμε την απόδοση του SOM, πρώτα πειραματιστήκαμε με διαφορετικά μεγέθη χάρτη και διανύσματος χρησιμοποιώντας τετράγωνους χάρτες και 1-grams.

Σε σχέση με τις 26 κλάσεις, παρατηρήσαμε ότι τα αποτελέσματα τείνουν να ακολουθούν το ίδιο μοτίβο, καθώς για κάθε μέγεθος χάρτη παρατηρήθηκαν μικρές διαφορές μεταξύ διαφορετικών μεγεθών διανυσμάτων. Για το ίδιο μέγεθος χάρτη, οι διαφορές μεταξύ των διαφορετικών μεγεθών διανυσμάτων κυμαίνονται από 0.05-0.2%. Η απόδοση του ταξινομητή ήταν γύρω στο 0.62 ή 62% για ένα χάρτη 16x16. Μπορούμε επίσης να παρατηρήσουμε ότι καθώς το μέγεθος του χάρτη αυξάνει, οι βαθμολογίες F1 Μίκρο-μέσου όρου τείνουν να μειώνονται φτάνοντας ως και 58% σε ένα πλέγμα 70x70. Αυτό είναι αναμενόμενο καθώς για μεγαλύτερα μεγέθη χάρτη τα δείγματα καθώς και οι αντίστοιχες ετικέτες τείνουν να διασπείρονται επάνω στο χάρτη. Για παράδειγμα, σε ένα πλέγμα 20x20 εάν εξετάσουμε ένα συγκεκριμένο κόμβο επάνω στο χάρτη, θα μπορούσαν π.χ. να συγκεντρώνονται πέντε (5) ετικέτες σε αυτό, ενώ αντίστοιχα σε ένα πλέγμα 40x40 στον ίδιο κόμβο στο χάρτη μπορεί πλέον να υπάρχει μόνο μια (1) ετικέτα. Θα πρέπει επίσης να εξετάσουμε την περίπτωση όπου δεν υπάρχει ετικέτα στον κόμβο, γεγονός που μπορεί να ισχύει για τα μεγαλύτερα πλέγματα, όπως θα παρουσιαστεί στον Πίνακα V. Αυτοί οι δύο παράγοντες εξηγούν τη μείωση των βαθμολογιών F1 για τα μεγαλύτερα πλέγματα, που είναι της τάξης 2 -4% από ό, τι στα μικρότερα πλέγματα. Οι βαθμολογίες Μίκρο-μέσου όρου για τις 26 κλάσεις χρησιμοποιώντας TF-IDF για την αναπαράσταση των πινάκων περιεχομένων των βιβλίων, και του αλγορίθμου Majority Voting για την επιλογή ετικετών για διαφορετικά μεγέθη διανυσμάτων παρουσιάζονται στην Εικόνα 22.

Αντίστοιχα για τις βαθμολογίες F1 Μάκρο-μέσου όρου, μπορούμε επίσης να εξάγουμε εύκολα ότι ο ταξινομητής μπορεί να προβλέψει καλύτερα την κλάση του κάθε βιβλίου όταν χρησιμοποιεί 3000 διαστάσεις ανά διάνυσμα, καθώς αποδίδει καλύτερα στην πλειονότητα των μεγεθών χαρτών με τα οποία πειραματιστήκαμε, όπως απεικονίζεται στην Εικόνα 23. Ο ταξινομητής επιτυγχάνει βαθμολογίες γύρω στο 27% για ένα χάρτη 40x40. Σε αυτή την περίπτωση η απόκλιση μεταξύ των διαφορετικών μεγεθών χάρτη κυμαίνεται γύρω από το 2%. Είναι εμφανές ότι τα αποτελέσματα του Μάκρο-μέσου όρου είναι σημαντικά χαμηλότερα σε σχέση με τα αντίστοιχα του Μίκρο-μέσου όρου, αλλά αυτό είναι αναμενόμενο καθώς είναι πιο δύσκολο για τον ταξινομητή να μπορέσει να προβλέψει καλά σε κλάσεις όπως οι “Music”, “Literature” και “Food” οι οποίες έχουν έναν μικρό αριθμό δειγμάτων όπως παρουσιάζεται αναλυτικά στον Πίνακα 21.



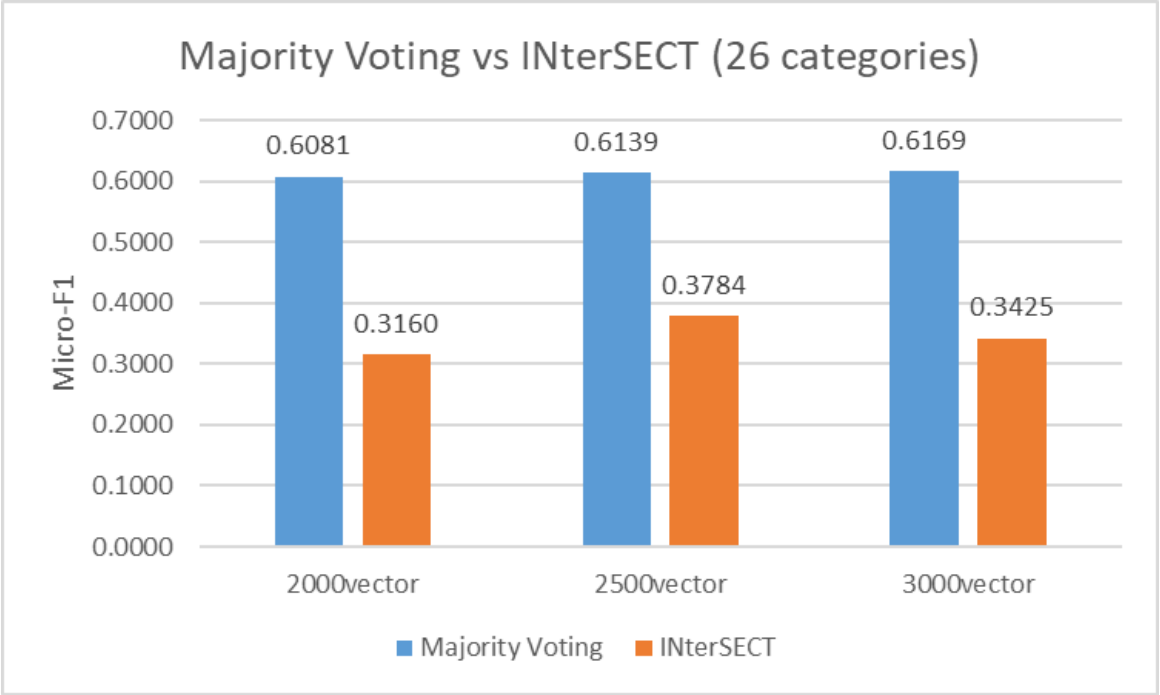
Εικόνα 22. Μίκρο-μέσος όρος βαθμολογίας F1 για τις 26 κλάσεις χρησιμοποιώντας TF-IDF για την αναπαράσταση των πινάκων περιεχομένων των βιβλίων, και του αλγορίθμου Majority Voting για την επιλογή ετικετών για διαφορετικά μεγέθη διανυσμάτων.



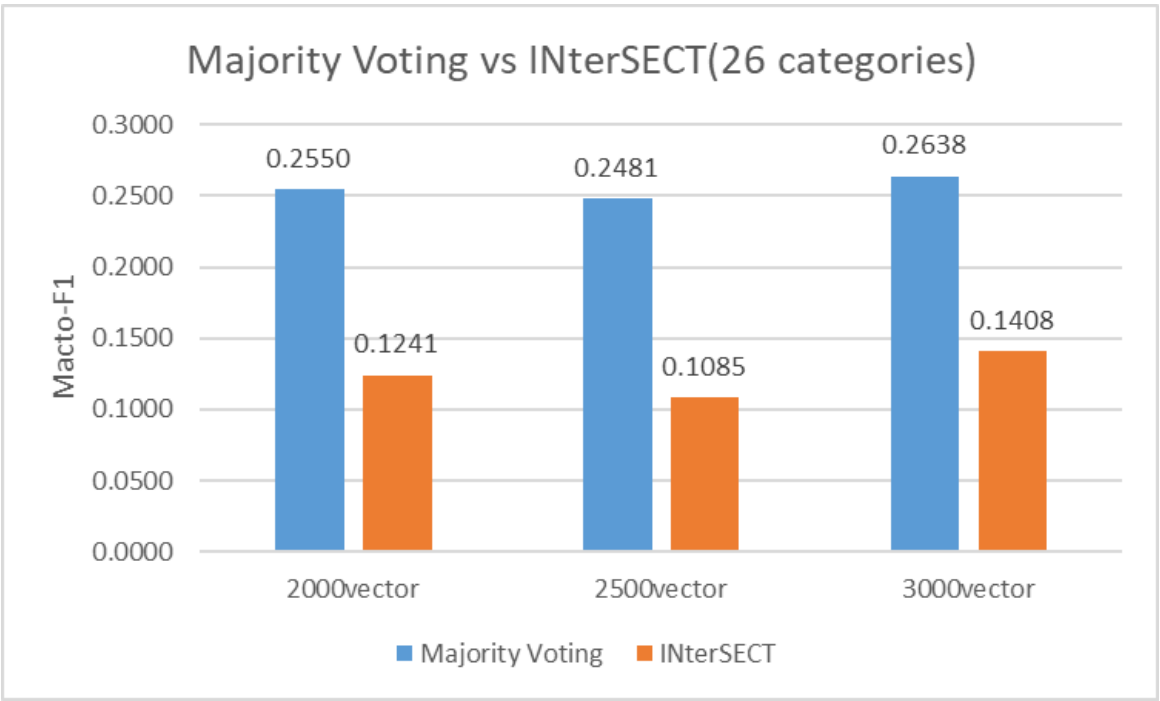
Εικόνα 23. Μάκρο-Μέσος Όρος βαθμολογίας F1 για τις 26 κλάσεις χρησιμοποιώντας TF-IDF για την αναπαράσταση των πινάκων περιεχομένων των βιβλίων, και του αλγορίθμου Majority Voting για την επιλογή ετικετών για διαφορετικά μεγέθη διανυσμάτων.

Η Εικόνα 24 παρουσιάζει τη σύγκριση μεταξύ των δύο διαφορετικών μεθόδων επιλογής ετικέτας για ένα συγκεκριμένο κόμβο επάνω στον χάρτη, συγκεκριμένα των αλγορίθμων Majority Voting και INterSECT για τις 26 κλάσεις. Όπως αναφέρθηκε και παραπάνω ο INterSECT είναι ένας αλγόριθμος προσαρμοσμένος για ταξινόμηση πολλαπλής ετικέτας και έτσι ήταν αναμενόμενο ότι η απόδοσή του θα ήταν χαμηλότερη σε ένα σενάριο ταξινόμησης πολλαπλών κλάσεων. Αυτό

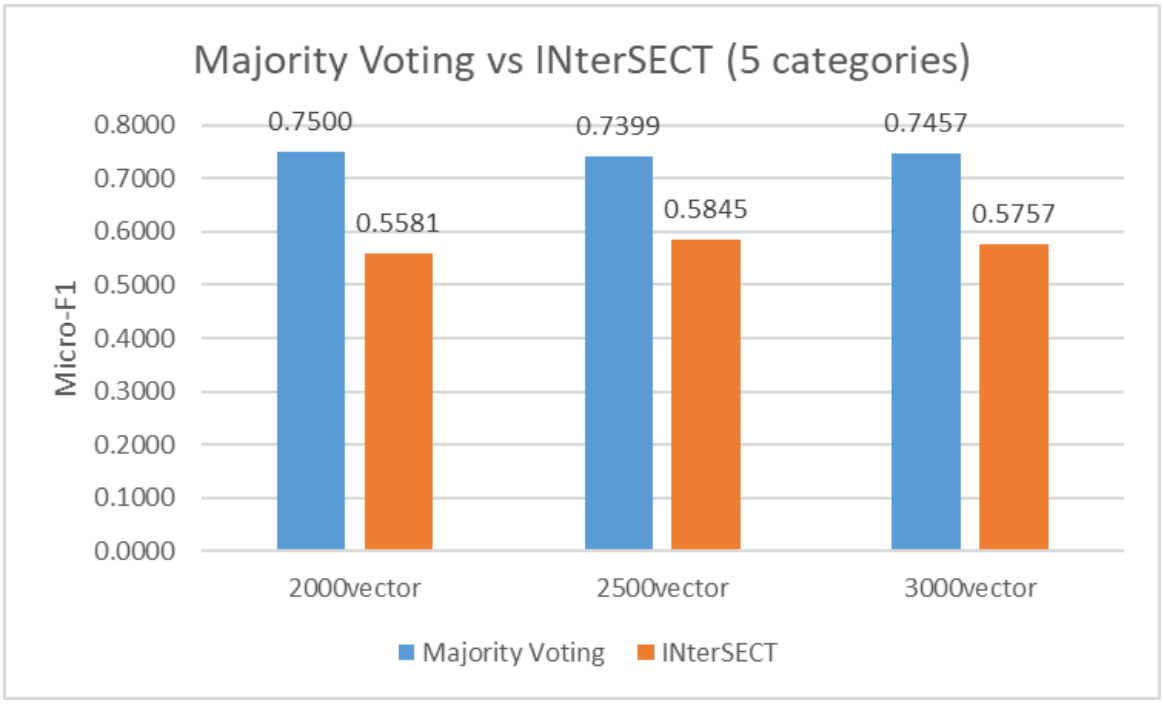
αντικατοπτρίζεται και στα αποτελέσματα όπως αυτά παρουσιάζονται στην Εικόνα 24 (Μίκρο) και Εικόνα 25 (Μάκρο) για τις 26 κλάσεις και στην Εικόνα 26 και Εικόνα 27, αντίστοιχα, για τις 5 κλάσεις. Από τα παραπάνω είναι εμφανές ότι ο αλγόριθμος "Majority Voting" παρουσιάζει καλύτερη επίδοση σε όλες τις περιπτώσεις σε σχέση με τον INterSECT. Η διαφορά αναλογίας μεταξύ του αλγορίθμου "Majority Voting" και του INterSECT είναι σχεδόν η ίδια (50% τοις εκατό χαμηλότερη για τον INterSECT) τόσο για τον Μίκρο- όσο και για τον Μάκρο-μέσο όρο.



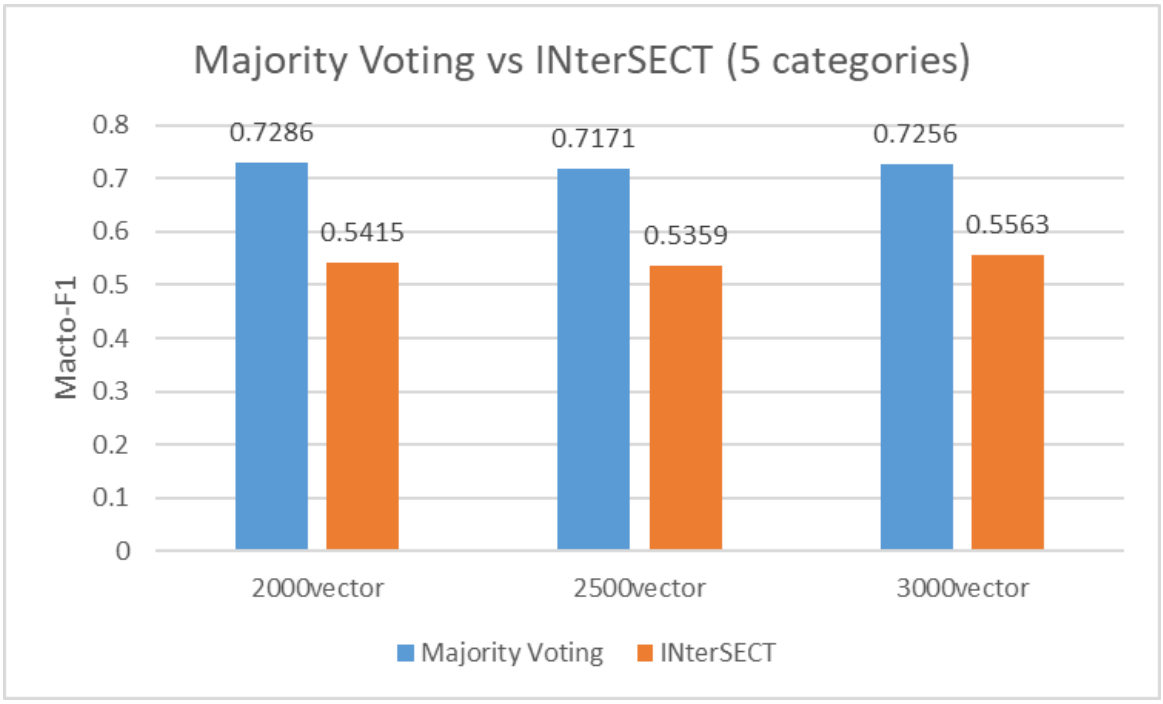
Εικόνα 24. Σύγκριση βαθμολογίας F1 Μίκρο-μέσου όρου για τις 2 διαφορετικές μεθόδους επιλογής ετικέτας για τις 26 κλάσεις.



Εικόνα 25. Σύγκριση βαθμολογίας F1 Μάκρο-μέσου όρου για τις 2 διαφορετικές μεθόδους επιλογής ετικέτας για τις 26 κλάσεις.



Εικόνα 26. Σύγκριση βαθμολογίας F1 Μίκρο-μέσου όρου για τις 2 διαφορετικές μεθόδους επιλογής ετικέτας για τις 5 κλάσεις.



Εικόνα 27. Σύγκριση βαθμολογίας F1 Μάκρο-μέσου όρου για τις 2 διαφορετικές μεθόδους επιλογής ετικέτας για τις 5 κλάσεις.

Τα καλύτερα αποτελέσματα που πέτυχαμε για κάθε μέγεθος χάρτη παρουσιάζονται στον Πίνακα 24. Αν παρατηρήσει κανείς τα αποτελέσματα για τις 5 κλάσεις, φαίνεται πως υπάρχει μια τάση στην απόδοση του ταξινομητή τόσο για τον Μίκρο- όσο και για τον Μάκρο-μέσο όρο. Παρόμοια με τις 26 κατηγορίες, οι βαθμολογίες F1 τείνουν επίσης να μειώνονται με μεγαλύτερα μεγέθη χάρτη. Μια άλλη παρατήρηση είναι ότι ειδικά για τις βαθμολογίες Μίκρο-μέσου όρου η διαφορά μεταξύ

διαφορετικών μεγεθών διανύσματος είναι αμελητέα και ως εκ τούτου τα αποτελέσματα για τα διαφορετικά μεγέθη διανυσμάτων παραλείπονται. Από τον Πίνακα 24 μπορεί να επίσης να εξαχθεί ότι ο χάρτης 16x16 δίνει τα καλύτερα αποτελέσματα με βαθμολογία F1 75% για Μίκρο- και 73% για Μάκρο-μέσο όρο αντίστοιχα.

Πίνακας 24. Βαθμολογίες F1 Μίκρο- και Μάκρο μέσου όρου για διαφορετικά μεγέθη χάρτη με χρήση TF-IDF και του αλγορίθμου Majority Voting (5 κλάσεις).

	Μέγεθος Χάρτη	QE	TE	Μίκρο-F1	Μάκρο-F1
5 κλάσεις	*10x10*	0.42	43.11	0.74	0.72
	16x16	0.24	42.65	**0.75**	**0.73**
	20x20	0.00	42.35	0.75	0.72
	30x30	0.00	41.52	0.74	0.71
	40x40	0.04	40.75	0.73	0.71
	50x50	0.10	40.25	0.72	0.70
	60x60	0.13	39.61	0.72	0.70
	70x70	0.00	39.05	0.71	0.68

Πίνακας 25. Βαθμολογίες F1 για κάθε μια από τις 5 κλάσεις για διαφορετικά μεγέθη χάρτη με χρήση TF-IDF και του αλγορίθμου Majority Voting.

	Κλάση	Ακρίβεια (Precision)	Ανάκληση (Recall)	Βαθμολογία F1
10x10	*Computer science*	0.73	0.85	0.79
	Engineering	0.67	0.52	0.58
	Mathematics	0.70	0.69	0.69
	Medicine	0.84	0.90	0.87
	Physics	0.74	0.63	0.68
	Μέσος Όρος	**0.74**	**0.75**	**0.74**
16x16	*Computer science*	0.75	0.84	0.79
	Engineering	0.61	0.55	0.58
	Mathematics	0.69	0.72	0.71
	Medicine	0.9	0.89	0.89
	Physics	0.74	0.59	0.66
	Μέσος Όρος	**0.74**	**0.75**	**0.74**
20x20	*Computer science*	0.76	0.83	0.79
	Engineering	0.63	0.58	0.61
	Mathematics	0.72	0.69	0.71
	Medicine	0.87	0.9	0.88
	Physics	0.7	0.6	0.65
	Μέσος Όρος	**0.75**	**0.75**	**0.75**
30x30	*Computer science*	0.76	0.81	0.78
	Engineering	0.62	0.58	0.60
	Mathematics	0.65	0.68	0.67
	Medicine	0.87	0.89	0.88
	Physics	0.69	0.58	0.63
	Μέσος Όρος	**0.73**	**0.74**	**0.73**
40x40	*Computer science*	0.72	0.83	0.77
	Engineering	0.64	0.53	0.58
	Mathematics	0.67	0.69	0.68
	Medicine	0.87	0.87	0.87
	Physics	0.71	0.59	0.64
	Μέσος Όρος	**0.73**	**0.73**	**0.73**

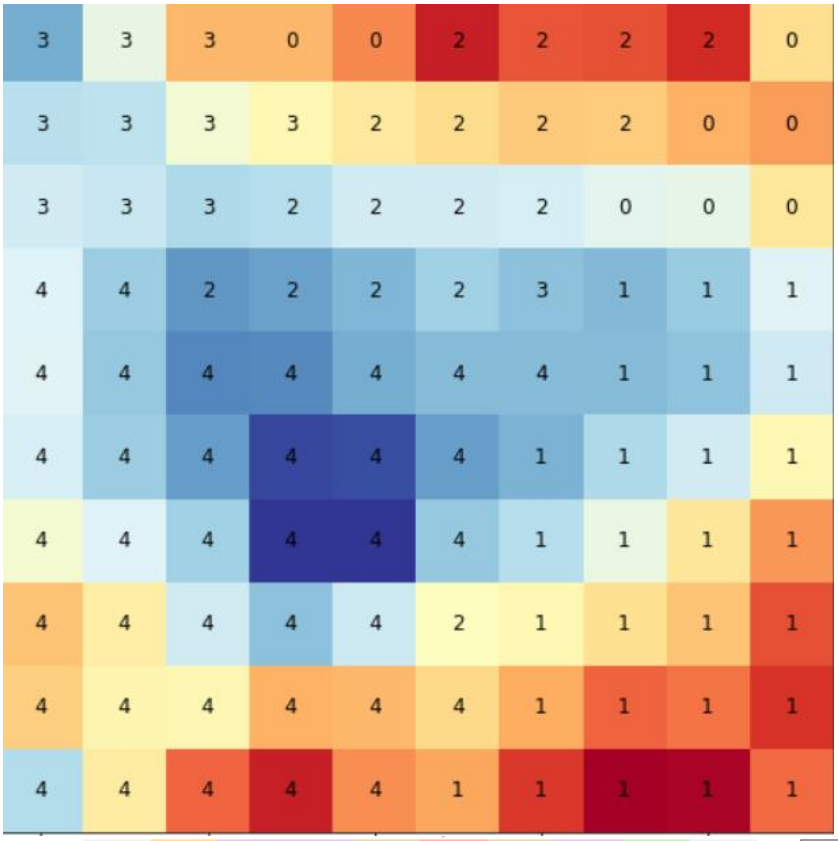
Ο Πίνακας 25 παρουσιάζει την Ακρίβεια (Precision), την Ανάκληση (Recall) και τις βαθμολογίες F1 που πετύχαμε ανά κατηγορία. Από τα αποτελέσματα βλέπουμε ότι ο ταξινομητής μπορεί να προβλέψει καλύτερα την κλάση 'Medicine' σε σύγκριση με τις υπόλοιπες. Η βαθμολογία F1 σε αυτή την περίπτωση κυμαίνεται από 87-89%. Αυτό μπορεί να εξηγηθεί καθώς η κλάση 'Medicine' είναι η 2η μεγαλύτερη κατηγορία σε αριθμό δειγμάτων και επίσης εννοιολογικά είναι περισσότερο διακριτή σε σύγκριση με τις κλάσεις 'Computer Science' και 'Mathematics'. Μια άλλη παρατήρηση είναι ότι ο ταξινομητής είναι πιο ικανός να προβλέψει την κλάση 'Computer science' από την κλάση 'Engineering' με τα επιτευχθέντα αποτελέσματα να κυμαίνονται από 77-79% και 58-61% αντίστοιχα. Αυτό οφείλεται στο γεγονός ότι η κατηγορία 'Engineering' σχετίζεται στενά με την κλάση 'Computer science' και έτσι κάποια δείγματα ελέγχου μπορεί να θεωρηθεί ψευδώς ότι ανήκουν στην κατηγορία 'Computer science' αντί της 'Engineering'. Επιπλέον, η κατηγορία 'Engineering' είναι μικρότερη κατά το ήμισυ σε αριθμό δειγμάτων σε σχέση με την κλάση 'Computer science'. Επιπλέον, ο ταξινομητής πέτυχε απόδοση περίπου 67-71% για την κλάση 'Mathematics' και 63-68% για την κλάση 'Physics', το οποίο μπορεί να θεωρηθεί αρκετά υψηλό με δεδομένο τον αριθμό των διαθέσιμων δειγμάτων για εκπαίδευση και έλεγχο, όπως αυτός παρουσιάζεται στον Πίνακα 21.

Η Εικόνα 28 παρουσιάζει διαφορετικούς τρόπους απεικόνισης ενός εκπαιδευμένου χάρτη SOM, συγκεκριμένα U-Matrices (Unified distance matrices – ενοποιημένους πίνακες απόστασης), Hitmaps και cluster maps για διαφορετικά μεγέθη πλέγματος για τις 5 κλάσεις. Όπως αναφέρθηκε προηγουμένως, κάθε κόμβος στο χάρτη αντιπροσωπεύεται από ένα αρχικό διάνυσμα (codebook), που έχει την ίδια διάσταση με τα διανύσματα που αναπαριστούν τους πίνακες περιεχομένων του συνόλου δεδομένων Springer. Ένα U-Matrix είναι ένα χρήσιμο εργαλείο για την οπτικοποίηση των αποστάσεων μεταξύ γειτονικών κόμβων στο χάρτη. Επιπλέον, βοηθά στην καλύτερη απεικόνιση της δομής των συστάδων (clusters) που έχουν δημιουργηθεί επάνω στον χάρτη. Παρατηρώντας τα χρώματα του χάρτη μπορεί κανείς να κάνει διάκριση μεταξύ συστάδων και των ορίων αυτών. Πιο συγκεκριμένα, οι κόμβοι που έχουν παρόμοιο χρώμα ανήκουν στην ίδια συστάδα ενώ οι κόμβοι με διαφορετικό ή σκούρο χρώμα υποδεικνύουν όρια συστάδων. Παρόμοια χρώματα υποδηλώνουν επίσης ομοιότητα μεταξύ διανυσμάτων στην ίδια περιοχή, ενώ τα σκούρα χρώματα υποδηλώνουν μεγάλη απόσταση και επομένως λιγότερο όμοιες τιμές των διανυσμάτων codebook στο χώρο εισόδου. Οι U-Matrices που απεικονίζονται στον Πίνακα VI έχουν αριθμούς επάνω σε κάθε έναν από τους κόμβους του χάρτη. Ο αριθμός κάθε κόμβου δείχνει την κλάση στην οποία αντιστοιχήθηκε ο συγκεκριμένος κόμβος μετά τη διαδικασία της εκπαίδευσης. Πιο συγκεκριμένα, θα μπορούσαμε να πούμε ότι δείχνει την ετικέτα που επιλέχθηκε με χρήση του αλγορίθμου "Majority Voting" όπου 0 – Physics, 1 – Medicine, 2 – Engineering, 3 – Mathematics, 4 – Computer science.

Επίσης, είναι εμφανές ότι οι χάρτες μικρότερου μεγέθους, τείνουν να έχουν περισσότερες ετικέτες σε κάθε κόμβο, ενώ οι μεγαλύτεροι χάρτες έχουν λιγότερες ή καθόλου. Αυτό μπορεί να απεικονιστεί καλύτερα χρησιμοποιώντας Hitmaps όπως φαίνεται στην ίδια Εικόνα. Σε αυτήν την προβολή, κάθε κόμβος έχει έναν αριθμό που αντιστοιχεί στον αριθμό των δειγμάτων που έχουν εκχωρηθεί στον κόμβο μετά την φάση εκπαίδευσης. Τα Hitmaps μπορούν να βοηθήσουν στην κατανόηση του τρόπου διασποράς των δειγμάτων πάνω στο πλέγμα. Τέλος, η τρίτη απεικόνιση της Εικόνας 27 για κάθε μέγεθος χάρτη δείχνει τα cluster maps. Πρέπει να σημειωθεί ότι σε αυτήν την περίπτωση έχουν δημιουργηθεί 5 ομάδες, 1 για κάθε κλάση. Ωστόσο, καθώς το SOM περιέχει μη οριοθετημένες ομάδες, ενδέχεται να υπάρχουν περισσότερες ομάδες που έτσι δεν είναι δυνατόν να απεικονιστούν με τη βοήθεια αυτής της προβολής. Για το λόγο αυτό, θεωρούμε πως η προβολή U-Matrix είναι ένας καλύτερος οδηγός για τον προσδιορισμό των συστάδων.

