Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Σύστημα συστάσεων για την πρόταση ειδησεογραφικών άρθρων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΔΩΡΟΘΕΑ Γ. ΚΑΛΛΙΩΡΑ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Γεώργιος Αλεξανδρίδης
Εργαστηριακό και Διδακτικό Προσωπικό Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Σύστημα συστάσεων για την πρόταση ειδησεογραφικών άρθρων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΔΩΡΟΘΕΑ Γ. ΚΑΛΛΙΩΡΑ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Γεώργιος Αλεξανδρίδης
Εργαστηριακό και Διδακτικό Προσωπικό Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 30η Σεπτεμβρίου 2021.

Ανδρέας Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2021

Δωροθέα Γ. Καλλιώρα

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Στη σημερινή εποχή όλο και περισσότεροι άνθρωποι ενημερώνονται για την επικαιρότητα από διαδικτυακές πηγές. Το πλήθος των διαδικτυακών πηγών καθώς και ο όγκος των άρθρων, παρ' όλη τη σφαιρική ενημέρωση που παρέχουν, καθιστούν δύσκολή και χρονοβόρα την ενημέρωση. Για την αντιμετώπιση αυτού του προβλήματος, έχουν δημιουργηθεί τα συστήματα συστάσεων ειδησεογραφικών άρθρων, που βοηθούν τους αναγνώστες στη διαχείριση του μεγάλου όγκου δεδομένων και τους προτείνουν άρθρα που μπορεί να τους ενδιαφέρουν.

Σκοπός της παρούσας διπλωματικής εργασίας είναι η δημιουργία ενός συστήματος σύστασης ειδησεογραφικών άρθρων βασισμένο σε λογισμικό συγκέντρωσης και αποθησαύρισης πηγών, το οποίο ομαδοποιεί άρθρα από διαφορετικές πηγές και μπορεί να προτείνει στους αναγνώστες άρθρα προς ανάγνωση. Με τη χρήση αυτού του συστήματος μειώνεται ο χρόνος που απαιτείται για την ενημέρωση, αφού τα άρθρα για κάθε θέμα και από πολλές ιστοσελίδες βρίσκονται σε μία πηγή.

Πιο συγκεκριμένα, η δημιουργία του συστήματος σύστασης ειδησεογραφικών άρθρων χωρίστηκε σε τρεις φάσεις. Αρχικά, στη συγκέντρωση και ομαδοποίηση ειδησεογραφικών άρθρων από διάφορες πηγές. Κατόπιν, στην υλοποίηση ενός αλγορίθμου εύρεσης προτάσεων για τους υποψήφιους αναγνώστες και τέλος στην κατασκευή της εν λόγω πλατφόρμας υπό τη μορφή ιστοσελίδας.

Για την συσταδοποίηση χρησιμοποιήθηκαν κλασικές τεχνικές, αλλά και τεχνικές που βασίζονται στην μηχανική μάθηση με τη βοήθεια νευρωνικών δικτύων, καθώς και στην αναγνώριση επώνυμων οντοτήτων, για μία περαιτέρω ομαδοποίηση των άρθρων σε γεγονότα. Για τον αλγόριθμο εύρεσης των προτάσεων μελετήθηκαν τεχνικές βαθειάς μάθησης και χρησιμοποιήθηκε ένας αλγόριθμος προτάσεων βασισμένος στο περιεχόμενο.

Λέξεις κλειδιά

ειδησεογραφικά νέα, συστήματα συστάσεων, συστήματα συλλογής ειδησεογραφικών άρθρων, συσταδοποίηση κειμένου, εύρεση γεγονότων, μηχανική μάθηση, βαθειά μάθηση, νευρωνικά δίκτυα, αναγνώριση επώνυμων οντοτήτων, τεχνικές βασισμένες στο περιεχόμενο

Abstract

Nowadays, more and more people tend to read news online. In order to help readers deal with the large number of online sources and articles, news recommendation systems have been developed. These systems can relieve the problem of information overload and suggest articles that the readers might be interested in.

The objective of the current diploma thesis is to create a news recommendation system, based on news aggregators that gather and cluster news from different online sources and can recommend articles to their users. Among others, the decrease of time needed to search news online, is one of the major benefits of a news aggregator.

In particular, the creation of the news recommendation system has been split in three phases. Initially, news articles are gathered from different online sources, and are clustered into groups. Following, an algorithm for suggesting news articles to readers is created and finally the overall platform is implemented in the form of a website.

For the clustering of the articles, baseline techniques as well as machine learning approaches based on neural networks have been used, along with named entity recognition, in order to further cluster news articles into events. For the recommendation algorithm, deep learning techniques have been studied and a content-based technique has been implemented.

Key words

news articles, recommender systems, news aggregators, text clustering, event detection, machine learning, deep learning, neural networks, named entity recognition, content-based techniques

Ευχαριστίες

Η παρούσα διπλωματική εργασία σηματοδοτεί την ολοκλήρωση των προπτυχιακών σπουδών μου. Συνεπώς, θα ήθελα να ευχαριστήσω θερμά όλους όσοι με υποστήριξαν στην εκπόνηση αυτής της διπλωματικής αλλά και συνολικά στις σπουδές μου.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα κ. Ανδρέα-Γεώργιο Σταφυλοπάτη, Καθηγητή Ε.Μ.Π., για την ευκαιρία που μου προσέφερε να εκπονήσω τη συγκεκριμένη εργασία καθώς και για την εμπιστοσύνη που μου έδειξε. Παράλληλα, θα ήθελα να ευχαριστήσω τον συνεπιβλέποντα κ. Γεώργιο Αλεξανδρίδη, Ε.ΔΙ.Π. Ε.Μ.Π., για τη συνεχή υποστήριξη και καθοδήγησή του. Οι πολύτιμες συμβουλές του και οι χρήσιμες συζητήσεις που πραγματοποιήσαμε με βοήθησαν να ξεπεράσω σημαντικά εμπόδια που συνάντησα κατά την εκπόνηση της εργασίας. Επιπροσθέτως, θα ήθελα να ευχαριστήσω τους κ.κ. Στέφανο Κόλλια και Γεώργιο Στάμου, Καθηγητές Ε.Μ.Π., που με τίμησαν με την παρουσία τους στην τριμελή επιτροπή εξέτασης. Ακόμα, θα ήθελα να πω ένα ευχαριστώ σε όλους όσους μπήκαν στην ιστοσελίδα που δημιουργήθηκε για την εκπόνηση της διπλωματικής μου εργασίας, για το χρόνο τους και τη βοήθεια τους.

Σε προσωπικό επίπεδο θα ήθελα πρωτίστως να ευχαριστήσω από τα βάθη της καρδίας μου τους γονείς μου Γιώργο και Γεωργία, και τις αδερφές μου Δάφνη και Δήμητρα για τη συνεχή ενθάρρυνση, υποστήριξη, και για τις πολύτιμες συμβουλές τους καθ' όλη τη διάρκεια των σπουδών μου. Τέλος, θα ήθελα να ευχαριστήσω τους φίλους μου που στέκονται στο πλευρό μου όλα αυτά τα χρόνια, για την στήριξη, την αγάπη και τις ωραίες στιγμές που έχουμε περάσει μαζί.

Δωροθέα Γ. Καλλιώρα,

Αθήνα, 30η Σεπτεμβρίου 2021

Περιεχόμενα

Κατάλογος πινάκων

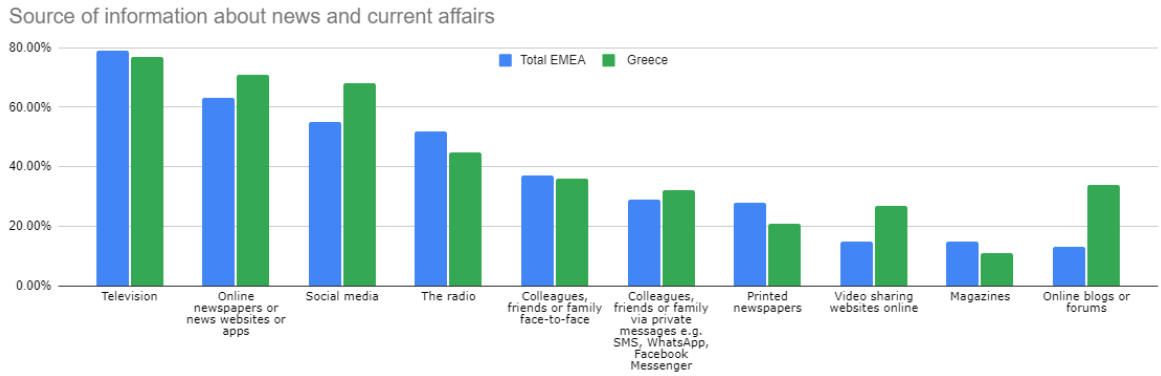
Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

1.1 Διαδικτυακές Εφημερίδες και ιστοσελίδες ειδησεογραφικών νέων

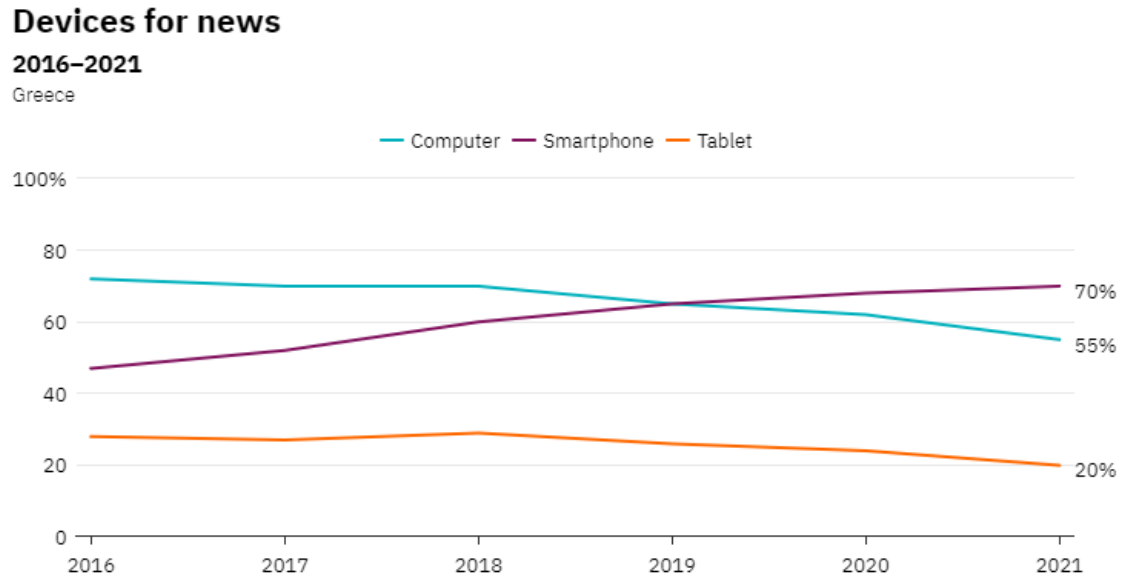
Τα τελευταία χρόνια όλο και περισσότεροι άνθρωποι τείνουν να ενημερώνονται για την επικαιρότητα από διαδικτυακές πηγές. Οι διαδικτυακές εφημερίδες, οι ιστοσελίδες νέων και εφαρμογών για την ανάγνωση άρθρων έρχονται δεύτερα στην κατάταξη πηγών πληροφορίας, σύμφωνα με την έρευνα που πραγματοποίησε η Ipsos MORI για την Google το 2020 [ipso20]. Στο Σχήμα 1.1 συγκρίνονται οι πηγές πληροφορίας που χρησιμοποιούνται στην Ελλάδα, με αυτές που χρησιμοποιούνται στις EMEA (Europe, Middle East and Asia). Πιο συγκεκριμένα, στην Ελλάδα παρατηρούμε ότι οι διαδικτυακές πηγές πληροφορίας καθώς και τα κοινωνικά δίκτυα επιλέγονται σε μεγαλύτερο ποσοστό από ότι στις υπόλοιπες χώρες, ενώ η τηλεόραση σε μικρότερο ποσοστό. Επίσης, τα έντυπα μέσα ενημέρωσης, όπως εφημερίδες και περιοδικά, έρχονται τελευταία στην κατάταξη.



Σχήμα 1.1: Έρευνα για τις πηγές ενημέρωσης από 6 Ιανουαρίου μέχρι 27 Μαρτίου 2020 σε 35.030 άτομα ηλικίας 16-70 ετών στις χώρες της EMEA και στην Ελλάδα (Πηγή Δεδομένων: [ipso20])

Υπάρχουν αρκετοί λόγοι που έχουν συμβάλει σε αυτή την εξέλιξη. Αρχικά, λόγω της αύξησης της αγοράς και χρήσης ηλεκτρονικών συσκευών, η ενημέρωση των χρηστών μπορεί να γίνει μέσω του κινητού ή του υπολογιστή τους. Συγκεκριμένα, σύμφωνα με την Έκθεση Ψηφιακών Ειδήσεων του Ινστιτούτου Reuters για την Μελέτη της Δημοσιογραφίας του Πανεπιστημίου της Οξφόρδης, στην Ελλάδα έχει παρατηρηθεί μία αύξηση 23%, από το 2016, στη χρήση κινητού για την ενημέρωση των πολιτών [Newm21]. Αντίθετα, η χρήση υπολογιστή έχει πέσει κατά 17% από το 2016, με αποτέλεσμα το 2021 το πρώτο μέσο διαδικτυακής ενημέρωσης στην Ελλάδα να είναι τα κινητά τηλέφωνα με ποσοστό 70%, δεύτερο οι υπολογιστές με 55% και τρίτο τα tablets με ποσοστό 20% (Σχήμα 1.2). Έτσι, μέσω του κινητού ο καθένας μπορεί να ενημερώνεται για έκτακτα γεγονότα, τοπικές ή παγκόσμιες ειδήσεις, ανεξαρτήτως χρονικής στιγμής και χωρίς χρηματική επιβάρυνση.

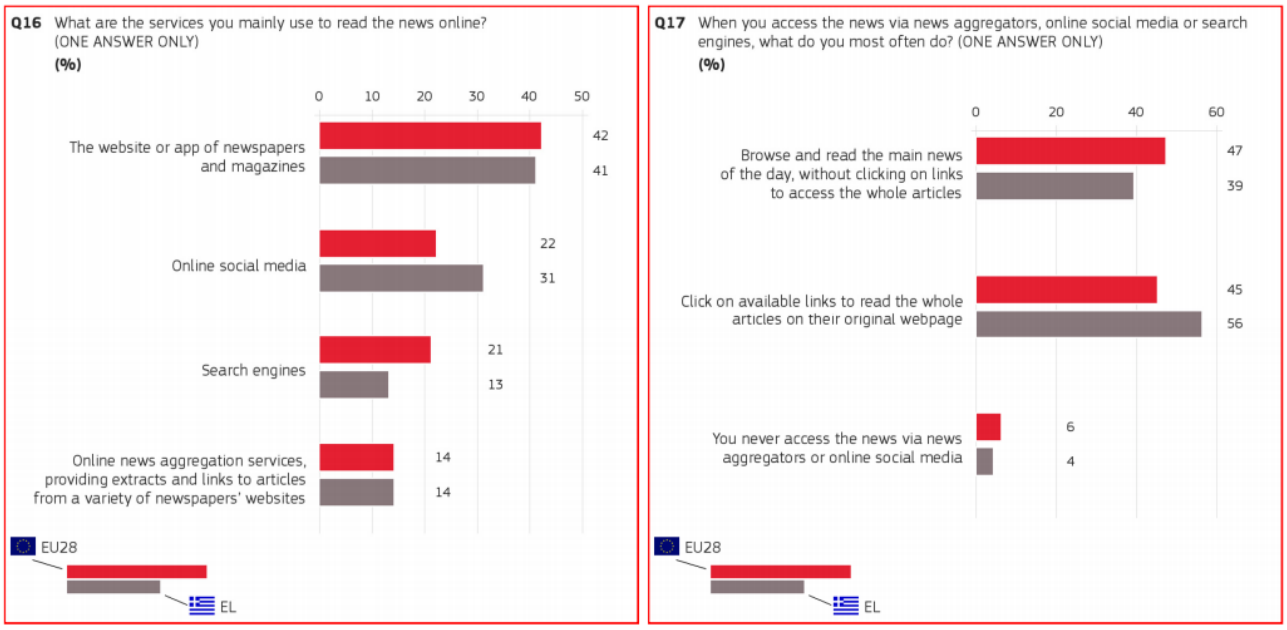
Ακόμα, η ψηφιοποίηση πολλών εφημερίδων έχει ωθήσει τους αναγνώστες τους στην διαδικτυακή ενημέρωση. Μία διαδικτυακή ιστοσελίδα νέων μπορεί να ενημερώνεται 24 ώρες το εικοσιτετράωρο σε αντίθεση μία εφημερίδα που αναφέρει τα νέα της προηγούμενης ημέρας. Επίσης, μέσω των ιστοσελίδων, ένας αναγνώστης μπορεί να διαβάσει νέα από διαφορετικές ιστοσελίδες χωρίς να χρειάζεται να αγοράσει όλες τις εφημερίδες που αναφέρονται σε κάποιο θέμα που τον ενδιαφέρει. Επιπλέον, στις



Σχήμα 1.2: Έρευνα για τις συσκευές που χρησιμοποιούνται στην Ελλάδα για την ανάγνωση άρθρων στο Διαδίκτυο (Πηγή: [Newm21])

διαδικτυακές εφημερίδες, εκτός από κείμενο, μπορεί κάποιο άρθρο να περιέχει εικόνες και βίντεο κάνοντας πιο σφαιρική και ενδιαφέρουσα την ενημέρωση των αναγνωστών.

Τέλος, η εξέλιξη της τεχνολογίας έχει συμβάλει σημαντικά στην ανάπτυξη συστημάτων για την εύκολη πρόσβαση και γρήγορη ενημέρωση των αναγνωστών. Τα συστήματα αυτά μπορούν να εντοπίσουν ψευδείς ειδήσεις, να κάνουν προσωποποιημένες προτάσεις στους αναγνώστες και να συγκεντρώνουν άρθρα από διάφορες πηγές για να διευκολύνουν τον χρήστη στην αντικειμενική ενημέρωση για κάποιο θέμα, χωρίς να χρειάζεται να επισκεφτεί απ' ευθείας τις πολλές ιστοσελίδες νέων. Σύμφωνα με έρευνα του TNS Political & Social Network για τη Γενική Διεύθυνση Τηλεπικοινωνιακών Δικτύων, Περιεχομένου και Τεχνολογίας της Ευρωπαϊκής Επιτροπής για 28 χώρες μέλη της Ευρωπαϊκής Ένωσης στο διάστημα 10 με 21 Μαρτίου του 2016 [euro21], φαίνεται ότι το 14% των Ελλήνων χρησιμοποιεί συστήματα συγκέντρωσης ηλεκτρονικών ειδήσεων για την ενημέρωση του (γνωστά και ως news aggregators - NA) και από αυτό το ποσοστό, το 56% ανοίγει τα άρθρα για να τα διαβάσει στην αρχική ιστοσελίδα, ενώ το 39% ενημερώνεται διαβάζοντας μόνο τους τίτλους. Τα ποσοστά αυτά καθώς και η σύγκριση της Ελλάδας με τις υπόλοιπες χώρες φαίνονται στο Σχήμα 1.3. Συμπερασματικά, η χρήση αυτών των συστημάτων όχι μόνο διευκολύνει τους χρήστες, μειώνοντας τον χρόνο ενημέρωσης τους, αλλά και αυξάνει την επισκεψιμότητα στις ιστοσελίδες των εφημερίδων.



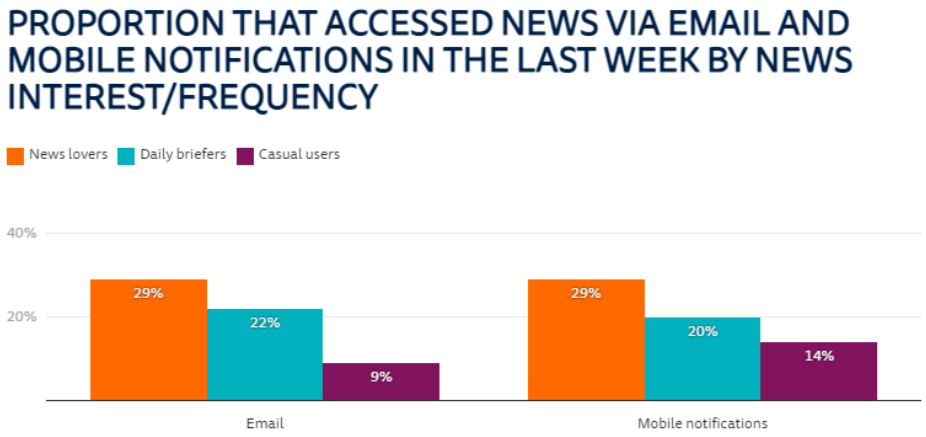
Σχήμα 1.3: Έρευνα για τις διαδικτυακές πηγές νέων που χρησιμοποιούνται στην Ευρώπη (Πηγή: [euro21])

1.2 Προσωποποιημένες προτάσεις άρθρων

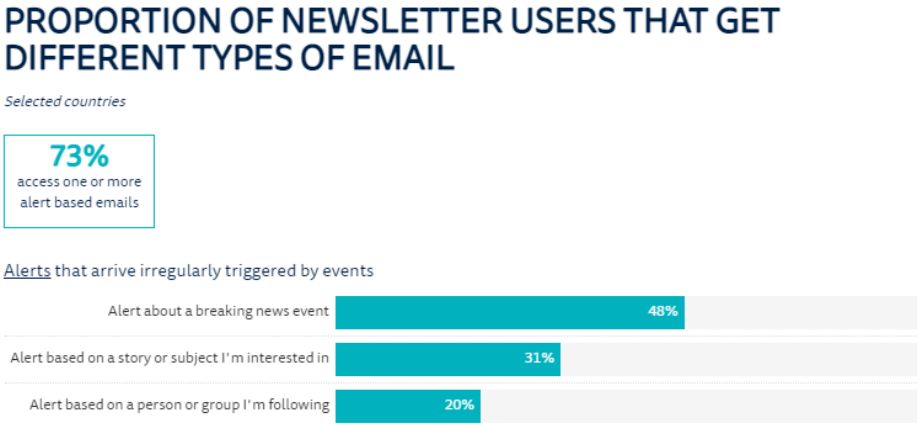
Οι γρήγοροι ρυθμοί ζωής της σύγχρονης κοινωνίας, οι διαφορετικές προτιμήσεις κάθε ατόμου και ο συνεχώς αυξανόμενος όγκος δεδομένων, σε συνδυασμό με την εξέλιξη της τεχνολογίας, έχουν συμβάλλει στην αύξηση της δημοτικότητας των *συστημάτων συστάσεων* (ΣΣ). Ένα ΣΣ μπορεί να χρησιμοποιηθεί σε διαδικτυακές πηγές νέων για την πρόταση άρθρων στους αναγνώστες. Με αυτόν τον τρόπο, οι χρήστες λαμβάνουν προσωποποιημένες προτάσεις πάνω σε θέματα που τους ενδιαφέρουν.

Τα άρθρα μπορούν να προταθούν με πολλούς τρόπους στους αναγνώστες. Πολλές εφημερίδες έχουν εντάξει στην ιστοσελίδα τους και ένα τμήμα με τίτλο «Προτάσεις για εσάς», το οποίο περιέχει έναν μικρό αριθμό προτεινόμενων άρθρων για κάθε χρήστη. Τα άρθρα αυτά προτείνονται με βάση το ιστορικό των χρηστών αν είναι εγγεγραμμένοι ή χρησιμοποιώντας προσεγγίσεις βασισμένες στη συνεδρία (session-based) αν είναι ανώνυμοι χρηστές. Ένας άλλος τρόπος προώθησης των προτάσεων είναι μέσω ειδοποιήσεων push ή των ενημερωτικών δελτίων (newsletters). Σύμφωνα με ερεύνα του Ινστιτούτου Reuters για την Μελέτη της Δημοσιογραφίας του Πανεπιστημίου της Οξφόρδης [Newm20], το ποσοστό των χρητσών που επιλέγουν να λαμβάνουν ειδοποιήσεις για νέα της επικαιρότητας μέσω e-mail ή ειδοποίησεων στο κινητό είναι περίπου το ίδιο, είτε διαβάζουν νέα συχνά είτε όχι, όπως φαίνεται στο Σχήμα 1.4. Από το ποσοστό των ατόμων που επιλέγουν τα email σαν μέσω ενημέρωσης για καινούργια άρθρα, το 31% θέλει να λαμβάνει email για γεγονότα που τον ενδιαφέρουν όπως φαίνεται στο Σχήμα 1.5. Ακόμα, το 73% αυτών των χρηστών ανοίγουν σίγουρα μία ή περισσότερες ειδοποιήσεις.

Άρα, η παροχή προσωποποιημένων προτάσεων αποτελεί ένα σημαντικό κομμάτι για την παροχή άρθρων στους χρήστες. Με την συμπερίληψη ΣΣ στις ιστοσελίδες νέων και τη χρήση των παραπάνω τρόπων παροχής των προτάσεων, μπορεί να αυξηθεί η αλληλεπίδραση των χρηστών με την ιστοσελίδα και να ωθήσει τους αναγνώστες να κάνουν εγγραφή ή και να αγοράσουν μία συνδρομή.



Σχήμα 1.4: Έρευνα για το ποσοστό των χρηστών που που χρησιμοποιούν το email και τα notifications για την ενημέρωση τους, με βάση το ενδιαφέρον τους στα νέα (Πηγή: [Newm20])



Σχήμα 1.5: Έρευνα για το περιεχόμενο των άρθρων που λαμβάνουν οι χρήστες μέσω των email (Πηγή: [Newm20])

1.3 Στόχος της εργασίας

Ο στόχος της παρούσας εργασίας είναι διττός: αφενός η δημιουργία ενός συστήματος συλλογής άρθρων νέων που να ομαδοποιεί άρθρα από διαφορετικές πηγές και αφετέρου η παραγωγή προτάσεων για άρθρα προς ανάγνωση στους χρήστες. Τα προτεινόμενα άρθρα θα πρέπει να ενδιαφέρουν τους χρήστες, έτσι ώστε να αυξάνεται η αλληλεπίδραση με το σύστημα, αλλά και να μειώνεται ο χρόνος που χρειάζονται οι αναγνώστες για την ενημέρωση τους για τα θέματα της επικαιρότητας.

Γενικότερα, το πρώτο βήμα υλοποίησης ενός ΝΑ είναι η συσταδοποίηση των άρθρων. Μέσω της ομαδοποίησης των άρθρων που αναφέρονται στο ίδιο θέμα, αλλά ανήκουν σε διαφορετικές ιστοσελίδες νέων, οι χρήστες μπορούν να έχουν μία σφαιρική και γρήγορη ενημέρωση πάνω σε θέματα που αφορούν διάφορους τομείς της καθημερινής τους ζωής. Έτσι, ένα μεγάλο μέρος της διπλωματικής εργασίας αφιερώθηκε στη μελέτη, εφαρμογή και σύγκριση διαφορετικών μεθόδων συσταδοποίησης κειμένου.

Επιπλέον, για την υλοποίηση ενός ΣΣ που θα παρέχει αξιόπιστες και ενδιαφέρουσες προτάσεις στους χρήστες, απαιτούνται δεδομένα (ιστορικό) χρήστη. Αυτά τα δεδομένα χρειάζονται χρόνο και αρκετές αλληλεπιδράσεις με το σύστημα, έτσι ώστε να μπορούν να χρησιμοποιηθούν στη δημιουργία αλγορίθμων για την παροχή προτάσεων. Μέσω της δημιουργίας του ΣΣΕΑ, αποκτήθηκαν κάποια αρχικά δεδομένα για το ιστορικό των χρηστών και τον τρόπο αλληλεπίδρασης τους με τις συστάσεις που τους γινόντουσαν.

Σε κάθε περίπτωση, για τη δημιουργία του ΣΣΕΑ αντιμετωπίστηκαν οι περισσότερες προκλήσεις των ΣΣ και μελετήθηκαν οι διαφορετικοί αλγόριθμοι προτάσεων ειδησεογραφικών άρθρων.

1.4 Δομή της εργασίας

Η υπόλοιπη εργασία διαρθρώνεται ως εξής. Στο Κεφάλαιο 2, γίνεται μία εισαγωγή στα συστήματα συστάσεων ειδησεογραφικών άρθρων. Συγκεκριμένα, αναφέρονται οι προκλήσεις που αντιμετωπίζουν, καθώς και κάποιες πιθανές λύσεις τους, και αναλύονται οι τεχνικές και στρατηγικές ανάλυσης που χρησιμοποιούνται για την υλοποίηση τους. Στη συνέχεια, στο Κεφάλαιο 3 περιγράφεται το θεωρητικό υπόβαθρο που χρησιμοποιήθηκε για την υλοποίηση της συσταδοποίησης των άρθρων. Αναφέρονται οι τεχνικές που χρησιμοποιήθηκαν καθώς και τα συμπεράσματα που προέκυψαν από τη χρήση της κάθε τεχνικής. Στο Κεφάλαιο 4 περιγράφεται η πρακτική υλοποίηση του ΣΣΕΑ. Πιο συγκεκριμένα, αναφέρονται ο τρόπος σχεδίασης της βάσης δεδομένων, ο αλγόριθμος που αναπτύχηθκε για την συσταδοποίηση των κειμένων, ο αλγόριθμος του ΣΣ καθώς και ο τρόπος που σχεδιάστηκε η ιστοσελίδα. Τέλος, στο Κεφάλαιο 5 παρουσιάζονται τα συμπεράσματα που προέκυψαν καθώς και στατιστικά από το ιστορικό των χρηστών, ενώ αναφέρονται και πιθανές μελλοντικές κατευθύνσεις για την βελτίωση του ΣΣΕΑ.

Κεφάλαιο 2

Συστήματα Συστάσεων

Τα σύστημα συστάσεων (recommender systems) είναι μια υποκατηγορία των συστημάτων φιλτραρίσματος πληροφορίας η οποία έχει σκοπό την πρόβλεψη της «βαθμολογίας» ή της «προτίμησης» που θα δείξει κάποιος χρήστης σε ένα αντικείμενο [Ricc10]. Τα αντικείμενα μπορεί να εντάσσονται σε διάφορες κατηγορίες, όπως βίντεο, μουσική, προϊόντα, περιεχομένου των συστημάτων κοινωνικής δικτύωσης ή ανοιχτού περιεχομένου του διαδικτύου [Gupt13, Bara18]. Τα συστήματα αυτά μπορούν να λειτουργήσουν χρησιμοποιώντας μία είσοδο όπως μουσική, ή πολλαπλές εισόδους όπως άρθρα, βιβλία και ερωτήματα αναζήτησης. Επίσης, υπάρχουν δημοφιλή ΣΣ για συγκεκριμένα θέματα, όπως εστιατόρια και διαδικτυακά ραντεβού. Έχουν αναπτυχθεί και ΣΣ τα οποία εξερευνούν επιστημονικά άρθρα και ειδικούς [Chen15], συνεργασίες [Chen11] και οικονομικές υπηρεσίες [Felf07].

2.1 Συστήματα Συστάσεων Ειδησεογραφικών Άρθρων

Τα ΣΣΕΑ αποτελούν μία υποκατηγορία των ΣΣ. Μέσω αυτών προτείνονται στους χρήστες ειδησεογραφικά νέα για ανάγνωση με βάση τις προτιμήσεις τους. Τα ΣΣΕΑ, έχουν πολλά κοινά με τα υπόλοιπα ΣΣ αλλά και αρκετές διαφορές. Αυτές οφείλονται στο γεγονός ότι τα ειδησεογραφικά νέα μεταβάλλονται με γρήγορους ρυθμούς. Οι ειδησεογραφικές ιστοσελίδες ανανεώνονται συνεχώς, και κάποια άρθρα μπορεί να θεωρηθούν σε μικρό χρονικό διάστημα ξεπερασμένα, λόγω κάποιου έκτακτου γεγονότος. Έτσι, απαιτείται η συχνή ενημέρωση του ΣΣ, ανάλογα με τις εξελίξεις που συμβαίνουν στον κόσμο. Ακόμα, το ενδιαφέρον των χρηστών στα ειδησεογραφικά άρθρα αλλάζει δυναμικά με βάση κάποιους παράγοντες όπως την ώρα της ημέρας, τη συσκευή που χρησιμοποιούν (τηλέφωνο ή υπολογιστή) ή και την τοποθεσία τους. [Kari18]

Λόγω της πρακτικής ουσίας του προβλήματος της πρότασης άρθρων προς ανάγνωση, καθώς και των προκλήσεων του, έχει γίνει σημαντική ερευνητική δουλειά τα τελευταία χρόνια. Μέσα από την ερεύνα έχουν προταθεί αρκετοί καινοτόμοι αλγόριθμοι οι οποίο παράγουν προσωποποιημένες προτάσεις. [Kari18] Οι κύριες προκλήσεις και οι στρατηγικές ανάλυσης των προβλημάτων των ΣΣΕΑ περιγράφονται στις Ενότητες 2.2 και 2.3 αντίστοιχα.

2.2 Κύριες προκλήσεις και πιθανές προσεγγίσεις

Στην παρούσα Ενότητα αναφέρονται οι κύριες προκλήσεις των ΣΣΕΑ, καθώς και πιθανές προσεγγίσεις τους. Κάποια από τα προβλήματα που αναφέρονται αφορούν όλα τα ΣΣ, όπως το πρόβλημα της ψυχρής εκκίνησης ή της έλλειψης δεδομένων, ενώ άλλα αφορούν μόνο τα ΣΣΕΑ, όπως η γρήγορη μεταβολή των άρθρων. Η βαρύτητα που δίνεται σε κάθε πλευρά πρέπει να είναι συνυφασμένη με τους στόχους του κάθε μοντέλου ΣΣ. Αυτό συμβαίνει διότι η βελτίωση κάποιων χαρακτηριστικών μπορεί να επιφέρει την χειροτέρευση κάποιων άλλων. Άρα, υπάρχει κάποιο σημείο ισορροπίας το οποίο θα πρέπει να λαμβάνεται κάθε φορά υπόψιν στη σχεδίαση του ΣΣ.