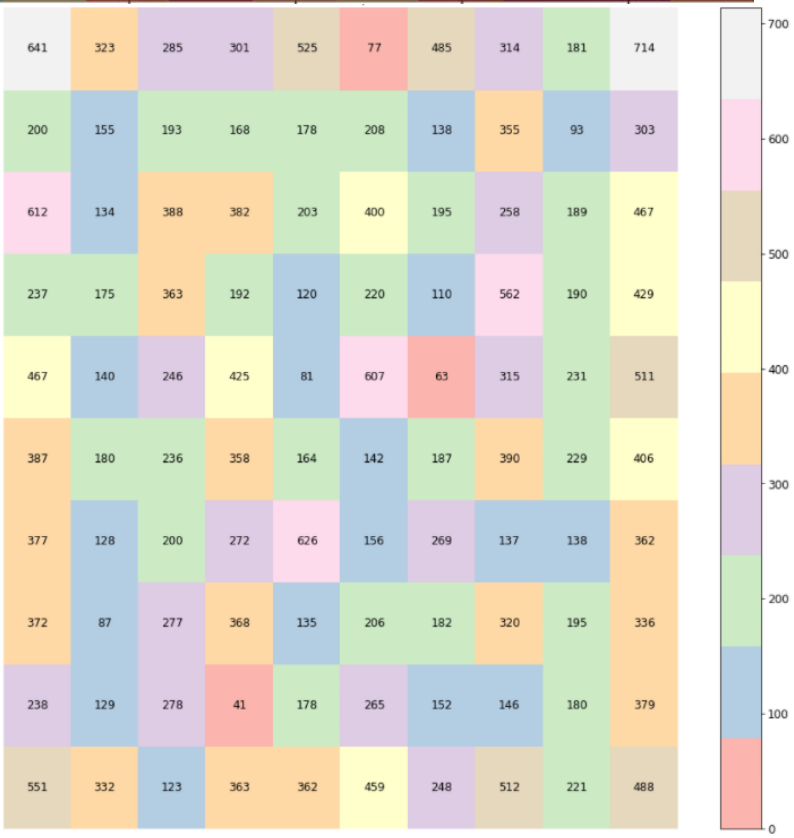
10x10

U-Matrix

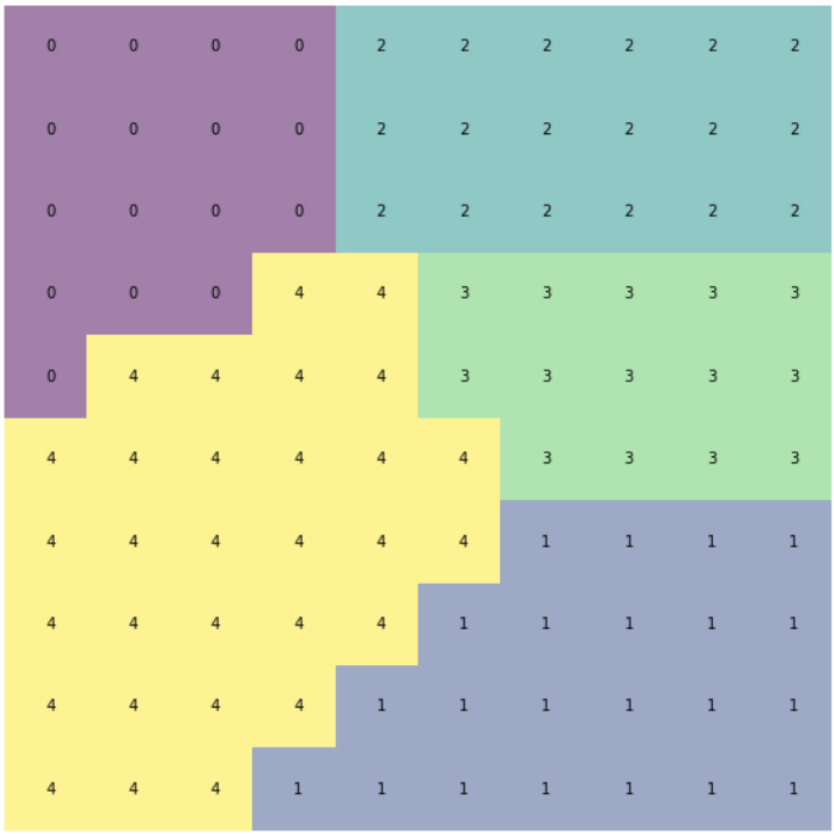


Legend:
0 – Physics,
1 – Medicine,
2 – Engineering,
3 – Mathematics,
4 – Computer science

Hitmap

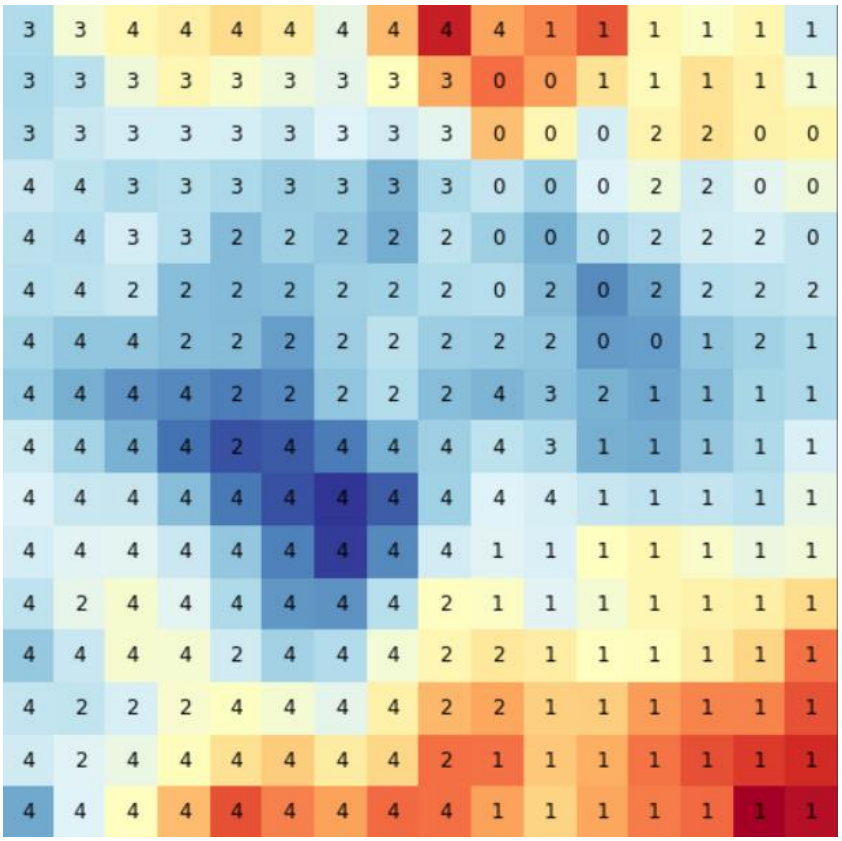


Cluster



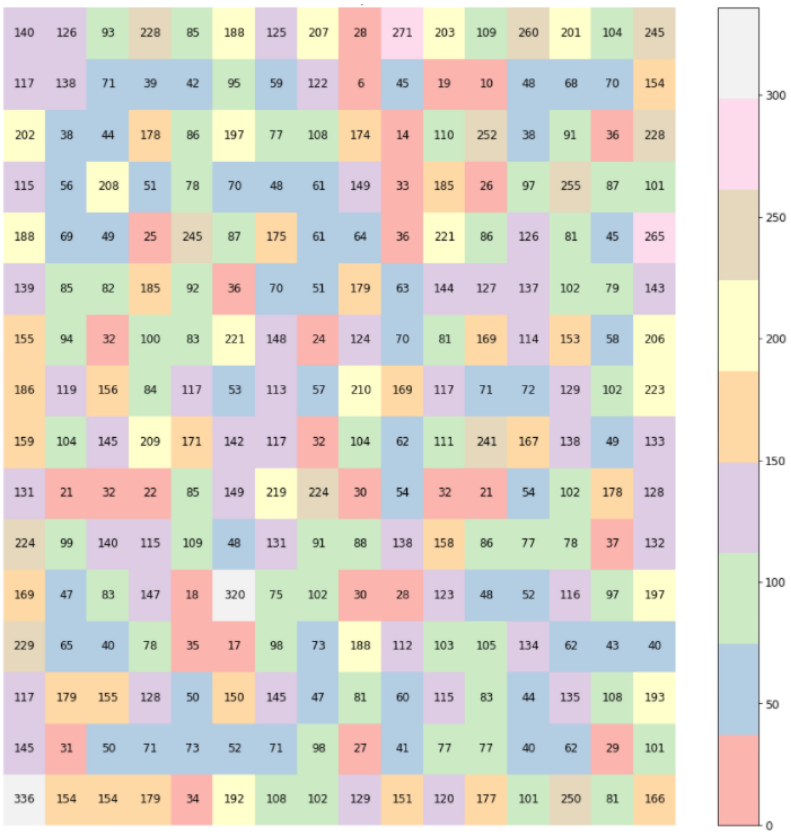
16x16

U-Matrix

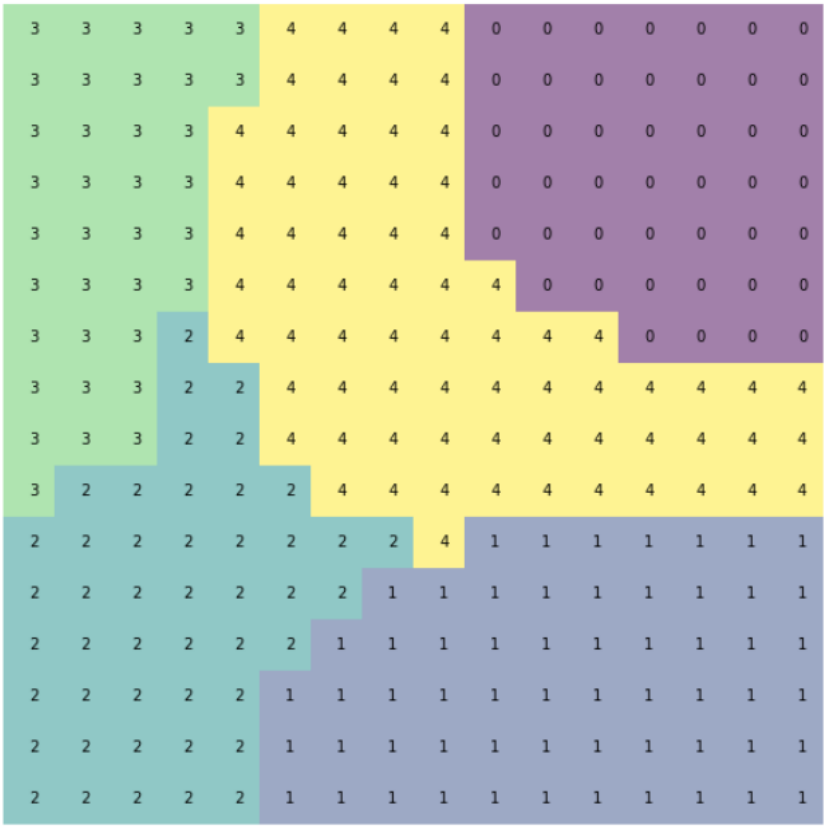


Legend:
0 – Physics,
1 – Medicine,
2 – Engineering,
3 – Mathematics,
4 – Computer science

Hitmap

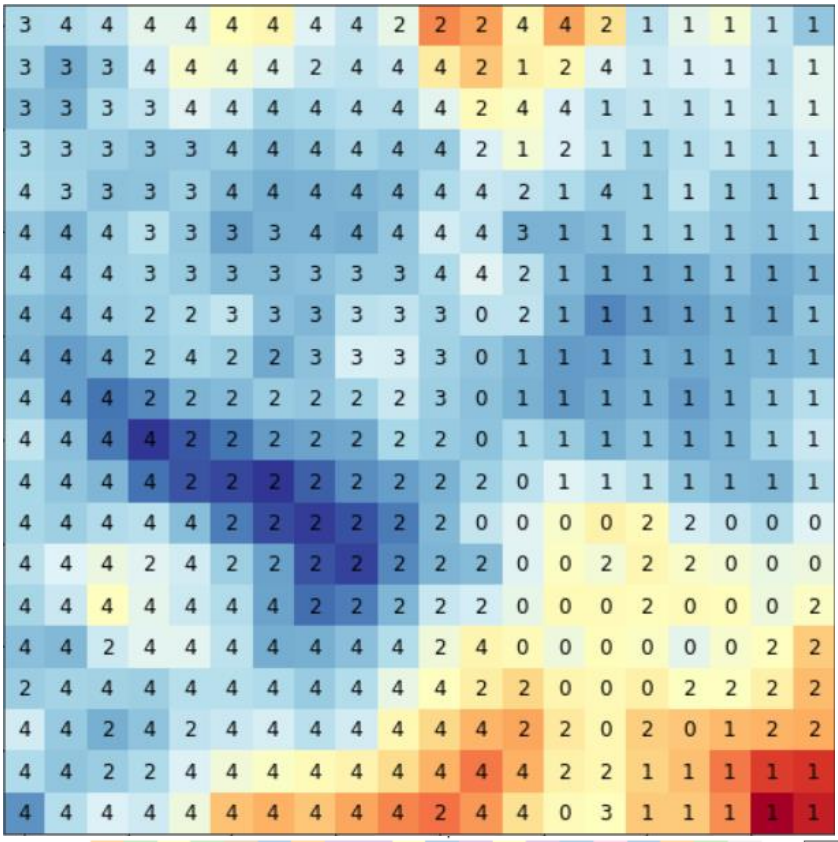


Cluster



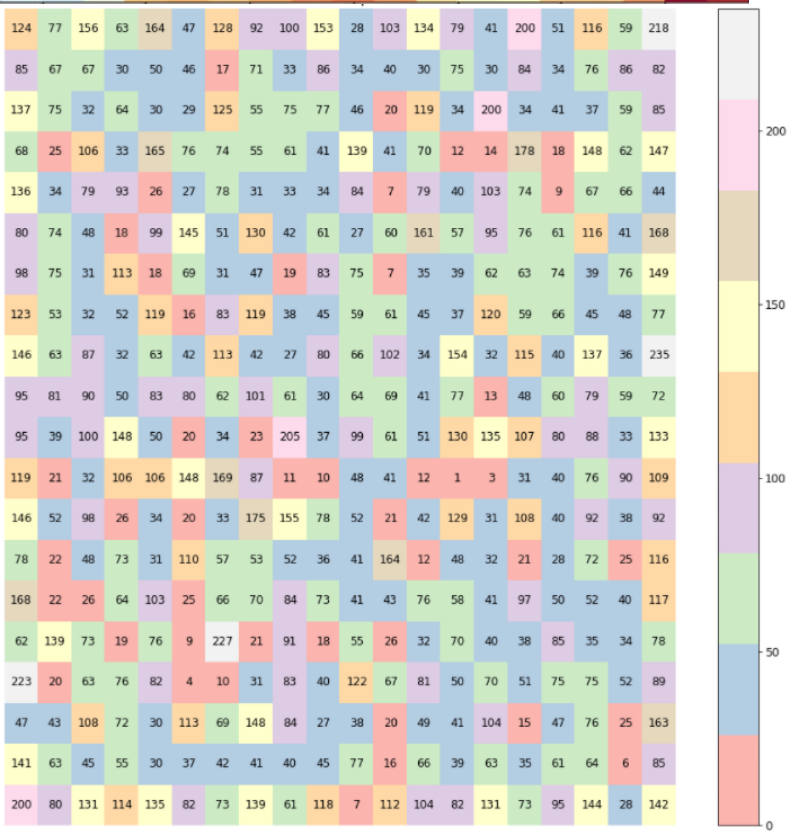
20x20

U-Matrix

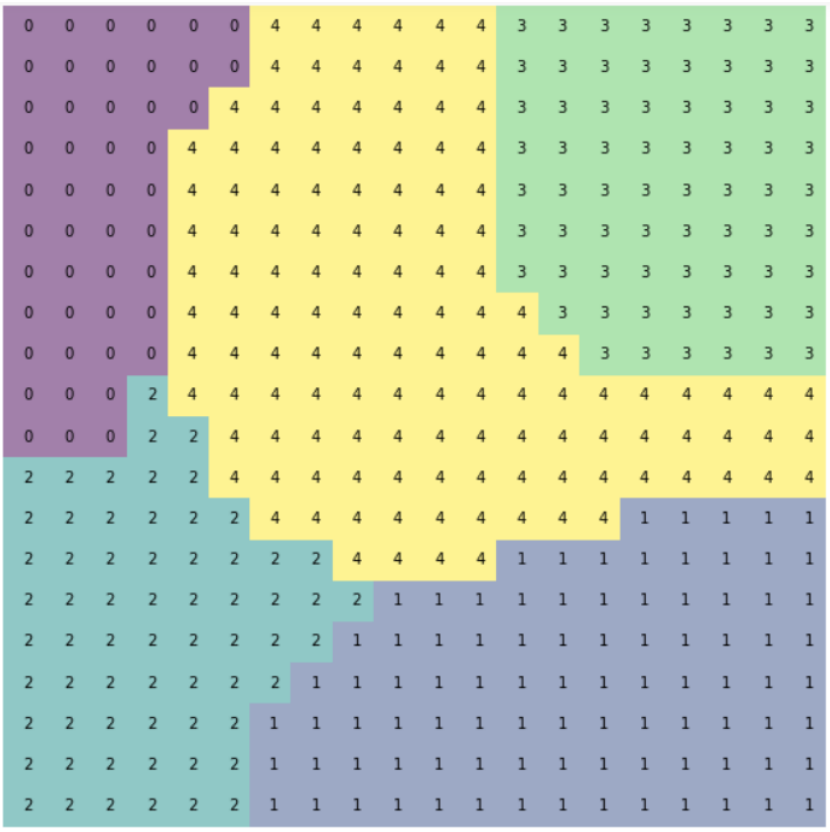


Legend:
0 – Physics,
1 – Medicine,
2 – Engineering,
3 – Mathematics,
4 – Computer science

Hitmap

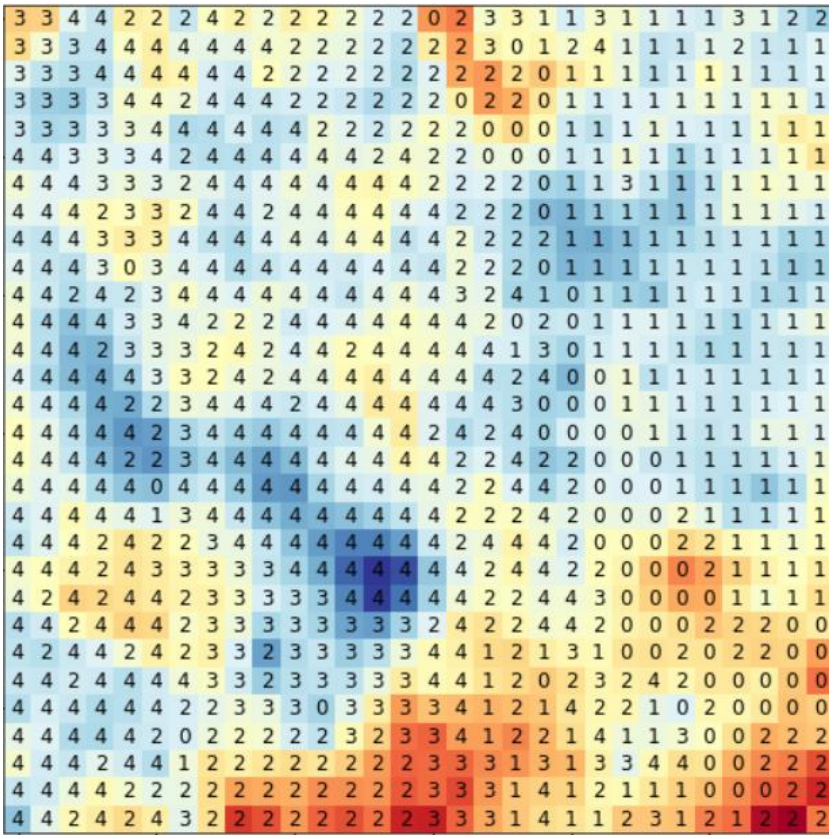


Cluster



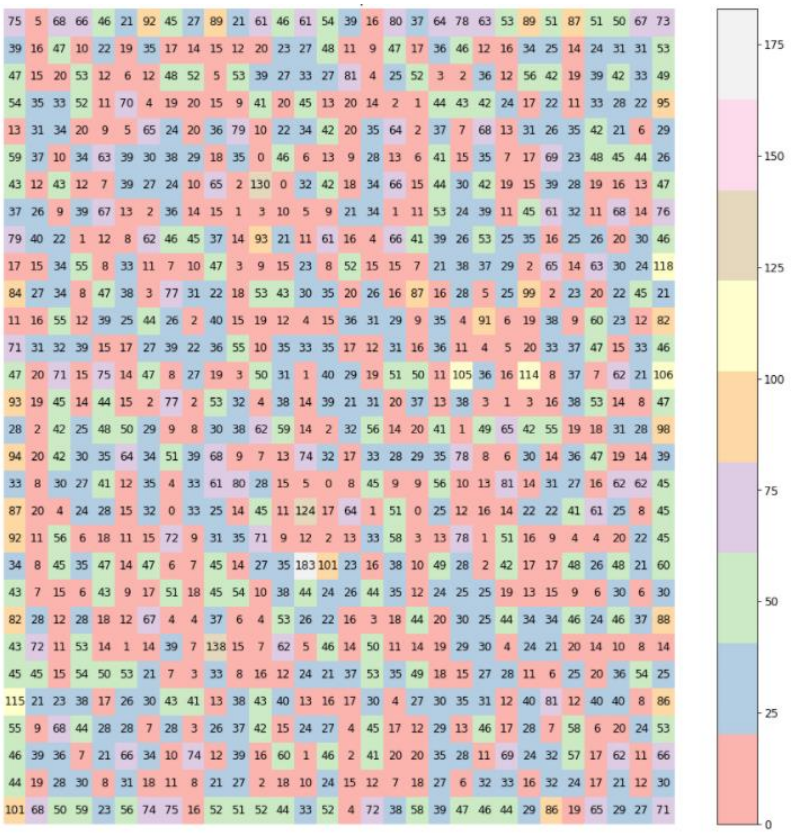
30x30

U-Matrix

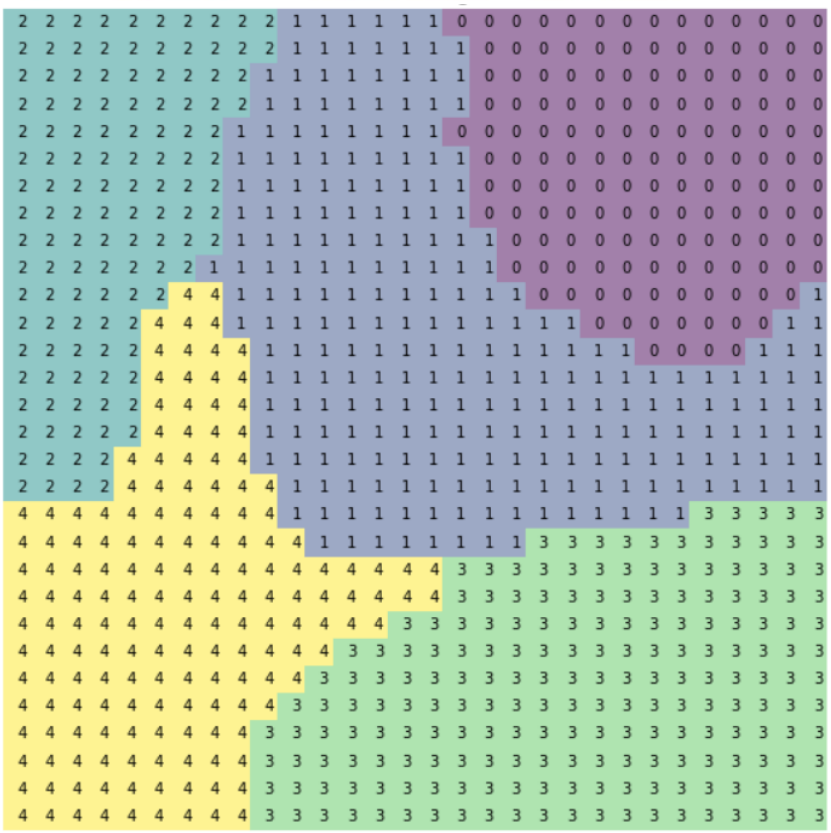


Legend:
0 – Physics,
1 – Medicine,
2 – Engineering,
3 – Mathematics,
4 – Computer science

Hitmap

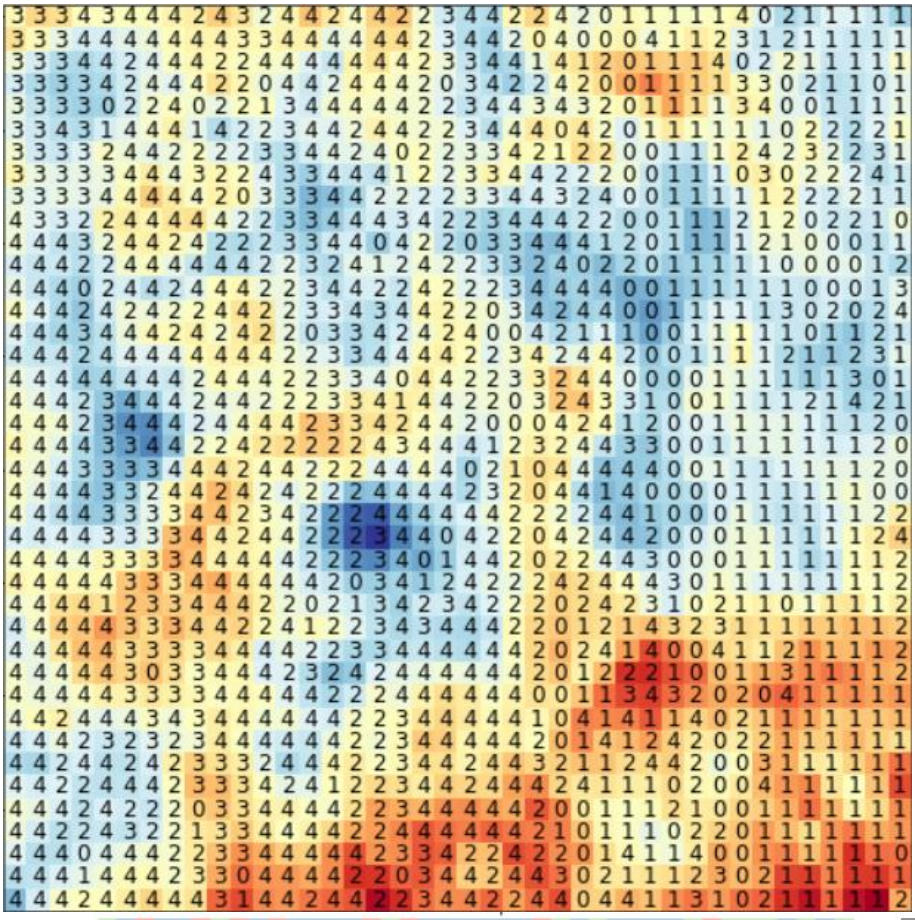


Cluster

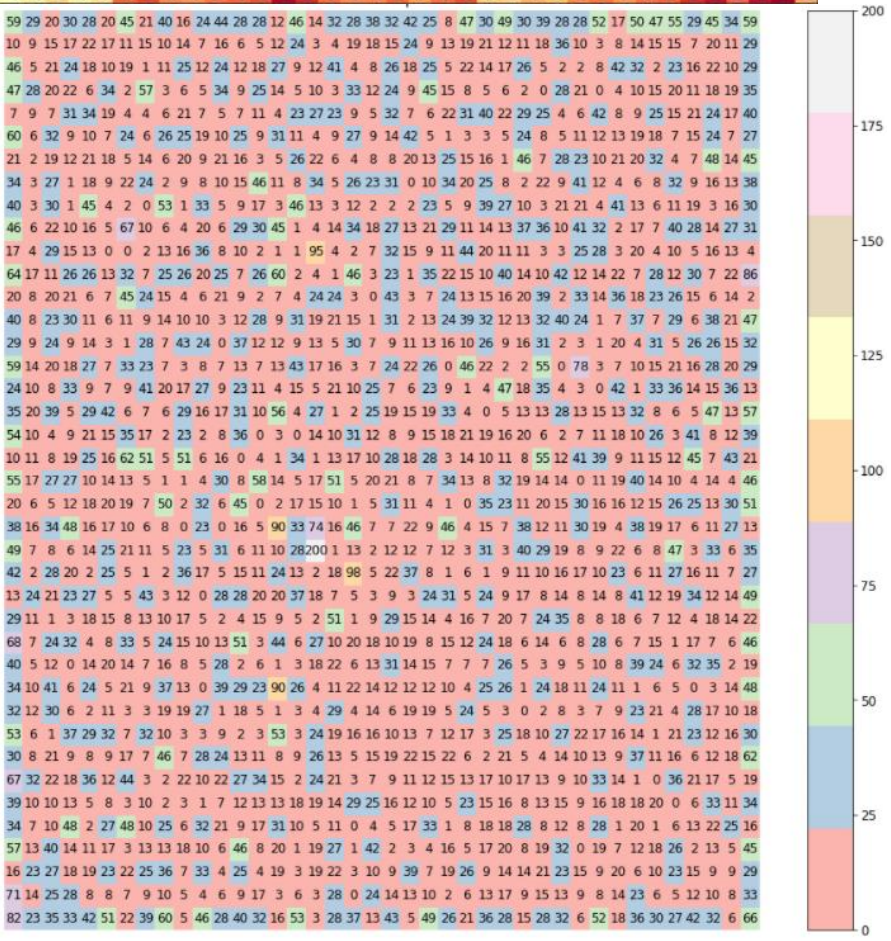


40x40

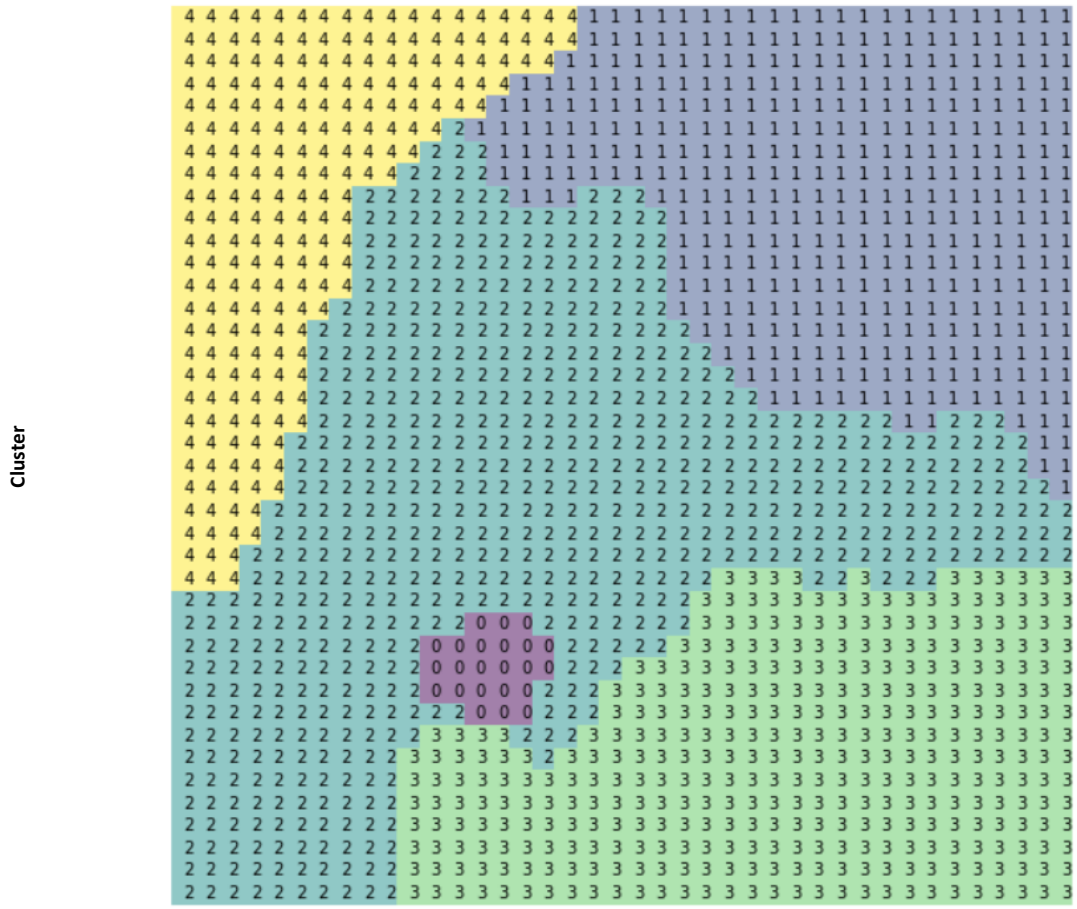
U-Matrix



Hitmap



Legend:
0 – Physics,
1 – Medicine,
2 – Engineering,
3 – Mathematics,
4 – Computer science



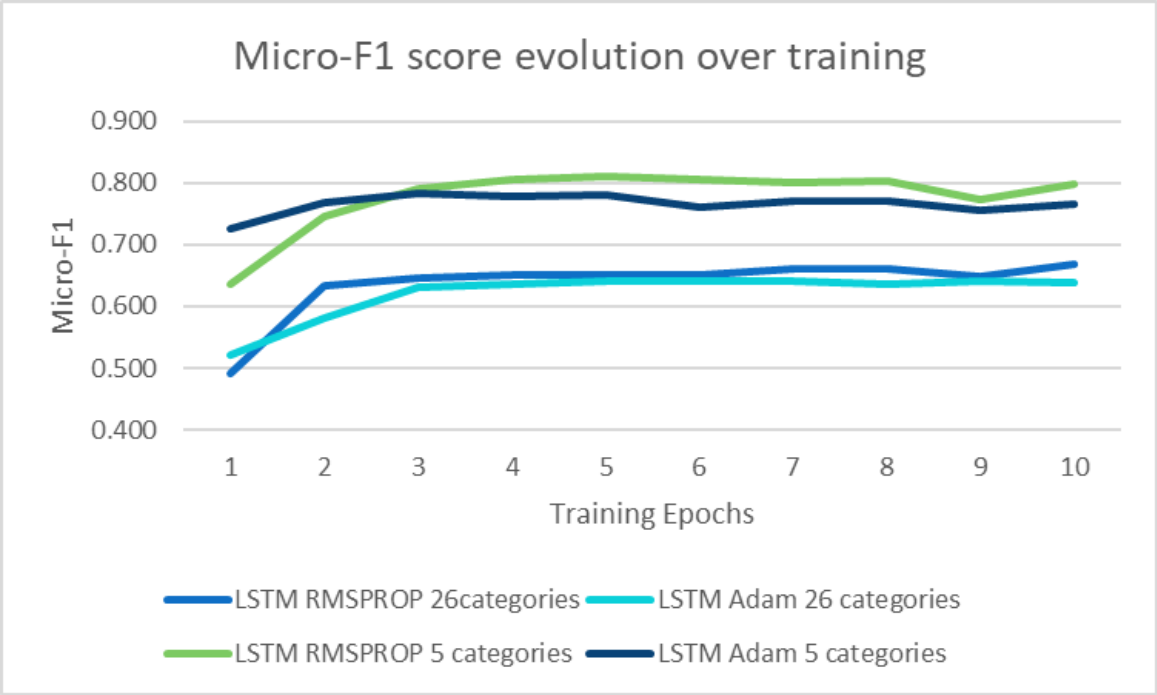
Εικόνα 28. Αναπαραστάσεις με χρήση U-MATRICES, HITMAPS και CLUSTER χάρτες με χρήση του σχήματος στάθμισης TF-IDF για διαφορετικά μεγέθη χάρτη (5 κλάσεις).

2ος γύρος: LSTM, CNN+LSTM – Keras Tokenizer

Στον δεύτερο γύρο πειραμάτων αντικαταστήσαμε το SOM με 2 διαφορετικά DNN, ένα LSTM και ένα CNN + LSTM. Όπως προαναφέρθηκε, εδώ χρησιμοποιήθηκε ο Keras Tokenizer για την αναπαράσταση των πινάκων περιεχομένων των βιβλίων ως διανυσμάτων. Τα αποτελέσματα που πέτυχε ο ταξινομητής χρησιμοποιώντας διαφορετικές αρχιτεκτονικές Νευρωνικών Δικτύων Βαθιάς Μάθησης συνοψίζονται παρακάτω.

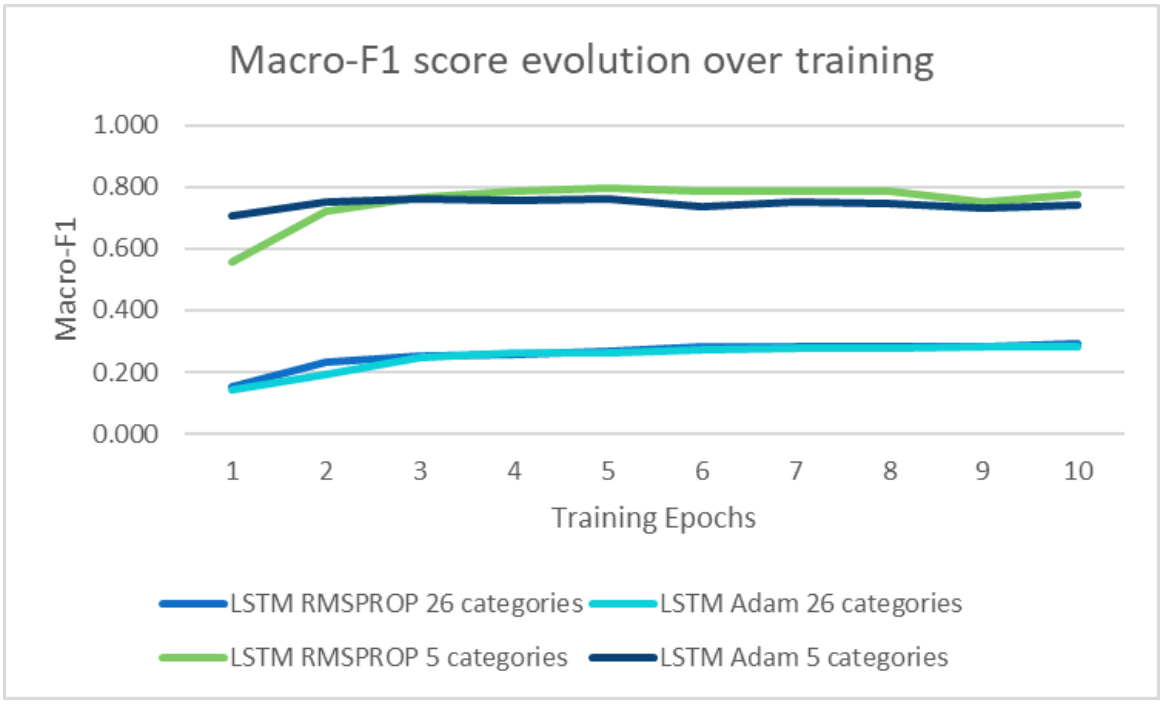
Η Εικόνα 29 απεικονίζει την εξέλιξη της βαθμολογίας F1 Μίκρο-μέσου όρου κατά τη διάρκεια της φάσης εκπαίδευσης. Είναι προφανές ότι τόσο τα πειράματα που πραγματοποιήσαμε χρησιμοποιώντας τις 26 κλάσεις όσο και εκείνα στα οποία χρησιμοποιήθηκαν οι 5 κλάσεις του συνόλου δεδομένων Springer, τα αποτελέσματα που πετυχαίνουμε είναι παρόμοια ανεξάρτητα από τη συνάρτηση βελτιστοποίησης που χρησιμοποιήθηκε (RMSProp ή Adam).

Επιπλέον, η συνάρτηση βελτιστοποίησης RMSProp παράγει καλύτερα αποτελέσματα από την Adam και στις δύο περιπτώσεις (26 και 5 κλάσεις) μετά την 3η εποχή εκπαίδευσης. Σε αυτή την περίπτωση, φαίνεται ότι οι βαθμολογίες δεν παρουσιάζουν διακυμάνσεις καθώς εξελίσσεται η εκπαίδευση. Το ίδιο συμπέρασμα μπορεί να εξαχθεί αν παρατηρήσει κανείς τα αποτελέσματα Μάκρο-μέσου όρου (Εικόνα 30), ωστόσο οι διαφορές μεταξύ των δύο συναρτήσεων βελτιστοποίησης σε αυτήν την περίπτωση είναι αμελητέες.

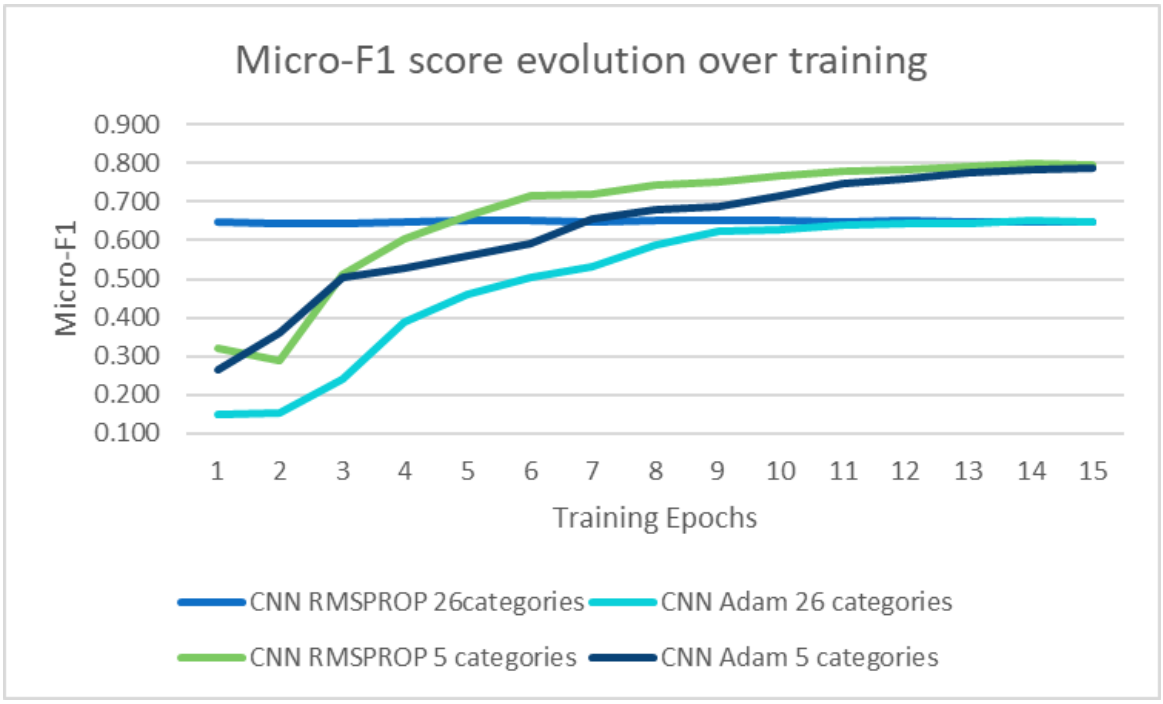


Εικόνα 29. Εξέλιξη βαθμολογίας F1 Μίκρο-μέσου όρου κατά τη διάρκεια των εποχών εκπαίδευσης για το Νευρωνικό Δίκτυο Βαθιάς Μάθησης LSTM με χρήση των συναρτήσεων βελτιστοποίησης RMSPROP και Adam.

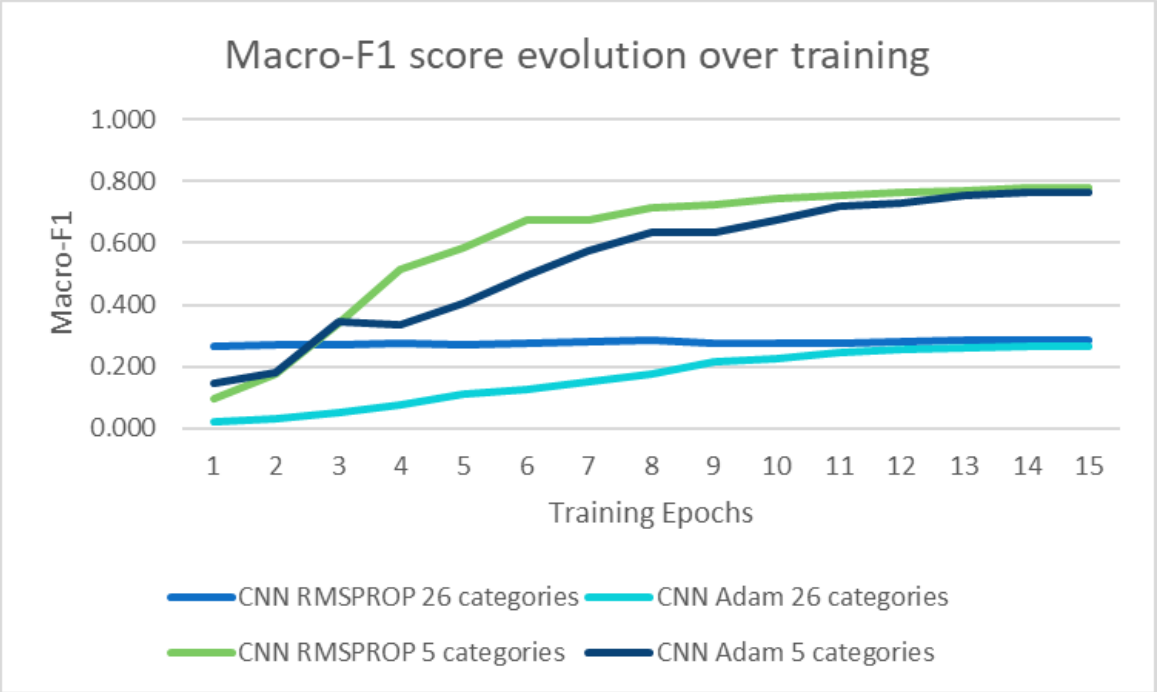
Αντίστοιχα η Εικόνα 31 και η Εικόνα 32 παρουσιάζουν τις βαθμολογίες F1 που πετύχαμε χρησιμοποιώντας την αρχιτεκτονική CNN+LSTM για Μίκρο- και Μάκρο μέσο όρο αντίστοιχα. Σε αυτή την περίπτωση, μπορεί κανείς να παρατηρήσει ότι οι βαθμολογίες F1 είναι χαμηλότερες για τις πρώτες εποχές εκπαίδευσης σε σύγκριση με την αρχιτεκτονική LSTM και τείνουν να σταθεροποιηθούν μετά την $9^{η}$ εποχή για τις 26 κλάσεις και μετά την $11^{η}$ εποχή για τις 5 κλάσεις. Η συνάρτηση βελτιστοποίησης RMSProp φαίνεται να παράγει καλύτερα αποτελέσματα και σε αυτήν την περίπτωση. Είναι επίσης προφανές ότι οι διαφορές στις βαθμολογίες που επιτεύχθηκαν σε κάθε περίοδο εκπαίδευσης μεταξύ των συναρτήσεων βελτιστοποίησης είναι πιο σημαντικές σε αυτήν την περίπτωση, αλλά στο τέλος οι βαθμολογίες τείνουν να συγκλίνουν.



Εικόνα 30. Εξέλιξη βαθμολογίας F1 Μάκρο-μέσου όρου κατά τη διάρκεια των εποχών εκπαίδευσης για το Νευρωνικό Δίκτυο Βαθιάς Μάθησης LSTM με χρήση των συναρτήσεων βελτιστοποίησης RMSPROP και Adam.



Εικόνα 31. Εξέλιξη βαθμολογίας F1 Μίκρο-μέσου όρου κατά τη διάρκεια των εποχών εκπαίδευσης για το Νευρωνικό Δίκτυο Βαθιάς Μάθησης CNN+LSTM με χρήση των συναρτήσεων βελτιστοποίησης RMSPROP και Adam.



Εικόνα 32. Εξέλιξη βαθμολογίας F1 Μάκρο-Μέσου Όρου κατά τη διάρκεια των εποχών εκπαίδευσης για το Νευρωνικό Δίκτυο Βαθιάς Μάθησης CNN+LSTM με χρήση των συναρτήσεων βελτιστοποίησης RMSPROP και Adam.

Οι καλύτερες βαθμολογίες που επιτεύχθηκαν για όλες τις μεθόδους συνοψίζονται στον Πίνακα 26. Για τις 26 κατηγορίες το νευρωνικό δίκτυο LSTM με τη συνάρτηση βελτιστοποίησης RMSProp παράγει τις καλύτερες βαθμολογίες, συγκεκριμένα 67% για Μίκρο- και 29% για Μάκρο-μέσου όρου ακολουθούμενο από το CNN + LSTM απόδοση 2% χαμηλότερα ενώ το νευρωνικό δίκτυο SOM πέτυχε απόδοση 3% χαμηλότερη.

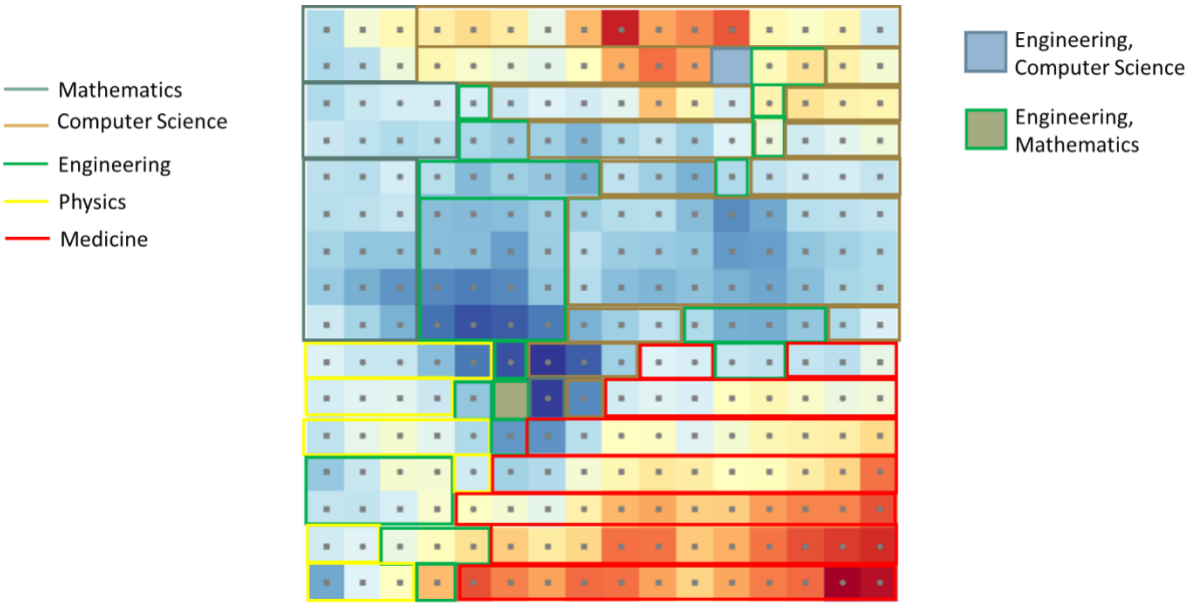
Για τις 5 κατηγορίες, το LSTM και το CNN + LSTM με RMSProp επιτυγχάνουν 80% στο Μίκρο-μέσο όρο, ενώ η συνάρτηση βελτιστοποίησης Adam δίνει 3% και 1% χαμηλότερα αντίστοιχα. Σε αυτήν την περίπτωση, το SOM επιτυγχάνει 5% χαμηλότερα (75%). Για το Μάκρο-μέσο όρο, η καλύτερη μέθοδος είναι το CNN + LSTM με RMSProp (79%) ακολουθούμενο από το LSTM με RMSProp (78%) ενώ η συνάρτηση βελτιστοποίησης Adam επιτυγχάνει 77% και 74% αντίστοιχα με το SOM να ακολουθεί με 73%. Είναι προφανές ότι οι διαφορές δεν είναι σημαντικές.

Πίνακας 26. Καλύτερες βαθμολογίες F1 Μίκρο- και Μάκρο Μέσου Όρου για TF-IDF με SOM, και KERAS TOKENIZER με LSTM και CNN+LSTM.

	Μέθοδος	Ακρίβεια (Precision)	Ανάκληση (Recall)	Μίκρο-F1	Ακρίβεια (Precision)	Ανάκληση (Recall)	Μάκρο-F1
26 κλάσεις	*TF-IDF / SOM*	0.62	0.62	0.62	0.27	0.27	0.26
	Keras / LSTM RMSPROP	**0.67**	**0.67**	**0.67**	**0.31**	**0.29**	**0.29**
	Keras / LSTM Adam	0.64	0.64	0.64	0.28	0.27	0.27
	Keras / CNN+LSTM RMSPROP	0.65	0.65	0.65	0.26	0.28	0.26
	Keras / CNN+LSTM Adam	0.65	0.65	0.65	0.27	0.27	0.27
5 κλάσεις	*TF-IDF / SOM*	0.75	0.75	0.75	0.74	0.73	0.73
	Keras / LSTM RMSPROP	**0.80**	**0.80**	**0.80**	**0.78**	**0.77**	**0.78**
	Keras / LSTM Adam	0.77	0.77	0.77	0.74	0.75	0.74

	Keras / CNN+LSTM RMSPROP	**0.80**	**0.80**	**0.80**	**0.79**	**0.77**	**0.78**
	Keras / CNN+LSTM Adam	0.79	0.79	0.79	0.77	0.77	0.77

Όπως αναφέρθηκε και στην εισαγωγή του κεφαλαίου η προτεινόμενη προσέγγιση θα μπορούσε να χρησιμοποιηθεί αποτελεσματικά για τη σύσταση συγγραμμάτων με σκοπό την υποστήριξη καθηγητών και φοιτητών στον εντοπισμό σχετικών πηγών. Πιο συγκεκριμένα, η χρήση μιας πιο λεπτομερούς περιγραφής του θέματος που αναζητά ο χρήστης (π.χ. η περιγραφή ενός μαθήματος ή ο πίνακας περιεχομένων ενός γνωστού βιβλίου) μπορεί να οδηγήσει σε πιο ακριβή αποτελέσματα. Ενδεικτικά παραδείγματα από αυτό το σενάριο χρήσης δίνονται παρακάτω. Αρχικά βλέπουμε τον εκπαιδευμένο χάρτη καθώς επίσης και τις ετικέτες που έχουν ανατεθεί σε κάθε κόμβο μετά τη φάση της εκπαίδευσης με χρήση του αλγορίθμου Majority Voting.



Εικόνα 33. Εκπαιδευμένος χάρτης όπου ξεχωρίζουν οι ετικέτες που έχουν ανατεθεί σε κάθε κόμβο με βάση τα χρώματα του υπομνήματος.

Από τον παραπάνω χάρτη μπορούν να βγουν συγκεκριμένα συμπεράσματα. Αρχικά, οι ΄δυο κλάσεις ‘Mathematics’ και ‘Medicine’ φαίνεται ότι είναι εντελώς διακριτές επάνω στον χάρτη, η κλάση ‘Mathematics’ καταλαμβάνει το επάνω αριστερό κομμάτι του χάρτη ενώ η κλάση ‘Medicine’ το κάτω δεξί. Αυτό που επίσης, μπορούμε να παρατηρήσουμε είναι ότι υπάρχει μια επικάλυψη ανάμεσα στις κλάσεις ‘Engineering’ και ‘Computer Science’ οι οποίες καταλαμβάνουν το επάνω δεξί κομμάτι του χάρτη. Αυτό είναι αναμενόμενο καθώς στην πραγματικότητα οι κλάσεις αυτές είναι αρκετά κοντινές και πολλές φορές δεν είναι εύκολα διακριτή η διαφοροποίηση μεταξύ τους. Τέλος, ένα ακόμα στοιχείο που αξίζει να επισημανθεί είναι ότι στον παραπάνω χάρτη υπάρχουν 2 κόμβοι στους οποίους υπάρχουν 2 ετικέτες ταυτόχρονα ακόμα και μετά την εφαρμογή του αλγορίθμου Majority Voting. Για το λόγο αυτό σημειώνονται επάνω στον χάρτη με διαφορετική χρωματική απεικόνιση σε σχέση με τους υπόλοιπους. Στον ένα κόμβο έχουν ανατεθεί οι ετικέτες ‘Engineering’ και ‘Computer Science’ ενώ στον άλλο οι ετικέτες ‘Engineering’ και ‘Mathematics’. Αυτό σημαίνει ότι το άθροισμα των δειγμάτων εκπαίδευσης από κάθε κλάση που ανατέθηκαν στον κόμβο είναι ακριβώς ίδια οπότε και ο αλγόριθμος δεν μπορεί να επιλέξει μια από τις 2 ετικέτες ως επικρατέστερη.

Στη συνέχεια παραθέτουμε ενδεικτικά παραδείγματα χρησιμοποιώντας τον εκπαιδευμένο χάρτη που παρουσιάστηκε παραπάνω και δίνοντας ως είσοδο πίνακες περιεχομένων βιβλίων που δεν χρησιμοποιήθηκαν κατά τη διάρκεια της φάσης εκπαίδευσης για να ελεγχθεί κατά πόσο μπορούν να ταξινομηθούν σωστά στην πράξη.

	Book_id	Title	Predicted BMU	BMU Label	Initial Label
1	99532726	E-Democracy Security Privacy and Trust in a Digital World	141	Engineering	Computer science
2	99556866	Exciting Interdisciplinary Physics	178	Physics	Physics
3	99537162	Machine Learning in Medical Imaging	14	Computer science	Computer science
4	99507449	Innovations in Computer Science and Engineering	30	Computer science	Engineering
5	99549359	Generalized Measure Theory	81	Mathematics	Mathematics
6	99564278	Nutrition and Gastrointestinal Disease	237	Medicine	Medicine
7	99558222	Dynamic Response of Linear Mechanical Systems	83	Engineering	Engineering
9	99539153	Experimental and Theoretical Investigations of Steel-Fibrous Concrete	133	Engineering	Engineering
10	99504750	Microbial Biomass Process Technologies and Management	154	Medicine	Medicine
11	99511433	Modeling Thermodynamic Distance, Curvature and Fluctuations	176	Physics	Engineering
12	99530893	Continuum Thermomechanics	192	Engineering	Physics
13	99546245	Temporal Information Processing Technology and Its Application	89	Computer science	Computer science
14	99519735	Magnetic Fields in Diffuse Media	178	Physics	Physics
15	99516165	Hot Carrier Degradation in Semiconductor Devices	101	Engineering	Engineering
16	99526147	Transition Towards Energy Efficient Machine Tools	133	Engineering	Engineering
17	99549034	Minimally Invasive Cancer Management	252	Medicine	Medicine
18	99562022	Advances in Electrical Engineering and Electrical Machines	101	Engineering	Engineering
19	99501204	The Rise of Radio Astronomy in the Netherlands	163	Physics	Physics

20	99513068	Textile Materials for Lightweight Constructions	227	Engineering	Engineering
21	99558598	An Introduction to the Theory of Point Processes	65	Mathematics	Mathematics
22	99539868	Constitutive Relations under Impact Loadings	209	Engineering	Engineering
23	99536779	Information Theoretic Security	9	Computer science	Computer science
24	99538621	Self-Normalized Processes	65	Mathematics	Mathematics
25	99507598	Generalized Metric Spaces and Mappings	81	Mathematics	Mathematics

Στη συνέχεια δίνουμε παραδείγματα στα οποία με βάση την BMU που αντιστοιχίστηκε κάθε βιβλίο του παραπάνω πίνακα, παρατίθενται επίσης βιβλία τα οποία έχουν ανατεθεί στην ίδια BMU.

	Book_id	Title	Predicted BMU	Other books in BMU
1	99532726	E-Democracy Security Privacy and Trust in a Digital World	141	99543623 - European Identity through Space 99538948 - Critical Information Infrastructures 99544685 - Yearbook on Space Policy 2006/2007 99536482 - The New Space Race 99553349 - Maritime Governance and Policy-Making 99517327 - Cybersecurity in Israel 99504438 - Blockchain Enabled Applications 99542847 - Yearbook on Space Policy 2007/2008 99547210 - Yearbook on Space Policy 2009/2010 99518993 - When China Goes to the Moon 99516600 - European Autonomy in Space
2	99556866	Exciting Interdisciplinary Physics	178	99523882 - Kinetic Simulations of Ion Transport in Fusion Devices 99560042 - Helioseismology Asteroseismology and MHD Connections 99549466 - The Electromagnetic Spectrum of Neutron Stars 99545996 - Light and Light Sources 99517415 - Asteroseismology of Stellar Populations in the Milky Way 99556020 - Atomic Processes in Basic and Applied Physics 99510024 - Multi-scale Structure Formation and Dynamics in Cosmic Plasmas 99553099 - The Interstellar Medium 99553399 - Plasma Astrophysics Part I 99542250 - Theory of Particle and Cluster Emission 99550971 - From Dust to Stars
3	99507449	Innovations in Computer	30	99556734 - Recent Progress in Data Engineering and Internet Technology

		Science and Engineering		99507375 - Proceedings of Fifth International Conference on Soft Computing for Problem Solving
				99554654 - Emerging Trends in Computing and Communication
				99557481 - Computational Intelligence in Multimedia Processing: Recent Advances
				99525062 - Software and Network Engineering
				99519468 - Proceedings of the 4th International Conference on Computer Engineering and Networks
				99507455 - Information Systems Design and Intelligent Applications
				99520002 - Intelligent Computing and Applications
				99546553 - Applications and Innovations in Intelligent Systems XV
				99511055 - Intelligent Systems and Applications
				99500183 - Advanced Computing and Systems for Security
4	99549034	Minimally Invasive Cancer Management	252	99554274 - Diagnostic Endosonography
				99512361 - Endoscopic Diagnosis of Superficial Gastric Cancer for ESD
				99530076 - Inflammation and Gastrointestinal Cancers
				99561692 - Pancreatic Cancer
				99519631 - Atlas of Head and Neck Cancer Surgery
				99507349 - Stem Cells, Pre-neoplasia, and Early Cancer of the Upper Gastrointestinal Tract
				99521876 - Atlas of Robotic Prostatectomy
				99557845 - Minimally Invasive Surgical Oncology
				99511477 - Multimodal Treatment of Recurrent Pelvic Colorectal Cancer
				99513461 - Laparoscopic Surgery
				99505512 - Adenocarcinoma of the Esophagogastric Junction
5	99549359	Generalized Measure Theory	81	99511376 - Diophantine Analysis
				99532152 - Eigenvalues Embeddings and Generalised Trigonometric Functions
				99523724 - Infinite-Horizon Optimal Control in the Discrete-Time Framework
				99512325 - Methods of Solving Sequence and Series Problems
				99513835 - Pi: The Next Generation
				99527080 - Fixed Point Theory in Distance Spaces
				99511462 - Approximation Methods in Probability Theory
				99556577 - Measure and Integration
				99504175 - Recent Developments in Fractals and Related Fields
				99523334 - Discrete Spectral Synthesis and Its Applications

				99544812 - Variational Topological and Partial Order Methods with Their Applications

4.7. Συμπεράσματα

Σε αυτό το Κεφάλαιο της παρούσας Διατριβής, ερευνήσαμε τη χρήση ενός νευρωνικού δικτύου μη εποπτευόμενης ΜΜ (SOM) και δύο αρχιτεκτονικών DNN, συγκεκριμένα ενός LSTM και ενός CNN+LSTM, με στόχο να εκτιμηθεί η ικανότητα ταξινόμησης των μεθόδων αυτών σε ένα σενάριο ταξινόμησης πολλαπλών κλάσεων. Η διαδικασία επιλογής χαρακτηριστικών σχεδιάστηκε έτσι ώστε να είναι όσο το δυνατόν πιο απλή και γενική, για να διασφαλιστεί η εφαρμογή της σε άλλα σύνολα δεδομένων με μικρές τροποποιήσεις. Η διαδικασία της κατασκευής διανυσμάτων για την αναπαράσταση των Πινάκων Περιεχομένων (ToC) αξιοποιεί τις πληροφορίες που εξάγονται από τους τελευταίους, χρησιμοποιώντας το σχήμα στάθμισης TF-IDF, στην περίπτωση του SOM και τον Keras Tokenizer, στην περίπτωση των DNN αρχιτεκτονικών. Συμπερασματικά, στο Κεφάλαιο αυτό δείξαμε ότι με τη χρήση της προτεινόμενης διαδικασίας επιλογής χαρακτηριστικών επιτυγχάνεται η αποδοτική αναπαράσταση των Πινάκων Περιεχομένων των βιβλίων καθώς η μέθοδος μπορεί να συλλάβει/ συγκεντρώσει τα πιο σημαντικά χαρακτηριστικά του συνόλου δεδομένων, διατηρώντας παράλληλα το μέγεθος των διανυσμάτων σχετικά μικρό, βελτιώνοντας έτσι την αποτελεσματικότητα της ταξινόμησης.