2.2.1 Πρόβλημα Ψυχρής Εκκίνησης

Το πρόβλημα της *ψυχρής εκκίνησης* (cold start problem) είναι ένα συχνό πρόβλημα των ΣΣ. Το πρόβλημα αυτό, αναφέρεται σε περιπτώσεις που χρειάζεται να γίνει σύσταση αντικειμένων στους χρήστες του ΣΣ τη στιγμή που δεν υπάρχουν αρκετές πληροφορίες στο σύστημα. Υπάρχουν τρεις περιπτώσεις ψυχρής εκκίνησης:

- **Νέα κοινότητα (New community)**: Το πρόβλημα της ψυχρής εκκίνησης της κοινότητας, αναφέρεται στην αρχή της δημιουργίας ενός ΣΣ. Στην περίπτωση αυτή, ενώ μπορεί να υπάρχουν αντικείμενα στη βάση δεδομένων του συστήματος, οι χρήστες, αν υπάρχουν, είναι ελάχιστοι και οι προτάσεις του συστήματος δεν είναι καλές, αφού δεν υπάρχουν αλληλεπιδράσεις των χρηστών με το σύστημα.

- **Νέος χρήστης (New user)**: Το πρόβλημα της ψυχρής εκκίνησης των χρηστών, αναφέρεται στους χρήστες που κάνουν εγγραφή και δεν έχουν ακόμα αλληλεπιδράσει με το ΣΣ. Έτσι, ενώ δεν υπάρχουν αρκετές πληροφορίες για τις προτιμήσεις του χρήστη, χρειάζεται να του προταθούν άρθρα προς ανάγνωση.

- **Νέο αντικείμενο (New item)**: το πρόβλημα της ψυχρής εκκίνησης των αντικείμενων αναφέρεται στις περιπτώσεις που προστίθεται κάποιο καινούργιο αντικείμενο στα δεδομένα. Συγκεκριμένα, για τα ΣΣΕΑ, το άρθρο που προστίθεται στο σύστημα, έχει μόνο πληροφορίες όπως τον τίτλο, το κείμενο, τον αρθρογράφο κλπ., αλλά δεν έχει πληροφορίες όπως πόσοι χρήστες το έχουν διαβάσει ή την βαθμολογία των χρηστών όπως έχουν τα υπόλοιπα άρθρα που είναι καιρό στο σύστημα.

Και οι τρεις περιπτώσεις του προβλήματος ψυχρής εκκίνησης έχουν κοινή την έλλειψη των δεδομένων αλληλεπίδρασης των χρηστών με τα αντικείμενα του ΣΣ. Για αυτό το λόγο, οι λύσεις που έχουν προταθεί παρουσιάζουν κάποια ομοιότητα μεταξύ τους και είναι ανεξάρτητες από την περίπτωση που εξετάζεται κάθε φορά.

Μία γενική ιδέα είναι να παρθούν περισσότερες πληροφορίες για τους χρήστες, για να μπορέσουν να χρησιμοποιηθούν στη διαδικασία πρότασης άρθρων, όταν οι χρήστες δεν έχουν δώσει ακόμα δεδομένα ανάδρασης (feedback) στο σύστημα. Ανάλογα με την ποσότητα πληροφορίας που χρειάζεται κάθε φορά, υπάρχουν και διαφορετικές τεχνικές απόκτησης της. Αυτές οι τεχνικές ονομάζονται *τεχνικές απόσπασης προτιμήσεων* [Elah14, Elah16]. Οι πληροφορίες μπορούν να παρθούν είτε άμεσα, ρωτώντας τον κάθε χρήστη, είτε έμμεσα, παρατηρώντας τη συμπεριφορά του. Και στις δύο περιπτώσεις, ο χρήστης θα πρέπει να αλληλεπιδράσει αρκετά με το σύστημα έτσι ώστε να μπορούν να του προταθούν αντικείμενα που όντως να τον ενδιαφέρουν. Ένα παράδειγμα απόκτησης άμεσης πληροφορίας για τους χρήστες δίνεται μέσα από την ιστοσελίδα προτάσεων ταινιών MovieLens [movi]. Σε αυτό το σύστημα, κάθε φορά που κάποιος χρήστης κάνει εγγραφή, του ζητείται να βαθμολογήσει κάποιες ταινίες. Έτσι, το σύστημα παίρνει πληροφορίες για τις προτιμήσεις του χρήστη, ώστε να μπορεί να του κάνει τις πρώτες προτάσεις. Η λύση αυτή, ενώ είναι εύκολη και πρακτική, έχει και κάποια μειονεκτήματα. Ένα από αυτά είναι η αύξηση του χρόνου εγγραφής. Λόγω αυτού, οι χρήστες μπορεί να δίνουν τυχαία βαθμολογία στα αντικείμενα έτσι ώστε να τελειώσουν πιο γρήγορα την εγγραφή στο σύστημα.

Λαμβάνοντας υπόψιν και το παραπάνω μειονέκτημα, μία άλλη λύση για το πρόβλημα ψυχρής εκκίνησης θα μπορούσε να είναι η δημιουργία του προφίλ του χρήστη μέσα από τις αλληλεπιδράσεις του με άλλες πλατφόρμες κοινωνικής δικτύωσης ή και του ιστορικού του στο διαδίκτυο. Έτσι, ένα ΣΣΕΑ θα μπορούσε να του προτείνει άρθρα ανάλογα με τις αλληλεπιδράσεις του στο Twitter.

Μία ακόμα προσέγγιση του προβλήματος ψυχρής εκκίνησης, είναι η χρήση τεχνικών *ενεργητικής μηχανικής μάθησης* (active machine learning). Ο στόχος αυτής της τεχνικής, είναι να καθοδηγήσει το χρήστη στη διαδικασία της εξαγωγής προτιμήσεων, έτσι ώστε να του ζητήσει να βαθμολογήσει μόνο τα αντικείμενα τα οποία έχουν κάποια αξία για το μοντέλο του ΣΣ. Η επιλογή των αντικειμένων γίνεται αναλύοντας τα διαθέσιμα δεδομένα και υπολογίζοντας το πόσο σημαντικά είναι τα δεδομένα των βαθμολογιών και των αλληλεπιδράσεων [Rube15].

Αντίστοιχη με την προσθήκη επιπλέον χαρακτηριστικών στα προφίλ των χρηστών είναι η χρήση κάποιων χαρακτηριστικών των άρθρων για την πρόταση τους στους χρήστες ψυχρής εκκίνησης, όπως για παράδειγμα το πόσο καινούργιο είναι το άρθρο ή το πόσο δημοφιλές είναι. Μια σχετική προσέγγιση είναι και η *διαφοροποίηση των βαρών ομαλοποίησης* (differentiating regularization weights). Η μέθοδος αυτή θέτει μικρότερους περιορισμούς στους λανθάνοντες παράγοντες οι οποίοι συνδέονται με τα αντικείμενα ή τους χρήστες που δίνουν περισσότερη πληροφορία (πχ στα δημοφιλή αντικείμενα και ενεργούς χρήστες) και μεγαλύτερους περιορισμούς στους υπόλοιπους (πχ λιγότερο δημοφιλή αντικείμενα και μη-ενεργούς χρήστες). [Chen19] . Αυτή η μέθοδος έχει δείξει ότι μπορεί να ωφελήσει πολλά μοντέλα ΣΣ, και μπορεί να χρησιμοποιηθεί παράλληλα με άλλες προσεγγίσεις για τη λύση του προβλήματος ψυχρής εκκίνησης.

2.2.2 Έλλειψη Δεδομένων

Σε πολλά ΣΣ, η μοντελοποίηση των χρηστών και των αντικειμένων γίνεται με τη βοήθεια ενός πίνακα που έχει στις στήλες του τα αντικείμενα και στις γραμμές τους χρήστες και σε κάθε κελί την βαθμολογία του χρήστη στο αντίστοιχο αντικείμενο. Συνήθως ο πίνακας αυτός είναι πολύ αραιός, είτε γιατί έχουμε πολλούς χρήστες και λίγα αντικείμενα (ή το αντίστροφο), ή γιατί οι χρήστες δεν βαθμολογούν πολλά αντικείμενα.

Έχουν προταθεί αρκετές λύσεις για την συμπλήρωση του αραίου πίνακα. Μία από αυτές είναι η εύρεση κάποιων (πχ δύο) υπερ-χρηστών οι οποίοι να έχουν βαθμολογήσει αρκετά αντικείμενα του ΣΣ. Με βάση τις βαθμολογίες τους, καθώς και ενός έμμεσου κοινωνικού δικτύου βασισμένο στο ιστορικό των χρηστών, προβλέπεται η βαθμολογία που ενδέχεται να δώσουν άλλοι χρήστες στα ίδια αντικείμενα και έτσι ο πινάκας γίνεται πιο πυκνός [Li14]. Άλλη λύση είναι η χρήση οντολογιών και τεχνολογίας του σημασιολογικού ιστού για την ένταξη πληροφοριών από τα έμμεσα δεδομένα των χρηστών στην αναπαράσταση των αλληλεπιδράσεων χρηστών και δεδομένων. Το πρόβλημα της έλλειψης των δεδομένων λύνεται με τη βοήθεια ενός γράφου, ο οποίος χρησιμοποιείται για συμπλήρωση του πίνακα αλληλεπιδράσεων χρηστών-αντικειμένων μέσω των σημασιολογικών σχέσεων του δικτύου του ΣΣ [Cant11]. Τέλος, μία πιο σύγχρονη λύση στο πρόβλημα του της έλλειψης δεδομένων είναι η χρήση τεχνικών βαθειάς μηχανικής μάθησης, όπως για παράδειγμα της μεταφοράς μάθησης.

2.2.3 Γρήγορη μεταβολή των άρθρων

Το πόσο πρόσφατο είναι κάποιο άρθρο, παίζει σημαντικό ρόλο στις προτιμήσεις του χρήστη, ανεξάρτητα από τον αλγόριθμο που χρησιμοποιείται για την παραγωγή των συστάσεων. Εκτός όμως από τα πρόσφατα άρθρα, ένας χρήστης μπορεί να ενδιαφέρεται και για παλαιότερα άρθρα, για παράδειγμα όταν θέλει να ενημερωθεί για την εξέλιξη ενός γεγονότος ή για άρθρα σχετικά με κάποιο πρόσφατο γεγονός.

Τεχνικά, το πόσο πρόσφατο είναι ένα άρθρο μπορεί να συμπεριληφθεί στην διαδικασία εύρεσης των προτάσεων με τρεις τρόπους:

1. Φιλτράρισμα των υποτιθέμενων ξεπερασμένων άρθρων πριν την εύρεση των προτάσεων ή πριν την ταξινόμηση των προτάσεων (pre-filtering). Για παράδειγμα στο [Desa14] διαλέγονται τα 100 πιο πρόσφατα άρθρα από τις επιλεγμένες ειδησεογραφικές ιστοσελίδες από το χρήστη και μόνο αυτά χρησιμοποιούνται στην εύρεση των προτάσεων.

2. Προσθήκη στα μοντέλα των ΣΣ (recency modeling). Σε αυτή την προσέγγιση, το πόσο πρόσφατο είναι το κάθε άρθρο λαμβάνεται υπόψιν από το μοντέλο, ταυτόχρονα με άλλα χαρακτηριστικά των άρθρων ή και του χρήστη. Η προσέγγιση αυτή έχει το πλεονέκτημα της ισοστάθμισης των διαφορετικών χαρακτηριστικών με έναν ενσωματωμένο τρόπο και όχι με απλό φιλτράρισμα των ξεπερασμένων άρθρων. [Kari18]

3. Φιλτράρισμα των προτάσεων ανάλογα με το πόσο πρόσφατες είναι (post-filtering). Μία προσέγγιση αναφέρεται στο [Zhen13], όπου τα άρθρα πρώτα διατάσσονται σύμφωνα με τη σχετι-

κότητα τους με τα ενδιαφέροντα της ομάδας που ανήκει ο κάθε χρήστης. Στη συνέχεια, αναδιατάσσονται σύμφωνα με το πόσο πρόσφατα και δημοφιλή είναι.

Το πόσο πρόσφατο είναι ένα άρθρο αποτελεί μία σημαντική παράμετρο των ΣΣΕΑ, αλλά εξαρτάται άμεσα από τις προτιμήσεις του κάθε χρήστη. Μπορεί να υπάρχουν χρήστες που να επιθυμούν μία γρήγορη ενημέρωση τα πρωινά πριν τη δουλειά τους και μία σφαιρική ενημέρωση για τα γεγονότα που τους ενδιαφέρουν το σαββατοκύριακο, ή χρήστες που επιθυμούν ένα από τα δύο. Έτσι, το ερώτημα είναι ποιο θα είναι το κατώφλι στη βάση του οποίου θα φιλτράρονται τα άρθρα, δηλαδή μετά από πόσο χρονικό διάστημα από την δημοσίευση ενός άρθρου, αυτό θα θεωρείται ξεπερασμένο. Το κατώφλι αυτό μπορεί να είναι ίσο με ώρες ή μέρες ανάλογα με την ανάδραση που δίνουν οι χρήστες. Εκτός από το κατώφλι, μπορεί να οριστεί και ένας παράγοντας απόσβεσης για τα άρθρα κάθε δευτερόλεπτο ή κάθε μία ώρα από τη στιγμή της δημοσίευσης. Ο συγκεκριμένος παράγοντας θα δίνει για παράδειγμα κάποια αρνητική βαθμολογία στα άρθρα και θα αποτελεί έναν παράγοντα επιλογής ή όχι του άρθρου για σύσταση σε κάποιο χρήστη.

2.2.4 Επεκτασιμότητα

Η *επεκτασιμότητα* (scalability) αποτελεί ακόμα μία πρόκληση για τα ΣΣΕΑ. Οι μεγάλες ειδησεογραφικές ιστοσελίδες έχουν εκατοντάδες εκατομμύρια χρήστες κάθε μήνα. Την ίδια στιγμή, ο αριθμός των άρθρων που μπορούν να αναζητηθούν και να προταθούν στους χρήστες μπορεί να είναι αρκετά μεγάλος, με χιλιάδες άρθρα να προβάλλονται κάθε μέρα [Kari18]. Παρατηρείται λοιπόν μία καθημερινή αύξηση του όγκου των δεδομένων που πρέπει να διαχειριστεί το ΣΣΕΑ. Μία κοινή πρακτική διαχείρισης αυτού του τεράστιου όγκου δεδομένων είναι η χρήση τεχνικών συσταδοποίησης ώστε να επιταχυνθούν οι υπολογισμοί. Μία άλλη λύση στο πρόβλημα της επεκτασιμότητας είναι τα κατανεμημένα συστήματα ή η χρήση πολλαπλών εξυπηρετητών. Με αυτές τις τεχνικές μπορούν οι υπολογισμοί να γίνονται παράλληλα και έτσι να μειωθεί ο χρόνος που απαιτείται ακόμα και αν αυξάνονται τα δεδομένα.

2.2.5 Ποικιλομορφία και Καινοτομία των άρθρων

Η *ποικιλομορφία* (diversity) και η *καινοτομία* (novelty) των άρθρων αποτελούν δύο σχετικές έννοιες των πεδίων της *ανάκτησης πληροφορίας* (information retrieval - IR) και των ΣΣ. Οι δύο αυτές έννοιες, έχουν προσεγγιστεί και από τα δύο πεδία και έτσι έχουν δημιουργηθεί διαφορετικά μοντέλα και μετρικές προσέγγισης τους. Όσον αφορά τα ΣΣΕΑ, οι έννοιες αυτές είναι διαφορετικές, αλλά ως ένα βαθμό και συνδεδεμένες μεταξύ τους. Η διαφορά τους είναι ότι η καινοτομία αναφέρεται στα άρθρα που οι χρήστες μπορεί να μην έχουν διαβάσει, ενώ η ποικιλομορφία στη διαφορετικότητα των θεμάτων που αναφέρονται τα προτεινόμενα άρθρα.