Επιπλέον, χρησιμοποιήθηκε ο αλγόριθμος INterSECT που προτάθηκε στο προηγούμενο Κεφάλαιο για την επιλογή ετικετών, με σκοπό να ερευνηθεί κατά πόσο αυτός μπορεί να χρησιμοποιηθεί με εξίσου καλά αποτελέσματα σε ένα σενάριο ταξινόμησης πολλαπλών κλάσεων αντί για ταξινόμηση πολλαπλής ετικέτας. Ο αλγόριθμος Majority Voting εφαρμόστηκε επίσης ως εναλλακτική μέθοδος επιλογής ετικέτας για κάθε κόμβο στην περίπτωση του SOM. Από τη σύγκριση των αποτελεσμάτων για τις δύο εναλλακτικές μεθόδους επιλογής ετικέτας καταλήξαμε πως ο αλγόριθμος Majority Voting υπερτερεί του INterSECT σε όλες τις περιπτώσεις με βάση τα πειράματα που διεξήχθησαν, καθιστώντας έτσι τον INterSECT πιο κατάλληλο για σενάρια ταξινόμησης πολλαπλής ετικέτας. Η προτεινόμενη προσέγγιση επαληθεύθηκε χρησιμοποιώντας διαφορετικές ρυθμίσεις διαμόρφωσης, για ένα σύνολο δεδομένων πολλαπλών κλάσεων, που περιλάμβανε περίπου 56000 Πίνακες Περιεχομένων ηλεκτρονικών βιβλίων (ebooks) του εκδοτικού οίκου Springer. Σε σύγκριση με την προσέγγιση που παρουσιάστηκε στο Κεφάλαιο 3 της παρούσας Διατριβής μπορούμε ασφαλώς να συμπεράνουμε ότι η προσέγγιση γενικεύεται, καθώς καταφέραμε να την εφαρμόσουμε εκ νέου σε ένα διαφορετικό σύνολο δεδομένων με μικρές αλλαγές. Εκμεταλλευόμενοι την ομοιότητα των βιβλίων που ανήκουν στην ίδια συστάδα, μπορέσαμε να προσδιορίσουμε παρόμοια βιβλία για κάθε κατηγορία που μπορούν να προταθούν ως εναλλακτικές λύσεις κατά την επιλογή βιβλίων για κάθε μάθημα. Σε αυτό το Κεφάλαιο επεκτείναμε την προσέγγιση του Κεφαλαίου 3, καθώς χρησιμοποιήσαμε επίσης δύο διαφορετικές αρχιτεκτονικές Νευρωνικών Δικτύων Βαθιάς Μάθησης ώστε να εκτιμήσουμε την απόδοση ταξινόμησής τους και να συγκρίνουμε τα αποτελέσματα με αυτά που επιτευχθήκαν όταν χρησιμοποιήσαμε το SOM.

Επιπλέον, δοκιμάσαμε δύο διαφορετικές συναρτήσεις βελτιστοποίησης για τα Νευρωνικά Δίκτυα Βαθιάς Μάθησης, συγκεκριμένα τις συναρτήσεις RMSProp και Adam. Τα αποτελέσματα υποδηλώνουν ότι και οι δύο αρχιτεκτονικές DNN επιτυγχάνουν καλύτερα αποτελέσματα κατά 5% σε σύγκριση με SOM. Ωστόσο, τα Νευρωνικά Δίκτυα Βαθιάς Μάθησης χρειάζονται πολύ περισσότερη υπολογιστική ισχύ και μνήμη, αλλά είναι ταχύτερα από το SOM όταν εκτελούνται σε GPU (Graphical

Processing Unit). Τα αποτελέσματα δείχνουν ότι τόσο η αρχιτεκτονική LSTM όσο και η CNN + LSTM αποδείχθηκαν εξίσου αποτελεσματικές στην ταξινόμηση ενός συνόλου δεδομένων πολλαπλών κλάσεων χρησιμοποιώντας πληροφορίες που έχουν εξαχθεί από τους Πίνακες Περιεχομένων των ηλεκτρονικών βιβλίων με ελάχιστη παραμετροποίηση. Επιπλέον, με δεδομένο ότι η αρχιτεκτονική CNN + LSTM είναι πιο περίπλοκη, και επομένως υπολογιστικά πιο ακριβή, σε σύγκριση με το αμφίδρομο LSTM και δεδομένου ότι και οι δύο αρχιτεκτονικές έχουν παρόμοια απόδοση, θεωρούμε την τελευταία καλύτερη.

Εφαρμόζοντας τη μεθοδολογία που παρουσιάζεται σε αυτό το Κεφάλαιο της παρούσας Διατριβής και αξιοποιώντας τις πληροφορίες από τους πίνακες περιεχομένων των βιβλίων, μπορούμε να αποκτήσουμε μια καλύτερη εικόνα του κάθε βιβλίου, ενώ είμαστε σε θέση να προσδιορίσουμε ποια βιβλία πραγματεύονται παρόμοια θέματα, παρόλο που ενδέχεται να μην μοιράζονται τις ίδιες λέξεις-κλειδιά στους πίνακες περιεχομένων τους ή ακόμα και την ίδια θεματική ετικέτα. Θα πρέπει να επισημανθεί ότι σε αυτό το πρόβλημα δεν περιμένουμε ακρίβεια ταξινόμησης 100%, καθώς οι ετικέτες που υπάρχουν για κάθε δείγμα εκχωρούνται από βιβλιοθηκονόμους και έτσι η διαδικασία επιλογής ετικέτας για κάθε βιβλίο είναι υποκειμενική. Για παράδειγμα, διαφορετικοί άνθρωποι μπορεί να δώσουν στο ίδιο βιβλίο διαφορετική θεματική ετικέτα. Επίσης, υπάρχουν στενά σχετιζόμενες κατηγορίες, για παράδειγμα η ετικέτα ‘Computer science’ και ‘Engineering’, και παρόλο που το σύστημα μπορεί να αποδώσει σε ένα βιβλίο την ετικέτα ‘Engineering’, αυτό μπορεί να ανήκει στην κατηγορία ‘Computer science’ και αντίστροφα.

4.8. Μελλοντικές επεκτάσεις

Η μεθοδολογία που περιγράφεται σε αυτό το Κεφάλαιο μπορεί να επεκταθεί περαιτέρω με διάφορους τρόπους. Μια πιθανή επέκταση αυτής θα ήταν η χρήση της ταξινόμησης για τον εντοπισμό της ιεραρχίας εντός της ίδιας της κατηγορίας χρησιμοποιώντας πιο συγκεκριμένες κατηγορίες (υπο-κατηγορίες) ως ετικέτες και επανεκπαιδεύοντας τα δίκτυα. Επιπλέον, τα Siamese Νευρωνικά Δίκτυα ή η μέθοδος Contrastive Supervised Learning θα μπορούσαν επίσης να αξιοποιηθούν για την αντιμετώπιση του προβλήματος της ανισορροπίας δειγμάτων μεταξύ των κλάσεων.

Επιπλέον, μπορούν να χρησιμοποιηθούν προ εκπαιδευμένα γλωσσικά μοντέλα όπως Transformers, BERT, XLNet και GPT-2. Μέσω αυτής της προσέγγισης, μπορούμε να αξιοποιήσουμε τις ήδη εκπαιδευμένες ενσωματώσεις διανυσμάτων αντί να διεξάγουμε δαπανηρές εργασίες εκπαίδευσης από το μηδέν. Τα δίκτυα αυτού του τύπου θα μπορούσαν να χρησιμοποιηθούν ως εναλλακτική λύση στα Siamese Νευρωνικά Δίκτυα ή στη μέθοδο Contrastive Supervisеded Learning για την αντιμετώπιση του προβλήματος της ανισορροπίας των δειγμάτων μεταξύ των διαφορετικών κλάσεων όπως αναφέρθηκε και παραπάνω, είτε στη συλλογή Springer ή και σε άλλα σύνολα δεδομένων πραγματικών εφαρμογών. Αυτό θα μπορούσε να πραγματοποιηθεί με την εφαρμογή της μεταφοράς μάθησης (Transfer Learning) από εποπτευόμενα δεδομένα. Από όσο γνωρίζουμε, οι Transformers δεν έχουν ακόμη εφαρμοστεί σε συστήματα προτάσεων περιεχομένου. Ως εκ τούτου, αυτή θα ήταν μια ενδιαφέρουσα επέκταση για την αξιολόγηση της δυνατότητας εφαρμογής τους στην ταξινόμηση πολλαπλών κλάσεων, χρησιμοποιώντας πληροφορίες από τους πίνακες περιεχομένων των βιβλίων. Τέλος, η προτεινόμενη μεθοδολογία μπορεί να εφαρμοστεί σε άλλα σύνολα δεδομένων που είναι διαθέσιμα μέσω της συνδρομής των Ελληνικών Ακαδημαϊκών Βιβλιοθηκών για να βοηθήσουν καθηγητές και φοιτητές να αναζητήσουν παρόμοιο περιεχόμενο.

5. Πύλες Διασυνδεδεμένων Δεδομένων με χρήση τεχνολογιών Σημασιολογικού Ιστού: Επισκόπηση και μια προτεινόμενη αρχιτεκτονική βασισμένη σε Συστήματα Διαχείρισης Περιεχομένου

5.1. Περίληψη

Στο συγκεκριμένο Κεφάλαιο γίνεται μια αναφορά σε συναφή ερευνητικά θέματα που έχουν εξεταστεί στο πλαίσιο της παρούσας Διατριβής. Αυτά αφορούν στο σχεδιασμό μιας Πύλης Διασυνδεμένων Δεδομένων με χρήση των τεχνολογιών Σημασιολογικού Ιστού και ενός Συστήματος Διαχείρισης Περιεχομένου (Content Management System - CMS). Σκοπός της προτεινόμενης Πύλης είναι η σημασιολογική επεξεργασία, διασύνδεση και αναζήτηση δεδομένων από διαφορετικές πηγές. Η προτεινόμενη προσέγγιση ενσωματώνει πολλαπλά κανάλια εισόδου για διαφορετικές ομάδες χρηστών αλλά και κίνητρα για επαναχρησιμοποίηση, αναφορά, σχολιασμό και κοινή χρήση του περιεχομένου.

Όπως αναφέρθηκε στα προηγούμενα κεφάλαια, οι λέξεις-κλειδιά και οι λίστες όρων που προκύπτουν από τις μεθόδους αυτόματης εξαγωγής πληροφορίας ταξινόμησης, μπορούν να χρησιμοποιηθούν στην κατηγοριοποίηση κειμένων, αλλά και στην ανάκτηση πληροφοριών που βασίζεται στο περιεχόμενο και στη θεματική αναζήτηση. Οι δύο τελευταίες εφαρμογές κρίνονται άκρως χρήσιμες στην περίπτωση της προτεινόμενης Πύλης Διασυνδεδεμένων Δεδομένων, καθώς η πληροφορία που εξάγεται θα επιτρέψει την αποτελεσματικότερη αναζήτηση και ανάκτηση πληροφοριών που βασίζονται στο περιεχόμενο. Η υλοποίηση της εφαρμογής βασίζεται στο Σύστημα Διαχείρισης Περιεχομένου Drupal, όμως οποιοδήποτε άλλο CMS μπορεί εναλλακτικά να χρησιμοποιηθεί.

5.2. Εισαγωγή

Ο Παγκόσμιος Ιστός (World Wide Web-WWW) στην εξέλιξή του άλλαξε ριζικά τον τρόπο διάθεσης της γνώσης, διευκολύνοντας τη δημοσίευση και την πρόσβαση σε αρχεία και δεδομένα [76], [121]. Ένα καινοτόμο βήμα προς αυτή την κατεύθυνση ήταν η εισαγωγή και χρήση υπερσυνδέσμων που επιτρέπουν στους χρήστες να ‘περιηγούνται’ μέσω του Διαδικτύου στα δημοσιευμένα κείμενα και δεδομένα χρησιμοποιώντας κατάλληλα προγράμματα περιήγησης. Επιπλέον, η ανάπτυξη και η εκτεταμένη χρήση των μηχανών αναζήτησης επέτρεψε την ευρετηρίαση των αρχείων και τη χρήση της ανάλυσης δομής των συνδέσμων για την εξαγωγή σχέσεων μεταξύ τους, με βάση τα ερωτήματα των χρηστών. Ο τελευταίος ήταν ο κύριος λόγος για την ανάπτυξη και την ευρεία αποδοχή του Ιστού. Ωστόσο, στη συντριπτική του πλειονότητα τα δεδομένα που δημοσιεύονται στον Παγκόσμιο Ιστό παρέχονται σε περίπλοκες και ιδιοταγείς μορφές, θυσιάζοντας έτσι την υποκείμενη δομή και τη σημασιολογία τους. Το γεγονός αυτό, σε συνδυασμό με τον τεράστιο όγκο του δημοσιευμένου υλικού, καθιστά επιτακτική την ανάγκη για νέους τρόπους οργάνωσης και αναζήτησης της πληροφορίας, σε βάση περισσότερο εννοιολογική και λιγότερο λεξικολογική/συντακτική. Δύο τεχνολογίες με σημαντική συνδρομή στην κατεύθυνση αυτή είναι τα **Ανοικτά Διασυνδεδεμένα Δεδομένα** (Linked Open Data - LOD) και οι **Οντολογίες** (Ontologies).

Τα LOD αποτελούν ένα σύνολο πρακτικών και κανόνων για τη δημοσίευση και διασύνδεση δομημένων δεδομένων στον Ιστό [1]. Τα LOD είναι δεδομένα που δημοσιεύονται στον Ιστό σε μορφή

αναγνώσιμη από μηχανή, η οποία υποδηλώνει τη σημασιολογία τους, ενώ συγχρόνως αλληλοσυνδέονται με άλλα σύνολα δεδομένων που διατίθενται σε αυτόν. Η δημοσίευση των δεδομένων στον Ιστό ως LOD έχει οδηγήσει στη δημιουργία ποικίλων υπηρεσιών προστιθέμενης αξίας. Απαραίτητη προϋπόθεση για την επιτυχή δημοσίευση των δεδομένων αποτελεί η τήρηση των αρχών που έχουν καθοριστεί ώστε αυτά να καταστούν διαθέσιμα ως τμήμα ενός συνολικού χώρου δεδομένων.

Ακριβώς όπως τα LOD, η χρήση οντολογιών στον Ιστό είναι μια ιδέα που υιοθετήθηκε πρόσφατα, οδηγώντας προς το όραμα του Σημασιολογικού Ιστού. Οι οντολογίες είναι δομημένα σύνολα στοιχείων μεταδεδομένων (metadata element sets) που παρέχουν την περιγραφή μιας θεματικής περιοχής ή ενός πεδίου εφαρμογής. Στόχος των οντολογιών είναι να μειωθεί η απόσταση μεταξύ του τρόπου που οι άνθρωποι και οι μηχανές διαχειρίζονται την πληροφορία. Η ανάπτυξη των οντολογιών γίνεται από ειδικούς στο εκάστοτε πεδίο και στοχεύουν στην οργανωμένη περιγραφή της υφιστάμενης γνώσης αλλά και στην ανακάλυψη και ανταλλαγή νέας γνώσης.

Τόσο οι τεχνολογίες του Σημασιολογικού Ιστού όσο και των Διασυνδεδεμένων Δεδομένων ενσωματώνονται σταδιακά στις εφαρμογές διεπαφής χρήστη [122], όπως σε ένα CMS [123], το οποίο χρησιμοποιείται για την ανάπτυξη διαφόρων ειδών πυλών που συγκεντρώνουν ποικίλους τύπους δεδομένων. Η τρέχουσα τάση στα LOD είναι η σύνδεση διαφορετικών πηγών δεδομένων που είναι διαθέσιμα σε ολόκληρο τον Ιστό [124]. Το πεδίο εφαρμογής μιας τέτοιας προσέγγισης είναι να προσφέρει σύνολα δεδομένων και να επιτρέψει σχετικές αναλύσεις σε διαφορετικά είδη χρηστών, όπως ερευνητές, επιχειρηματίες, δημοσιογράφους και πολίτες [125]. Αυτός ο στόχος μπορεί να επιτευχθεί με τη δημιουργία βιώσιμων ροών δεδομένων από ανοικτές πηγές. Ως επόμενο βήμα, κάποιος μπορεί να οραματιστεί τη σύνδεση αυτών των διαφορετικών πηγών δεδομένων, προκειμένου να παράσχει μηχανισμούς σημασιολογικής αναζήτησης και συμπερασμάτων στο σύνολο των πηγών αυτών. Απαραίτητη προϋπόθεση για την επιτυχία των μηχανισμών σημασιολογικής αναζήτησης και συμπερασμάτων είναι η ύπαρξη κατάλληλων μηχανισμών αυτόματης εξαγωγής πληροφορίας ταξινόμησης για τον εμπλουτισμό των μεταδεδομένων των πηγών που θα διασυνδεθούν.

Όλες οι προαναφερθείσες ανάγκες μπορούν να καλυφθούν επαρκώς από τεχνολογίες και εργαλεία Σημασιολογικού Ιστού, τα οποία θα παράσχουν ένα σημείο αναφοράς για τη δημιουργία μιας **Πύλης Διασυνδεδεμένων Δεδομένων** που θα επιτρέπει τη συσσώρευση δεδομένων, καθώς και τη δυνατότητα σημασιολογικής αναζήτησης και συμπερασμού για διαφορετικές πηγές προερχόμενες από διαφορετικά πεδία γνώσης. Στο συγκεκριμένο Κεφάλαιο της Διατριβής παρουσιάζεται η αρχιτεκτονική για την κατασκευή μιας τέτοιας σημασιολογική Πύλης με εξειδικευμένο στόχο τη φιλοξενία του περιεχομένου μιας Ψηφιακής Βιβλιοθήκης. Η προτεινόμενη Πύλη χρησιμοποιεί ένα δημοφιλές CMS, το Drupal, το οποίο επιτρέπει στους χρήστες να οργανώνουν, να διαχειρίζονται και να δημοσιεύουν εύκολα το περιεχόμενό τους, παρέχοντας υψηλό επίπεδο παραμετροποίησης. Πρέπει όμως να τονιστεί ότι η εφαρμογή της αρχιτεκτονικής δεν περιορίζεται στο Drupal αλλά μπορεί να επεκταθεί σε οποιοδήποτε Σύστημα Διαχείρισης Περιεχομένου, όπως το Joomla, το WordPress ή με χρήση του Elasticsearch Stack (ELK)[30].

Η διάρθρωση του κεφαλαίου αυτού έχει ως εξής: Αρχικά, γίνεται μια σύντομη αναφορά στο όραμα του Σημασιολογικού Ιστού και στα Ανοικτά Διασυνδεδεμένα Δεδομένα. Ακολουθεί μια

[30] https://www.elastic.co/what-is/elk-stack

επισκόπηση υφιστάμενων πρακτικών και παραδειγμάτων διάθεσης των δεδομένων ενός Ιδρύματος ή ενός Οργανισμού ως LOD. Τέλος, παρουσιάζεται η πρωτότυπη αρχιτεκτονική και η αναλυτική μεθοδολογία υλοποίησης της προτεινόμενης Πύλης.

5.3. Σημασιολογικός Ιστός, Διασυνδεδεμένα Ανοικτά Δεδομένα και η χρήση Οντολογιών

Το 'όραμα' του Σημασιολογικού Ιστού, όπως αποκαλείται, είναι να μπορεί να αναπαραστήσει το περιεχόμενο του Παγκόσμιου Ιστού σε μια μορφή επεξεργάσιμη και ερμηνεύσιμη από τις μηχανές και να επιτρέπει έξυπνες και αποτελεσματικές τεχνικές χρήσης της εξαγόμενης πληροφορίας [126]. Ο σκοπός, φυσικά, δεν είναι να δημιουργηθεί εκ νέου ο Ιστός αλλά να μετατραπεί ο υφιστάμενος σε Σημασιολογικό Ιστό. Βασική προϋπόθεση για αυτό είναι η δημιουργία και χρήση εργαλείων αυτόματης εξαγωγής πληροφορίας θεματικής ταξινόμησης και επισημείωσης των διαθέσιμων πηγών δεδομένων. Ένα ακόμα στοιχείο που πρέπει να αναφερθεί στο σημείο αυτό είναι πως πρωτεύοντα ρόλο στην επιτυχία των υπηρεσιών Σημασιολογικού Ιστού, αποτελεί η ποιότητα των πληροφοριών που εξάγονται καθώς ποιοτικοί και ακριβείς περιγραφείς (descriptors) θα υποβοηθήσουν την αποτελεσματικότερη αναζήτηση σε ημι-δομημένες πηγές δεδομένων. Ο Σημασιολογικός Ιστός αποτελεί το όραμα του εκφράστηκε από τον εμπνευστή του Παγκόσμιου Ιστού Tim Berners-Lee [125], με τη μετάβαση από έναν Ιστό εγγράφων, όπου το έγγραφο συνιστά τη βασική δομική μονάδα, σε έναν Ιστό αντικειμένων (Web of Data or Web of Objects), όπου η βαρύτητα μετατοπίζεται στις έννοιες που ενυπάρχουν πίσω από κάθε τέτοιο αντικείμενο ή πηγή δεδομένων. Ο Ιστός αυτός συνοδεύεται από μια στοίβα τεχνολογιών η οποία καθορίζει τις αλληλεπιδράσεις μεταξύ των επιμέρους τεχνολογιών αλλά και τη μορφή που θα πρέπει να έχουν τα δεδομένα έτσι ώστε να δημοσιευθούν για να είναι δυνατή η σημασιολογική ερμηνεία τους από τα εργαλεία.

Η έννοια του Σημασιολογικού Ιστού μπορεί να εφαρμοστεί σε διάφορους τομείς, όπως στη διαχείριση γνώσης, το ηλεκτρονικό εμπόριο τόσο μεταξύ επιχειρήσεων και καταναλωτών (B2C) όσο και μεταξύ επιχειρήσεων (B2B), κ.ά. Η διαχείριση γνώσης περιλαμβάνει την απόκτηση, πρόσβαση και διατήρηση της γνώσης μέσα σε έναν οργανισμό [126]. Η χρηματοπιστωτική και οικονομική κρίση απαιτεί τη σύλληψη και την εφαρμογή εργαλείων που στοχεύουν στην ενίσχυση της ανταλλαγής γνώσεων, της δημιουργικότητας και της καινοτομίας, καθώς τα παραδοσιακά συστήματα διαχείρισης της γνώσης δεν καταφέρνουν να καταστήσουν τις επιχειρήσεις ανταγωνιστικές σε διακρατικό επίπεδο [127]. Η χρήση των τεχνολογιών Σημασιολογικού Ιστού μπορεί να οδηγήσει στη δημιουργία προηγμένων συστημάτων διαχείρισης γνώσης τα οποία παρέχουν: εννοιολογική οργάνωση γνώσης, αυτοματοποιημένα εργαλεία για την εξαγωγή γνώσης, δυνατότητες σημασιολογικής αναζήτησης σε ένα ή περισσότερα έγγραφα και ανάκτηση, εξαγωγή και διατύπωση γνώσης με φιλικό προς τον χρήστη τρόπο [126]. Ένα άλλο δημοφιλές στοιχείο του Ιστού που μπορεί να ωφεληθεί σίγουρα από τη χρήση σημασιολογικών τεχνολογιών είναι τα συστήματα που βασίζονται σε Wiki. Η βασική ιδέα ενός σημασιολογικού Wiki [128] είναι να παράσχει μια σημασιολογική δομή για το υποκείμενο wiki, προκειμένου να είναι προσβάσιμη σε μηχανήματα που μπορούν να προσφέρουν υπηρεσίες πέρα από την απλή πλοήγηση, παρέχοντας σημασιολογική επισήμανση στα δεδομένα [150]. Τέτοια σημασιολογικά Wiki μπορεί να χρησιμοποιηθούν σε πολλούς διαφορετικούς τομείς, από ψηφιακές βιβλιοθήκες έως και κυβερνητικά δεδομένα [129].

Όπως προαναφέρθηκε, τα σημερινά συστήματα μπορούν να ωφεληθούν πλήρως από το όραμα του Σημασιολογικού Ιστού ενσωματώνοντας διάφορες τεχνικές και τεχνολογίες για την επίτευξη των προαναφερθέντων στόχων. Οι δύο πιο συχνά χρησιμοποιούμενες τεχνολογίες είναι τα

LOD και οι Οντολογίες. Η έννοια των LOD περιλαμβάνει τη χρήση του Ιστού για τη σύνδεση δεδομένων από διαφορετικές πηγές. Από τεχνική άποψη, τα LOD είναι δεδομένα που δημοσιεύονται στον Ιστό σε μορφή αναγνώσιμη από μηχανή, η οποία υποδηλώνει τη σημασιολογία τους, ενώ συγχρόνως αλληλοσυνδέονται με άλλα σύνολα δεδομένων που διατίθενται σε αυτόν. Η υιοθέτηση των LOD πρακτικών και η εφαρμογή τους στον Παγκόσμιο Ιστό έχει οδηγήσει στην επέκτασή του, έτσι ώστε ο τελευταίος να περιέχει δεδομένα από πολλές διαφορετικές πηγές, όπως ηλεκτρονικά βιβλία, μουσική, ανθρώπους και οργανισμούς που μπορεί να καλύπτουν εντελώς διαφορετικούς τομείς. Αυτή η ποικιλία των δεδομένων επιτρέπει τη δημιουργία ποικίλων υπηρεσιών προστιθέμενης αξίας. Προκειμένου να δημοσιεύονται στον Ιστό δεδομένα ως Διασυνδεδεμένα Δεδομένα, υπάρχει ένα σύνολο αρχών που θα πρέπει να ακολουθηθούν ώστε αυτά να καταστούν διαθέσιμα ως τμήμα ενός συνολικού χώρου δεδομένων.

Η ενσωμάτωση των LOD σε έναν ιστότοπο εξαρτάται σε μεγάλο βαθμό από πρότυπα όπως το RDF μοντέλο [130] και η OWL [131] που αποτελούν γλώσσες επισημείωσης πηγών, με διαφορετικό επίπεδο εκφραστικότητας η κάθε μια. Το μοντέλο RDF, λόγω της απλότητας που προσφέρει στην περιγραφή δεδομένων, αποτελεί βασικό μοντέλο αναπαράστασης του Σημασιολογικού Ιστού. Ωστόσο, αντί απλώς να συνδέει δεδομένα, το RDF μπορεί επίσης να παρέχει έναν μηχανισμό δημιουργίας δηλώσεων που συνδέουν αυθαίρετα αντικείμενα στον Ιστό με αποτέλεσμα τη δημιουργία ενός Ιστού Δεδομένων (Web of Data). Ως πρώτο βήμα προς αυτή την κατεύθυνση, μπορεί κανείς να θεωρήσει τη χρήση των URI (Uniform Resource Identifier) ως αναγνωριστικά για τις οντότητες που υπάρχουν στον Ιστό [1]. Επιπλέον, τα URI πρέπει να χρησιμοποιούνται μέσω του πρωτοκόλλου HTTP, προκειμένου να μπορούν οι χρήστες να αναζητούν αυτά τα ονόματα. Όταν κάποιος ψάχνει για ένα URI, παρέχεται χρήσιμη δέσμη πληροφοριών σχετικά με τους υποκείμενους πόρους χρησιμοποιώντας το RDF ως γλώσσα αναπαράστασης των δεδομένων. Για την αναζήτηση των δεδομένων που έχουν εκφραστεί με χρήση της RDF, δημιουργήθηκε η γλώσσα SPARQL [132]. Τέλος, οι σύνδεσμοι με άλλους πόρους χρησιμοποιούνται για να επιτρέπουν την ανακάλυψη άλλων αντικειμένων με τον ίδιο τρόπο. Αυτά τα πρότυπα παρέχουν έναν τρόπο για τη δημοσίευση και τη σύνδεση των δεδομένων στον Ιστό, διατηρώντας την αρχιτεκτονική του ενώ ανταποκρίνεται πλήρως στα παραπάνω χαρακτηριστικά.

Μια ακόμα τεχνολογία που είναι άρρηκτα συνδεδεμένη με το Σημασιολογικό Ιστό και τα Διασυνδεδεμένα Δεδομένα είναι οι **οντολογίες** (ontologies). Οι οντολογίες είναι δομημένα σύνολα στοιχείων μεταδεδομένων (metadata element sets) [133] που παρέχουν την περιγραφή μιας θεματικής περιοχής ή ενός πεδίου εφαρμογής με έναν αυστηρό τρόπο σε μια προσπάθεια να μειωθεί η απόσταση μεταξύ του τρόπου που οι άνθρωποι και οι μηχανές διαχειρίζονται την πληροφορία. Οι οντολογίες αποτελούν τη «ραχοκοκαλιά» του Σημασιολογικού Ιστού και μπορεί να χρησιμοποιηθούν αποτελεσματικά σε διαφορετικά πεδία εφαρμογής. Ο λόγος πίσω από την ανάπτυξη μιας οντολογίας είναι, πέρα από την οργανωμένη περιγραφή της υφιστάμενης γνώσης, η ανακάλυψη και η ανταλλαγή νέας γνώσης [123]. Έτσι, έχουν αναπτυχθεί διαφορετικές οντολογίες σε διαφορετικά πεδία εφαρμογής. Η χρήση αλλά και ο συνδυασμός υποσυνόλων από διαφορετικές οντολογίες παρέχει ένα μέσο για τη δημιουργία διαφορετικών μοντέλων με ενοποιημένο τρόπο. Το ενδιαφέρον για την εξέλιξη των οντολογιών οφείλεται στην ανάγκη καθορισμού και σύνδεσης δεδομένων που παρέχονται μέσω του Διαδικτύου με τρόπο που οι οντολογίες να μπορούν να χρησιμοποιηθούν αποτελεσματικά για την ανακάλυψη, αυτοματοποίηση, ανάπτυξη και επαναχρησιμοποίηση δεδομένων εντός εφαρμογών. Λόγω των πλεονεκτημάτων που παρέχουν, οι οντολογίες έχουν ενσωματωθεί σε μια ποικιλία εφαρμογών σε διαφορετικά πεδία εφαρμογής [123], [134], [135].

Εκτός από τα σύνολα στοιχείων μεταδεδομένων έχουν αναπτυχθεί και τα λεγόμενα **λεξιλόγια τιμών** (value vocabularies) τα οποία μπορούν να χρησιμοποιηθούν για την περιγραφή και επισημείωση οποιασδήποτε πηγής δεδομένων, όπως ατόμων, γεωγραφικών τοποθεσιών, προϊόντων κλπ. Στην κατηγορία των λεξιλογίων τιμών εμπίπτουν επίσης οι **θησαυροί**, τα **ελεγχόμενα λεξιλόγια** (controlled vocabularies) και τα λεξιλόγια **καθιερωμένων όρων** (subject headings), όπως οι όροι της Βιβλιοθήκης του Κογκρέσου (LCSH) και το Διεθνές Αρχείο Καθιερωμένων Όρων (VIAF). Αντίστοιχα, τα ελεγχόμενα λεξιλόγια αποτελούν μια προσεκτικά επιλεγμένη αλληλουχία από λέξεις και φράσεις οι οποίες χρησιμοποιούνται για την επισημείωση και την αναζήτηση, τόσο στον τομέα των Ψηφιακών Βιβλιοθηκών όσο και γενικότερα στο πεδίο της Επιστήμης της Πληροφορίας. Για την επισημείωση δεδομένων οι μηχανισμοί αυτόματης εξαγωγής πληροφορίας ταξινόμησης έχουν κομβικό ρόλο, καθώς όσο καλύτερα είναι τα αποτελέσματα αυτής της ταξινόμησης τόσο πιο επισταμένη είναι η επισημείωση, με αποτέλεσμα την αυξημένη ακρίβεια στην ανάκτηση της ζητούμενης πληροφορίας. Τα ελεγχόμενα λεξιλόγια χρησιμοποιούνται ευρέως στους παραπάνω τομείς καθώς η χρήση τους έχει οδηγήσει στον περιορισμό του προβλήματος της αμφισημίας και της πολυσημίας, όπου η ίδια έννοια μπορεί να αποδίδεται με περισσότερους από έναν τρόπους, ενισχύοντας έτσι την συνέπεια του περιεχομένου ενός εγγράφου. Όπως αναφέρθηκε και στα προηγούμενα κεφάλαια τα ελεγχόμενα λεξιλόγια μπορούν με τη σειρά τους να ενσωματωθούν στους μηχανισμούς εξαγωγής πληροφορίας ταξινόμησης, καθώς εγγενώς ενσωματώνουν ένα είδος ιεραρχίας. Αυτό είναι ένα πολύ σημαντικό χαρακτηριστικό καθώς αντιστοιχίζοντας τους όρους από ένα ελεγχόμενο λεξιλόγιο στην πληροφορία ταξινόμησης που εξάγεται, είναι δυνατή η αναζήτηση σε διαφορετικά επίπεδα, χρησιμοποιώντας είτε κάποιον γενικό είτε κάποιον ειδικότερο όρο. Έτσι, είναι δυνατή η αποτελεσματικότερη και ταχύτερη αναζήτηση σε μια πηγή δεδομένων.

Επανερχόμενοι στις οντολογίες, αυτές, όπως προείπαμε, αποτελούν έναν από τους πυλώνες του Σημασιολογικού Ιστού, καθώς μπορεί να θεωρηθούν ως ένας ειδικός τύπος λεξιλογίου ή μερικές φορές ακόμα και ως μια συλλογή από URIs [136], τα οποία μπορεί να οδηγούν σε μια δομημένη περιγραφή. Οι οντολογίες συνήθως ακολουθούνται από έγγραφα γραμμένα σε κάποια γλώσσα περιγραφής οντολογιών όπως η RDFS [137] και η OWL. Ανάλογα με το πεδίο εφαρμογής έχουν αναπτυχθεί διαφορετικές οντολογίες όπως η FOAF (Friend Of A Friend) για την περιγραφή ατόμων, το λεξιλόγιο AGROVOC [138] για τον τομέα της αγροτικής παραγωγής ενώ στον τομέα των βιβλιοθηκών χρησιμοποιούνται εκτενέστατα οι οντολογίες DC (Dublin Core), DC terms (Dublin Core terms), BIBO (Bibliographic Ontology) και SKOS (Simple Knowledge Organization System).

Η χρήση των παραπάνω τεχνολογιών έχει ως στόχο να υποβοηθήσει την αναζήτηση και την ανάκτηση περισσότερων και ταυτόχρονα πιο πλούσιων σημασιολογικά πληροφοριών. Με αυτό τον τρόπο, μια τέτοια προσέγγιση θα μπορούσε να υπερκεράσει τους περιορισμούς που θέτει μια κλασική αναζήτηση με λέξεις-κλειδιά και ταυτόχρονα να καλύψει την ανάγκη από την πλευρά των χρηστών για περισσότερο ευέλικτες αναζητήσεις, που θα επιστρέφουν πιο σχετικά και ποιοτικά αποτελέσματα. Σημαντικό ερευνητικό θέμα αποτελεί η δημιουργία πυλών στον Ιστό οι οποίες θα αξιοποιούν τις παραπάνω τεχνολογίες και θα παρέχουν καινοτόμες υπηρεσίες αναζήτησης και παροχής περιεχομένου ανοικτής πρόσβασης προς τους τελικούς χρήστες.

Για τους λόγους που αναλύθηκαν παραπάνω, οι τεχνολογίες του Σημασιολογικού Ιστού ενσωματώνονται σταδιακά σε εφαρμογές και διεπαφές χρήστη [1] όπως Συστήματα Διαχείρισης Περιεχομένου (Content Management Systems - CMS) [140], τα οποία χρησιμοποιούνται για την ανάπτυξη διαφόρων τύπων Πυλών (portals). Τέτοιες πύλες χρησιμοποιούνται για τη συνάθροιση δεδομένων/πληροφορίας που προέρχονται από ετερογενείς πηγές, όπως κατανεμημένες βάσεις

δεδομένων, ιστότοπους και αρχεία. Η «τάση» αυτή τη στιγμή στον Ιστό είναι η διασύνδεση διαφορετικών πηγών δεδομένων οι οποίες βρίσκονται διάσπαρτες [1], με σκοπό την παροχή συνόλων δεδομένων που θα επιτρέψουν την ανάλυση πάνω στα δεδομένα αυτά και τελικά θα παρέχουν διαφορετικές οπτικές της ίδιας πληροφορίας, για διαφορετικές κατηγορίες χρηστών, όπως ερευνητές, επιχειρηματίες, πολίτες κλπ. [139].

Ο παραπάνω στόχος μπορεί να επιτευχθεί μέσα από την κατάρτιση ροών δεδομένων από πηγές δεδομένων ελεύθερης πρόσβασης οι οποίες βρίσκονται διάσπαρτες στον Ιστό. Το επόμενο βήμα, είναι η διασύνδεση αυτών των ετερογενών πηγών δεδομένων έτσι ώστε να καταστεί δυνατή μια καθολική σημασιολογική αναζήτηση αλλά και η εφαρμογή τεχνικών συλλογιστικής πάνω στα δεδομένα, με σκοπό την παραγωγή νέας γνώσης. Όλες οι ανάγκες που αναφέρθηκαν παραπάνω, μπορεί να καλυφθούν μέσα από τη χρήση των τεχνολογιών του Σημασιολογικού Ιστού αλλά και των εργαλείων που ο τελευταίος παρέχει, με σκοπό τη δημιουργία ενός σημείου αναφοράς/εισόδου, όπως μιας Πύλης Διασυνδεδεμένων Δεδομένων η οποία θα επιτρέπει την συνάθροιση πληροφοριών από τις ετερογενείς πηγές, σημασιολογική αναζήτηση, και δυνατότητα εξαγωγής πληροφορίας με χρήση τεχνικών Μηχανικής και Βαθιάς Μάθησης από διάφορα πεδία γνώσης.

5.4. Εφαρμογή των τεχνολογιών Σημασιολογικού Ιστού σε πύλες

Ένα CMS είναι ένα Web-based σύστημα που χρησιμοποιείται για τη διαχείριση τις παραγωγής και τις διανομής περιεχομένου τόσο σε ενδοδίκτυα (intranet) όσο και σε ιστοσελίδες του Διαδικτύου. Οι τεχνολογίες XML έχουν βασικό ρόλο σε αυτά τα συστήματα προκειμένου να διαχωριστεί το ακατέργαστο περιεχόμενο από την αναπαράστασή του [140]. Τα CMS είναι πολύ χρήσιμα τις χρήστες του Διαδικτύου, είτε πρόκειται για οργανισμούς είτε για μεμονωμένους χρήστες, με διάφορους τρόπους, καθώς παρέχουν πληροφορίες που παρουσιάζονται και ενημερώνονται δυναμικά με φιλικό τις το χρήστη τρόπο. Οι πληροφορίες αυτές μπορούν να προέρχονται από διάφορες πηγές και να παρουσιάζονται μέσω μιας πύλης με ενιαίο τρόπο, παρέχοντας έτσι το απαραίτητο επίπεδο αφαίρεσης για τον χρήστη. Ο βαθμός στον οποίο εμφανίζεται το περιεχόμενο με αυτόν τον τρόπο εξαρτάται από το ρόλο του χρήστη καθώς και από τις άδειες που χορηγούνται ανά ρόλο. Οι δικτυακές πύλες πληροφοριών μπορούν να παρέχουν ενοποιημένη πρόσβαση είτε σε δομημένες πληροφορίες ενός συγκεκριμένου τομέα είτε σε συγκεντρωμένες δομές από ένα ευρύτερο σύνολο τομέων. Στη συνέχεια, γίνεται αναφορά σε υφιστάμενα παραδείγματα δικτυακών πυλών με τεχνολογίες Σημασιολογικού Ιστού.

Υπάρχουν πύλες πληροφοριών που έχουν σχεδιαστεί για να υποστηρίζουν και να διευκολύνουν τις δραστηριότητες μιας συγκεκριμένης κοινότητας χρηστών (Community Information Portals). Η καινοτομία που παρέχουν είναι η ενεργός συμμετοχή των χρηστών τους στη δημιουργία των πληροφοριών που παρουσιάζονται από την Πύλη. Υπάρχουν δύο τρόποι αλληλεπίδρασης σε αυτές τις πύλες, είτε με άμεση υποβολή (δηλ. μέσω μιας φόρμας ή μέσω της δημοσίευσης πληροφοριών σε κάποιο εργαλείο συνεργασίας) είτε με την προσφορά ειδήσεων και πληροφοριών σε μια ομάδα χρηστών [140], [127]. Το KRC (KnowInG Resource Center) είναι μια έξυπνη crowdsourcing πλατφόρμα που συγκεντρώνει και παρέχει πρόσβαση και ανταλλαγή υπηρεσιών, πληροφοριών και γνώσεων, οργανωμένων με ενοποιημένο τρόπο. Μια crowdsourcing πλατφόρμα είναι ένα "σύστημα συλλογικής νοημοσύνης" που αποτελείται από τρία στοιχεία: α. μια οργάνωση που εκμεταλλεύεται το έργο του πλήθους, β. το ίδιο το πλήθος και γ. μια πλατφόρμα που ενεργεί ως ενδιάμεσος ανάμεσα στο πλήθος και την οργάνωση. Ο σκοπός μιας πλατφόρμας crowdsourcing είναι να εμπλέξει το πλήθος και τους ενδιαφερόμενους φορείς στον εντοπισμό προβλημάτων και αναγκών.

Περιστασιακά αναφέρουμε ότι η προτεινόμενη δική μας προσέγγιση, η οποία θα παρουσιαστεί σε επόμενες παραγράφους, είναι παρόμοια με το crowdsourcing σύστημα κατά το ότι στοχεύει στη δέσμευση των χρηστών να συνεισφέρουν στις πληροφορίες που παρουσιάζονται από την Πύλη, λαμβάνοντας υπόψη τις ανάγκες τους.

Στο [140], περιγράφεται η έννοια μιας Σημασιολογικής Πύλης Πληροφοριών που εκμεταλλεύεται τα πρότυπα του Σημασιολογικού Ιστού με σκοπό τη βελτίωση τις δομής, τις επεκτασιμότητας, τις προσαρμογής και τις βιωσιμότητας τις. Μια άλλη προσέγγιση είναι η Πύλη TWC LOGD [141] που έχει σχεδιαστεί για να χρησιμεύει ως πόρος τόσο για τις Ηνωμένες Πολιτείες (ΗΠΑ) όσο και για τις παγκόσμιες κοινότητες ανοικτών κυβερνητικών δεδομένων (LOGD). Αυτό το έργο καταδεικνύει τις πρακτικές εφαρμογές των Διασυνδεδεμένων Δεδομένων στην δημοσίευση και κατανάλωση ανοικτών κυβερνητικών δεδομένων (Open Government Data – OGD) και αντιπροσωπεύει τις την πρώτη πρωτοβουλία που ενσωματώνει πρότυπα Σημασιολογικού Ιστού που υποστηρίζουν αποτελεσματικά τις ανοικτές κυβερνητικές δραστηριότητες των ΗΠΑ.

Το HealthFinland [129] είναι μια σημασιολογική δικτυακή Πύλη για τη διαχείριση πληροφοριών υγείας. Χρησιμοποιεί τεχνολογίες Σημασιολογικού Ιστού για να προσαρμόσει τις πληροφορίες υγείας που είναι διαθέσιμες στον Ιστό ανάλογα με τις ανάγκες ενός συγκεκριμένου χρήστη. Το όραμα της προσέγγισης αυτής είναι η επαναχρησιμοποίηση περιεχομένου από διαδικτυακές πύλες άλλων οργανισμών με χρήση τεχνολογιών Σημασιολογικού Ιστού, ώστε να επισημάνει το περιεχόμενο τοπικά με σημαντικά μεταδεδομένα βασισμένα σε κοινά χρησιμοποιούμενες οντολογίες και παρέχοντας υπηρεσίες διαδικτυακής πρόσβασης σε ένα παγκόσμιο αποθετήριο.

Το SEAL (Semantic portAL) είναι μια προσέγγιση ανεξάρτητη πεδίου για την ανάπτυξη σημασιολογικών πυλών [142]. Αυτή η προσέγγιση ενσωματώνει τη σημασιολογία για την πρόσβαση και την παροχή πληροφοριών στην Πύλη με αποτελεσματικό τρόπο. Ασχολείται επίσης με ζητήματα όπως η κατασκευή και η συντήρηση της Πύλης. Η αρχιτεκτονική της SEAL ενσωματώνει μια σειρά από τμήματα λογισμικού που έχουν χρησιμοποιηθεί και σε άλλες εφαρμογές, όπως το Ontobroker (για την πλοήγηση και την εκτέλεση ερωτημάτων).

Το ONTOVIEWS [143] παρέχει έναν τρόπο δημοσίευσης των δεδομένων RDF στον Ιστό, ενσωματώνοντας στην αρχιτεκτονική του δύο βασικά συστατικά τα OntoDella και OntoGator. Ο πρώτος είναι ένας διακομιστής συστήματος για σύσταση υπερσυνδέσμων (link recommendation system server), ενώ ο τελευταίος είναι ένας διακομιστής μηχανών αναζήτησης με βάση το περιεχόμενο (content-based search engine server). Η βασική ιδέα αυτού του συστήματος είναι να συνδυάσει ένα μοντέλο αναζήτησης πολλαπλών κριτηρίων, το οποίο αναπτύχθηκε μέσα στην ερευνητική κοινότητα ανάκτησης πληροφοριών, με οντολογίες RDF αλλά και να επεκτείνει την λειτουργία αναζήτησης ενσωματώνοντας έναν μηχανισμό σημασιολογικής περιήγησης που βασίζεται σε οντολογική συλλογιστική (ontological reasoning).

Η προσέγγιση που παρουσιάζεται στο [144] στοχεύει στην παροχή φιλικών στους το χρήστη αναπαραστάσεων δεδομένων, αλλά και δυνατότητες περιήγησης και εκτέλεσης ερωτημάτων που εκμεταλλεύονται στους τεχνολογίες του Σημασιολογικού Ιστού. Αυτό επιτυγχάνεται μέσω στους δημιουργίας πολύπλευρων διεπαφών πλοήγησης που επιτρέπουν στους χρήστες να αναζητούν και να προβάλλουν αποτελεσματικά RDF τριπλέτες.

Μια ακόμα προσέγγιση που πλέον έχει ξεπεράσει το στάδιο του πρωτοτύπου, παρουσιάζεται από τους Joubert και συνεργάτες [145], και αναφέρεται στη μοντελοποίηση και την υλοποίηση πυλών διαδικτύου οι οποίες έχουν ως στόχο την παροχή πρόσβασης σε πιστοποιημένες, υψηλής ποιότητας πληροφορίες στον τομέα της ιατρικής. Η συγκεκριμένη υλοποίηση βασίζεται στο UMLS[31] (Unified Medical Language System) και στο έργο ARIANE [124]. Στη συγκεκριμένη προσέγγιση σημασιολογικές συσχετίσεις χρησιμοποιούνται για την αυτόματη μετάφραση ερωτημάτων που μπορούν να εκτελεστούν στις βάσεις δεδομένων του PubMed, του Health on the Net (HON), του CISMeF, ή άλλων πηγών.

Όπως προαναφέρθηκε, οι τεχνολογίες του Σημασιολογικού Ιστού ενισχύουν τη χρήση αυτών των πυλών με πολλούς διαφορετικούς τρόπους. Παρακάτω αναλύεται το σύστημα MILD (Meaningful Integrated Linked Data), το οποίο παρουσιάστηκε στην εργασία [146] και έχει ως στόχο τη σημασιολογική επεξεργασία και διασύνδεση δεδομένων από διαφορετικές πηγές αλλά και την ενσωμάτωση πολλαπλών καναλιών εισόδου για διάφορες ομάδες χρηστών. Ακόμα, στοχεύει στην παροχή κινήτρων για την επαναχρησιμοποίηση και διαμοιρασμό του περιεχομένου. Η υλοποίηση του συστήματος γίνεται με βάση ένα CMS. Στη συγκεκριμένη περίπτωση χρησιμοποιήθηκε το Drupal αλλά ενναλακτικά μπορεί να χρησιμοποιηθεί οποιοδήποτε σύστημα διαχείρισης περιεχομένου χωρίς η προσέγγιση να μεταβληθεί ουσιαστικά.

Ο συνδυασμός του MILD με τις τεχνικές εξαγωγής πληροφορίας ταξινόμησης που αναλύθηκαν στα προηγούμενα κεφάλαια μπορούν να οδηγήσουν σε πιο πλούσιες αναπαραστάσεις των δεδομένων και συνεπώς στην παροχή υπηρεσιών προστιθέμενης αξίας για τους τελικούς χρήστες. Το προτεινόμενο σύστημα συνδυάζει διαφορετικές υπηρεσίες που έχουν ως σκοπό την παροχή προβολών επάνω στα δεδομένα τόσο για ειδικούς όσο και μη ειδικούς χρήστες. Έτσι, παρέχεται η δυνατότητα στους χρήστες να δημιουργήσουν τις δικές τους όψεις επάνω στα δεδομένα, μια διαφοροποίηση σε σχέση με τις κλασσικές προσεγγίσεις που είτε παρέχουν προβολές των συνδυασμένων δεδομένων είτε απλώς παρέχουν μια διεπαφή αναζήτησης.

5.5. Η προτεινόμενη προσέγγιση

Όπως είναι φανερό από τα παραπάνω, παρόλο που οι τεχνολογίες του Σημασιολογικού Ιστού παρέχουν διάφορους τρόπους για αποτελεσματική απεικόνιση των δεδομένων που δημοσιεύονται στον Ιστό, ταυτόχρονα έχουν και ελλείψεις, όπως οι περιορισμοί στον τρόπο προβολής, περιήγησης και αναζήτησης σημασιολογικά συνδεδεμένων δεδομένων [147]. Έτσι, είναι απαραίτητες κατάλληλες διεπαφές, φιλικές προς το χρήστη και ταυτόχρονα αποτελεσματικές στην πλοήγηση και την αναζήτηση, όπως η αρχιτεκτονική MILD που παρουσιάζεται εδώ. Μια Ψηφιακή Βιβλιοθήκη θα μπορούσε να χρησιμοποιήσει την προτεινόμενη προσέγγιση για την αναβάθμιση των υπηρεσιών που παρέχονται στους φοιτητές και τα άλλα μέλη της κοινότητας.

Με βάση την εκτεταμένη βιβλιογραφική ανασκόπηση που παρουσιάστηκε στην προηγούμενη παράγραφο, μπορούν να εντοπιστούν τόσο τα κοινά σημεία όσο και εκείνα που διαφοροποιούν τις υφιστάμενες προσεγγίσεις σε σχέση με την προτεινόμενη. Συγκριτικά με το [140], η προτεινόμενη προσέγγισή επικεντρώνεται στην πλευρά των δεδομένων και όχι στον τρόπο με τον οποίο θα υλοποιηθεί η Πύλη. Η χρήση ενός CMS και όχι η υλοποίηση μιας προσαρμοσμένης πύλης με βάση τα δεδομένα, αντικατοπτρίζει αυτήν την επιλογή. Η προτεινόμενη προσέγγιση στοχεύει στην επαναχρησιμοποίηση δεδομένων από διάφορους οργανισμούς, δημόσιους ή ιδιωτικούς όπως και

[31] https://www.nlm.nih.gov/research/umls/

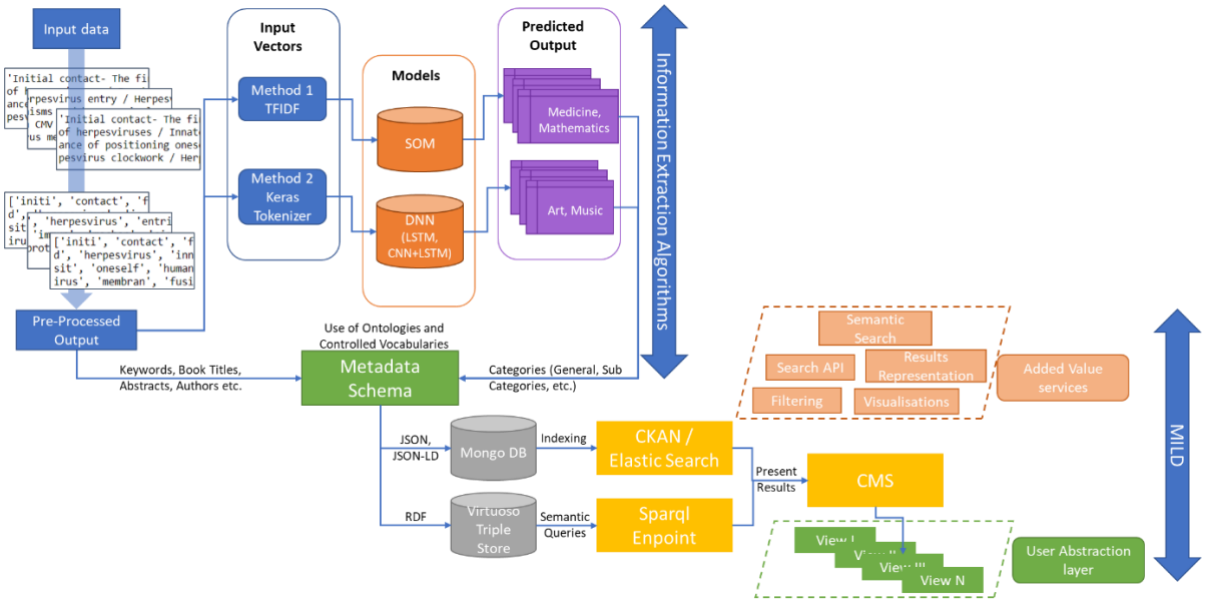
στην περίπτωση του Health Finland, προκειμένου να παράσχει με αποτελεσματικό τρόπο όχι μόνο τα σημασιολογικά σχολιασμένα μεταδεδομένα αλλά και περιεχόμενο που προέρχεται από τα υποκείμενα δεδομένα.

Ακόμα, το σύστημα ONTOVIEWS επικεντρώνεται στις δραστηριότητες αναζήτησης και συλλογιστικής, ενώ η δική μας προσέγγιση επεκτείνεται σε άλλες λειτουργίες όπως η ηλεκτρονική μάθηση, παρέχοντας έτσι μια ολοκληρωμένη εμπειρία χρήστη. Το [144] σε αντίθεση με την προσέγγισή μας, επικεντρώνεται στη δημιουργία, επεξεργασία και οπτικοποίηση οντολογιών μέσα στο σύστημα, ενώ η δική μας προσέγγιση στην παροχή και απεικόνιση δεδομένων που έχουν σημασία για τους τελικούς χρήστες. Τέλος, η SEAL σε αντίθεση με την προσέγγισή μας, εστιάζει στην εισαγωγή λειτουργιών σημασιολογικής κατάταξης (semantic ranking) και ανίχνευσης (crawling), για πρόσβαση σε μια τοποθεσία Web.

5.5.1. Η προτεινόμενη αρχιτεκτονική MILD

Ο κύριος στόχος της προσέγγισης MILD είναι η σημασιολογική επεξεργασία και η διασύνδεση δεδομένων από διαφορετικές πηγές. Αυτές οι πηγές μπορεί να προέρχονται από διαφορετικά/ανόμοια πεδία. Το επιθυμητό αποτέλεσμα αυτής της διαδικασίας θα ήταν να συνδυάσει δεδομένα από αυτές τις πηγές, προκειμένου να προσφέρει υπηρεσίες προστιθέμενης αξίας (Εικόνα 35). Τα δεδομένα που προέρχονται από αυτές τις πηγές μπορεί να διατίθενται σε διαφορετικά μορφότυπα όπως XLS / CSV και να είναι αποθηκευμένα σε βάσεις δεδομένων, όπως MySQL ή Oracle, αλλά και σε αρχεία κειμένου. Τα δεδομένα που περιέχονται σε κάθε πηγή πρέπει να επισημειωθούν, χρησιμοποιώντας το υποκείμενο RDF μοντέλο. Το μοντέλο αυτό είναι αποτέλεσμα μιας διαδικασίας μοντελοποίησης εκτός σύνδεσης (offline), μοναδική για κάθε πεδίο εφαρμογής. Το κοινό στοιχείο αυτής της διαδικασίας μοντελοποίησης είναι η χρήση οντολογιών και ελεγχόμενων λεξιλογίων ή θησαυρών. Στην Επιστήμη της Πληροφορίας και των Βιβλιοθηκών ένα ελεγχόμενο λεξιλόγιο, είναι μια δομημένη λίστα όρων ή φράσεων που χρησιμοποιούνται για την επισήμανση τμημάτων πληροφοριών (τεκμηρίων) έτσι ώστε αυτά να μπορούν να ανακτηθούν ευκολότερα και ακριβέστερα σε μια αναζήτηση. Το αποτέλεσμα αυτής της διαδικασίας είναι η επισημείωση των δεδομένων με χρήση RDF, έτσι ώστε να δημιουργηθούν δηλώσεις που να διασυνδέουν πόρους που διατίθενται στον Ιστό. Ο γράφος RDF που προκύπτει, φορτώνεται στη συνέχεια σε μια βάση τριπλετών όπως το Virtuoso[32]. Η γενική αρχιτεκτονική του συστήματος αλλά και η διασύνδεση του με τις τεχνικές που αναλύθηκαν στα Κεφάλαια 3 και 4 δίνεται στην παρακάτω εικόνα.

[32] http://virtuoso.openlinksw.com/



Εικόνα 34. Αρχιτεκτονική MILD και ένα παράδειγμα διασύνδεσης με τις τεχνικές που αναλύθηκαν στα Κεφάλαια 3 και 4.

Σύμφωνα με τις ανάγκες των χρηστών, οι οποίες ορίζονται στη φάση ανάλυσης απαιτήσεων, γράφονται και επαληθεύονται τα αντίστοιχα ερωτήματα SPARQL. Στη συνέχεια, τα αποτελέσματα εγγράφονται σε ένα αρχείο JSON και μεταφορτώνονται σε ένα σύστημα διαχείρισης δεδομένων όπως το CKAN[33] (Comprehensive Knowledge Archive Network), ένα ισχυρό σύστημα διαχείρισης δεδομένων που παρέχει εργαλεία για τη βελτιστοποίηση της δημοσίευσης, της κοινής χρήσης, της εύρεσης και του διαμοιρασμού δεδομένων. Το CKAN απευθύνεται σε παρόχους δεδομένων, όπως εθνικές και περιφερειακές κυβερνήσεις, εταιρείες και οργανισμούς ή ιδρύματα που επιθυμούν να κάνουν τα δεδομένα τους ανοικτά και διαθέσιμα. Ένα άλλο χαρακτηριστικό που κάνει το CKAN κατάλληλη επιλογή για το σύστημα διαχείρισης δεδομένων στην παρούσα προσέγγιση είναι ότι τα μεταδεδομένα του συνόλου δεδομένων του CKAN δημοσιεύονται στη μορφή JSON. Το JSON είναι ένα εύχρηστο μορφότυπο που επιλέγεται από οργανισμούς για την εξαγωγή των δεδομένων τους. Όσον αφορά τις προσπάθειες τυποποίησης λογισμικού, το CKAN πειραματίζεται επίσης με την κωδικοποίηση RDF χρησιμοποιώντας το λεξιλόγιο καταλόγου δεδομένων του DERI (dcat). Εναλλακτικά, αντί για το CKAN μπορεί να χρησιμοποιηθεί το ELK για την διαχείριση των δεδομένων. Μέσω του ELK είναι δυνατή η ευρετηρίαση όλων των δεδομένων που θα γίνουν διαθέσιμα από το σύστημα. Έτσι, είναι δυνατή η γρήγορη αναζήτηση και ανάκτηση των αποτελεσμάτων. Το σύστημα διαχείρισης δεδομένων του MILD είναι το ενδιάμεσο μεταξύ των πραγματικών δεδομένων που θα εμφανιστούν στην Πύλη και της ίδιας της Πύλης. Το MILD, μέσω των διαφορετικών επιπέδων που ενσωματώνει, παρέχει ένα επίπεδο αφαίρεσης στους χρήστες της Πύλης, με διαφορετικές προβολές, ανάλογα με τον ρόλο που αποδίδεται σε κάθε χρήστη.

Η προσέγγιση αυτή συνδυάζει διάφορες υπηρεσίες οι οποίες αποσκοπούν στην παροχή όψεων στα δεδομένα που μπορεί να ενδιαφέρουν τόσο τους ειδικούς όσο και τους μη ειδικούς χρήστες (παρέχοντάς τους την ευκαιρία να δημιουργήσουν τις δικές τους οπτικοποιήσεις όπως στο [144], αντί να παρέχουν απλώς συγκεντρωτικά δεδομένα), καθώς και ένα API αναζήτησης ή συμπερασμού (reasoning). Ένα απαραίτητο στοιχείο που παρέχει αυτή τη λειτουργία είναι ένα API σημασιολογικής αναζήτησης (Semantic Search API) που χρησιμοποιείται ως σωρευτής, παρέχοντας ενοποιημένα

[33] http://ckan.org/

αποτελέσματα αναζήτησης από διαφορετικές πηγές δεδομένων. Στην συνέχεια μπορούν να δημιουργηθούν φίλτρα έτσι ώστε ο τελικός χρήστης να μπορεί να πλοηγηθεί γρήγορα αλλά και να εντοπίσει την πληροφορία που τον ενδιαφέρει φιλτράροντας ουσιαστικά μη σχετικά αποτελέσματα. Κατά το σχεδιασμό της Πύλης μπορεί να συμπεριληφθεί μια ενότητα «Νέων» προκειμένου να παρέχονται ειδήσεις όπως η διάθεση νέων συνόλων δεδομένων. Άλλες προσεγγίσεις, όπως η [148], περιλαμβάνουν μια ενότητα οπτικοποίησης που στην περίπτωσή μας θα μπορούσε να παρέχει συνδυασμένες πληροφορίες, στατιστικά στοιχεία και γραφικές παραστάσεις των δεδομένων. Τέλος, η προσέγγισή μπορεί να περιλαμβάνει ένα τμήμα ηλεκτρονικής μάθησης που μπορεί να χρησιμοποιηθεί αποτελεσματικά προκειμένου να ενισχυθεί η ευαισθητοποίηση του κοινού μέσα στην ακαδημαϊκή κοινότητα, αποτελούμενη κυρίως από φοιτητές, καθηγητές και ερευνητές, εξυπηρετώντας έτσι τους σκοπούς διάδοσης της Πύλης.

5.5.2. Μεθοδολογία

Το MILD, τις υποδηλώνεται από την αρχιτεκτονική που παρουσιάστηκε προηγουμένως, είναι μια προσέγγιση που χρησιμοποιείται για την αναζήτηση και την οπτικοποίηση των δεδομένων κατά τρόπο ουσιαστικό και φιλικό τις τον χρήστη. Σε αυτή την ενότητα θα αναπτυχθεί η μεθοδολογία που βρίσκεται πίσω από την αρχιτεκτονική του (Εικόνα 35), η οποία συνίσταται σε τρία (3) στάδια. Κάθε στάδιο είναι ξεχωριστό και περιλαμβάνει συγκεκριμένες ενέργειες, διαθέτοντας παράλληλα τις απαραίτητες πληροφορίες από τα προηγούμενα στάδια.

Φάση Ανάλυσης Δεδομένων (Data Analysis Phase): Η κύρια λειτουργία αυτής της φάσης είναι η ανάλυση των διαθέσιμων δεδομένων, η οποία εκτελείται συνήθως από κάποιον ειδικό στον τομέα. Οι στόχοι και τα αποτελέσματα αυτής της φάσης εξαρτώνται από τον υπό εξέταση τομέα.

Φάση Μοντελοποίησης (Modelling Phase): Αυτή η φάση στοχεύει να προσδιορίσει ένα κατάλληλο μοντέλο RDF για τα υποκείμενα δεδομένα. Το προτεινόμενο μοντέλο RDF εξαρτάται από το αποτέλεσμα του προηγούμενου σταδίου. Αυτό το μοντέλο παράγεται από τον ειδικό του τομέα σε συνεργασία με τον προγραμματιστή καθώς χρειάζεται να αποφασίσουν από κοινού για τις οντολογίες που θα χρησιμοποιηθούν. Το τελικό μοντέλο μετά από μια φάση βελτίωσης δίδεται ως είσοδος στην φάση υλοποίησης.