2.2.5.1 Ποικιλομορφία

Οι χρήστες των ΣΣΕΑ μπορεί να έχουν προτίμηση για διαφορετικές θεματικές κάθε φορά που επισκέπτονται μια ειδησεογραφική ιστοσελίδα. Έτσι, το ΣΣ θα πρέπει να μπορεί να προτείνει στους χρήστες μια ποικιλία άρθρων, για να ικανοποιεί κάθε αναγνώστη. Εμπειρικές έρευνες έχουν δείξει ότι η ποικιλομορφία στα προτεινόμενα άρθρα βελτιώνει αρκετά την ποιότητα των συστάσεων [Ekst14]. Συνήθως ορίζεται ως το ποσοστό της διαφορετικότητας μεταξύ των προτεινόμενων άρθρων [Raza21] και υπολογίζεται στο τελικό στάδιο, δηλαδή μετά την παραγωγή των συστάσεων. Υπάρχουν διάφορες μετρικές στα ΣΣΕΑ για την μέτρηση της ποικιλομορφίας όπως η κανονικοποιημένη ποικιλομορφία, η χρονική ποικιλομορφία [Lath10], αλλά η πιο συνηθισμένη είναι η ποικιλομορφία λίστας συστάσεων που υπολογίζεται στη βάση της ομοιότητας των προτεινόμενων άρθρων (intra-list similarity - ILS). Πιο συγκεκριμένα, δοθείσης μίας μετρικής ομοιότητας $sim()$ που λαμβάνει τιμές στο $[-1, +1]$, η ILS για μια λίστα προτεινόμενων άρθρων RL του ΣΣ υπολογίζεται σύμφωνα με την Εξίσωση 2.1

παρακάτω:

$$ILS(RL) = \frac{1}{2} \sum_{i \in RL} \sum_{j \in RL} sim(i, j) \tag{2.1}$$

Η μετρική αυτή υπολογίζεται για κάθε πιθανή λίστα προτάσεων, και κρατείται η λίστα με την μικρότερη τιμή, αφού αυτή θα έχει και τη μεγαλύτερη ποικιλομορφία. Η ILS μπορεί να υπολογιστεί ανάμεσα στις θεματικές, στις κατηγορίες, στις ετικέτες ή στις οντότητες του κάθε άρθρου. Ένα μειονέκτημα της ILS είναι ότι ο υπολογισμός της είναι αρκετά χρονοβόρος, αφού υπολογίζεται για κάθε πιθανή λίστα κάθε πιθανού χρήστη του ΣΣ.

2.2.5.2 Καινοτομία

Εκτός από την ποικιλομορφία στα προτεινόμενα άρθρα, οι χρήστες επιθυμούν και την καινοτομία. Τα άρθρα σε ένα ΣΣΕΑ προέρχονται από διαφορετικές πηγές και έτσι μπορεί να υπάρχει πλεονάζουσα πληροφορία. Για αυτό το λόγο η εύρεση της σημαντικής πληροφορίας είναι ένα απαραίτητο κομμάτι των ΣΣΕΑ, το οποίο επιτυγχάνεται μέσω της καινοτομίας. Καινοτομία ονομάζεται το ποσοστό της άγνωστης και γνωστής πληροφορίας για κάθε χρήστη [Sara17]. Ο εντοπισμός της καινοτομίας των άρθρων γίνεται σε συνδυασμό με τα άρθρα που έχει διαβάσει ήδη ο χρήστης από τα προτεινόμενα άρθρα. Αυτό συμβαίνει γιατί ο χρήστης κάθε φορά που διαβάζει ένα καινούργιο άρθρο ενημερώνεται περισσότερο για ένα θέμα και έτσι πολλά από τα άρθρα του ΣΣ δεν θα είναι πλέον καινότομα για αυτόν. Με βάση αυτά, από την ανάδραση που δίνει ο χρήστης στο ΣΣ θα πρέπει να λαμβάνονται υπόψιν τα γεγονότα για τα οποία είναι ενήμερος, καθώς και το πόσο ενημερωμένος είναι για αυτά. Έτσι, κάθε λίστα προτάσεων θα μπορεί να βαθμολογείται από το ΣΣ ως προς την καινοτομία που έχει για κάθε χρήστη, ανάλογα με την ανάδρασή του. Τυπικά, η καινοτομία μιας λίστας προτεινόμενων άρθρων RL ορίζεται από την Εξίσωση 2.2:

$$Novelty(RL) = \frac{\sum_{i \in RL} - \log_2 p(i)}{|RL|} \tag{2.2}$$

, όπου το $p(i)$ είναι η δημοφιλία του άρθρου και ορίζεται από την Εξίσωση 2.3:

$$p(i) = \frac{|u \in U, r_{ui} \neq \emptyset|}{|U|} \tag{2.3}$$

δηλαδή από τον λόγο των χρηστών $u \in U$ που έχουν προσπελάσει ένα άρθρο i [Kami16].

2.2.6 Επικαιρότητα

Η *επικαιρότητα* (timeliness) έχει την έννοια του να γίνεται κάτι την κατάλληλη στιγμή. Έτσι, στα ΣΣΕΑ μπορεί να εκφράσει αν κάποιο άρθρο είναι επίκαιρο ή όχι. Γενικότερα, κάθε αντικείμενο των ΣΣ έχει τον κύκλο ζωής του. Κάποια αντικείμενα έχουν μικρό κύκλο ζωής, για παράδειγμα τα άρθρα σε μία ειδησεογραφική ιστοσελίδα, και κάποια μεγάλο, όπως μια δημοφιλής ταινία. Έτσι, οι χρήστες των ΣΣΕΑ τείνουν να επιλέγουν κάποιο καινούργιο άρθρο να διαβάσουν, ή κάποιο παλαιότερο αλλά ακόμα δημοφιλές άρθρο. Για να βρεθούν άρθρα με αυτά τα χαρακτηριστικά έχουν προταθεί μοντέλα τα οποία εντάσσονται στις παρακάτω κατηγορίες:

- **Μοντέλα απόσβεσης χρόνου**: Τα μοντέλα αυτά δίνουν περισσότερη βαρύτητα στα αντικείμενα που είναι πιο πρόσφατα σε ένα ΣΣ. Αυτό γίνεται με τη βοήθεια ενός μοντέλου που υπολογίζει τις βραχυπρόθεσμες προτιμήσεις του χρήστη και προβλέπει επίκαιρα άρθρα τα οποία θα διαβάσει αμέσως ο χρήστης. Το χρονικό παράθυρο στο οποίο κάποιο άρθρο θεωρείται επίκαιρο ή όχι, μπορεί να πάρει διαφορετικές τιμές. Γενικά, ένα μεγάλο χρονικό παράθυρο μπορεί να οδηγήσει σε φαινόμενα *ολίσθησης πλαισίου* (concept drift), ενώ ένα μικρό χρονικό παράθυρο μπορεί να μην έχει αρκετά δεδομένα να παρέχει στο μοντέλο. Έτσι, το χρονικό παράθυρο θα πρέπει να οριστεί ανάλογα με τους σκοπούς του ΣΣ.

- **Γραφοθεωρητικά μοντέλα**: Τα μοντέλα αυτά αναπαριστούν τις αλληλεπιδράσεις του χρήστη με τα αντικείμενα του ΣΣ με τη βοήθεια διμερών γράφων. Τα άρθρα και οι χρήστες αναπαρίστανται με τη βοήθεια κόμβων και οι αλληλεπιδράσεις μεταξύ τους μέσω ακμών με ή χωρίς βάρη. Τα μοντέλα επικεντρώνονται στα ακολουθιακά μοτίβα που δίνει ο χρήστης ως ανάδραση στο σύστημα για να προβλέψουν το επόμενο αντικείμενο που θα του προταθεί. Όσο αυξάνονται τα δεδομένα του συστήματος, τα γραφοθεωρητικά αδυνατούν να βρουν τα πολύπλοκα μοτίβα που σχηματίζουν τα δεδομένα και έτσι δεν μπορούν να δώσουν καλές προτάσεις στους πολλούς χρήστες ενός ΣΣ.

- **Μοντέλα δημοφιλίας**: Τα μοντέλα αυτά βασίζονται στο πόσο δημοφιλή είναι τα άρθρα. Χρειάζεται προσοχή στη σχεδίαση τους διότι μπορεί κάποιες καλές προτάσεις για τον χρήστη να μην πραγματοποιηθούν γιατί δεν είναι δημοφιλείς σε άλλους χρήστες. Επίσης, δεν υπάρχει εγγύηση ότι οι προτάσεις του μοντέλου είναι πραγματικά δημοφιλείς και αξιόπιστες.

2.2.7 Μοντελοποίηση χρηστών

Η μοντελοποίηση των χρηστών σε ένα ΣΣ είναι απαραίτητη για να μπορέσει το σύστημα να τους προτείνει αντικείμενα που τους ενδιαφέρουν. Για να γίνει η μοντελοποίηση των χρηστών χρειάζεται να υπάρχουν δεδομένα για κάθε χρήστη. Τα δεδομένα μπορεί να είναι είτε άμεσα, πχ η βαθμολογία που έδωσε ο χρήστης σε κάποιο άρθρο, ή έμμεσα, πχ κλικ σε κάποιο σύνδεσμο. Συγκεκριμένα, τα άμεσα δεδομένα είναι αυτά που μπορεί να μετρηθούν, ενώ τα έμμεσα είναι τα δεδομένα που παρέχει ο χρήστης στην εφαρμογή κατά τη χρήση της. Τα έμμεσα δεδομένα είναι αρκετά χρήσιμα γιατί κάποιος χρήστης, ενώ μπορεί να διαβάσει ένα άρθρο ενδέχεται να μην το βαθμολογήσει, οπότε έτσι να μην δώσει άμεση ανάδραση στο σύστημα. Άλλα παραδείγματα έμμεσων δεδομένων είναι το ιστορικό περιήγησης, ο χρόνος ανάγνωσης ενός άρθρου, το ποσοστό των άρθρων στα οποία κύλισε την μπάρα μέχρι το τέλος (scrolling) κλπ.

Η μοντελοποίηση χρηστών χρειάζεται να αντιμετωπίσει αρκετά ζητήματα. Μερικά είναι η ανωνυμία των χρηστών, ποιες θα είναι οι πληροφορίες που θα συνθέτουν το προφίλ των εγγεγραμμένων χρηστών, η ανάγνωση άρθρων χωρίς ανάδραση, καθώς και η μοντελοποίηση των βραχυχρόνιων και μακροχρόνιων προτιμήσεων των χρηστών.

Για την μοντελοποίηση χρηστών έχουν προταθεί πολλές τεχνικές, οι οποίες μπορούν να χωριστούν στις παρακάτω κατηγορίες [Raza21]:

- **Μοντελοποίηση χρηστών βασισμένη στα χαρακτηριστικά**: Σε αυτή την περίπτωση η μοντελοποίηση του χρήστη γίνεται με βάση το ιστορικό του, δηλαδή τα άρθρα που έχει διαβάσει. Τα άρθρα έχουν κάποια συγκεκριμένα χαρακτηριστικά, όπως τίτλο, κείμενο, κατηγορία, ετικέτες κλπ. Αυτά τα χαρακτηριστικά μπορούν να συνδυαστούν και αν κάποιο άρθρο έχει παρόμοια χαρακτηριστικά με αυτά του προφίλ του χρήστη, να του προταθεί προς ανάγνωση. Το προφίλ που δημιουργείται με αυτό τον τρόπο εκφράζει τις μακροχρόνιες προτιμήσεις του χρήστη μέσα από λέξεις κλειδιά που παίρνει από το ιστορικό των κειμένων που έχει διαβάσει ή από έμμεση ανάδραση που δίνει στο σύστημα. Η μέθοδος αυτή έχει αρκετούς περιορισμούς. Αρχικά, τις περισσότερες φορές είναι δύσκολο να δημιουργηθεί το μοντέλο του χρήστη με αυτό τον τρόπο για δύο λόγους. Πρώτον, οι προτιμήσεις του χρήστη μπορεί να αλλάξουν με την πάροδο του χρόνου, αφού αλλάζουν και τα γεγονότα της επικαιρότητας. Δεύτερον, οι περισσότεροι χρήστες είναι ανώνυμοι και το σύστημα δεν μπορεί να αποθηκεύει το ιστορικό τους. Ακόμα, η χρήση μόνο των χαρακτηριστικών για την υλοποίηση του προφίλ των χρηστών, μπορεί να οδηγήσει σε ελλειπή αναπαράσταση των δεδομένων, επανάληψη στις συστάσεις και υψηλή διάσταση στα δεδομένα. Έτσι, οι συστάσεις προς τον χρήστη δεν θα είναι καινοτόμες και διαφορετικές, αλλά και θα δημιουργηθεί πρόβλημα στο σύστημα με τον μεγάλο όγκο των δεδομένων που θα χρειάζεται για κάθε προφίλ.

- **Συνεργατική διήθηση**: Σε αυτή τη μέθοδο, η δημιουργία του μοντέλου του χρήστη βασίζεται στην εύρεση παρόμοιων αντικειμένων ή χρηστών. Για τη δημιουργία του προφίλ για κάποιο

χρήστη, συλλέγονται κοινά χαρακτηριστικά από χρήστες με κοινά ενδιαφέροντα και αποθηκεύονται στο ιστορικό του χρήστη για κάποιο χρονικό διάστημα. Με βάση αυτά τα ενδιαφέροντα, προβλέπει το σύστημα τα άρθρα που θα προτείνει στον κάθε χρήστη.

- **Μοντελοποίηση χρήστη βασισμένη στα συστήματα γνώσης**: Η χρήση βάσεων γνώσης για τη μοντελοποίηση των χρηστών γίνεται προσθέτοντας σημασιολογικό περιεχόμενο, οντολογίες ή άλλο περιεχόμενο στις προτιμήσεις του χρήστη.

- **Microblogging**: Για την μοντελοποίηση του χρήστη χρησιμοποιούνται οι αλληλεπιδράσεις που έχει με άλλες πλατφόρμες κοινωνικής δικτύωσης όπως λ.χ. το Twitter. Για παράδειγμα, ανάλογα με το περιεχόμενο με το οποίο αλληλεπιδρούν οι χρήστες στο Twitter, το ΣΣΕΑ τους προτείνει και σχετικά άρθρα να διαβάσουν. Η λύση αυτή ενώ έχει αρκετά θετικά, όπως την αντιμετώπιση του προβλήματος ψυχρής εκκίνησης και την παροχή αρκετών δεδομένων για την δημιουργία του προφίλ του χρήστη, απαιτεί την χρήση επιπλέον μετρικών για την διαπίστωση της εγκυρότητας της πληροφορίας.

Ένα παράδειγμα μοντελοποίησης χρήστη που λαμβάνει υπόψιν και τις μακροχρόνιες και τις βραχυχρόνιες προτιμήσεις του χρήστη, περιγράφεται στο [Vian16]. Πιο συγκεκριμένα, το μακροπρόθεσμο προφίλ του χρήστη δημιουργείται με βάση το ιστορικό του χρησιμοποιώντας τις ετικέτες και τη συχνότητα που εμφανίζονται αυτές στα άρθρα που έχει διαβάσει. Αυτά αποτελούν τα άμεσα δεδομένα του μοντέλου. Στη συνέχεια, το βραχυπρόθεσμο προφίλ δημιουργείται χρησιμοποιώντας τις αλληλεπιδράσεις των χρηστών με τα πρόσφατα άρθρα που διάβασαν και μπορεί να μην λαμβάνονται ακόμα υπόψιν από το μακροπρόθεσμο μοντέλο. Τα έμμεσα δεδομένα που χρησιμοποιούνται από το βραχυπρόθεσμο μοντέλο είναι τρία: αν διάβασαν ή όχι κάποιο άρθρο, πχ με το αν το κοιτούν ή όχι για πάνω από 10 δευτερόλεπτα, αν βαθμολόγησαν ένα άρθρο που τους άρεσε και αν μοιράζονται τα άρθρα σε κάποιο μέσο κοινωνικής δικτύωσης. Στα έμμεσα δεδομένα δίνεται και διαφορετικό βάρος, πχ η ανάγνωση έχει βάρος 1, η βαθμολογία βάρος 2 και ο διαμοιρασμός βάρος 3, αφού για να το κοινοποιήσουν σημαίνει ότι τους ενδιαφέρει αρκετά το γεγονός. Αναλύοντας αυτή τη μέθοδο βλέπουμε ότι χρησιμοποιεί τη μοντελοποίηση χρηστών βασισμένη στα χαρακτηριστικά για το μακροχρόνιο προφίλ και το microblogging για το βραχυχρόνιο προφίλ, με χρήση των έμμεσων δεδομένων.