Φάση Υλοποίησης (Implementation Phase): Αυτή η φάση περιλαμβάνει δύο ξεχωριστές εργασίες: α. Επισημείωση δεδομένων (Data Annotation) και β. τη δημιουργία του RDF μοντέλου. Αυτές οι εργασίες εκτελούνται είτε εκτός σύνδεσης είτε σε ένα VM (Virtual Machine) που εκτελεί το προτεινόμενο μοντέλο για κάθε πεδίο εφαρμογής. Η διαδικασία αυτή μπορεί να επαναλαμβάνεται εάν υπάρχουν νέα σύνολα δεδομένων προκειμένου αυτά να γίνουν διαθέσιμα μέσω της Πύλης. Στη συνέχεια, μπορούν να χρησιμοποιηθούν οι μηχανισμοί εξαγωγής έτσι ώστε να ανακτηθούν χρήσιμες πληροφορίες από το σύνολο δεδομένων που με τη σειρά τους θα δοθούν ως είσοδος στο υποσύστημα που είναι υπεύθυνο για την επισημείωση των δεδομένων. Κατά την εργασία της επισημείωσης, κάθε τμήμα των δεδομένων κατηγοριοποιείται είτε ως πόρος (resource) είτε ως ιδιότητα (property), σύμφωνα με την κατανομή που έγινε κατά τη διάρκεια της μοντελοποίησης, πάντα κατά το πρότυπο RDF. Το επόμενο βήμα αυτής της φάσης είναι η παραγωγή του γράφου RDF που περιέχει τριπλέτες με σημασιολογικά επισημειωμένα δεδομένα. Μετά την κατασκευή του γράφου RDF και την εισαγωγή των τριπλετών σε μια βάση (triple store), η επόμενη εργασία είναι η υλοποίηση των απαραίτητων SPARQL ερωτημάτων για την ανάδειξη της σημασιολογίας των υποκείμενων δεδομένων. Τα αποτελέσματα των ερωτημάτων αυτών θα τροφοδοτήσουν το CMS με

τα κατάλληλα δεδομένα που θα παρουσιαστούν σε κάθε ομάδα χρηστών, σύμφωνα με τους ρόλους που τους έχουν ανατεθεί.

5.5.3. Λειτουργικότητα

Η **λειτουργική δομή** της Πύλης μπορεί να παρουσιαστεί αναλύοντας τη δομή του μενού της. Ενδεικτικά, θα μπορούσε να περιλαμβάνει τα εξής: μια αρχική σελίδα, μια σελίδα θεματικών περιοχών, μια σελίδα που θα παρουσιάζει διαφορετικές όψεις δεδομένων και των υποκείμενών πηγών αυτών, και τέλος τον πίνακα ελέγχου της Πύλης. Η αρχική σελίδα χρησιμεύει ως σημείο εισόδου, με στόχο να παρέχει στους επισκέπτες μια επισκόπηση των δυνατοτήτων της Πύλης καθώς και να λειτουργεί ως σελίδα επιστροφής. Στη σελίδα θεματικών περιοχών μπορούν να φιλοξενηθούν οι διάφοροι θεματικοί τομείς του υπό εξέταση πεδίου. Για παράδειγμα, εάν η περιοχή ενδιαφέροντος είναι οι Ακαδημαϊκές Βιβλιοθήκες, οι θεματικές περιοχές θα μπορούσαν να είναι είτε διακριτές Επιστημονικές Περιοχές (Μαθηματικά, Φυσική, Πληροφορική, Ανθρωπιστικές Επιστήμες κοκ), είτε διακριτές Συλλογές (Ιδρυματικό Αποθετήριο -Institutional Repository, Κύρια Συλλογή και Συλλογή «Εύδοξος» -συμπεριλαμβανομένων των συγγραμμάτων για κάθε σχολή του Ιδρύματος, Συλλογή Ανοικτών Συγγραμμάτων «Κάλλιπος»). Η προτεινόμενη προσέγγιση μπορεί να υλοποιηθεί εύκολα λόγω του γεγονότος ότι βασίζεται σε ένα CMS, το οποίο μπορεί να παραμετροποιηθεί και να χρησιμοποιηθεί ακόμα και από μη εξειδικευμένους χρήστες. Ένα άλλο καινοτόμο χαρακτηριστικό της προτεινόμενης προσέγγισης είναι ότι μπορεί εύκολα να επεκταθεί και σε άλλες θεματικές περιοχές ή ακόμα και να εφαρμοστεί σε ένα εντελώς διαφορετικό πεδίο εφαρμογής. Επιπλέον, η προσέγγιση είναι ανεξάρτητη από τα υποκείμενα δεδομένα ή το μοντέλο που ακολουθούν αυτά τα δεδομένα. Το τμήμα που παρέχει διαφορετικές όψεις πάνω στα δεδομένα της Πύλης μπορεί να είναι πολύ χρήσιμο για την απεικόνιση διαφορετικών διαστάσεων των δεδομένων αυτών, όπως είναι η χρονική (ημερήσια, εβδομαδιαία, μηνιαία ετήσια κλπ.), η γεωχωρική (ταχυδρομικός κώδικας, δήμος, περιφέρεια, πόλη, χώρα) και η θεματική, σε διαφορετικές ομάδες χρηστών. Η ενότητα των πηγών δεδομένων θα παρουσιάζει τις διαθέσιμες πηγές δεδομένων για αναζήτηση. Θα μπορούσε επίσης να παρουσιάσει τη δομή των υποκείμενων δεδομένων για κάθε πηγή δεδομένων, και τη θεματική τους περιοχή, κλπ.

Αυθεντικοποίηση χρήστη (User Authentication): Οι προδιαγραφές της προσέγγισης επιβάλλουν τη χρήση τριών τουλάχιστον διαφορετικών επιπέδων αυθεντικοποίησης για τους εμπλεκόμενους χρήστες. Το πρώτο επίπεδο πρόσβασης παρέχεται σε κάθε χρήστη καθώς αποτελεί τη δημόσια προβολή της πύλης. Το δεύτερο επίπεδο παρέχεται μόνο σε εγγεγραμμένους χρήστες που εξυπηρετούν συγκεκριμένο σκοπό ή έχουν ειδικές ανάγκες, όπως σπουδαστές ή ερευνητές. Τέλος, το τρίτο επίπεδο εξουσιοδότησης παρέχεται μόνο σε εγγεγραμμένους χρήστες που εξυπηρετούν ειδικούς/διαχειριστικούς σκοπούς. Προκειμένου να γίνει διάκριση μεταξύ των διαφορετικών λογικών όψεων για τα χορηγούμενα επίπεδα εξουσιοδότησης, η προτεινόμενη προσέγγιση παρέχει μια περιοχή σύνδεσης. Αυτή η περιοχή επιτρέπει την πρόσβαση στο δεύτερο και στο τρίτο επίπεδο εξουσιοδότησης εκχωρώντας τις επιπλέον δυνατότητες σε όλους τους χρήστες.

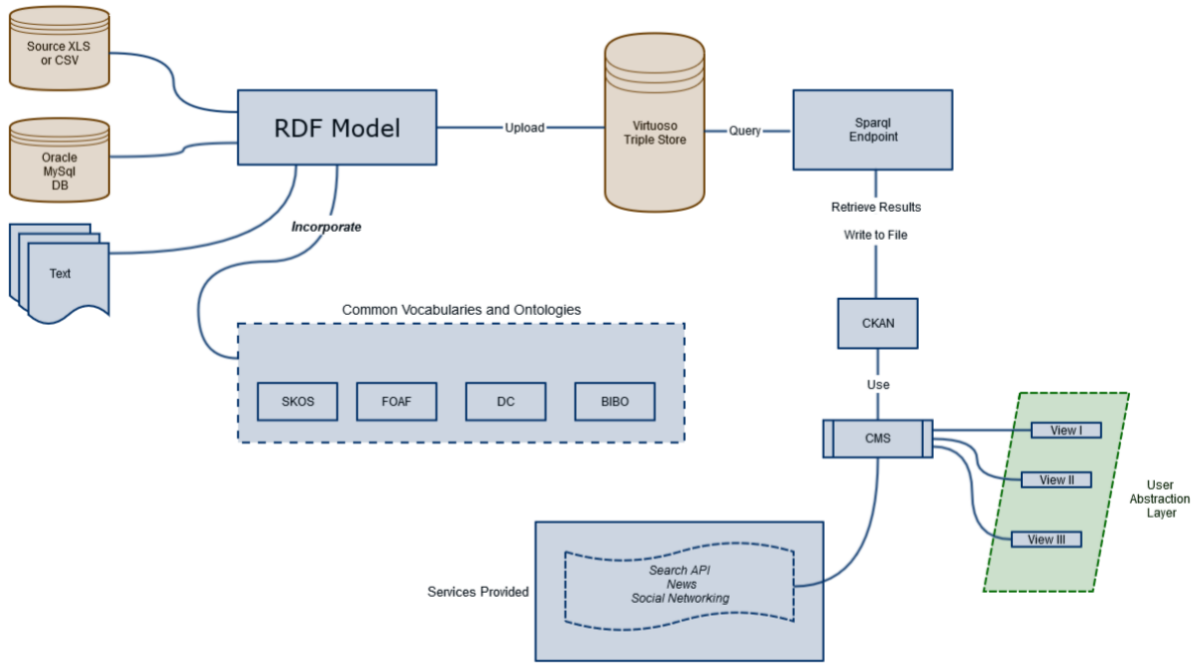
Αναπαράσταση Αποτελεσμάτων (Result Representation): Ένα βασικό χαρακτηριστικό της προτεινόμενης προσέγγισης είναι η παροχή συνδυασμένων πληροφοριών, όπως των αποτελεσμάτων με βάση την αναζήτηση που πραγματοποίησε ο χρήστης, φίλτρων για την πλοήγηση στα δεδομένα αλλά και πρόσβαση σε υπηρεσίες προστιθέμενης αξίας που αφορούν τα υποκείμενα δεδομένα. Οι υπηρεσίες αυτές μπορεί να περιλαμβάνουν παρουσίαση του πλήρους κειμένου για ένα από τα αποτελέσματα αναζήτησης, παροχή στατιστικών, καθώς και γραφικών παραστάσεων των

δεδομένων, όμως γράφων διασύνδεσης με άλλα αντικείμενα κλπ. Ακόμα μπορεί να χρησιμοποιηθούν εργαλεία επισήμανσης (highlighting tools) στο πλήρες κείμενο με βάση τα κριτήρια αναζήτησης που έθεσε ο χρήστης έτσι ώστε να είναι δυνατός ο εντοπισμός της πληροφορίας. Τέλος μπορεί να παρέχεται πρόσβαση στα μεταδεδομένα που έχουν εξαχθεί για το συγκεκριμένο αντικείμενο με χρήση των μηχανισμών εξαγωγής πληροφορίας ταξινόμησης. Όλες οι προαναφερθείσες πληροφορίες μπορούν να παρέχονται μέσω όμως εργαλείου απεικόνισης (dashboard), η επιλογή του οποίου εξαρτάται από το υποκείμενο CMS. Το εργαλείο αυτό χρησιμεύει ως ενδιάμεσος μεταξύ των μη επεξεργασμένων δεδομένων και του χρήστη έτσι ώστε να μπορεί ο τελευταίος να παρέχει διαφορετικές προβολές των δεδομένων, ανάλογα με όμως επιλογές του. Μια άλλη υπηρεσία προστιθέμενης αξίας του εργαλείου αυτού είναι η παρουσίαση των διαθέσιμων πληροφοριών σε μεγαλύτερο επίπεδο λεπτομέρειας ανάλογα με τις ανάγκες κάθε χρήστη. Ένα σύστημα διαχείρισης δεδομένων είναι ένα ουσιαστικό στοιχείο, για την παροχή τέτοιας λειτουργικότητας. Όμως αναφέρθηκε προηγουμένως, το CKAN είναι το σύστημα διαχείρισης δεδομένων που χρησιμοποιείται από το MILD στα πλαίσια όμως της εργασίας, εναλλακτικά όμως μπορεί να χρησιμοποιηθεί και το ELK.

Ενίσχυση της ευαισθητοποίησης του κοινού (Enhancing Public Awareness): Όπως προαναφέρθηκε, προκειμένου να ενισχυθεί η ευαισθητοποίηση του κοινού θα πρέπει να συμπεριληφθεί ένας μικρο-ιστότοπος (micro site) ασύγχρονης ηλεκτρονικής μάθησης συνδεδεμένος με την Πύλη. Επιλέξαμε να υλοποιήσουμε το micro site χρησιμοποιώντας το Opigno[34], επειδή είναι ένα ισχυρό, ανοικτού κώδικα, σύστημα διαχείρισης μάθησης (Learning Management System) βασισμένο στο Drupal. Το γεγονός αυτό έπαιξε καθοριστικό ρόλο στην επιλογή του Opigno, καθώς το Drupal είναι το επιλεγμένο CMS για την περίπτωσή μας και ως εκ τούτου διασφαλίζεται η συμβατότητα με την υπόλοιπη Πύλη. Ωστόσο, πρέπει να σημειωθεί ότι θα μπορούσε να επιλεγεί οποιοδήποτε άλλο σύστημα. Τα κύρια χαρακτηριστικά του επιλεγμένου περιλαμβάνουν: βίντεο (video gallery), εκδηλώσεις (events), ανακοινώσεις (announcements), φόρουμ (forum), έρευνες (surveys), τηλεδιάσκεψη, στατιστικά στοιχεία, μηνύματα και συνομιλία (chat). Στο μέλλον θα μπορούσε να συμπεριληφθεί στην Πύλη μια επιπλέον λειτουργικότητα που να προσφέρει σύγχρονη ηλεκτρονική μάθηση, ανεξάρτητα από το υποκείμενο CMS. Με τον τρόπο αυτό θα μπορούσαμε να παρέχουμε σε απευθείας σύνδεση σεμινάρια για την ενημέρωση και την ενίσχυση της ευαισθητοποίησης του κοινού και τη δημιουργία συνεργατικής γνώσης, όπως συμβαίνει στις διαδικτυακές πύλες κοινοτήτων (Community Web Portals) [123].

Διεπαφή Αναζήτησης (Search API): Μια διεπαφή σημασιολογικής αναζήτησης παίζει βασικό ρόλο στην προτεινόμενη προσέγγιση καθώς μπορεί να χρησιμοποιηθεί ως σωρευτής δεδομένων, παρέχοντας ενοποιημένα αποτελέσματα αναζήτησης από διαφορετικές πηγές. Υπάρχει εκτεταμένη έρευνα σε αυτό το πεδίο με διαφορετικές προτεινόμενες προσεγγίσεις όπως οι [123], [141]. Το κοινό χαρακτηριστικό στις περισσότερες προσεγγίσεις είναι ότι οι χρησιμοποιούμενες διεπαφές αναζήτησης βασίζονται σε μια μηχανή ευρετηρίασης όπως η Lucene[35] [144] ή η ELK.

[34] https://www.opigno.org/en
[35] http://lucene.apache.org/core/



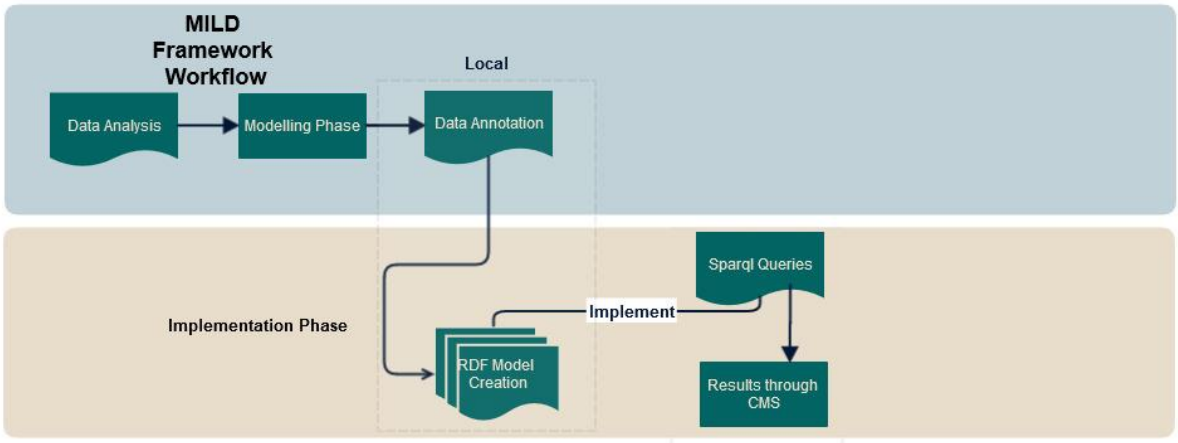
Εικόνα 35. Αρχιτεκτονική MILD.

Ροή Ειδήσεων (News Feed): Ένας δείκτης μέτρησης της επιτυχίας της Ροής Ειδήσεων είναι η συχνότητα με την οποία ενημερώνονται τα δεδομένα που παρουσιάζονται από την Πύλη. Η λειτουργία αυτή, εκτός από την παροχή ενημερώσεων σχετικά με νέα διαθέσιμα δεδομένα ή την ενσωμάτωση νέων συνόλων δεδομένων από διαφορετικές πηγές, θα μπορούσε επίσης να λειτουργήσει για να προσελκύσει νέους χρήστες στην Πύλη, αυξάνοντας έτσι τη δημοτικότητά της.

Ζήτηση για σύνολα δεδομένων (Demand for datasets): Μια τάση όσον αφορά τα διαθέσιμα σύνολα δεδομένων στον Ιστό είναι η διαρκής ανάγκη για παροχή νέων δεδομένων. Αυτό εισάγεται επίσημα υπό την έννοια των αιτήσεων (petition). Μια τέτοιου τύπου αίτηση, είναι ένα επίσημο γραπτό αίτημα, υπογεγραμμένο συνήθως από πολλούς ανθρώπους, απευθυνόμενο σε μια αρχή, σε σχέση με μια συγκεκριμένη αιτία, σύμφωνα με το λεξικό της Οξφόρδης[36]. Παρόμοιες δράσεις έχουν ήδη εγκριθεί στις ΗΠΑ[37], το Ηνωμένο Βασίλειο[38] και την ΕΕ[39], οι οποίες επιτρέπουν στους πολίτες να ψηφίζουν για τα υπό εξέταση αιτήματα ή να βλέπουν τα ολοκληρωμένα αιτήματα, δίνοντάς τους την ευκαιρία να εκφράσουν τη γνώμη τους. Επιπλέον, τα αιτήματα που δημιουργούνται μέσω αυτού του μηχανισμού σε μια χώρα μπορούν να προωθήσουν θέματα ενδιαφέροντος για συζήτηση από αρμόδια θεσμικά όργανα, όπως για παράδειγμα το Κοινοβούλιο της χώρας. Υπάρχουν επίσης πρωτοβουλίες που προσφέρουν ελεύθερα τον κώδικα τέτοιου τύπου λογισμικού, όπως για παράδειγμα το Drupal, το οποίο τροφοδοτεί το "We The People" για τον Λευκό Οίκο των Η.Π.Α.[40] και τη βάση δεδομένων για την υπηρεσία e-petitions της Κυβέρνησης του Ηνωμένου Βασιλείου[41]. Με αυτόν τον τρόπο οι χρήστες έχουν την ευκαιρία να αιτηθούν δημοσίως για ανοικτά δεδομένα τόσο από δημόσιους όσο και από ιδιωτικούς φορείς. Τα βασικά χαρακτηριστικά μιας τέτοιας υπηρεσίας θα μπορούσαν να είναι η δημιουργία ποικίλων αιτημάτων δωρεάν από τους χρήστες, παρέχοντας

[36] http://www.oxforddictionaries.com/definition/english/petition
[37] https://petitions.whitehouse.gov/
[38] http://epetitions.direct.gov.uk/
[39] http://www.europarl.europa.eu/committees/en/peti/home.html
[40] https://github.com/WhiteHouse/petitions
[41] https://github.com/alphagov/e-petitions

ευκολότερες διεπαφές σε σύγκριση με τις υπάρχουσες προσεγγίσεις (π.χ. αιτήσεις μέσω ηλεκτρονικού ταχυδρομείου - e-mail petitions, ανεπίσημες αιτήσεις μέσω φόρουμ- informal web forum-based petitions). Ορισμένες τέτοιες υπηρεσίες παρέχουν επίσης εργαλεία για την προώθηση των αιτημάτων που δημιουργούν οι χρήστες. Στην προτεινόμενη προσέγγιση σχεδιάσαμε να δώσουμε στους χρήστες που έχουν ειδικά προνόμια (2ο και 3ο επίπεδο αδειοδότησης) την ευκαιρία να δημιουργήσουν αιτήματα σχετικά με θέματα που ενδιαφέρουν την υποκείμενη κοινότητα χρηστών. Για παράδειγμα, αιτήματα για νέα βιβλία ή κείμενα που δεν διατίθενται προς το παρόν σε μια Ψηφιακή Βιβλιοθήκη ή άλλα θέματα που αφορούν διαφορετικές ομάδες χρηστών π.χ. την ερευνητική κοινότητα ενός ιδρύματος. Αυτό θα επιτευχθεί με τη δημιουργία μιας υπηρεσίας αναφορών που θα ενσωματωθεί στο CMS.



Εικόνα 36. Ροή εργασίας της προσέγγισης.

5.5.4. Λογική Πλοήγησης

Κατά τη διάρκεια της φάσης ανάλυσης απαιτήσεων, αποφασίστηκε ότι η Πύλη θα πρέπει να έχει ιεραρχική δομή ως προς το περιεχόμενό της. Αυτή η προσέγγιση είναι φιλική προς τους χρήστες, δεδομένου αυτοί μπορούν εύκολα να εντοπίσουν που ακριβώς βρίσκονται κάθε στιγμή. Επίσης η σύνδεση των δεδομένων καθίσταται προσαρμόσιμη στις ανάγκες του χρήστη. Με άλλα λόγια, δύο διαφορετικές ομάδες χρηστών θα μπορούσαν να παράγουν δύο διαφορετικές όψεις. Όπως αναφέρθηκε προηγουμένως, υπάρχουν διαφορετικά επίπεδα εξουσιοδότησης για τους χρήστες της Πύλης. Αυτό επηρεάζει τον μηχανισμό απόφασης που συνδέεται με τη λογική πλοήγησης της Πύλης. Η «πρόσβαση επισκεπτών» παρέχεται σε όλους τους χρήστες, δίνοντάς τους τη δυνατότητα να βλέπουν συγκεκριμένες πληροφορίες σε σελίδες όπως η Αρχική σελίδα, τα δεδομένα ή η σελίδα πηγών δεδομένων.

Το δεύτερο επίπεδο εξουσιοδότησης παρέχεται μόνο σε εγγεγραμμένους χρήστες. Με αυτόν τον τρόπο, οι χρήστες που ανήκουν σε αυτή την κατηγορία μπορούν να δουν όλα όσα μπορεί να δει ένας απλός χρήστης, αλλά μπορούν επίσης να εκτελέσουν ερωτήματα σχετικά με τα διαθέσιμα δεδομένα αλλά και να δουν τα αποτελέσματα που παρέχει η Πύλη. Αυτή η ομάδα μπορεί επίσης να συμμετέχει στη διαδικασία ηλεκτρονικής μάθησης όπως περιεγράφηκε προηγουμένως ή να παράσχει νέα δεδομένα. Ακόμα, η ομάδα αυτή μπορεί να έχει πρόσβαση σε ένα προσωπικό πίνακα ελέγχου μέσω του οποίου ο κάθε χρήστης μπορεί να έχει πρόσβαση στις πληροφορίες του λογαριασμού του, να ανακτήσει τις τελευταίες αναζητήσεις που έκανε μέσω της Πύλης αλλά και να αποθηκεύσει τα αποτελέσματα μιας αναζήτησης με σκοπό να ανατρέξει σε αυτά σε δεύτερο χρόνο.

Τέλος, το τρίτο επίπεδο εξουσιοδότησης δίνεται σε εγγεγραμμένους χρήστες με ειδικά προνόμια. Οι χρήστες αυτής της κατηγορίας έχουν πρόσβαση σε μερικές πιο προηγμένες λειτουργίες, όπως η οργάνωση των αιτημάτων (petitions), ή παίρνουν αποφάσεις σχετικά με τα δεδομένα που θα γίνουν διαθέσιμα μέσω της Πύλης. Επιπλέον, τέτοιοι χρήστες θα μπορούσαν να ζητήσουν συγκεκριμένα δεδομένα ή συγκεκριμένα αποτελέσματα που έχουν προκύψει από την ανάλυση των δεδομένων. Ακόμα, οι χρήστες αυτοί θα μπορούσαν να επιθεωρήσουν τα μεταδεδομένα που έχουν προκύψει από τη διαδικασία της επισημείωσης αλλά και να κάνουν διορθώσεις σε αυτά, καθώς μπορεί να έχουν προκύψει λάθη από τη διαδικασία της αυτόματης επισημείωσης.

Η έννοια της ιεραρχικής δομής της Πύλης έγκειται στην απαίτηση ότι το σύστημα πλοήγησης πρέπει να επιτρέπει στον επισκέπτη να πλοηγηθεί γρήγορα στον ιστότοπο ξεκινώντας από οποιαδήποτε σελίδα. Αυτό επικεντρώνεται στην παροχή διαφορετικών όψεων των δεδομένων, παρέχοντας ταυτόχρονα το απαιτούμενο επίπεδο αφαίρεσης ανάλογα με το ρόλο του χρήστη. Για παράδειγμα, στο τρίτο επίπεδο εξουσιοδότησης, οι χρήστες θα μπορούσαν να ταξινομήσουν τους οργανισμούς που παρέχουν τα δεδομένα, δηλαδή τις πηγές δεδομένων. Με αυτόν τον τρόπο, παρέχεται στους χρήστες αποκλειστική πρόσβαση στα συνδυαστικά δεδομένα που παράγονται μέσω των υπηρεσιών της Πύλης.

5.5.5. Απαιτήσεις υποδομής (back-end)

Στην προηγούμενη ενότητα, εξετάστηκαν οι υπηρεσίες που παρέχονται από την προτεινόμενη Πύλη. Σε αυτήν την ενότητα θα αναφερθούν οι απαιτήσεις της προτεινόμενης προσέγγισης σε λειτουργίες υποδομής. Πέρα από τη βάση δεδομένων που υποστηρίζει το υποκείμενο CMS, όσον αφορά τα εξωτερικά δεδομένα της Πύλης, έχει επιλεγεί μια NoSQL προσέγγιση. Η NoSQL περιλαμβάνει μια μεγάλη ποικιλία διαφορετικών τεχνολογιών βάσεων δεδομένων που αναπτύχθηκαν ως απάντηση στην εκθετική αύξηση του διαθέσιμου όγκου δεδομένων, της συχνότητας πρόσβασης σε αυτά αλλά και την μείωση της επίδοσης των παραδοσιακών συστημάτων SQL σε αυτές τις συνθήκες [149]. Αρχικά, οι σχεσιακές βάσεις δεδομένων δεν σχεδιάστηκαν για να αντιμετωπίσουν τις προκλήσεις της κλίμακας και της ευελιξίας που δημιουργούν οι σύγχρονες εφαρμογές, ούτε κατασκευάστηκαν για να επεξεργαστούν τόσο μεγάλο όγκο δεδομένων. Αυτό μπορεί να οδηγήσει σε κακή επίδοση, καθιστώντας μερικές φορές το όλο το σύστημα ακόμα και αδύνατο να χρησιμοποιηθεί. Υπάρχουν αρκετές NoSQL προσεγγίσεις, όπως βάσεις δεδομένων εγγράφων, Graph stores, Key value stores ή Wide-column stores [165] που χρησιμοποιούνται συνήθως. Ένα από τα σημαντικότερα οφέλη των NoSQL βάσεων δεδομένων που τις καθιστούν ανώτερες σε σύγκριση με τις κλασσικές σχεσιακές είναι το γεγονός ότι είναι περισσότερο κλιμακώσιμες και έχουν καλύτερες επιδόσεις. Επιπλέον, το NoSQL μοντέλο δεδομένων αντιμετωπίζει αρκετά ζητήματα που το σχεσιακό μοντέλο δεν έχει σχεδιαστεί για να αντιμετωπίσει, όπως μεγάλους όγκους δομημένων, ημι-δομημένων ή και αδόμητων δεδομένων.

Δύο παραδείγματα συστημάτων που υποστηρίζουν την προαναφερθείσα προσέγγιση είναι οι CouchDB[42] και MongoDB[43]. Η CouchDB στοχεύει στη δημιουργία μιας νέας βάσης που μπορεί να αποθηκεύσει δεδομένα ως έγγραφα σε μορφή JSON. Αυτή η υλοποίηση παρέχει έναν μοναδικό τρόπο πρόσβασης σε έγγραφα και ερωτηματολόγια μέσω HTTP με τη χρήση ενός προγράμματος περιήγησης ιστού. Άλλες λειτουργίες παρεχόμενες από την CouchDB είναι η ευρετηρίαση των δεδομένων, ο συνδυασμός και η μετατροπή εγγράφων χρησιμοποιώντας JavaScript, μετασχηματισμό

[42] http://couchdb.apache.org/
[43] https://www.mongodb.com/x

εγγράφων και ειδοποιήσεις αλλαγών σε πραγματικό χρόνο. Από την άλλη πλευρά, η MongoDB, είναι μια άλλη βάση δεδομένων εγγράφων ανοιχτού κώδικα, που είναι ευέλικτη και κλιμακώσιμη, παρέχει υποστήριξη αποθήκευσης εγγράφων με πλήρη ευρετήριο, όπως και ευρετηρίαση με οποιοδήποτε χαρακτηριστικό θέλει ο χρήστης. Επιπλέον, παρέχει υπηρεσίες δημιουργίας αντιγράφων (replication) και υψηλή διαθεσιμότητα των δεδομένων, καθώς και ο αυτόματος διαμοιρασμός, η εκτέλεση ερωτημάτων και οι ενημερώσεις είναι μερικά από τα χαρακτηριστικά της MongoDB που την καθιστούν μοναδική. Τέλος, η MongoDB παρέχει μια υπηρεσία διαχείρισης, με στόχο την παρακολούθηση και την παροχή δυνατοτήτων δημιουργίας αντιγράφων ασφαλείας στη βάση δεδομένων.

Στην προσέγγισή μας, η CouchDB χρησιμεύει ως βάση δεδομένων για έγγραφα σε μορφή JSON καθώς υποστηρίζει εγγενώς τα δεδομένα σε αυτή τη μορφή, και η πλειονότητα των διαθέσιμων δεδομένων ήταν σε αυτή τη μορφή. Εναλλακτικά, θα μπορούσε να χρησιμοποιηθεί η MongoDB εάν τα διαθέσιμα σύνολα δεδομένων διατίθενται σε διαφορετικά μορφότυπα. Τα κριτήρια επιλογής σε αυτή την περίπτωση εξαρτώνται από τα διαθέσιμα δεδομένα.

5.6. Συμπεράσματα και μελλοντικές επεκτάσεις

Η προσέγγιση που παρουσιάστηκε σε αυτή την ενότητα στοχεύει στη σημασιολογική επεξεργασία και τη σύνδεση δεδομένων από διαφορετικές πηγές. Σχετικές προσεγγίσεις και μεθοδολογίες έχουν ληφθεί υπόψη από την αρχική σύλληψη της προτεινόμενης αρχιτεκτονικής, ώστε να ενσωματωθούν οι τεχνικές και οι αρχές του Σημασιολογικού Ιστού και των Ανοικτών Διασυνδεδεμένων Δεδομένων.

Η αρχιτεκτονική που παρουσιάστηκε σε αυτό το Κεφάλαιο της Διατριβής ορίζει τη βάση για περαιτέρω έρευνα. Ξεκινώντας από την ανάπτυξη της προσέγγισης που παρουσιάστηκε εδώ, τα μελλοντικά βήματα περιλαμβάνουν τη χρησιμοποίηση ανοικτών δεδομένων από την Ψηφιακή Βιβλιοθήκη του ΕΜΠ ή και άλλων για την πιλοτική φάση υλοποίησης. Οι ομάδες ενδιαφέροντος θα μπορούσαν να είναι φοιτητές, καθηγητές, ερευνητές και υπάλληλοι του ΕΜΠ., ομάδες που μπορεί να ενδιαφέρονται πραγματικά για την σημασιολογία που βρίσκεται πίσω από τα δεδομένα. Έτσι θα είναι δυνατόν να αξιολογηθεί η αποτελεσματικότητα της συγκεκριμένης προσέγγισης, και με στόχο την παροχή υπηρεσιών προστιθέμενης αξίας.