2.3 Τεχνικές και Στρατηγικές Ανάλυσης

Η δημιουργία ενός ΣΣΕΑ αποτελεί ένα δύσκολο πρόβλημα λόγω του δυναμικού περιβάλλοντος των ειδησεογραφικών νέων. Αυτό φαίνεται και από τον μεγάλο αριθμό προκλήσεων που παρουσιάστηκαν στην Ενότητα 2.2. Έτσι, οι αλγόριθμοι συστάσεων θα πρέπει να αρχικά να αντιμετωπίσουν το συνεχώς μεταβαλλόμενο περιβάλλον των άρθρων σε πραγματικό χρόνο, σε συνδυασμό με τις προκλήσεις των ΣΣ. Γενικότερα, οι αλγόριθμοι συστάσεων εντάσσονται στις ακόλουθες πέντε κατηγορίες:

1. Συνεργατικής διήθησης (collaborative filtering - CF)

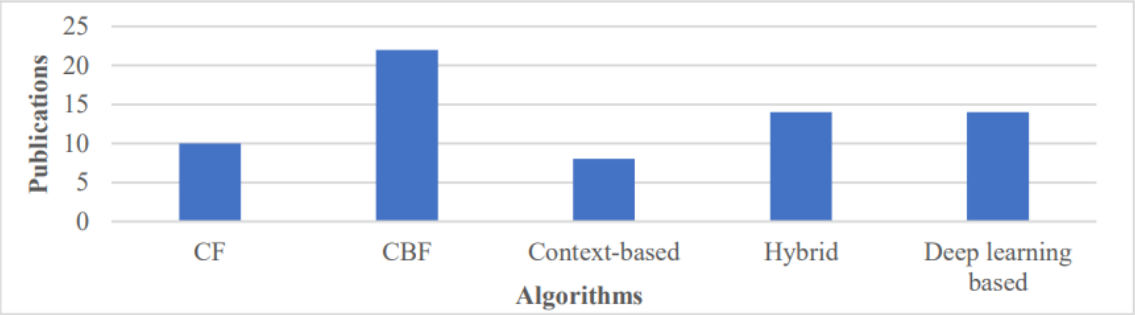
2. Συστήματα με βάση το περιεχόμενο (content-based systems - CB)

3. Συστήματα βασισμένα στη γνώση (knowledge-based - KB)

4. Υβριδικά συστήματα (hybrid systems)

5. Συστήματα βαθειάς μηχανικής μάθησης (Deep Learning-DL)

Από τις προαναφερόμενες κατηγορίες, οι περισσότερα αλγόριθμοι ΣΣΕΑ εντάσσονται στην CB, ενώ έπονται τα υβριδικά συστήματα, τα DL και τέλος τα CF και KB [Raza21]. Στο Σχήμα 2.1 συνοψίζονται οι κατηγορίες που εντάσσονται οι περίπου 70 εργασίες που μελετήθηκαν στο πλαίσιο της παρούσας διπλωματικής εργασίας. Οι DL τεχνικές έχουν αρχίσει να γίνονται δημοφιλείς από το 2016



Σχήμα 2.1: Πλήθος εργασιών ανά κατηγορία ΣΣΕΑ σύμφωνα με το [Raza21]

και μετά, ενώ από ότι φαίνεται όλο και περισσότερα ΣΣ τις χρησιμοποιούν. Πιο συγκεκριμένα, οι εργασίες που μελετήθηκαν κατανέμονται μεταξύ του 2012 και του 2019, όπου οι DL τεχνικές έρχονται δεύτερες στην κατάταξη από το 2016 και μετά.

Κάθε τεχνική χειρίζεται διαφορετικά τα άμεσα και έμμεσα δεδομένα που αποθηκεύει το σύστημα, στην προσπάθεια της να παράγει ενδιαφέρουσες προτάσεις για τους χρήστες. Ακόμα, μπορεί να χρησιμοποιεί διαφορετικά τα δεδομένα που παίρνει για κάθε άρθρο. Έτσι, κάθε τεχνική έχει πλεονεκτήματα αλλά και μειονεκτήματα σε σύγκριση με τις υπόλοιπες. Κάθε μία από τις πέντε τεχνικές που αναφέρθηκαν περιγράφεται αναλυτικότερα στις παρακάτω ενότητες.

2.3.1 Συνεργατική Διήθηση

Η συνεργατική διήθηση είναι μία τεχνική η οποία χρησιμοποιεί σχέσεις μεταξύ των χρηστών ή των αντικειμένων για να προτείνει αντικείμενα στους χρήστες. Συγκεκριμένα, επιλέγει χρήστες ή αντικείμενα που έχουν μεγάλη ομοιότητα με τον χρήστη ή αντικείμενο στόχο, εκτιμώντας και χρησιμοποιώντας τα χαρακτηριστικά τους (πχ βαθμολογίες, σχόλια κλπ) για να κάνει προτάσεις. Τα CF συστήματα χωρίζονται σε τέσσερις μεγάλες κατηγορίες:

- **Μνημονικά CF**: Σε αυτή την κατηγορία χρησιμοποιείται η βαθμολογία των χρηστών για την εύρεση ομοιοτήτων μεταξύ των χρηστών ή των αντικειμένων. Υπάρχουν δύο είδη συστημάτων τα **βασισμένα στο χρήστη** και τα **βασισμένα στο αντικείμενο**.
  Βασισμένα στο χρήστη
  Σε αυτή την περίπτωση σαν σύνολο χρηστών χρησιμοποιούνται οι χρήστες που έχουν δώσει την ίδια βαθμολογία σε ένα αντικείμενο. Στη συνέχεια, προβλέπεται η βαθμολογία του χρήστη για κάποιο αντικείμενο ανάλογα με την βαθμολογία που έχουν δώσει οι άλλοι χρήστες σε αυτό. Το σημαντικό κομμάτι του αλγορίθμου είναι η εύρεση του καλύτερου συνόλου χρηστών, για κάθε χρήστη. Αρχικά, βρίσκεται η ομοιότητα μεταξύ των χρηστών, λ.χ. με τη βοήθεια του συντελεστή συσχέτισης Pearson (Εξίσωση 2.4):

$$\text{simil}(x,y) = \frac{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r_x})(r_{y,i} - \bar{r_y})}{\sqrt{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r_x})^2} \sqrt{\sum_{i \in I_{xy}} (r_{y,i} - \bar{r_y})^2}} \tag{2.4}$$

  ή της ομοιότητας συνημιτόνου (Εξίσωση 2.5):

$$\text{simil}(x,y) = \cos(\vec{x},\vec{y}) = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| \times ||\vec{y}||} = \frac{\sum_{i \in I_{xy}} r_{x,i} r_{y,i}}{\sqrt{\sum_{i \in I_x} r_{x,i}^2} \sqrt{\sum_{i \in I_y} r_{y,i}^2}} \tag{2.5}$$

  όπου $r_{x,i}$ και $r_{y,i}$ οι βαθμολογίες των χρηστών x, y στο αντικείμενο i. Στη συνέχεια, για κάθε χρήστη βρίσκεται το καλύτερο σύνολο χρηστών με τη βοήθεια της τεχνικής των k πλησιέστε-

ρων γειτόνων. Τέλος, προβλέπεται η βαθμολογία που θα δώσει ο χρήστης σε κάποιο αντικείμενο, υπολογίζοντας τη μέση βαθμολογία που έχουν δώσει οι χρήστες του συνόλου χρηστών, που έχουν κοινά ενδιαφέροντα με αυτόν.

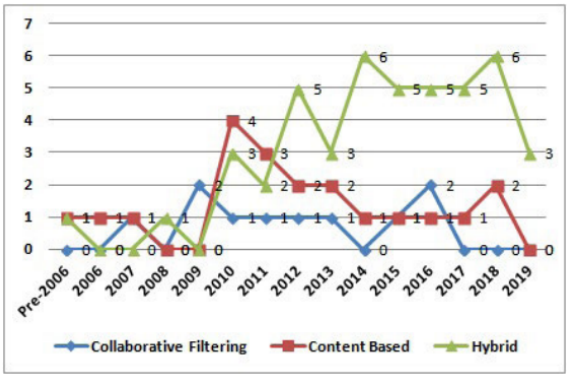
Βασισμένα στο αντικείμενο

Σε αυτή την περίπτωση συγκρίνεται η ομοιότητα που έχουν τα διαφορετικά αντικείμενα που έχει βαθμολογήσει ο χρήστης μεταξύ τους έτσι ώστε να βρεθεί η βαθμολογία του χρήστη για κάποιο παρόμοιο αντικείμενο από τις ήδη υπάρχουσες βαθμολογίες. Η ομοιότητα των αντικειμένων μπορεί να προσεγγιστεί από την ομοιότητα συνημιτόνου (Εξίσωση 2.5). Στη συνέχεια, χρησιμοποιείται πάλι η τεχνική των k πλησιέστερων γειτόνων, αυτή τη φορά για την εύρεση γειτόνων του αντικειμένου i στο σύνολο τον παρόμοιων αντικειμένων του για κάποιο χρήστη u. Τέλος, η βαθμολογία του αντικειμένου i για τον χρήστη u δίνεται από το μέσο όρο των βαθμολογιών των γειτονικών αντικειμένων του i [Dou16].

- **CF βασισμένα σε μοντέλο**: Σε αυτή την προσέγγιση έχουν αναπτυχθεί μοντέλα τα οποία χρησιμοποιούν τεχνικές εξόρυξης δεδομένων και μηχανικής μάθησης για να προβλέψουν τη βαθμολογία των χρηστών στα μη-βαθμολογημένα αντικείμενα του ΣΣ. Υπάρχουν πολλά μοντέλα για αυτή την προσέγγιση, όπως τα μπεϋζιανά δίκτυα, τα μοντέλα συσταδοποίησης, τα λανθάνοντα σημασιολογικά μοντέλα όπως η ανάλυση λανθανουσών τιμών, τα μοντέλα πιθανοτικής λανθάνουσας σημασιολογικής ανάλυσης και τέλος τα Μαρκοβιανά μοντέλα [Su09].

- **Υβριδική CF**: Στην υβριδική περίπτωση , συνδυάζονται οι δύο προηγούμενες μέθοδοι (μνημονικές και βασισμένες σε μοντέλο). Με αυτό τον τρόπο, μειώνονται οι περιορισμοί κάθε μεθόδου και βελτιώνεται η ακρίβεια των συστάσεων. Ωστόσο, στα υβριδικά συστήματα αυξάνεται η πολυπλοκότητα ενώ είναι ακριβή η υλοποίηση τους.

Τα συστήματα συνεργατικής διήθησης, παρά την απλότητα τους, συνδέονται άρρηκτα με τη διάθεση του χρήση να βαθμολογήσει αντικειμενικά τα αντικείμενα του ΣΣ, δηλαδή με την άμεση ανάδραση στο σύστημα. Αυτό δημιουργεί αρκετές προκλήσεις στη σχεδίαση ενός ΣΣ και κυρίως στη σχεδίαση ενός ΣΣΕΑ. Αυτό συμβαίνει γιατί οι χρήστες σε ένα ΣΣΕΑ, δεν τείνουν να βαθμολογούν όλα τα άρθρα που διαβάζουν. Σύμφωνα με πρόσφατη έρευνα, τα τελευταία 14 χρόνια μόνο 8 από τις 81 (13%) εργασίες που μελετήθηκαν έχουν βασιστεί σε τεχνικές CF (Σχήμα 2.2) [Feng20]. Παρόλα αυτά, η εν λόγω τεχνική αυτή χρησιμοποιείται ακόμα σε περιπτώσεις προσωποποίησης ειδησεογραφικών νέων και σε ΣΣΕΑ που είναι σχεδιασμένα για κινητές συσκευές.



Σχήμα 2.2: Τεχνικές για τα ΣΣΕΑ μέχρι το 2019 (Πηγή: [Feng20])

2.3.2 Συστήματα βασισμένα στο περιεχόμενο

Τα CB συστήματα αποτελούν την πιο δημοφιλή προσέγγιση στα ΣΣΕΑ. Σε αντίθεση με τα CF, τα CB προτείνουν στο χρήστη νέα αντικείμενα λαμβάνοντας υπόψιν τα προηγούμενα αντικείμενα που του άρεσαν και την έμμεση ανάδραση που αυτός δίνει στο σύστημα. Έτσι, οι προτάσεις γίνονται πιο

υποκειμενικές, αφού δεν στηρίζονται πλέον στις προτιμήσεις χρηστών με παρόμοια ενδιαφέροντα, αλλά στις ίδιες τις προτιμήσεις του κάθε χρήστη. Αυτή η μέθοδος χρησιμοποιείται κυρίως όταν υπάρχουν αρκετά δεδομένα για τα αντικείμενα του ΣΣ και όχι τόσα πολλά για τον χρήστη. Έτσι, με βάση τα χαρακτηριστικά των αντικειμένων που έχει αλληλεπιδράσει ο χρήστης στο παρελθόν, δημιουργείται το προφίλ του και του προτείνονται καινούργια αντικείμενα. Στα ΣΣΕΑ, υπάρχουν αρκετές πληροφορίες για τα άρθρα, πχ τίτλος, κείμενο, κατηγορία, ετικέτες, οντότητες κλπ, γεγονός που καθιστά εύκολη την σύγκριση μεταξύ τους για την εύρεση καλών προτάσεων για κάποιο χρήστη.

Οι προτάσεις για κάθε χρήστη βρίσκονται συνήθως ακολουθώντας τα εξής βήματα:

1. Αναπαράσταση των άρθρων σε ένα διανυσματικό χώρο με βάση τα χαρακτηριστικά τους

2. Δημιουργία του προφίλ του χρήστη με βάση το ιστορικό του (αναπαράσταση στο ίδιο διανυσματικό χώρο με τα άρθρα)

3. Επιλογή των καλύτερων προτάσεων για το χρήστη συγκρίνοντας τα αποτελέσματα των βημάτων 1 και 2

Η αναπαράσταση των άρθρων στο διανυσματικό χώρο μπορεί να γίνει με διάφορους τρόπους. Ο πιο γνωστός είναι η χρήση της μετρικής TF-IDF (Ενότητα 3.1.2) για την αναπαράσταση των επιλεγμένων χαρακτηριστικών. Οι πιο σύγχρονες τεχνικές είναι η χρήση των *ενσωματώσεων* (embeddings) λέξεων (Ενότητα 3.1.6) ή προτάσεων (Ενότητα 3.1.7). Στη δεύτερη περίπτωση, κάθε άρθρο αναπαρίσταται από τις ενσωματώσεις του τίτλου ή του κειμένου του, στις οποίες μπορούν να συμπεριληφθούν οι κατηγορίες, οι οντότητες και οι ετικέτες του. Αντίστοιχα, το προφίλ των χρηστών μπορεί να δημιουργηθεί είτε με το μέσο όρο του TF-IDF των άρθρων του ιστορικού του, ή με το μέσο όρο των ενσωματώσεων των άρθρων αντίστοιχα. Επίσης, για τη μοντελοποίηση του χρήστη μπορεί να χρησιμοποιηθεί η αγαπημένη του κατηγορία και οι πιο συχνές οντότητες και ετικέτες των άρθρων που διαβάζει, για να συνενωθούν οι ενσωματώσεις τους με τις αντίστοιχες του κειμένου ή του τίτλου, ενισχύοντας και διαφοροποιώντας έτσι το προφίλ του κάθε χρήστη. Το τελευταίο βήμα είναι η παραγωγή των συστάσεων. Έχοντας αναπαραστήσει τα άρθρα και τους χρήστες στον ίδιο διανυσματικό χώρο, μπορούμε να χρησιμοποιήσουμε την ομοιότητα συνημιτόνου για να κατατάξουμε τα άρθρα με βάση την ομοιότητα τους ως προς το προφίλ του χρήστη. Με βάση αυτή τη λίστα προτείνονται στο χρήστη άρθρα προς ανάγνωση.

Η μέθοδος αυτή, εκτός από τα πλεονεκτήματα, όπως την επέκταση της μεθόδου σε πολλούς χρήστες, αφού το σύστημα δεν χρειάζεται πληροφορίες για άλλους χρήστες, καθώς και την πρόταση πιο κατάλληλων άρθρων για κάθε χρήστη, έχει και μειονεκτήματα, τα οποία συνοψίζονται παρακάτω:

- Το μοντέλο μπορεί να είναι τόσο καλό όσο και η επεξεργασία των χαρακτηριστικών των άρθρων και του ιστορικού του χρήστη. Έτσι, η επεξεργασία των χαρακτηριστικών θα πρέπει να εμπεριέχει τις προτιμήσεις του χρήστη ώστε να του προτείνονται ενδιαφέροντα αντικείμενα.

- Το μοντέλο μπορεί να κάνει καλές προτάσεις όταν υπάρχουν αρκετά δεδομένα για στο ιστορικό του χρήστη. Αυτό σημαίνει ότι δεν λύνεται το πρόβλημα της ψυχρής εκκίνησης και έτσι στους χρήστες που μόλις εγγράφονται θα πρέπει να γίνονται προτάσεις με άλλο μοντέλο μέχρι να συλλεχθούν τα απαραίτητα δεδομένα.

- Η αναπαράσταση του χρήστη καθώς και των άρθρων στο διανυσματικό χώρο μπορεί να καταλήξει σε διανύσματα πολλών διαστάσεων τα οποία αυξάνουν το χρόνο και την πολυπλοκότητα των πράξεων.

Για την αντιμετώπιση αυτών των ζητημάτων, η ερευνητική δραστηριότητα σχετικά με τα ΣΣΕΑ έχει στραφεί προς τις υβριδικές τεχνικές.

2.3.3 Υβριδικά συστήματα

Στη προσπάθεια αντιμετώπισης των προβλημάτων που προέκυψαν από τις τεχνικές CF και CB, δημιουργήθηκαν τα υβριδικά συστήματα. Αυτά μπορούν να συνδυάζουν τις τεχνικές CB και CF αλλά και άλλες τεχνικές με τέτοιο τρόπο ώστε να εκμεταλλεύονται τα προτερήματα της κάθε μεθόδου. Υπάρχουν πολλοί τρόποι να συνδυαστούν οι μέθοδοι σε ένα υβριδικό σύστημα. Για παράδειγμα, μπορούν να φτιαχτούν δύο CF και CB συστήματα ξεχωριστά και στη συνέχεια να συνδυαστούν οι προτάσεις τους, έτσι ώστε να ληφθούν οι πιο ενδιαφέρουσες για κάθε χρήστη. Ακόμα, ο συνδυασμός μπορεί να γίνει κατά τη δημιουργία του μοντέλου. Για παράδειγμα, σε ένα CF μοντέλο μπορούν να προστεθούν χαρακτηριστικά ενός CB μοντέλου και αντίστροφα.

Έρευνες έχουν δείξει ότι οι προτάσεις των υβριδικών μοντέλων είναι πιο ακριβείς από τις προτάσεις των CB και CF μοντέλων. Αυτός είναι και ο λόγος που τα τελευταία χρόνια το επιστημονικό ενδιαφέρον έχει κινηθεί προς αυτά τα συστήματα, όπως φαίνεται και από τα Σχήματα 2.1 και 2.2. Από το πρώτο Σχήμα φαίνεται ότι η τεχνική αυτή είναι δεύτερη στην κατάταξη, και από το δεύτερο ότι τα τελευταία χρόνια έχει αυξηθεί αρκετά η χρήση τους σε σχέση με τα μοντέλα CF και CB. Η χρήση υβριδικών συστημάτων μπορεί να αντιμετωπίσει κάποιες κοινές προκλήσεις των ΣΣ, όπως το πρόβλημα της ψυχρής εκκίνησης, της αραιότητας των δεδομένων καθώς και το πρόβλημα της γνωσιακής «συμφόρησης» στα συστήματα που βασίζονται στη γνώση.

Μια γνωστή υπηρεσία που χρησιμοποιεί υβριδικό σύστημα συστάσεων είναι το Netflix [Gome16]. Πρόκειται για τη συνδρομητική υπηρεσία οπτικοακουστικού περιεχομένου η οποία επιτρέπει στους χρήστες της να βλέπουν ταινίες, σειρές κλπ χωρίς διαφημίσεις στο ίντερνετ. Επειδή ο όγκος των ταινιών που έχει το σύστημα είναι μεγάλος, η υπηρεσία κάνει και προτάσεις στους χρήστες της για ταινίες που μπορεί να τους ενδιαφέρουν. Για να το κάνει αυτό συγκρίνει τις ταινίες που είδαν και έψαξαν χρήστες με τα ίδια ενδιαφέροντα με κάποιο χρήστη (CF) και τις συνδυάζει με ταινίες που έχουν κοινά χαρακτηριστικά με ταινίες που έχει βαθμολογήσει υψηλά ο χρήστης (CB).

Στο πεδίο των ΣΣΕΑ έχουν δημιουργηθεί υβριδικά συστήματα συστάσεων ειδησεογραφικών άρθρων. Μερικά από αυτά είναι τα *NewsPer SCENE* και το *PEN* [Raza21].

2.3.4 Συστήματα βασισμένα στη γνώση

Τα συστήματα αυτά χρησιμοποιούν τις πληροφορίες που τους παρέχει ένας *γράφος γνώσης* (knowledge graph - KG) σε συνδυασμό με το μοντέλο που έχει επιλεχθεί για να κάνουν προτάσεις στους χρήστες. Ένας γράφος γνώσης είναι ένας ετερογενής γράφος, όπου οι κόμβοι του αποτελούν τις οντότητες και οι ακμές τις αλληλεπιδράσεις μεταξύ τους. Με τη βοήθεια ενός KG μπορούν να αναπαρασταθούν τα αντικείμενα και τα χαρακτηριστικά τους, καθώς και οι χρήστες και οι πληροφορίες που υπάρχουν για αυτούς. Έτσι, μέσω του γράφου θα καθίσταται δυνατή η αναπαράσταση των αλληλεπιδράσεων των χρηστών με τα αντικείμενα, καθώς και των προτιμήσεων του χρήστη. Ένας γνωστός KG είναι η DBpedia [Auer07], ο οποίος δημιουργείται εξάγοντας δομημένα δεδομένα από λήμματα της Wikipedia [Wiki], σε διάφορες γλώσσες. Η πιο πρόσφατη έκδοση της, DBpedia Diamond, έχει 220 εκατομμύρια οντότητες και 1,45 δισεκατομμύρια τριπλέτες. Με βάση τη DBpedia, μπορεί να δημιουργηθεί κάποιος γράφος γνώσης για ΣΣ. Υπάρχουν τρεις μέθοδοι να χρησιμοποιηθούν οι KG στα ΣΣ: η μέθοδος που βασίζεται στις ενσωματώσεις, η μέθοδος που βασίζεται στα μονοπάτια του γράφου και τέλος ο συνδυασμός τους [Guo20].

- **Μέθοδοι που βασίζονται στα μονοπάτια του γράφου**: Αυτές οι μέθοδοι δημιουργούν ένα γράφο για τους χρήστες και τα αντικείμενα και λαμβάνουν υπόψιν τα μοτίβα των συνδέσεων των οντοτήτων στο γράφο για να κάνουν προτάσεις. Οι εν λόγω γράφοι είναι γνωστοί και ως *δίκτυα ετερογενών πληροφοριών* (heterogeneous information networks - HINs). Γενικότερα, στις μεθόδους που βασίζονται σε HINs συνδυάζονται τεχνικές παραγοντοποίησης πινάκων με μεταδιαδρομές σε ένα HIN, χρησιμοποιώντας τα μονοπάτια στο γράφο για να κανονικοποιήσουν ή να ενισχύσουν τις αναπαραστάσεις των χρηστών ή και των αντικειμένων.

- **Μέθοδοι βασισμένοι στις ενσωματώσεις του γράφου γνώσης**: Τα μοντέλα αυτά, ενισχύουν

την αναπαράσταση των κειμένων και των χρηστών σε ένα ΣΣΕΑ, με εξωτερική πληροφορία η οποία εμπεριέχεται στους γράφους γνώσης. Η πληροφορία αυτή κωδικοποιείται με την βοήθεια των *ενσωματώσεων γράφου γνώσης* (knowledge graph embeddings - KGE). Οι αλγόριθμοι που χρησιμοποιούνται για την εύρεση των KGEs χωρίζονται σε αυτούς που λαμβάνουν υπόψιν την απόσταση των οντοτήτων σε ένα γράφο (translation distance models) και σε αυτούς που λαμβάνουν υπόψιν τη σημασιολογία (semantic matching models).

- **Μέθοδοι που συνδυάζουν και τις δύο τεχνικές**: Σε αυτή την περίπτωση χρησιμοποιούνται από κοινού οι δύο προαναφερόμενες τεχνικές, με στόχο την παραγωγή καλύτερων συστάσεων.

Στα ΣΣΕΑ οι ΚΒ προσεγγίσεις χρησιμοποιούνται για την εύρεση της ομοιότητας αναμέσα στα νέα, βελτιώνοντας έτσι την ακρίβεια των προτάσεων. Τα βήματα για τη δημιουργία ενός KG για ένα ΣΣΕΑ είναι αρχικά η εξαγωγή των οντοτήτων από τον τίτλο ή το κείμενο των άρθρων. Στη συνέχεια, δημιουργούνται υπογράφοι εξάγοντας τους γείτονες των οντοτήτων από κάποια υπάρχουσα βάση γνώσης. Αυτοί οι υπογράφοι χρησιμοποιούνται για την εξαγωγή των προτάσεων.

2.3.5 Συστήματα βαθειάς μάθησης

Τα τελευταία χρόνια η χρήση αλγορίθμων βαθειάς μηχανικής μάθησης για την υλοποίηση ΣΣΕΑ έχει γίνει όλο και πιο δημοφιλής. Οι DL τεχνικές έχουν συμβάλει σημαντικά στην βελτιστοποίηση πολλών ζητημάτων των προηγούμενων μεθόδων, όπως η έλλειψη της δυνατότητας του ΣΣ να μάθει από δεδομένα μεγάλων διαστάσεων, η γραμμικότητα των μοντέλων, η αδυναμία διαχείρισης διαδοχικών εργασιών καθώς και το πρόβλημα της ψυχρής εκκίνησης και της έλλειψης δεδομένων. Ορισμένα χαρακτηριστικά των συστημάτων DL τα οποία συμβάλουν στην αντιμετώπιση των ζητημάτων που προαναφέρθηκαν είναι τα ακόλουθα: [Raza21]

- **Εκμάθηση αναπαραστάσεων** (representation learning) για τα δεδομένα με τη βοήθεια μη-γραμμικών μετασχηματισμών. Τα DL μοντέλα μπορούν να μάθουν τις υποκείμενες αλληλεπιδράσεις μεταξύ των χρηστών και των αντικειμένων, καθώς και τις αναπαραστάσεις τους από ένα μεγάλο σύνολο δεδομένων εισόδου σε ένα ΣΣ.

- **Μη-γραμμικότητα**: Τα DL μοντέλα μπορούν να εντοπίσουν και να μοντελοποιήσουν καλύτερα τις μη-γραμμικές αλληλεπιδράσεις που εμφανίζονται στα δεδομένα. Έτσι, μπορούν να παράγουν πιο σωστές προτάσεις στους χρήστες από τα υπόλοιπα ΣΣ.

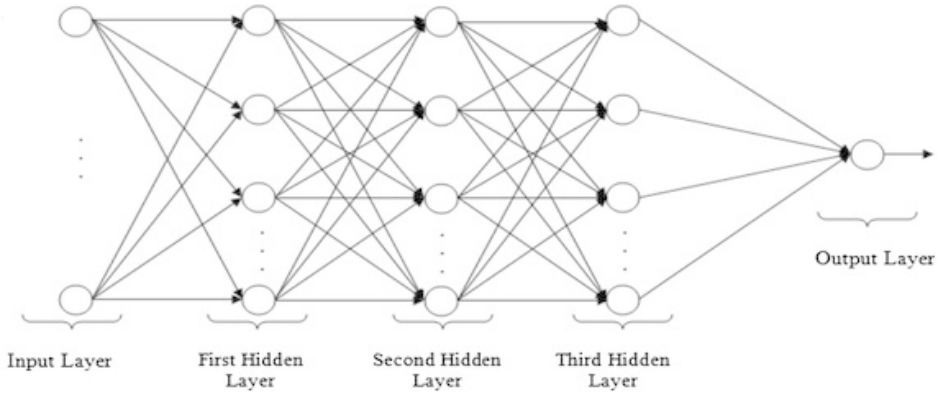
- **Ακολουθιακή μοντελοποίηση**: Τα DL μοντέλα έχουν μεγάλη απόδοση στα διαδοχικά προβλήματα όπως στη σύσταση του επόμενου αντικειμένου και στις συστάσεις καλαθιού αγοράς.

- **Μεταφορά γνώσης** από ένα πεδίο σε κάποιο άλλο σχετικό πεδίο. Η τεχνική αυτή είναι πολύ χρήσιμη σε περιπτώσεις που υπάρχουν προβλήματα ψυχρής εκκίνησης και έλλειψης δεδομένων, αφού το DL έχει ήδη μάθει τις αναπαραστάσεις των δεδομένων από κάποιο άλλο παρόμοιο task.

Για τη δημιουργία των ΣΣΕΑ έχουν αναπτυχθεί αρκετά DL μοντέλα, τα βασικότερα εκ των οποίων παρουσιάζονται στις επόμενες υπο-ενότητες [Raza21].

2.3.5.1 Πολυεπίπεδα Perceptron

Τα *πολυεπίπεδα Perceptrons* (Multi-layer Perceptrons - MLP) είναι ένα νευρωνικά δίκτυα αποτελούμενα από πολλά *επίπεδα* (layers) επεξεργασίας μεταξύ της εισόδου τους και της εξόδου τους. Κάθε κόμβος του δικτύου είναι ένας νευρώνας, ο οποίος ενεργοποιείται από μη-γραμμική συνάρτηση (λ.χ. την ημι-γραμμική). Για την εκπαίδευση του δικτύου χρησιμοποιείται ο αλγόριθμος της προς τα πίσω διάδοσης του σφάλματος (back-propagation). Ένα τυπικό MLP φαίνεται στο Σχήμα 2.3.

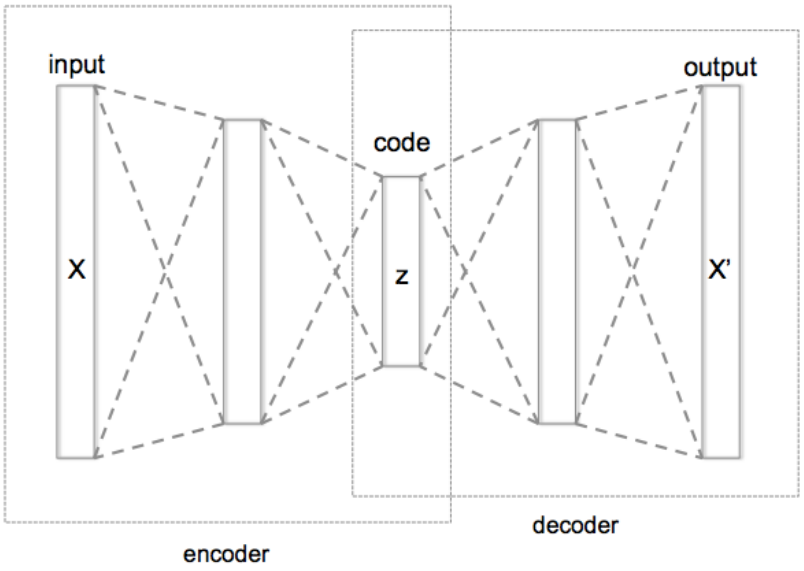


Σχήμα 2.3: MLP με τρία κρυφά επίπεδα (Πηγή: [mlp])

Το MLP μπορεί να χρησιμοποιηθεί για την προσθήκη μη-γραμμικών αλληλεπιδράσεων σε τεχνικές όπως το CF. Για παράδειγμα, στην τεχνική της παραγοντοποίησης πινάκων,το MLP μπορεί να μάθει τις αλληλεπιδράσεις των χρηστών με τα αντικείμενα και να μην χρειαστεί να υπολογιστούν με το εσωτερικό γινόμενο των χαρακτηριστικών. Αυτή η τεχνική ονομάζεται neural collaborative filtering (NCF) [He17]. Γενικότερα, η χρήση των MLP έχει δώσει αρκετά καλά αποτελέσματα στην αναπαράσταση των χαρακτηριστικών των άρθρων σε μικρής κλίμακας ΣΣΕΑ.

2.3.5.2 Αυτοκωδικοποιητής

Ο *αυτοκωδικοποιητής* (autoencoder - AE) είναι μία τεχνική μη-επιβλεπόμενης μάθησης η οποία μαθαίνει πως να συμπιέζει αποτελεσματικά τα δεδομένα εισόδου σε ένα διάνυσμα μικρότερης διάστασης και στη συνέχεια πως να αναπαράγει όσο καλύτερα γίνεται, την είσοδο από το συμπιεσμένο διάνυσμα. Ένα διάγραμμα ενός AE φαίνεται στο Σχήμα 2.4. Στα δεξιά είναι ο κωδικοποιητής που απεικονίζει την είσοδο στο μικρότερης διάστασης επίπεδο κωδικοποίησης και στα αριστερά ο αποκωδικοποιητής που από το επίπεδο κωδικοποίησης προσπαθεί να αναπαράγει την είσοδο στην έξοδο.



Σχήμα 2.4: Σχεδιάγραμμα ενός AE (Πηγή: [ae])

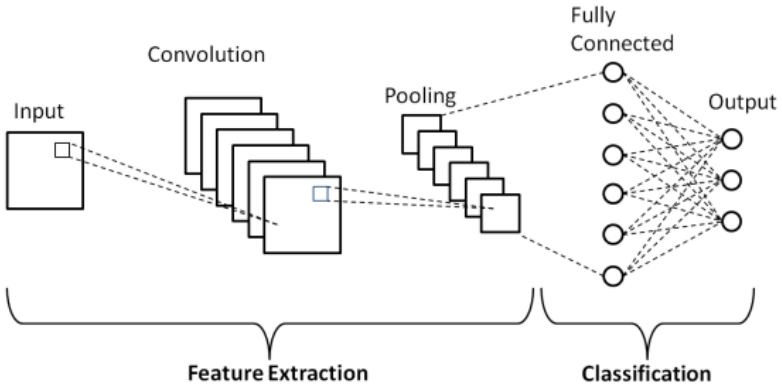
Γενικά, στα ΣΣ, ο AE μπορεί να χρησιμοποιηθεί για να μάθει τη δομή των δεδομένων και για να μπορεί να ανακατασκευάζει τις βαθμολογίες των χρηστών με βάση τις βαθμολογίες που υπάρχουν στο ιστορικό τους. Στα ΣΣΕΑ, όμως, είναι ιδιαίτερα σημαντικό το σύστημα να μαθαίνει ακριβείς αναπαραστάσεις των χρηστών και των κειμένων. Έτσι, το να αναπαραχθεί στην έξοδο του AE αυτούσια

η είσοδος δεν προσφέρει κανενός είδους γνώση. Αυτό το πρόβλημα μπορεί να λυθεί με την προσθήκη κάποιου θορύβου στην είσοδο του ΑΕ. Ο ΑΕ, προσπαθώντας να αφαιρέσει το θόρυβο, μαθαίνει τα πιο σημαντικά χαρακτηριστικά της εισόδου. Αυτή η τεχνική ονομάζεται *απαλοιφή θορύβου*. Με αυτό το τρόπο, σε ένα ΣΣΕΑ μπορούν να αναπαρασταθούν οι χρήστες και τα άρθρα με βάση το ιστορικό, από το διάνυσμα κωδικοποίησης. Λόγω της μη-γραμμικότητας του, ο ΑΕ αποτελεί καλύτερη τεχνική μείωσης διαστάσεων σε σύγκριση με άλλες (λχ. ανάλυση κυρίων συνιστωσών ή PCA), αφού μπορεί να μαθαίνει μικρής διάστασης διανύσματα από μεγάλης διάστασης δεδομένα άρθρων.