Ένα άλλο βήμα προς την κατεύθυνση αυτού του στόχου θα ήταν η χρήση άλλων CMS ως βάση δοκιμών για την αξιολόγηση της φορητότητας της προτεινόμενης προσέγγισης. Επιπλέον, προκειμένου να δοκιμαστεί πλήρως η ενσωμάτωση των αρχών των Ανοικτών Διασυνδεδεμένων Δεδομένων, είναι δυνατό να χρησιμοποιηθούν νέες οντολογίες, λεξιλόγια ή θησαυροί για την προώθηση της σύνδεσης δεδομένων με άλλες πηγές, διαθέσιμες στον Ιστό. Ακόμα, θα πρέπει να υλοποιηθεί η ενσωμάτωση των μηχανισμοί εξαγωγής πληροφορίας ταξινόμησης οι οποίοι αναλύθηκαν στα προηγούμενα κεφάλαια και η παροχή τους με τη μορφή υπηρεσιών στο σύστημα διαχείρισης δεδομένων. Τέλος, για την αξιολόγησή της μπορεί να είναι απαραίτητη η υλοποίηση μιας Πύλης στον ίδιο τομέα εφαρμογής με και χωρίς τη χρήση τεχνολογιών Σημασιολογικού Ιστού για να διερευνηθεί συγκριτικά η αποτελεσματικότητά τους.

6. Συμπεράσματα και μελλοντικές επεκτάσεις

Στο παρόν Κεφάλαιο συνοψίζονται οι άξονες έρευνας, τα κυριότερα συμπεράσματα, καθώς και οι μελλοντικές κατευθύνσεις της παρούσας Διατριβής.

Κατ' αρχήν, στο Κεφάλαιο 2 έγινε μια σύντομη εισαγωγή και επισκόπηση των μεθόδων αυτόματης εξαγωγής πληροφορίας από κείμενα και της συνακόλουθης ταξινόμησής τους, καθώς και των σχετικών εφαρμογών αυτών. Από την ανάλυση της σχετικής βιβλιογραφίας καταδεικνύεται σαφώς η αυξανόμενη ενσωμάτωση των εν λόγω μεθόδων σε ένα πλήθος εφαρμογών, όπως ο εμπλουτισμός μεταδεδομένων των τεκμηρίων, που μπορεί στη συνέχεια να υποβοηθήσουν την πλοήγηση και την αναζήτηση αυτών. Ένα ακόμα θέμα που αναλύθηκε στο Κεφάλαιο 2 ήταν ο τρόπος αξιολόγησης των μεθόδων, μια διαδικασία αρκετά δύσκολη (και σε ορισμένες περιπτώσεις ανέφικτη), λόγω της έλλειψης μιας καθιερωμένης μεθοδολογίας αξιολόγησης και σύγκρισης αυτών.

Κοινό συμπέρασμα το οποίο προκύπτει από την ανάλυση των υφιστάμενων μεθόδων στο υπό μελέτη πεδίο, είναι ότι οι περισσότερες αφορούν κυρίως έναν πολύ συγκεκριμένο τομέα εφαρμογών, χωρίς την παροχή συνοδευτικού λογισμικού, είτε ανοικτού είτε κλειστού κώδικα. Ακόμα, πολλές αποτελούν επιστημονικά ευρήματα χωρίς άμεση πρακτική εφαρμογή, ενώ ελάχιστες εξ' αυτών συντηρούνται ενεργά. Επιπρόσθετα, στην πλειονότητά τους οι εν λόγω μέθοδοι δεν είναι πλήρως αυτοματοποιημένες και απαιτούν τη χειροκίνητη εξαγωγή όρων από ειδικούς, κατά κύριο λόγο για τα δεδομένα εκπαίδευσης που χρησιμοποιούνται από τους αλγορίθμους ΜΜ. Ακόμα, πολλές μέθοδοι στα πλαίσια της αξιολόγησής τους τείνουν να συγκρίνουν τα αποτελέσματά τους με αυτά των ειδικών του αντίστοιχου πεδίου. Είναι προφανές πως ο ανθρώπινος παράγοντας είναι περισσότερο ικανός να εντοπίσει αντιπροσωπευτικούς όρους εντός ενός κειμένου σε σχέση με οποιαδήποτε αυτοματοποιημένη μέθοδο. Ως εκ τούτου, οι ημι-αυτόματες μέθοδοι τείνουν να δίνουν καλύτερα αποτελέσματα εξαγωγής σε σχέση με τις πλήρως αυτοματοποιημένες. Οι τελευταίες, βεβαίως, επειδή εμπεριέχουν λιγότερη χειρωνακτική εργασία, είναι λιγότερο χρονοβόρες και περισσότερο ανθεκτικές στα λάθη.

Στην πλειονότητα των προσεγγίσεων εφαρμόζονται μέθοδοι ΜΜ, οι οποίες συνδυάζονται και με άλλες τεχνικές, όπως οντολογίες ή ελεγχόμενα λεξιλόγια, στατιστικές αλλά και ευρετικές μεθόδους για να πετύχουν το επιθυμητό αποτέλεσμα. Με βάση την ανάλυση που πραγματοποιήθηκε, παρατηρήσαμε πως η εξαγωγή πληροφορίας σπάνια αποτελεί τον τελικό στόχο ενός συστήματος, αλλά τα αποτελέσματα της εξαγωγής και ο τρόπος εφαρμογής τους είναι που προσδίδουν προστιθέμενη αξία στην προσέγγιση, καθώς αυτά μπορούν να χρησιμοποιηθούν άμεσα από τους τελικούς χρήστες. Οι προσεγγίσεις που ανήκουν στη συγκεκριμένη κατηγορία και χρησιμοποιούν αποκλειστικά ευρετικές μεθόδους είναι αρκετά χρήσιμες μεν για τη μοντελοποίηση ενός συγκεκριμένου πεδίου, όμως έχουν το μειονέκτημα πως δεν μπορούν να εφαρμοστούν και σε άλλα πεδία χωρίς μικρότερες ή μεγαλύτερες τροποποιήσεις. Για παράδειγμα, προσεγγίσεις που εφαρμόζονται στο πλήρες κείμενο μιας επιστημονικής δημοσίευσης είτε σε κάποιο υποσύνολο αυτής, δεν μπορούν να εφαρμοστούν με τον ίδιο τρόπο σε άλλους τύπους δεδομένων, όπως άρθρα ειδήσεων ή άλλο περιεχόμενο. Πολλές φορές οι προσεγγίσεις εξαγωγής επιλέγονται με γνώμονα το αν υπάρχουν διαθέσιμα και μπορούν να χρησιμοποιηθούν κάποια λεξιλόγια, οντολογίες ή δεδομένα θεματικών επικεφαλίδων. Ένα σημαντικό κριτήριο είναι η ποιότητα και η δυνατότητα επαναχρησιμοποίησης των παραγόμενων όρων που είναι και το ζητούμενο σε τέτοιες εφαρμογές. Αντίθετα, οι προσεγγίσεις ανάθεσης που αναλύθηκαν στο Κεφάλαιο 2 τείνουν να κατηγοριοποιούν

τις φράσεις σε δύο βασικές ομάδες, με βάση το εάν ο υποψήφιος όρος είναι ή όχι αντιπροσωπευτικός για ένα κείμενο. Βασική πρόκληση σε αυτή την κατεύθυνση αποτελεί είτε η διεύρυνση αυτών των ομάδων με σκοπό τη βελτιστοποίηση της διαδικασίας ή την εξαγωγή των διαφορετικών ομάδων με αυτόματο τρόπο, καθώς στην πλειονότητα των προσεγγίσεων αυτές παρέχονται έπειτα από εμπειρική παρατήρηση. Επίσης, σε αυτή την κατηγορία εντοπίζεται έλλειψη συστημάτων που να μπορούν να χρησιμοποιήσουν ουσιαστικά τις προτεινόμενες προσεγγίσεις, όπως για παράδειγμα μέσα από την παροχή μιας διεπαφής, την οποία ακόμα και μη ειδικοί θα μπορούσαν να χρησιμοποιήσουν με σκοπό την ανακάλυψη νέας γνώσης.

Στο Κεφάλαιο 3 παρουσιάστηκε η εργασία της αυτόματης ταξινόμησης ενός συνόλου δεδομένων πολλαπλής ετικέτας με χρήση αυτο-οργανούμενων χαρτών. Εκεί, ερευνήσαμε τη χρήση τετραγωνικών και ορθογωνικών SOMs για την αξιολόγηση της ικανότητας ταξινόμησης του συγκεκριμένου Νευρωνικού Δικτύου με εφαρμογή σε ένα σύνολο δεδομένων ειδήσεων. Βασικός πυλώνας της προσέγγισης αποτέλεσε η διαδικασία επιλογής χαρακτηριστικών, η οποία σχεδιάστηκε να είναι όσο το δυνατόν ελαφρύτερη και γενικότερη, προκειμένου να εξασφαλιστεί η δυνατότητα εφαρμογής της σε άλλα σύνολα δεδομένων με μικρές τροποποιήσεις. Επίσης, δείξαμε ότι με τη χρήση της προτεινόμενης διαδικασίας επιλογής χαρακτηριστικών μπορούμε να εντοπίσουμε τα πιο σημαντικά χαρακτηριστικά του συνόλου δεδομένων, διατηρώντας παράλληλα το μέγεθος των διανυσμάτων σχετικά μικρό, βελτιώνοντας έτσι την αποδοτικότητα της ταξινόμησης. Χρησιμοποιώντας τον αλγόριθμο SOM, μπορέσαμε να δείξουμε ότι τα έγγραφα που ανήκουν στην ίδια ή παρόμοια (από άποψη περιεχομένου) κατηγορία τοποθετούνται σε γειτονικούς κόμβους του χάρτη. Αυτό δικαιολογεί την επιλογή του SOM έναντι άλλων παραδοσιακών προσεγγίσεων, καθώς το SOM είναι ικανό να απεικονίσει τέτοιες σχέσεις εγγενώς.

Με την επιλογή ενός μη εποπτευόμενου Νευρωνικού Δικτύου (όπως είναι το SOM) αποφύγαμε την μετατροπή του προβλήματος σε ένα σύνολο ανεξάρτητων δυαδικών ταξινομητών, που δεν θα λάμβαναν υπόψη τη συσχέτιση μεταξύ των ετικετών, όπως συμβαίνει με άλλες υφιστάμενες προσεγγίσεις. Επίσης, μετατρέψαμε το πρόβλημα ταξινόμησης πολλαπλής ετικέτας σε πρόβλημα πολλαπλών κλάσεων, χρησιμοποιώντας μια προσέγγιση διαφορετικών σταδίων για την κατασκευή των διανυσμάτων έτσι ώστε να μπορούν να διαχειριστούν από το SOM. Μετά την κατασκευή των διανυσμάτων, κάθε αντικείμενο του συνόλου δεδομένων που είχε πολλαπλές ετικέτες χωρίστηκε σε περισσότερα αντικείμενα, καθένα από τα οποία έχει μία ετικέτα από το αρχικό σύνολο ετικετών, διατηρώντας όμως το τμήμα του διανύσματος άθικτο. Αυτό είχε ως αποτέλεσμα τη διατήρηση όλης της αρχικής πληροφορίας που αφορούσε τα διανύσματα χαρακτηριστικών και τις ετικέτες που είχαν ανατεθεί αρχικά σε κάθε ένα από αυτά.

Μια ακόμα συνεισφορά του Κεφαλαίου 3 είναι ο αλγόριθμος INterSECT για την επιλογή των ετικετών που ανατίθενται σε κάθε κόμβο του χάρτη, ο οποίος λαμβάνει υπόψη όχι μόνο τις ετικέτες που υπάρχουν στον ίδιο τον κόμβο αλλά και τις ετικέτες που έχουν ανατεθεί μεταξύ της BMU και των γειτονικών κόμβων αυτής, δίνοντας ως έξοδο την τομή των συνόλων. Με βάση τα αποτελέσματα που συγκεντρώθηκαν φαίνεται ότι ο INterSECT παράγει καλύτερα αποτελέσματα ειδικότερα όμως στις βαθμολογίες F1 Μάκρο-μέσου όρου. Επιπλέον, χρησιμοποιώντας μια ευρετική διαδικασία, για τον υπολογισμό του μεγέθους του SOM, που περιγράφεται αναλυτικά στο Κεφάλαιο αυτό, προκρίνεται η επιλογή ορθογωνικών πλεγμάτων. Τα αποτελέσματα που ελήφθησαν υποδηλώνουν ότι η χρήση ορθογωνικών πλεγμάτων μπορεί να έχει καλύτερη απόδοση σε σύγκριση με τα τετραγωνικά πλέγματα για το εν λόγω σύνολο δεδομένων και τα μεγέθη διανυσμάτων με τα οποία

πειραματιστήκαμε. Στις δοκιμές που πραγματοποιήθηκαν παρατηρήθηκε πως τα ορθογωνικά πλέγματα υπερτερούν σε απόδοση σε σχέση με τα τετραγωνικά για αντίστοιχο σύνολο κόμβων.

Επιπλέον, το σχήμα στάθμισης που χρησιμοποιήθηκε για τη διαδικασία επιλογής χαρακτηριστικών βασίστηκε σε στατιστικά λέξεων και ειδικότερα στο σχήμα στάθμισης TF-IDF. Στα πειράματα που διενεργήθηκαν στο επιλεγμένο σύνολο δεδομένων (Reuters), τα συμφραζόμενα είτε δεν χρησιμοποιούνταν καθόλου είτε χρησιμοποιήθηκε ένα μικρό παράθυρο γύρω από κάθε λέξη εύρους 1-2 λέξεων. Δεδομένου ότι το Reuters είναι ένα σύνολο ειδήσεων, η χρήση ενός εκτεταμένου παραθύρου γύρω από μια λέξη για την εύρεση των συμφραζόμενων δεν δικαιολογείται, καθώς τα ειδησεογραφικά στοιχεία έχουν περιορισμένο μήκος και τα συμφραζόμενα διαφέρουν σημαντικά από στοιχείο σε στοιχείο για να μπορέσει να επωφεληθεί η προσέγγιση από ένα τέτοιο παράθυρο. Τα αποτελέσματα που παρουσιάστηκαν υποστηρίζουν περαιτέρω αυτήν την ένδειξη, καθώς η χρήση των συμφραζόμενων λέξεων δεν επηρεάζει σημαντικά την απόδοση της ταξινόμησης. Αξίζει να τονιστεί ότι η απόδοση ταξινόμησης βελτιώνεται όταν χρησιμοποιούνται 1-grams (δηλαδή χωρίς τη χρήση συμφραζομένων) για κατηγορίες που έχουν μικρό αριθμό δειγμάτων.

Η προτεινόμενη προσέγγιση δοκιμάστηκε εκτενώς υπό διαφορετικές ρυθμίσεις διαμόρφωσης χρησιμοποιώντας το υποσύνολο Mod’ Apte των δεδομένων Reuters, ένα από τα πιο συχνά χρησιμοποιούμενα για τη συγκριτική αξιολόγηση μεθόδων ταξινόμησης πολλαπλής ετικέτας. Από το σύνολο δεδομένων χρησιμοποιήθηκαν τόσο οι 90 όσο και οι 10 κλάσεις με τα περισσότερα δείγματα. Επιτεύχθηκε αύξηση στην ακρίβεια ταξινόμησης κατά 10% για τις κατηγορίες με μικρό αριθμό δειγμάτων, σε σχέση με άλλες προσεγγίσεις, και καλύτερη απόδοση από άποψη χρόνου και πόρων για την εκπαίδευση του μοντέλου. Εκτός αυτού, χρησιμοποιήθηκε λεξιλόγιο με μέγεθος που κυμαίνεται από 250 έως 3000 και αντίστοιχες διαστάσεις ανά διάνυσμα (περίπου 34 φορές μικρότερο σε σχέση με το βασικό σενάριο σύγκρισης), καθώς μεγαλύτερα μεγέθη διανυσμάτων θα προκαλούσαν σημαντικά αυξημένες απαιτήσεις χρόνου υπολογισμού και επεξεργασίας. Ωστόσο, πετύχαμε συγκρίσιμα αποτελέσματα όταν χρησιμοποιήσαμε όλες τις (90) κλάσεις για τις βαθμολογίες Μίκρο-μέσου όρου και καλύτερα αποτελέσματα στις βαθμολογίες Μάκρο-μέσου όρου.

Επιπλέον, χρησιμοποιήσαμε την batch έκδοση του SOM που βάση της βιβλιογραφικής ανασκόπησης που πραγματοποιήθηκε είναι πιο σταθερή και τείνει να συγκλίνει ταχύτερα. Μια άλλη διαφορά σε σχέση με το βασικό σενάριο σύγκρισης έγκειται στη διαδικασία επιλογής της ετικέτας. Στο βασικό σενάριο δοκιμάστηκε η μέθοδος Majority Voting, η οποία μετρά το πλήθος των ετικετών που έχουν ανατεθεί σε κάθε κόμβο για κάθε κλάση, και στη συνέχεια αναθέτει στον κόμβο την πιο συχνή ετικέτα. Η προσέγγιση αυτή μειονεκτεί διότι τελικά ανατίθεται σε κάθε κόμβο μόνο μια ετικέτα, ενώ το πρόβλημα αναφέρεται σε αυτόματη ταξινόμηση πολλαπλής ετικέτας και δεύτερον λόγω του ότι δεν λαμβάνονται υπόψη για τον υπολογισμό της ετικέτας που θα ανατεθεί σε κάθε κόμβο, οι γειτονικοί του κόμβοι επάνω στο χάρτη. Εξαίρεση αποτελεί η περίπτωση στην οποία κατά τη διαδικασία εκπαίδευσης του SOM δεν έχει ανατεθεί στον κόμβο καμία ετικέτα, Σε αυτή την περίπτωση εξετάζονται και οι γειτονικοί κόμβοι για να καθοριστεί η ετικέτα του κόμβου. Για να ξεπεράσουμε αυτή την αδυναμία σχεδιάσαμε και υλοποιήσαμε τον δικό μας αλγόριθμο για την επιλογή ετικετών με βάση τις ετικέτες που ανατίθενται όχι μόνο στον κόμβο αλλά και στους γειτονικούς του κόμβους, αναθέτοντας όμως όλες τις ετικέτες που δίνει σαν έξοδο ο αλγόριθμος.

Στο πλαίσιο των μελλοντικών ερευνητικών κατευθύνσεων του Κεφαλαίου 3, προτάθηκε η χρήση και άλλων συνόλων δεδομένων, ώστε να καταδειχθεί η γενίκευση της προτεινόμενης προσέγγισης στο σύνολό της, καθώς και ο πειραματισμός με άλλους αλγορίθμους με εποπτευόμενης

Μηχανικής Μάθησης ή Βαθιάς Μάθησης, προκειμένου να εκτιμηθεί η απόδοσή τους σε σύγκριση με το SOM υπό τις ίδιες πειραματικές συνθήκες. Η ερευνητική αυτή επέκταση υλοποιήθηκε στο **Κεφάλαιο 4** της παρούσας Διατριβής για την **αυτόματη ταξινόμηση πολλαπλών κλάσεων μιας συλλογής ηλεκτρονικών βιβλίων με χρήση του πίνακα περιεχομένων των βιβλίων**. Πιο συγκεκριμένα, στο Κεφάλαιο αυτό ερευνήθηκε η εφαρμογή ενός Νευρωνικού Δικτύου μη εποπτευόμενης ΜΜ (SOM) και δύο αρχιτεκτονικών DNN (συγκεκριμένα ενός LSTM και ενός CNN+LSTM), έτσι ώστε να εκτιμηθεί η ικανότητα ταξινόμησης των μεθόδων αυτών σε ένα σενάριο πολλαπλών κλάσεων. Η προτεινόμενη προσέγγιση εφαρμόστηκε με διαφορετικές ρυθμίσεις διαμόρφωσης σε ένα σύνολο πολυ-επιστημονικών τίτλων ηλεκτρονικών βιβλίων του εκδοτικού οίκου Springer (περίπου 56000 τον αριθμό) με χρήση των πινάκων περιεχομένων. Η διαδικασία επιλογής χαρακτηριστικών είχε την ίδια βάση με αυτήν που αναλύθηκε στο Κεφάλαιο 3 και επεκτάθηκε στο Κεφάλαιο 4 για να καλύψει τις ανάγκες των δυο αρχιτεκτονικών DNN που χρησιμοποιήθηκαν για σύγκριση. Κοινός τόπος των δύο αυτών διαδικασιών ήταν ο απλός σχεδιασμός τους έτσι ώστε να διασφαλιστεί η εφαρμογή τους σε άλλα σύνολα δεδομένων με μικρές τροποποιήσεις, αποφεύγοντας έτσι την εκτεταμένη κατασκευή χαρακτηριστικών, μια διαδικασία χρονοβόρα, που απαιτεί την εκτέλεσή της από ειδικούς στο εκάστοτε πεδίο αλλά και είναι δύσκολα εφαρμόσιμη αν το υποκείμενο σύνολο μεταβληθεί.

Η διαδικασία της κατασκευής διανυσμάτων για την αναπαράσταση των πινάκων περιεχομένων αξιοποιεί τις πληροφορίες που εξάγονται από τους τελευταίους χρησιμοποιώντας το σχήμα στάθμισης TF-IDF, στην περίπτωση του SOM, και τον Keras Tokenizer, στην περίπτωση των DNN αρχιτεκτονικών. Συμπερασματικά, στο Κεφάλαιο αυτό δείξαμε ότι με τη χρήση της προτεινόμενης διαδικασίας επιλογής χαρακτηριστικών επιτυγχάνεται η αποδοτική αναπαράσταση των πινάκων περιεχομένων των βιβλίων καθώς η μέθοδος μπορεί να συλλάβει τα πιο σημαντικά χαρακτηριστικά του συνόλου δεδομένων, διατηρώντας παράλληλα το μέγεθος των διανυσμάτων σχετικά μικρό. Επιπλέον, χρησιμοποιήθηκε και πάλι ο αλγόριθμος INterSECT για την επιλογή ετικετών, με σκοπό να ερευνηθεί κατά πόσο μπορεί να χρησιμοποιηθεί με εξίσου καλά αποτελέσματα σε ένα σενάριο ταξινόμησης πολλαπλών κλάσεων, αντί για ταξινόμηση πολλαπλής ετικέτας. Ο αλγόριθμος Majority Voting εφαρμόστηκε επίσης ως εναλλακτική μέθοδος επιλογής ετικέτας για κάθε κόμβο στην περίπτωση του SOM. Από τα αποτελέσματα των πειραμάτων που διεξήχθησαν έγινε φανερό πως ο αλγόριθμος Majority Voting είναι προτιμητέος για το συγκεκριμένο πρόβλημα, καθώς παρουσιάζει 2-φορές καλύτερη απόδοση σε σχέση με τον INterSECT.

Βασικό συμπέρασμα του Κεφαλαίου 4 είναι ότι η προσέγγιση που αναλύθηκε στο Κεφάλαιο 3 γενικεύεται, καθώς μπόρεσε να εφαρμοστεί εκ νέου σε ένα διαφορετικό σύνολο δεδομένων με μικρές μόνο αλλαγές. Σε αυτό το Κεφάλαιο επεκτείναμε επίσης την προσέγγιση με την εφαρμογή δύο διαφορετικών αρχιτεκτονικών Νευρωνικών Δικτύων Βαθιάς Μάθησης, έτσι ώστε να εκτιμήσουμε την απόδοση ταξινόμησής τους σε σύγκριση τόσο μεταξύ τους όσο και με τη μέθοδο SOM. Τα αποτελέσματα δείχνουν ότι οι δύο αρχιτεκτονικές DNN είναι εξίσου αποτελεσματικές στην ταξινόμηση ενός συνόλου δεδομένων πολλαπλών κλάσεων χρησιμοποιώντας τους πίνακες περιεχομένων των ηλεκτρονικών βιβλίων, με ελάχιστη παραμετροποίηση. Σε σύγκριση με το SOM, επιτυγχάνουν καλύτερα αποτελέσματα (κατά 5% περίπου). Όσον αφορά στους απαιτούμενους πόρους, τα Νευρωνικά Δίκτυα Βαθιάς Μάθησης χρειάζονται πολύ περισσότερη υπολογιστική ισχύ και μνήμη, είναι όμως ταχύτερα από το SOM όταν εκτελούνται σε GPU. Τέλος, με δεδομένο ότι η αρχιτεκτονική CNN + LSTM είναι πιο περίπλοκη σε σχέση με το αμφίδρομο LSTM, καταλήγουμε πως το τελευταίο είναι προτιμητέο για το συγκεκριμένο πρόβλημα.

Συμπερασματικά, εφαρμόζοντας τη μεθοδολογία που παρουσιάστηκε αναλυτικά στο Κεφάλαιο 4 της παρούσας Διατριβής αξιοποιήσαμε τις πληροφορίες από τους πίνακες περιεχομένων των βιβλίων, σε αντίθεση με τις υφιστάμενες στη βιβλιογραφία μεθόδους οι οποίες συνήθως κάνουν χρήση των περιλήψεων, των τίτλων και των υπότιτλων ή ακόμα και του πλήρους κειμένου για την εξαγωγή χαρακτηριστικών. Με χρήση της προτεινόμενης προσέγγισης δείξαμε ότι μπορούμε να προσδιορίσουμε ποια βιβλία πραγματεύονται παρόμοια θέματα, παρόλο που ενδέχεται να μην μοιράζονται τις ίδιες λέξεις-κλειδιά στους πίνακες περιεχομένων τους ή ακόμα και την ίδια θεματική ετικέτα. Ένα ακόμα θέμα που αναδείχθηκε στο Κεφάλαιο αυτό είναι ότι σε καμία περίπτωση δεν περιμένουμε ακρίβεια ταξινόμησης 100%. Αυτό οφείλεται στο γεγονός πως η διαδικασία επιλογής ετικέτας για κάθε βιβλίο είναι υποκειμενική καθώς βασίζεται στον ανθρώπινο παράγοντα (διαφορετικοί άνθρωποι μπορεί να δώσουν στο ίδιο βιβλίο διαφορετική θεματική ετικέτα). Επίσης, υπάρχουν στενά σχετιζόμενες ή αλληλεπικαλυπτόμενες κατηγορίες και, παρόλο που το σύστημα μπορεί να αποδώσει σε ένα βιβλίο ετικέτα της μιας κατηγορίας, αυτό μπορεί να ανήκει στην άλλη.

Τέλος, στο Κεφάλαιο 4 παρουσιάστηκαν διάφορες **μελλοντικές κατευθύνσεις** για την επέκταση και βελτίωση της προσέγγισης. Ένα βήμα προς αυτή την κατεύθυνση θα μπορούσε να είναι η εύρεση όχι μόνο της γενικής κατηγορίας αλλά και της «ειδικότερης» υποκατηγορίας με βάση την γενική. Αυτό προϋποθέτει την ύπαρξη ιεραρχίας στις ετικέτες που θα χρησιμοποιηθούν ή τη χρήση ενός ελεγχόμενου λεξιλογίου. Αυτό θα μπορούσε να επιτευχθεί σαν μια διαδικασία 2 σταδίων, όπου αρχικά βρίσκεται η γενική ετικέτα και στην συνέχεια, με χρήση την ιεραρχίας και μέσω της επανεκπαίδευσης των δικτύων με μόνο δείγματα της συγκεκριμένης κατηγορίας, εκχωρούνται στα δείγματα οι ειδικότερες κατηγορίες (υπό-κατηγορίες) ως ετικέτες.

Επιπλέον, τα Siamese Νευρωνικά Δίκτυα ή η μέθοδος Contrastive Supervised Learning θα μπορούσαν να αξιοποιηθούν για την αντιμετώπιση του προβλήματος της ανισορροπίας δειγμάτων μεταξύ των κλάσεων. Μια ακόμα επέκταση της προσέγγισης που παρουσιάστηκε στο Κεφάλαιο 4 θα μπορούσε να είναι η ενσωμάτωση προ εκπαιδευμένων γλωσσικών μοντέλων όπως οι Transformers ή τα BERT, XLNet και GPT-2. Οι προσεγγίσεις αυτές με βάση και την βιβλιογραφική ανασκόπηση που πραγματοποιήθηκε, χαίρουν της ευρείας αποδοχής της ερευνητικής κοινότητας λόγω του ότι με χρήση τέτοιων προσεγγίσεων, αξιοποιούνται οι ήδη εκπαιδευμένες ενσωματώσεις διανυσμάτων αντί να διεξάγονται δαπανηρές εργασίες εκπαίδευσης από το μηδέν. Και αυτά μπορούν να χρησιμοποιηθούν εναλλακτικά για να δώσουν λύση στο πρόβλημα της ανισορροπίας των δειγμάτων μεταξύ των κλάσεων. Ακόμα, ένας παράγοντας που τα κάνει κατάλληλα για ενσωμάτωση στη συγκεκριμένη προσέγγιση είναι η δυνατότητα μεταφοράς μάθησης από εποπτευόμενα δεδομένα. Τέλος, από όσο γνωρίζουμε, οι Transformers δεν έχουν ακόμη εφαρμοστεί σε συστήματα προτάσεων περιεχομένου. Ως εκ τούτου, αυτή θα ήταν μια ενδιαφέρουσα επέκταση για την αξιολόγηση της δυνατότητας εφαρμογής τους στην ταξινόμηση πολλαπλών κλάσεων, χρησιμοποιώντας τους πίνακες περιεχομένων των βιβλίων.

Στο Κεφάλαιο 5 παρουσιάστηκαν άλλα συναφή ερευνητικά θέματα, τα οποία είχαν εξεταστεί στο πλαίσιο της παρούσας Διατριβής και αφορούσαν τη σημασιολογική επεξεργασία και τη σύνδεση δεδομένων από διαφορετικές πηγές. Αυτό πραγματοποιήθηκε με τη σχεδίαση και υλοποίηση μιας Πύλης Διασυνδεμένων Δεδομένων (Linked Data Portal) με χρήση των τεχνολογιών του Σημασιολογικού Ιστού επάνω σε ένα Σύστημα Διαχείρισης Περιεχομένου (CMS). Η Πύλη αυτή μπορεί να ενσωματώσει τους μηχανισμούς εξαγωγής πληροφορίας ταξινόμησης που αναπτύχθηκαν στα προηγούμενα Κεφάλαια και των αποτελεσμάτων αυτών, με απώτερο στόχο τον εμπλουτισμό των μεταδεδομένων, για αποτελεσματικότερη αναζήτηση και ανάκτηση πληροφοριών από τους τελικούς

χρήστες μιας ψηφιακής βιβλιοθήκης. Πραγματοποιήθηκε εκτεταμένη βιβλιογραφική ανασκόπηση σχετικά με τις τεχνολογίες Σημασιολογικού Ιστού αλλά και τους τρόπους οργάνωσης της πληροφορίας που βρίσκεται διάσπαρτη στον Ιστό, με σκοπό να μπορούν οι εφαρμογές και κατ' επέκταση οι τελικοί χρήστες τους να επωφεληθούν από τις τεχνολογίες αυτές.

Βασικός στόχος της προσέγγισης που αναλύθηκε στο Κεφάλαιο 5 είναι η σημασιολογική επεξεργασία και διάθεση δεδομένων από διαφορετικές πηγές. Οι πηγές αυτές μπορεί να προέρχονται από εντελώς διαφορετικά πεδία ενώ το προσδοκώμενο αποτέλεσμα της διαδικασίας είναι ο συνδυασμός δεδομένων από τις πηγές, για την παροχή υπηρεσιών προστιθέμενης αξίας. Με χρήση των μεθόδων εξαγωγής ποιοτικής πληροφορίας που αναπτύχθηκαν, η πληροφορία που εξάγεται μπορεί να χρησιμοποιηθεί αποδοτικά κατά την αναζήτηση ή και την πλοήγηση μέσω της Πύλης. Με χρήση των μηχανισμών αυτόματης ή ημιαυτόματης εξαγωγής πληροφορίας ταξινόμησης μειώνεται η ανάγκη χειροκίνητης επισημείωσης του περιεχομένου από ειδικούς στο αντικείμενο, μιας διαδικασίας χρονοβόρας και μη εφικτής όσο το περιεχόμενο της Πύλης αυξάνεται. Οι λέξεις ή οι φράσεις που εξάγονται μπορούν ακόμα να χρησιμοποιηθούν για λόγους ευρετηρίασης στην Πύλη. Σε αυτή την περίπτωση ένας χρήστης είναι δυνατόν να βρει την πληροφορία που καλύπτει τις ανάγκες του μέσα σε μικρό χρονικό διάστημα.

Επίσης, είναι δυνατή η χρήση της λίστας λέξεων/φράσεων κατά τη διάρκεια μιας αναζήτησης. Οι φράσεις αυτές μπορούν να αποτελέσουν την είσοδο μιας μηχανής αναζήτησης αντί για μεμονωμένες λέξεις κλειδιά έτσι ώστε η τελευταία να επιστέψει πιο σχετικά ή/και πιο ακριβή αποτελέσματα. Τα αποτελέσματα της αναζήτησης μπορούν να βελτιστοποιηθούν περαιτέρω εάν η λίστα αυτή συνδυαστεί (ή κάποιες φράσεις που ανήκουν στη λίστα αντιστοιχιστούν) με όρους που προέρχονται από κάποιο ελεγχόμενο λεξιλόγιο. Έτσι μπορεί να καταστεί δυνατή η ανάκτηση περισσότερο σχετικών ή πιο γενικών ή ακόμα και πιο ειδικών αποτελεσμάτων, λόγω του ότι τα ελεγχόμενα λεξιλόγια ή οι θησαυροί ενσωματώνουν συνήθως ένα είδος ιεραρχίας. Απόρροια αυτού θα είναι μια μικρότερη σε μέγεθος και ταυτόχρονα πιο σχετική λίστα αποτελεσμάτων με περισσότερο ποιοτικά αποτελέσματα. Τέλος, η πληροφορία που εξάγεται θα μπορούσε να ενσωματωθεί στην Πύλη με τη μορφή φίλτρων τα οποία θα μπορούσαν να χρησιμοποιηθούν από τους χρήστες για την μείωση των αποτελεσμάτων που επιστρέφονται στα πλαίσια μιας αναζήτησης.

Το προτεινόμενο σύστημα συνδυάζει διαφορετικές υπηρεσίες που έχουν ως σκοπό την παροχή προβολών επάνω στα δεδομένα τόσο για ειδικούς όσο και μη ειδικούς χρήστες. Έτσι, παρέχεται η δυνατότητα στους χρήστες να δημιουργήσουν τις δικές τους όψεις στα δεδομένα, μια διαφοροποίηση σε σχέση με τις κλασσικές προσεγγίσεις που είτε παρέχουν προβολές των συνδυασμένων δεδομένων είτε απλώς παρέχουν μια διεπαφή αναζήτησης. Προκειμένου να αξιολογηθεί η αποτελεσματικότητα της προσέγγισης αυτής, και με στόχο την παροχή υπηρεσιών προστιθέμενης αξίας, θα μπορούσαν να χρησιμοποιηθούν άλλα σύνολα ανοικτών δεδομένων από Ψηφιακές Βιβλιοθήκες.

Η εργασία που παρουσιάστηκε και σε αυτό το Κεφάλαιο (5) της Διατριβής ορίζει τη βάση για περαιτέρω έρευνα. Ένα βήμα θα ήταν η χρήση άλλων CMS ως βάση δοκιμών για την αξιολόγηση της προτεινόμενης προσέγγισης. Επιπλέον, προκειμένου να δοκιμαστεί πλήρως η ενσωμάτωση των αρχών των Διασυνδεδεμένων Ανοικτών Δεδομένων, είναι δυνατό να χρησιμοποιηθούν νέες οντολογίες, λεξιλόγια ή θησαυροί για την προώθηση της σύνδεσης δεδομένων με άλλες πηγές, διαθέσιμες στον Ιστό. Τέλος, για την αξιολόγησή της μπορεί να είναι απαραίτητη η υλοποίηση μιας

πύλης στον ίδιο τομέα εφαρμογής με και χωρίς τη χρήση τεχνολογιών Σημασιολογικού Ιστού για να διερευνηθεί κατά πόσο μπορούν να εξαχθούν συγκεκριμένα συμπεράσματα.

Αναφορές

[1] Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts, 205-227.

[2] Hasan, K. S., & Ng, V. (2014, June). Automatic Keyphrase Extraction: A Survey of the State of the Art. In ACL (1) (pp. 1262-1273).

[3] Tan, Y. F., Kan, M. Y., & Lee, D. (2006, June). Search engine driven author disambiguation. In Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (pp. 314-315). ACM.

[4] Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., & Fox, E. A. (2003, May). Automatic document metadata extraction using support vector machines. In Digital Libraries, 2003. Proceedings. 2003 Joint Conference on (pp. 37-48). IEEE.

[5] Feather, J., & Sturges, P. (Eds.). (2003). International encyclopedia of information and library science. Routledge.

[6] Turney, P. (1999). Learning to extract keyphrases from text.

[7] Barker, K., & Cornacchia, N. (2000). Using noun phrase heads to extract document keyphrases. In Advances in Artificial Intelligence (pp. 40-52). Springer Berlin Heidelberg.

[8] Mahoui, M. (2000). Hierarchical document clustering using automatically extracted Keyphrase. In proceedings of the third international Asian conference on digital libraries (pp. 113-20).

[9] Hulth, A. (2004). Combining machine learning and natural language processing for automatic keyword extraction. Department of Computer and Systems Sciences [Institutionen för Data-och systemvetenskap], Univ.

[10] Bdewilde.github.io. (2017). Intro to Automatic Keyphrase Extraction. [online] Available at: http://bdewilde.github.io/blog/2014/09/23/intro-to-automatic-keyphrase-extraction/ [Accessed 7 Feb. 2017].

[11] Liu, Z., Li, P., Zheng, Y., & Sun, M. (2009, August). Clustering to find exemplar terms for keyphrase extraction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1 (pp. 257-266). Association for Computational Linguistics.

[12] Zimmermann, A., Lopes, N., Polleres, A., & Straccia, U. (2012). A general framework for representing, reasoning and querying with annotated semantic web data. Web Semantics: Science, Services and Agents on the World Wide Web, 11, 72-95.

[13] Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999, August). KEA: Practical automatic Keyphrase extraction. In Proceedings of the fourth ACM conference on Digital libraries (pp. 254-255). ACM.

[14] Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999, July). Domain-specific Keyphrase extraction. In IJCAI (Vol. 99, pp. 668-673).

[15] Park, Y., Byrd, R. J., & Boguraev, B. K. (2002, August). Automatic glossary extraction: beyond terminology identification. In Proceedings of the 19th international conference on Computational Linguistics-Volume 1 (pp. 1-7). Association for Computational Linguistics.

[16] Nguyen, T. D., & Kan, M. Y. (2007). Keyphrase extraction in scientific publications. In Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers (pp. 317-326). Springer Berlin Heidelberg.

[17] Kim, S. N., & Kan, M. Y. (2009, August). Re-examining automatic Keyphrase extraction approaches in scientific articles. In Proceedings of the workshop on multiword expressions: Identification,

interpretation, disambiguation and applications (pp. 9-16). Association for Computational Linguistics.
[18] Turney, P. (2003). Coherent Keyphrase extraction via web mining.
[19] Hulth, A., & Megyesi, B. B. (2006, July). A study on automatically extracted keywords in text categorization. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 537-544). Association for Computational Linguistics.
[20] Mendelyan, O., & Witten, I. H. (2006, June). Thesaurus based automatic Keyphrase indexing. In Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (pp. 296-297). ACM.
[21] Sugiyama, K., Kumar, T., Kan, M. Y., & Tripathi, R. C. (2010, March). Identifying citing sentences in research papers using supervised learning. In Information Retrieval & Knowledge Management, (CAMP), 2010 International Conference on (pp. 67-72). IEEE.
[22] Luong, M. T., Nguyen, T. D., & Kan, M. Y. (2012). Logical structure recovery in scholarly articles with rich document features. Multimedia Storage and Retrieval Innovations for Digital Library Systems, 270.
[23] Esser, D., Schuster, D., Muthmann, K., Berger, M., & Schill, A. (2012, January). Automatic indexing of scanned documents: a layout-based approach. In IS&T/SPIE Electronic Imaging (pp. 82970H-82970H). International Society for Optics and Photonics.
[24] Do, H. H. N., Chandrasekaran, M. K., Cho, P. S., & Kan, M. Y. (2013, July). Extracting and matching authors and affiliations in scholarly documents. In Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries (pp. 219-228). ACM.
[25] Lin, S., Ng, J. P., Pradhan, S., Shah, J., Pietrobon, R., & Kan, M. Y. (2010, June). Extracting formulaic and free text clinical research articles metadata using conditional random fields. In Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents (pp. 90-95). Association for Computational Linguistics.
[26] Mendelyan, O. (2005). Automatic Keyphrase indexing with a domain-specific thesaurus. Master's thesis, Albert-Ludwigs University.
[27] Wan, X., & Xiao, J. (2008, August). CollabRank: towards a collaborative approach to single-document Keyphrase extraction. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 (pp. 969-976). Association for Computational Linguistics.
[28] S.N. Kim, T. Baldwin, M.Y. Kan, „The use of topic representative words in text categorization," in *Proc. of the 14th Australasian Document Computing Symposium (ADCS)*, Sydney, Australia, 2009, pp. 75–81.
[29] Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3), 130-137.
[30] Lovins, J. B. (1968). Development of a stemming algorithm (p. 65). Cambridge: MIT Information Processing Group, Electronic Systems Laboratory.
[31] VRL, N. (2009, December). An unsupervised approach to domain-specific term extraction. In Australasian Language Technology Association Workshop 2009 (p. 94).
[32] Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.
[33] Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955.
[34] Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Advances in neural information processing systems (pp. 2177-2185).
[35] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*(Aug), 2493-2537.

[36] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

[37] Baroni, M., Dinu, G., & Kruszewski, G. (2014, June). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)* (pp. 238-247).

[38] Bosnić, I., Verbert, K., & Duval, E. (2010). Automatic keywords extraction–a basis for content recommendation. In Workshop Proceedings (p. 51).

[39] L. Bungum and B. Gambäck, "Self-Organizing Maps for Classification of a Multi-Labelled Corpus," in *12th Int. Conf. on NLP*, 2015, p. 35.

[40] T. Kohonen, "Essentials of the self-organizing map," *Neural Netw*, vol. *37*, pp. 52-65, 2013.

[41] S. Vidhya, DAAG. Singh, EJ. Leavline, "Feature Extraction for Document Classification", Int. J. Innov. Res. Sci. Eng. Technol., vol. 4, no. 6, pp. 50-56, 2015.

[42] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1-47, 2002.

[43] K. Gayathri, and A. Marimuthu, "Text document pre-processing with the KNN for classification using the SVM," in *7th Int. Conf. on ISCO,* Jan. 2013, pp. 453-457.

[44] V. Bijalwan, V. Kumar, P. Kumari, J. Pascual, "KNN based machine learning approach for text and document mining," *Int. J. of Dat. Th. and Appl.*, vol. 7, no. 1, pp.61-70, Feb. 2014, DOI: 10.14257/ijdta.2014.7.1.06.

[45] Daviddlewis.com, *Reuters-21578 Text Categorization Test Collection.* [Online]. Available: http://www.daviddlewis.com/resources/testcollections/reuters21578/

[46] X. Luo and A.N. Zincir-Heywood, "Evaluation of two systems on multi-class multi-label document classification," in *Int. Symp. on Method. for Intel. Sys.,* Springer, Berlin, Heidelberg, May 2005, pp. 161-169.

[47] O. Dekel, and O. Shamir, "Multiclass-multilabel classification with more classes than examples," *in Proc. of the 13th Int. Conf. on AI and Stat*. 2010.

[48] H. Zhiyang, W. Ji, L. Tao, "Label Correlation Mixture Model: A Supervised Generative Approach to Multilabel Spoken Document Categorization", *IEEE Trans. on Emerg. Topics in Comp (TETC)*, vol.3, no.2, pp. 235-245, 2015.

[49] B. Altınel, and M.C. Ganiz, "A new hybrid semi-supervised algorithm for text classification with class-based semantics," *Knowl-Based Sys*, vol. 108, pp. 50-64, Sep. 2016, DOI: 10.1016/j.knosys.2016.06.021.

[50] L. Lenc, and P. Král, "Deep neural networks for Czech multi-label document classification," 2017, arXiv preprint arXiv:1701.03849.

[51] J. Li, and H. Liu, "Challenges of Feature Selection for Big Data Analytics," *IEEE Intell Syst*, vol.32, no.2, pp.9-15, 2017.

[52] J. Dai, Z. He, F. Hu, "A High-performance algorithm for text feature automatic selection," in *Proc. of Int. Symp. on Information Processing (ISIP),* p. 414*,* 2009.

[53] J. Gui, Z. Sun, S. Ji, D. Tao, T. Tan, "Feature Selection Based on Structured Sparsity: A Comprehensive Study," *IEEE Trans. Neural Netw. Learn. Syst*, vol. 28, no. 7, pp.1490-1507, 2017.

[54] Y. Yang, and J.O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML*, vol. 97, pp. 412-420, 1997.

[55] E. Brill, "A simple rule-based part of speech tagger," in *Proc of the workshop on Speech and Natural Language*, Association for Computational Linguistics, Feb. 1992, pp. 112-116.

[56] R. Chandrasekar and B. Srinivas, "Using syntactic information in document filtering: A comparative study of part-of-speech tagging and super tagging," in *Computer-Assisted Information Searching on Internet,* LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, June 1997, pp. 531-545.

[57] C.S. Lim, K.J. Lee, G.C. Kim, "Multiple sets of features for automatic genre classification of web documents," *Information processing & management*, vol. 41, no. 5, pp. 1263-1276, 2005.

[58] T. Brychcín and P. Král, "Novel unsupervised features for Czech multi-label document classification," in *Mexican International Conference on Artificial Intelligence*, Springer, Cham, Nov. 2014, pp. 70-79.

[59] A. Burkovski, W. Kessler, G. Heidemann, H. Kobdani, H. Schütze, "Self-Organizing Maps in NLP: Exploration of Coreference Feature Space," in *International Workshop on Self-Organizing Maps (WSOM)*, Springer, Berlin, Heidelberg, 2011, pp. 228-237.

[60] Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464-1480.

[61] X. Luo and A.N. Zincir-Heywood, "A comparison of som based document categorization systems," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, IEEE, Chicago, 2003, vol. 3, pp. 1786-1791.

[62] S. Kaski, "Dimensionality reduction by random mapping: Fast similarity computation for clustering." in *Neural networks proc, 1998. ieee world congress on comp intelligence. the 1998 ieee int joint conf on*, vol. 1, IEEE, May 1998, pp. 413-418.

[63] Liu, Y. C., Wu, C., & Liu, M. (2011). Research of fast SOM clustering for text information. *Expert Systems with Applications*, *38*(8), 9325-9333.

[64] H.C. Yang, C.H. Lee, H.W. Hsiao, "Incorporating self-organizing map with text mining techniques for text hierarchy generation," *Appl Soft Comput*, vol. 34, pp. 251-259, Sep. 2015, DOI: 10.1016/j.asoc.2015.05.005.

[65] Y. Ding and X. Fu, "The Research of Text Mining Based on Self-Organizing Maps," *Procedia Engineering*, vol. *29*, pp. 537-541, 2012, DOI: 10.1016/j.proeng.2011.12.757.

[66] Y.C. Liu, M. Liu, X.L. Wang, "Application of self-organizing maps in text clustering: a review, *Applications of Self-Organizing Maps*, InTech, 2012.

[67] D. Isa, V.P. Kallimani, L.H. Lee, "Using the self-organizing map for clustering of text documents," *Expert Systems with Applications*, vol. 36, no. 5, pp. 9584-9591, July 2009, DOI: 10.1016/j.eswa.2008.07.082.

[68] D. D. Lewis, Y. Yang, T.G. Rose, F. Li, "Rcv1: A new benchmark collection for text categorization research," *J. Machine Learning Res.* vol.5, pp. 361-397, Apr. 2004.

[69] Nltk.org. (2018). *Natural Language Toolkit — NLTK 3.2.5 documentation*. [Online]. Available: http://www.nltk.org/

[70] C.D. Manning, P. Raghavan, H. Schütze, *Introduction to information retrieval,* vol. 39, Cambridge: Cambridge university press, 2008.

[71] Scikit-learn.org. (2018). *sklearn.feature_extraction.text.TF-IDFVectorizer — scikit-learn 0.19.1 documentation*. [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TF-IDFVectorizer.html

[72] V. Moosavi, S. Packmann, L. Vallés, "SOMPY: A Python Library for Self-Organizing Map (SOM)." GitHub. [Online]. Available: https://github.com/sevamoo/SOMPY

[73] T. Kohonen and H. Xing, "Contextually self-organized maps of Chinese words," in *International Workshop on Self-Organizing Maps,* pp. 16-29, Springer, Berlin, Heidelberg, June 2011.

[74] Scikit-learn.org. (2018). *Precision-Recall — scikit-learn 0.19.1 documentation*. [Online]. Available: http://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

[75] Scikit-learn.org. (2018). *Precision Recall Fscore — scikit-learn 0.19.1 documentation*. [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html

[76] M. Ostendorff, T. Ruas, M. Schubotz, G. Rehm, and B. Gipp, "Pairwise Multi-Class Document Classification for Semantic Relations between Wikipedia Articles," 2020. [Online] Available: arXiv:2003.09881.

[77] B.A. Rabut, A.C. Fajardo, and R.P. Medina, "Multi-class Document Classification Using Improved Word Embeddings," in Proc. ICCBD 2019 pp. 42-46.

[78] H. T. Sueno, B. D. Gerardo, and R. P. Medina, "Multi-class Document Classification using Support Vector Machine (SVM) Based on Improved Naïve Bayes Vectorization Technique", Int J, vol. 9, no. 3, 2020.

[79] P. Tiwari, and M. Melucci, "Multi-class classification model inspired by quantum detection theory," 2018. [Online]. Available: arXiv:1810.04491.

[80] M. Attia, Y. Samih, A. Elkahky, and L. Kallmeyer, "Multilingual multi-class sentiment classification using convolutional neural networks," in Proc. 11th Int. Conf. Lang. Res. Eval., May, 2018.

[81] V. Gupta, A. Saw, P. Nokhiz, H. Gupta and P. Talukdar, "Improving Document Classification with Multi-Sense Embeddings," 2019. [Online] Available: arXiv:1911.07918.

[82] K. Kowsari, D.E. Brown, M. Heidarysafa, K.J. Meimandi, M.S. Gerber, and L.E. Barnes, "Hdltex: Hierarchical deep learning for text classification," in 16th ICMLA, IEEE, 2017, pp. 364-371.

[83] E. Giannopoulou, and N. Mitrou. "Extensive Experimental Evaluation of Self-Organizing Maps for Automatic Classification of a Multi-Class Multi-Label Corpus," IEEE Access vol. 6, pp. 67385-67403, Sep. 2018.

[84] D. Kim, D. Seo, S. Cho, and P. Kang, "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec," Inf. Sci. vol. 477, pp.15-29, Oct. 2018.

[85] M. M. Mirończuk, and J. Protasiewicz, "A recent overview of the state-of-the-art elements of text classification," Expert Syst. Appl. vol. 106, pp. 36-54, 2018.

[86] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," Inf. vol. 10, no. 4, pp. 150, 2019.

[87] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Trans. Assoc. Comput. Linguist. vol. 5, pp. 135-146, 2017.

[88] Y. Goldberg and L. Omer, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," 2014. [Online] Available: arXiv:1402.3722.

[89] L. Quoc and T. Mikolov, "Distributed representations of sentences and documents," in ICML, 2014, pp. 1188-1196.

[90] J. Pennington, R. Socher, and C.D. Manning, "Glove: Global vectors for word representation," in Proc. of the EMNLP, 2014, pp. 1532-1543.

[91] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: arXiv:1810.04805.

[92] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le., "Xlnet: Generalized autoregressive pretraining for language understanding," in Adv. NeurIPS, 2019, pp. 5753-5763.

[93] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI blog 1, no. 8, 2019, pp. 9.

[94] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," in Adv. NeurIPS, 2017, pp. 5998-6008.

[95] H. A. M. Hassan, G. Sansonetti, F. Gasparetti, A. Micarelli, and J. Beel, "BERT, ELMo, USE and InferSent Sentence Encoders: The Panacea for Research-Paper Recommendation?," in RecSys (Late-Breaking Results), 2019, pp. 6-10.

[96] J.C. Lamirel, P. Cuxac, A.S. Chivukula, K. Hajlaoui, "Optimizing text classification through efficient feature selection based on quality metric," J Intell Inform Syst (JIIS), vol. 45, no. 3, pp. 379-396, Dec. 2015, DOI: 10.1007/s10844-014-0317-4.

[97] X. Deng, Y. Li, J. Weng, and J. Zhang, "Feature selection for text classification: A review," Multimed. Tools. Appl., vol. 78, no. 3, pp. 3797-3816, 2019.

[98] P. Somol, and J. Novovicova, "Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality," IEEE Trans. Pattern Anal. Mach. Intell. vol. 32, no. 11, pp. 1921-1939, 2010.

[99] T. Sabbah, A. Selamat, M. H. Selamat, F. S. Al-Anzi, E. H. Viedma, O. Krejcar, and H. Fujita, "Modified frequency-based term weighting schemes for text classification," Appl. Soft Comput. vol. 58, pp. 193-206, 2017.

[100] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi, "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods," Appl. Soft Comput. vol. 86, pp. 105836, 2020.

[101] H. Peng, and Y. Fan, "Feature selection by optimizing a lower bound of conditional mutual information," Inf. Sci. vol. 418, pp. 652-667, 2017.

[102] P. Singh, and A. Mukerjee, "Words are not equal: Graded weighting model for building composite document vectors," 2015. [Online] Available: arXiv:1512.03549.

[103] D. Mekala, V. Gupta, B. Paranjape and H. Karnick, "SCDV: Sparse Composite Document Vectors using soft clustering over distributional representations," 2016. [Online] Available: arXiv:1612.06778.

[104] N. Kalchbrenner, E. Grefenstette and P. Blunsom, "A convolutional neural network for modelling sentences," 2014. [Online] Available: arXiv:1404.2188.

[105] Y. Kim, "Convolutional neural networks for sentence classification," 2014. [Online] Available: arXiv:1408.5882.

[106] F.A. Gers, N.N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," J Mach Learn Res, no. 3, pp. 115-143, Aug. 2002.

[107] M. Morchid, M. Bouallegue, R. Dufour, G. Linarès, D. Matrouf, R. De Mori, "Compact Multiview Representation of Documents Based on the Total Variability Space," IEEE/ACM Trans. Audio, Speech, Language Process. (TSLP), vol.23, no.8, 2015.

[108] H. Amiri and M. Mohtarami, "Vector of Locally Aggregated Embeddings for Text Representation," in Proc. NAACL, vol. 1, June 2019, pp. 1408-1414.

[109] M.F Porter, "Snowball: A language for stemming algorithms," 2001.

[110] T. Kohonen, "The self-organizing map," Neurocomput., vol. 21, no. 1, pp.1-6, 1998.

[111] N. Saini, S. Saha and P. Bhattacharyya, "Cascaded Som: an improved technique for automatic email classification," in IJCNN, IEEE, pp. 1-8), July 2018.

[112] G.E. Hinton and R.R. Salakhutdinov, "Reducing the dimensionality of data with neural networks. science," Science vol. 313, no. 5786, pp.504-507, 2006.

[113] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," 2014. [Online] Available: arXiv:1412.1058.

[114] J.Y. Lee, and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," 2016. [Online] Available: arXiv:1603.03827.

[115] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in Adv Neural Inf Process Syst, pp. 649-657, 2015.

[116] L. Medsker and L.C. Jain, "Recurrent neural networks: design and applications," CRC press, 1999.

[117] G. Pölzlbauer, "Survey and comparison of quality measures for self-organizing maps," pp. 67-82, 2004.

[118] D. Scherer, A. Muller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in ICANN, pp. 92–101, 2010.

[119] G. Klambauer, T. Unterthiner, A. Mayr, S. Hochreiter, "Self-Normalizing Neural Networks," 2017. [Online]. Available: https://arxiv.org/abs/1706.02515.

[120] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: arXiv:1412.6980.

[121] Daconta, M. C., Obrst, L. J., & Smith, K. T. (2003). The semantic web: a guide to the future of XML, web services, and knowledge management. John Wiley & Sons.

[122] Quan, D., Huynh, D., & Karger, D. R. (2003). Haystack: A platform for authoring end user semantic web applications. In The semantic web-ISWC 2003 (pp. 738-753). Springer Berlin Heidelberg.

[123] Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Maedche, A., ... & Sure, Y. (2000, May). Ai for the web-ontology-based community web portals. In AAAI/IAAI (pp. 1034-1039).

[124] Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarrousse, É., Grabar, N., ... & Darmoni, S. (2005). UMLF: a unified medical lexicon for French. International Journal of Medical Informatics, 74(2), 119-124.

[125] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific American, 284(5), 28-37.

[126] S. Edition, G. Antoniou, and F. Van Harmelen, A Semantic Web Primer, Second edition. The MIT Press, Cambridge, England, 2008.

[127] F. Ferri, P. Grifoni, M.C Caschera, A. D'Ulizia, C. Praticò. "KRC: KnowInG crowdsourcing platform supporting creativity and innovation", *AISS: Advances in Information Sciences and Service Sciences*, vol. 5, no. 16, Nov. 2013, pp. 1 -15.

[128] M. Krötzsch, D. Vrandečić, M. Völkel, "Semantic mediawiki", *The Semantic Web-ISWC 2006*, pp. 935-942, Springer Berlin Heidelberg, 2006.

[129] Suominen, O., Hyvönen, E., Viljanen, K., & Hukka, E. (2009). HealthFinland—A national semantic publishing network and portal for health information. Web Semantics: Science, Services and Agents on the World Wide Web, 7(4), 287-297.

[130] World Wide Web Consortium. (2014). RDF 1.1 Primer.

[131] Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F. & Rudolph, S. (2012). OWL 2 Web Ontology Language Primer. Available online at: www.w3.org/TR/owl2-primer/.

[132] Harris, S., Seaborne, A., & Prud'hommeaux, E. (2013). SPARQL 1.1 query language. W3C Recommendation, 21.

[133] Library Linked Data Incubator Group: Final Report. World Wide Web Consortium (W3C), 2011.

[134] Q. Rajput and S. Haider, "BNOSA: A Bayesian network and ontology based semantic annotation framework," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 9, no. 2, Jul. 2011, pp. 99–112.

[135] A. Kalyanpur, B. Parsia, E. Sirin, B. Grau, and J. Hendler, "Swoop: A Web Ontology Editing Browser," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 4, no. 2, Jun. 2006, pp. 144–153.

[136] Sauermann, L., Cyganiak, R., & Völkel, M. (2011). Cool URIs for the semantic web.

[137] Brickley, D., & Guha, R. V. (2014). RDF Schema 1.1. W3C Recommendation, 25, 2004-2014.

[138] Thesaurus, F. M. A. (1995). Food and Agricultural Organization of the United Nations.

[139] do Amaral, M. B., Roberts, A., & Rector, A. L. (2000). NLP techniques associated with the OpenGALEN ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent linguistic and logical constructs. In Proceedings of the AMIA Symposium (p. 76). American Medical Informatics Association.

[140] Reynolds, D., Shabajee, P., & Cayzer, S. (2004, May). Semantic information portals. In Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters (pp. 290-291). ACM.

[141] Ding, L., Lebo, T., Erickson, J. S., DiFranzo, D., Williams, G. T., Li, X., ... & Flores, J. (2011). TWC LOGD: A portal for linked open government data ecosystems. Web Semantics: Science, Services and Agents on the World Wide Web, 9(3), 325-333.

[142] N. Stojanovic, A. Maedche, H. Straße, and S. Staab, "SEAL — A Framework for Developing, Semantic PortALs", Proceedings of the *1st International Conference on Knowledge Capture, ACM*, 2001, pp. 155–162.

[143] M. Eetu, E. Hyvönen, S. Saarela, and K. Viljanen, "Ontoviews - A Tool for Creating Semantic Web Portals," *Semant. Web – ISWC, vol. LNCS 3298*, 2004, pp. 797–811.

[144] S. Lee, M. Lee, P. Kim, H. Jung, & K. W. Sung. OntoFrame s3: semantic web-based academic research information portal service empowered by STAR-WIN. In *The semantic web: Research and applications*, pp. 401-405. Springer Berlin Heidelberg, 2010.

[145] Joubert, M., Dufour, J. C., Aymard, S., Falco, L., & Fieschi, M. (2005). Designing and implementing health data and information providers. International journal of medical informatics, 74(2), 133-140.

[146] Giannopoulou, E., Mitrou, N., Chimos, K., Karvounidis, T., & Douligeris, C. (2014, November). A Semantic Web Approach in the implementation of a Linked Data Portal using a CMS. In Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on (pp. 164-171). IEEE.

[147] Y. Ding, et al. "Semantic web portal: a platform for better browsing and visualizing semantic data." *Active Media Technology*, vol.6335, pp. 448-460. Springer Berlin Heidelberg, 2010.

[148] B. Nowack, "Paggr: Linked Data widgets and dashboards," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 7, no. 4, Dec. 2009, pp. 272–277.

[149] N. Leavitt, N, "Will NoSQL databases live up to their promise?", *Computer*, vol. 43, no.2, 2010, pp. 12-14.

[150] Oren, E., Delbru, R., & Decker, S. (2006). Extending faceted navigation for RDF data. In the Semantic Web-ISWC 2006 (pp. 559-572). Springer Berlin Heidelberg.

[151] Berners-Lee, T. (2006). Linked data-design issues.

[152] Deliot, C. (2014). Publishing the British National Bibliography as Linked Open Data. Catalogue & Index, (174), 13-18.

[153] German National Library (2014). The Linked Data Service of the German National Library: Modelling of bibliographic data. Available online at: http://www.dnb.de/SharedDocs/Downloads/EN/DNB/service/linkedDataModellierungTiteldaten.pdf.

[154] Zervanou, K., Korkontzelos, I., Van Den Bosch, A., & Ananiadou, S. (2011, June). Enrichment and structuring of archival description metadata. In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (pp. 44-53). Association for Computational Linguistics.

[155] Honkela, T. (1997, September). Self-organizing maps of words for natural language processing applications. In *Proceedings of the International ICSC Symposium on Soft Computing* (pp. 401-407).

[156] Mayer, R. (2011, June). Analysing the similarity of album art with self-organising maps. In *International Workshop on Self-Organizing Maps* (pp. 357-366). Springer Berlin Heidelberg.

[157] Sjöberg, M., & Laaksonen, J. (2011, June). Analysing the structure of semantic concepts in visual databases. In *International Workshop on Self-Organizing Maps* (pp. 338-347). Springer Berlin Heidelberg.

[158] Liu, Y., Wang, X., & Wu, C. (2008). ConSOM: A conceptional self-organizing map model for text clustering. *Neurocomputing*, *71*(4), 857-862.

[159] Liu, Y., Wang, X., & Liu, M. (2009). V-SOM: A Text Clustering Method based on Dynamic SOM Model. Journal of Computational Information Systems, 5(1), 141-145.

[160] Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* (pp. 2177-2185).

[161] Villmann, T., & Bauer, H. U. (1998). Applications of the growing self-organizing map. *Neurocomputing*, *21*(1), 91-100.

[162] Berglund, E., & Sitte, J. (2006). The parameter less self-organizing map algorithm. *IEEE Transactions on neural networks*, *17*(2), 305-316.

[163] Dittenbach, M., Merkl, D., & Rauber, A. (2000). The growing hierarchical self-organizing map. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on* (Vol. 6, pp. 15-19). IEEE.

[164] Kuremoto, T., Komoto, T., Kobayashi, K., & Obayashi, M. (2010). Parameter less-Growing-SOM and its application to a voice instruction learning system. *Journal of Robotics*, *2010*.

[165] R. Cattell, “Scalable SQL and NoSQL data stores”, *ACM SIGMOD Record*, vol.39, no.4, 2011, pp.12-27.