Με βάση τα παραπάνω λοιπόν, ο κωδικοποιητής του ΑΕ χρησιμοποιείται για τη μείωση των διαστάσεων αναπαράστασης των χαρακτηριστικών των χρηστών ή των αντικειμένων, ενώ ο αποκωδικοποιητής για την ανασκευή της πληροφορίας εισόδου από το επίπεδο κωδικοποίησης της πληροφορίας.

2.3.5.3 Συνελικτικά νευρωνικά δίκτυα

Τα *συνελικτικά νευρωνικά δίκτυα* (convolutional neural networks - CNNs) είναι νευρωνικά δίκτυα με συνελικτικά επίπεδα. Χρησιμοποιούνται σε ένα ΣΣ για τον εντοπισμό χαρακτηριστικών από εικόνες, κείμενο, φωνή, βίντεο και άλλες μορφές δεδομένων και την κατάταξή τους σε κατηγορίες. Ένα παράδειγμα ενός συνελικτικού δικτύου φαίνεται στο Σχήμα 2.5.



Σχήμα 2.5: Δίκτυο με συνελικτικά επίπεδα (Πηγή: [cnn])

Μετά τα συνελικτικά επίπεδα ακολουθούν αυτά τις *υποδειγματοληψίας* (pooling), όπως για παράδειγμα η υποδειγματοληψία μεγίστου, η οποία έχει ως αποτέλεσμα το μέγιστο αριθμό σε ένα χάρτη χαρακτηριστικών, με βάση ένα παράθυρο στον δισδιάστατο διανυσματικό χώρο. Στις αναπαραστάσεις κειμένων, ένα μεγάλο παράθυρο μπορεί να εντοπίσει μεγάλα μοτίβα στις προτάσεις, ενώ ένα ένα μικρό παράθυρο πιο κοντινά μοτίβα.

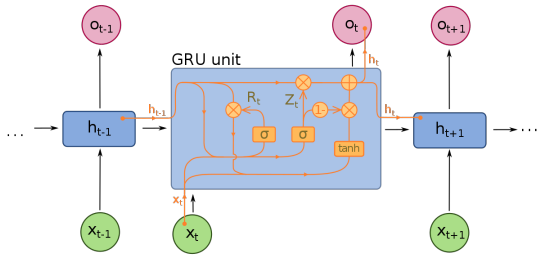
Στα ΣΣΕΑ, τα CNN μπορεί να χρησιμοποιηθεί στην εξαγωγή χαρακτηριστικών από τους τίτλους ή το κείμενο των άρθρων. Αυτό μπορεί να γίνει σε επίπεδο λέξεων ή προτάσεων και κειμένων. Στη συνέχεια, μπορούν να χρησιμοποιηθούν μετρικές ομοιότητας για να βρεθούν άρθρα με όμοια χαρακτηριστικά με αυτά που έχει δει ο χρήστης. Αυτά τα άρθρα θα αποτελούν και τις προτάσεις του ΣΣΕΑ για κάθε χρήστη. Σε σύγκριση με τις τεχνικές NCF, τα CNN έχουν δώσει καλύτερα αποτελέσματα στην εκμάθηση αλληλεπιδράσεων μεταξύ χρηστών και αντικειμένων.

Συμπερασματικά, το CNN είναι ένα καλό μοντέλο για την εκμάθηση αναπαραστάσεων σε επίπεδο λέξεων, προτάσεων ή ακόμα και κειμένων, καθώς και για την εξαγωγή χαρακτηριστικών από δεδομένα που συσχετίζονται με κάθε άρθρο.

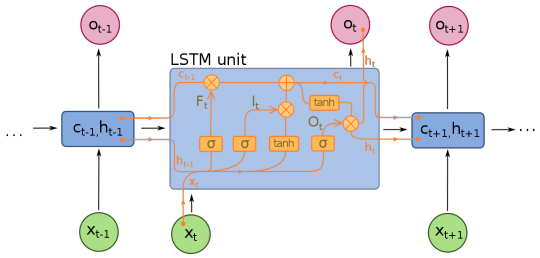
2.3.5.4 Αναδρομικά νευρωνικά δίκτυα

Τα *αναδρομικά νευρωνικά δίκτυα* (recurrent neural networks - RNNs) χρησιμοποιούνται για τη μοντελοποίηση ακολουθιακών δεδομένων. Τα δίκτυα αυτά σε κάθε βήμα κρατάνε ένα διάνυσμα από *μονάδες ενεργοποίησης* (activation units) και μαθαίνουν περίπλοκες αλληλεπιδράσεις στα δεδομένα. Υπάρχουν δύο διαφορετικά RNN μοντέλα, το Long-Short-Term Memory (LSTM) και το Gate

Recurrent Unit (GRU). Η σχεδίαση αυτών των μοντέλων φαίνεται στα Σχήματα 2.6 και 2.7. Η διαφορά μεταξύ τους είναι ότι το GRU, σε αντίθεση με το LSTM, δεν χρειάζεται μονάδες μνήμης και για αυτό εκπαιδεύεται πιο γρήγορα.



Σχήμα 2.6: GRU δίκτυο (Πηγή:[gru])



Σχήμα 2.7: LSTM δίκτυο (Πηγή: [lstm])

Στα ΣΣΕΑ, τα RNN χρησιμοποιούνται για την εκμάθηση αναπαραστάσεων των χρηστών με βάση το ιστορικό τους. Συγκρίνοντας τα αποτελέσματα των δύο μοντέλων, το GRU φαίνεται να δίνει καλύτερα αποτελέσματα από το LSTM, αλλά και τα δύο δίνουν καλύτερα αποτελέσματα από τα παραδοσιακά μοντέλα. Η χρήση των RNNs πρέπει να γίνεται με προσοχή γιατί επηρεάζονται από το πρόβλημα του μεγέθους των *κλίσεων* (gradients) κατά τη διαδικασία εκπαίδευσης του δικτύου με τον αλγόριθμο της προς τα πίσω διάδοσης του σφάλματος. Το πρόβλημα αυτό εμφανίζεται όταν η κλίση της συνάρτησης σφάλματος κατά την ενημέρωση κάποιου βάρους του νευρωνικού δικτύου γίνει πάρα πολύ μικρή ή πάρα πολύ μεγάλη, εμποδίζοντας την ενημέρωση των βαρών. Έτσι, το δίκτυο δεν εκπαιδεύεται σωστά αφού από τις πρώτες εποχές εκπαίδευσης τα βάρη του μοντέλου παραμένουν στις ίδιες τιμές. Για την αντιμετώπιση αυτού του προβλήματος χρησιμοποιείται η τεχνική του *«κουρέματος» των κλίσεων* (gradient clipping).

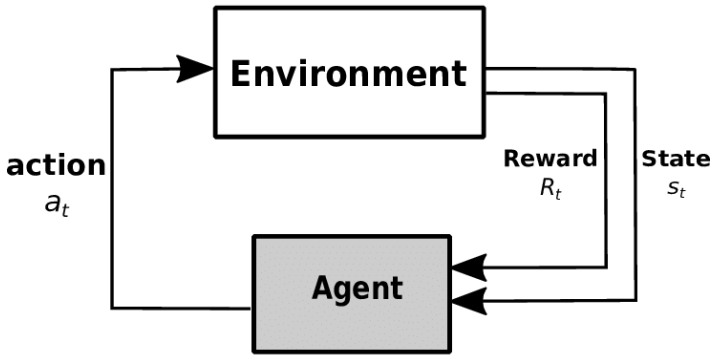
Όσον αφορά τα ΣΣΕΑ, είναι γνωστό ότι κάποιες προτιμήσεις του χρήστη μπορεί να μένουν σταθερές για μια μεγάλη χρονική περίοδο, και άλλες να αλλάζουν αρκετά γρήγορα. Τα RNNs μπορούν να χρησιμοποιηθούν για να εντοπίσουν αυτές τις βραχυχρόνιες και τις μακροχρόνιες προτιμήσεις, εντοπίζοντας ακολουθιακά μοτίβα από το ιστορικό του χρήστη. Για παράδειγμα, μπορούν να εντοπίσουν τις ημερήσιες ή και τις εβδομαδιαίες αλλαγές στις προτιμήσεις του χρήστη. Ακόμα, χρησιμοποιώντας ένα LSTM δίκτυο δύο κατευθύνσεων (bidirectional LSTM), ένα ΣΣΕΑ μπορεί να εντοπίζει και τις βραχυχρόνιες και τις μακροχρόνιες προτιμήσεις του χρήστη, αφού το δίκτυο μπορεί να διαβάσει την ακολουθία εισόδου από τα πιο παλιά άρθρα στα πιο πρόσφατα (forward pass) και αντίστροφα (backward pass).

Συμπερασματικά, τα RNNs μπορούν να χρησιμοποιηθούν για τη μοντελοποίηση ακολουθιακών δεδομένων, να εντοπίσουν ακολουθιακά μοτίβα στο ιστορικό των χρηστών και αντιμετωπίσουν τις δυναμικές αλλαγές στις προτιμήσεις των χρηστών σε περιβάλλοντα συστάσεων βασισμένα στη συνεδρία.

2.3.5.5 Αυτοενισχυόμενη μάθηση

Η *αυτοενισχυόμενη μάθηση* (reinforcement learning - RL) έχει δώσει καλά αποτελέσματα σε πολλούς τομείς όπως παιχνίδια, ρομποτική, οικονομικά, ΣΣ κ.λ.π. Ένα μοντέλο RL αποτελείται από πέντε δομικά στοιχεία και πιο συγκεκριμένα τους *πράκτορες* (agents), το *περιβάλλον* (environment), τις *καταστάσεις* (states), τις *δράσεις* (actions) και τις *ανταμοιβές* (rewards). Ο τρόπος με τον οποίο αλληλεπιδρούν αυτά τα δομικά στοιχεία φαίνεται στο Σχήμα 2.8. Συγκεκριμένα, οι πράκτορες λαμβάνουν αποφάσεις σε ένα περιβάλλον έτσι ώστε να μεγιστοποιηθεί η ανταμοιβή τους και να πάνε στην επόμενη κατάσταση, η οποία είναι η καλύτερη για το σύστημα.

Η αυτοενισχυόμενη μάθηση έχει βρει μικρή εφαρμογή στα ΣΣΕΑ. Για παράδειγμα. στο [Zhen18] περιγράφεται η χρήση ενός αλγορίθμου Deep Q-Learning για την μοντελοποίηση των δυναμικών αλλαγών στις προτιμήσεις ενός χρήστη στα άρθρα.



Σχήμα 2.8: Παράδειγμα ενός RL δικτύου (Πηγή: [rl])

2.3.5.6 Neural Attention

Οι μηχανισμοί neural attention χρησιμοποιούνται για τον εντοπισμό σημαντικών χαρακτηριστικών σε μία είσοδο μεγάλης διάστασης. Στα ΣΣΕΑ, ο μηχανισμός αυτός μπορεί να εντοπίσει σημαντικά σημεία στις ενσωματώσεις της εισόδου, ώστε να μπορεί το μοντέλο σε αυτά να δώσει διαφορετικά βάρη ανάλογα με τη σημασία τους.

Στα ΣΣΕΑ, ο μηχανισμός *προσοχής* (attention) μπορεί να χρησιμοποιηθεί στην εκμάθηση αναπαραστάσεων για τα άρθρα με περισσότερο νόημα, από τις ενσωματώσεις του τίτλου, του κειμένου ή και από τις κατηγορίες και τη θεματολογία τους. Αντίστοιχα, μπορεί να χρησιμοποιηθεί και για την εκμάθηση καλύτερων αναπαραστάσεων των χρηστών με βάση το ιστορικό τους. Για παράδειγμα, στο [Wang18], η αναπαράσταση του χρήστη γίνεται με το ζυγισμένο άθροισμα των ενσωματώσεων των άρθρων που έχει δει ο χρήστης. Τα βάρη κάθε άρθρου δίνονται από το neural attention μοντέλο, το οποίο μοντελοποιεί τις αλληλεπιδράσεις μεταξύ των υποψήφιων νέων, των προτάσεων προς το χρήστη και του ιστορικού του.

Κεφάλαιο 3

Επεξεργασία Κειμένου

3.1 Επεξεργασία Φυσικής Γλώσσας

Η *επεξεργασία φυσικής γλώσσας* (ΕΦΓ ή natural language processing - NLP) είναι ένας διεπιστημονικός κλάδος της επιστήμης της πληροφορικής, της τεχνητής νοημοσύνης και της υπολογιστικής γλωσσολογίας που ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών και των ανθρώπινων (φυσικών) γλωσσών. Κατά συνέπεια, η ΕΦΓ συνδέεται στενά με την αλληλεπίδραση ανθρώπου-υπολογιστή. Οι κυριότερες προκλήσεις στην ΕΦΓ περιλαμβάνουν την κατανόηση φυσικής γλώσσας, δηλαδή την προσπάθεια να καταστούν ικανοί οι υπολογιστές να εξάγουν νοήματα από ανθρώπινα ή γλωσσικά δεδομένα, αλλά και την παραγωγή φυσικής γλώσσας. Για την αναπαράσταση της φυσικής γλώσσας σε δεδομένα εισόδου για τον υπολογιστή, έχουν προταθεί διαφορετικές μέθοδοι, οι κυριότερες εκ των οποίων περιγράφονται στις παρακάτω υπό-ενότητες.

3.1.1 Μοντέλο «σάκου» λέξεων

Το μοντέλο *«σάκου» λέξεων* (bag-of-words - BoW) είναι μια απλή αναπαράσταση, η οποία χρησιμοποιείται στην ΕΦΓ. Σε αυτό το μοντέλο, το κείμενο αναπαρίσταται σαν σάκος (πολυσύνολο) των λέξεων που περιέχει, χωρίς να λαμβάνει υπόψιν τη γραμματική και τη σειρά τους, αλλά μόνο την συχνότητα εμφάνισης τους [Zhan10].

	she	loves	pizza	is	delicious	a	good	person	people	are	the	best
She loves pizza, pizza is delicious	1	1	2	1	1	0	0	0	0	0	0	0
She is a good person	1	0	0	1	0	1	1	1	0	0	0	0
good people are the best	0	0	0	0	0	0	1	0	1	1	1	1

Σχήμα 3.1: Παράδειγμα για το BoW

Στο Σχήμα 3.1 βλέπουμε τον τρόπο που αναπαρίσταται το κείμενο με την συγκεκριμένη μέθοδο. Αρχικά, βρίσκεται το λεξιλόγιο των κειμένων, δηλαδή όλες οι διαφορετικές λέξεις που εμφανίζονται στα κείμενα. Στη συνέχεια, δημιουργείται ένα διάνυσμα που περιέχει αυτές τις λέξεις. Τέλος, για κάθε λέξη του διανύσματος του λεξιλογίου, μετράται το πλήθος των φορών που εμφανίζεται η λέξη σε κάθε κείμενο. Το πλήθος αποθηκεύεται στην αντίστοιχη θέση στο διάνυσμα του κάθε κειμένου. Έτσι, σύμφωνα με το παράδειγμα, στο πρώτο κείμενο η λέξη pizza εμφανίζεται 2 φορές ενώ στα υπόλοιπα δύο καμία φορά.

3.1.2 Συχνότητα όρου - αντίστροφη συχνότητα κειμένου

Όπως αναφέρθηκε και στην Ενότητα 3.1.1, η *συχνότητα όρου - αντίστροφη συχνότητα κειμένου* (term frequency-inverse document frequency ή TF-IDF), είναι μία στατιστική μετρική η οποία εκφράζει πόσο σημαντική είναι μία λέξη για κάποιο κείμενο ενός συνόλου κειμένων. Συνήθως, χρησιμοποιείται σαν σταθμικός παράγοντας σε προβλήματα ανάκτησης πληροφορίες, εξόρυξης γνώσης από κείμενο και μοντελοποίησης χρήστη. Επίσης χρησιμοποιείται για να δώσει περισσότερη βαρύτητα σε λέξεις σε αλγορίθμους μηχανικής μάθησης για την ΕΦΓ. Το βάρος κάθε λέξης προκύπτει από την Εξίσωση 3.1

$$tf - idf(t, d, D) = tf(t, d) * idf(t, D) \tag{3.1}$$

όπου t όρος του κειμένου, d πλήθος κειμένων που περιέχουν τη λέξη, D πλήθος κειμένων του συνόλου δεδομένων. Πιο συγκερκιμένα, το TF-IDF προκύπτει ως το γινόμενο δύο ποσοτήτων, της συχνότητας του κάθε όρου σε ένα κείμενο (Εξίσωση 3.2) και της αντίστροφης συχνότητας εμφάνισης του συγκερκιμένου όρου στη συλλογή των κειμένων (Εξίσωση 3.3).

$$tf(t, d) = log(1 + freq(t, d)) \tag{3.2}$$

$$idf(t, D) = log(\frac{|D|}{count(d \in D : t \in d)}) \tag{3.3}$$

Η συχνότητα εμφάνισης ενός όρου σε κάθε κείμενο δεν είναι από μόνη της αρκετή για να υπολογιστεί ένα σωστό βάρος για κάθε λέξη, διότι υπάρχουν λέξεις που μπορεί να εμφανίζονται περισσότερες φορές από άλλες, αλλά να μην προσδίδουν στο νόημα του κειμένου. Έτσι, για τη στάθμιση αυτού του φαινομένου, χρησιμοποιείται και η αντίστροφη συχνότητα εμφάνισης του όρου στη συλλογή των κειμένων, μειώνοντας κατ' αυτό τον τρόπο το βάρος των λέξεων που εμφανίζονται πολλές φορές και αυξάνοντας το βάρος των λέξεων που εμφανίζονται πιο σπάνια, αφού οι σπάνιες λέξεις ενδέχεται να δίνουν περισσότερο νόημα στο κείμενο. Η αντίστροφη συνχότητα εμφάνισης του όρου στη συλλογή των κειμένων εκφράζεται ως το κλάσμα του πλήθους των κειμένων του συνόλου δεδομένων διά το πλήθος των κειμένων που εμφανίζεται ο εκάστοτε όρος [Samm10].

3.1.3 Ενσωματώσεις λέξεων

Στην ΕΦΓ, οι *ενσωματώσεις λέξεων* (word embeddings) χρησιμοποιούνται για την αναπαράσταση των λέξεων στην ανάλυση κειμένου. Τυπικά μια ενσωμάτωση λέξης έχει τη μορφή ενός διανύσματος πραγματικών αριθμών, το οποίο κωδικοποιεί την έννοια μίας λέξης. Το διάνυσμα κάθε λέξης δημιουργείται έτσι ώστε να έχει μικρή απόσταση από διανύσματα άλλων λέξεων που εκφράζουν παρόμοιες έννοιες.

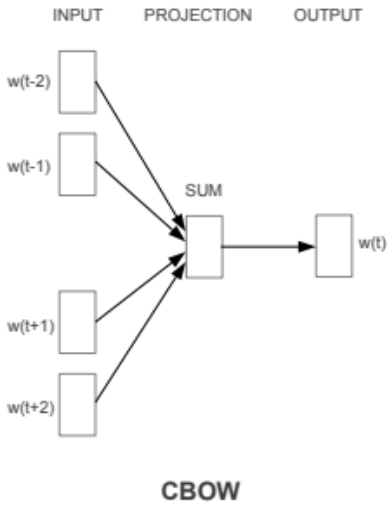
Οι ενσωματώσεις λέξεων δημιουργούνται χρησιμοποιώντας ένα σύνολο από τεχνικές μοντελοποίησης γλώσσας και από τεχνικές εκμάθησης χαρακτηριστικών, όπου οι λέξεις από το λεξιλόγιο ενός κειμένου, αντιστοιχίζονται σε διανύσματα πραγματικών αριθμών. Οι μέθοδοι που χρησιμοποιούνται για την δημιουργία των ενσωματώσεων περιλαμβάνουν τα νευρωνικά δίκτυα, τη μείωση διαστάσεων στον πίνακα συνύπαρξης των λέξεων (co-occurrence matrix), πιθανοτικά μοντέλα, μοντέλα βάσεων γνώσης και άμεσες αναπαραστάσεις σε όρους ανάλογα με την έννοια που αναφέρονται οι λέξεις. Η χρήση των ενσωματώσεων λέξεων φαίνεται να αυξάνει την αποδοτικότητα στις εφαρμογές της ΕΦΓ όπως η συντακτική ανάλυση και η ανάλυση συναισθήματος.

Έχουν προταθεί πολλά μοντέλα αναπαράστασης φυσικής γλώσσας με τη χρήση ενσωματώσεων λέξεων, όπως το **WordVec** [Miko13], το **FastText** [Boja16] και το **BERT** [Devl19]. Στα πλαίσια της παρούσας διπλωματικής εργασίας εξετάστηκαν αυτά τα μοντέλα με τη χρήση προεκπαιδευμένων ενσωματώσεων λέξεων. Πιο συγκερκιμένα, για το BERT χρησιμοποιήθηκαν οι προεκπαιδευμένες ενσωματώσεις λέξεων του Greek-BERT [Kout20], που έχει αναπτυχθεί από ερευνητές του Οικονομικού Πανεπιστημίου Αθηνών.

3.1.4 Word2Vec

Το Word2Vec [Miko13] είναι μία τεχνική εκμάθησης ενσωματώσεων λέξεων της ΕΦΓ η οποία χρησιμοποιεί ένα μοντέλο νευρωνικού δικτύου για την εκμάθηση συσχετίσεων μεταξύ των λέξεων ενός μεγάλου συνόλου κειμένων. Όταν το μοντέλο εκπαιδευτεί μπορεί να εντοπίσει συνώνυμες λέξεις ή να προτείνει λέξεις για να συνεχιστεί μία πρόταση. Το μοντέλο αποτελείται δύο στρώματα, τα οποία εκπαιδεύονται για να ανακατασκευάζουν γλωσσικά πλαίσια λέξεων. Παίρνει σαν είσοδο ένα μεγάλο σύνολο κειμένων και παράγει έναν διανυσματικό χώρο συνήθως εκατοντάδων διαστάσεων. Κάθε μοναδική λέξη του συνόλου των κειμένων αντιστοιχίζεται σε ένα διάνυσμα στο χώρο (ενσωμάτωση λέξης). Τα διανύσματα δύο ή περισσότερων λέξεων, είναι τοποθετημένα κοντά στο χώρο, αν οι λέξεις αναφέρονται στην ίδια έννοια.

Υπάρχουν δύο διαφορετικές αρχιτεκτονικές που μπορούν να χρησιμοποιηθούν για να παραχθεί μία κατανεμημένη αναπαράσταση των λέξεων, η continuous bag-of-words (CBOW) και η continuous skip gram. Στην πρώτη (Σχήμα 3.2) προβλέπεται η λέξη την χρονική στιγμή t, με βάση τις λέξεις ενός χρονικού παραθύρου (πχ $[t-2, t+2]$). Η σειρά των λέξεων δεν επηρεάζει την πρόβλεψη, όπως έχει αναφερθεί και στην προηγούμενη Ενότητα 3.1.1 για το μοντέλο BOW.



Σχήμα 3.2: Αρχιτεκτονική του μοντέλου CBOW (Πηγή: [Miko13])

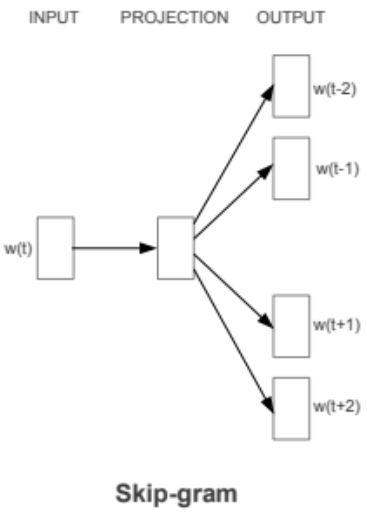
Στην περίπτωση του continuous skip gram, το μοντέλο χρησιμοποιεί την λέξη την χρονική στιγμή t για να βρει το παράθυρο σχετικών λέξεων που περιβάλουν τη λέξη. Από το Σχήμα 3.3, βλέπουμε ότι την χρονική στιγμή t έχουμε την ενσωμάτωση λέξης $w(t)$ και βρίσκουμε το πάραθυρο που σε αυτή την περίπτωση είναι το $[t-2, t+2]$. Το μοντέλο αυτό δίνει περισσότερο βάρος στις λέξεις που είναι πιο κοντά σημασιολογικά. Συγκρίνοντας τα δύο μοντέλα, το μοντέλο CBOW είναι πιο γρήγορο, ενώ το continuous skip gram δουλεύει καλύτερα για τις λέξεις που εμφανίζονται λιγότερο συχνά [Miko13].

¨

3.1.5 FastText

Το FastText είναι μία βιβλιοθήκη για την εκμάθηση ενσωματώσεων λέξεων και για την κατηγοριοποίηση κειμένου, η οποία έχει δημιουργηθεί από το AI Research Lab του Facebook και έχει κάνει διαθέσιμα προεκπαιδευμένα διανύσματα σε 294 γλώσσες [Boja16]. Αποτελεί επέκταση του μοντέλου Word2Vec, χρησιμοποιώντας τα ίδια νευρωνικά μοντέλα που περιγράφηκαν στην Ενότητα 3.1.4.

Η διαφορά των δύο μοντέλων έγκειται στην αναπαράσταση των λέξεων σε n-grams, δηλαδή σε μία συνεχόμενη αλληλουχία n χαρακτήρων μίας συμβολοσειράς. Στην συγκεκριμένη περίπτωση, κάθε λέξη των κειμένων του συνόλου των δεδομένων, χωρίζεται σε n τμήματα. Για παράδειγμα, αν το $n = 3$ και θέλουμε να χωρίσουμε τη λέξη «παιδί», θα έχουμε τον εξής διαχωρισμό: <πα, παι, αιδ,



Σχήμα 3.3: Αρχιτεκτονική του μοντέλου Skip-Gram (Πηγή: [Miko13])

ιδί, δί> . Τα $<, >$ προστίθενται για να ξεχωρίζεται το n-gram μίας λέξης από την ίδια την λέξη. Αυτό βοηθάει στο να μπορεί να διατηρηθεί το νόημα μικρότερων λέξεων που μπορεί να εμφανιστούν σε n-grams άλλων λέξεων. Επίσης, βοηθάει στο να γίνεται κατανοητό το νόημα των προθεμάτων και των καταλήξεων των λέξεων.

Έτσι, στο μοντέλο FastText, κάθε λέξη αναπαρίσταται από το σύνολο των n-gram που την αποτελούν, στο οποίο προστίθεται και η λέξη με τα σύμβολα $<, >$ στην αρχή και στο τέλος αντίστοιχα. Με αυτή την αναπαράσταση, δημιουργούνται ενσωματώσεις λέξεων που εμπεριέχουν πληροφορίες στο επίπεδο της *υπό-λέξης* (sub-word).

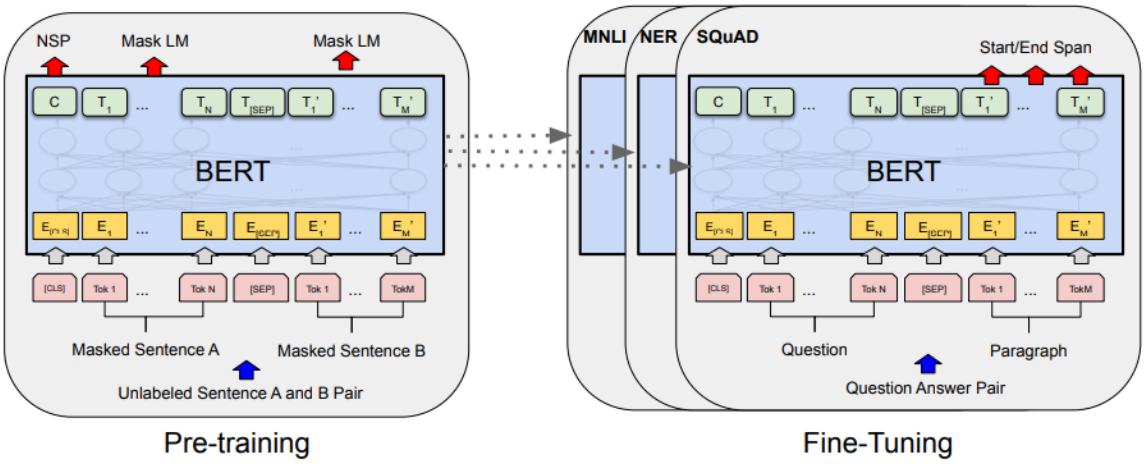
Οι ενσωματώσεις FastText φαίνεται να είναι πιο ακριβείς από τις αντίστοιχες του Word2vec, ως προς διάφορες μετρικές. Το χαρακτηριστικό των n-grams είναι η πιο σημαντική βελτίωση, η οποία μπορεί να λύσει το πρόβλημα των λέξεων εκτός λεξικού, αφού ακόμα και αν μια λέξη δεν υπάρχει στο λεξιλόγιο, αν μοιράζεται κοινό πρόθεμα με κάποια από τις υπάρχουσες λέξεις, τότε το νευρωνικό θα θεωρήσει ότι οι λέξεις είναι κοντά στον διανυσματικό χώρο. Επίσης με τη χρήση των n-grams λαμβάνεται υπόψιν η εσωτερική δομή των λέξεων κατά τη μάθηση των ενσωματώσεων λέξης.

3.1.6 BERT

Το μοντέλο BERT (bidirectional encoder representations from transformers) [Devl19] είναι σχεδιασμένο να προ-εκπαιδεύει βαθιές αμφίδρομες αναπαραστάσεις από κείμενα χωρίς ετικέτα, λαμβάνοντας υπόψιν το νόημα δεξιά και αριστερά από την εκάστοτε θέση στο κείμενο, σε κάθε επίπεδο. Τα προεκπαιδευμένα BERT μοντέλα μπορούν να βελτιστοποιηθούν προσθέτοντας ένα επιπλέον επίπεδο εξόδου. Με αυτό τον τρόπο, μπορούν να δημιουργηθούν αποδοτικά μοντέλα για διάφορα προβλήματα όπως λ.χ. η *απάντηση ερωτήσεων* (question answering), χωρίς σημαντικές αλλαγές στην αρχιτεκτονική. Την παρούσα χρονική στιγμή, το BERT εμφανίζει τα καλύτερα αποτελέσματα σε έντεκα τομείς της ΕΦΓ.

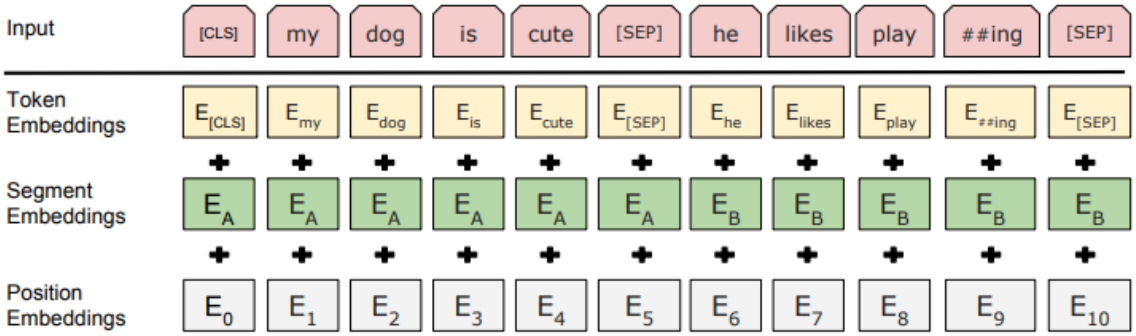
Το BERT βελτιώνει την απόδοσή του απαλείφοντας τον περιορισμό της μονοκατευθυντικότητας κατά την κατασκευή των ενσωματώσεων λέξεων. Αυτό επιτυγχάνεται με την χρήση ενός *γλωσσικού μοντέλου μάσκας* (masked language model - MLM), το οποίο κρύβει τυχαία κάποια σύμβολα της εισόδου, με τον στόχο να είναι να προβλεφθούν τα κρυμμένα σύμβολα από τα συμφραζόμενα. Σε αντίθεση με τα γλωσσικά μοντέλα που βασίζονται στην επεξεργασία των λέξεων από τα αριστερά προς τα δεξιά, στόχος του MLM είναι να επιτρέπει στην αναπαράσταση του κειμένου να συγχωνεύει το νόημα των λέξεων και από τα αριστερά και από τα δεξιά τους. Εκτός από το MLM, το BERT μπορεί να χρησιμοποιηθεί και για την πρόβλεψη της επόμενης πρότασης.

Η λειτουργία του BERT βασίζεται σε δύο στάδια, αυτό της *προ-εκπαίδευσης* και αυτό της *προσαρμογής*, όπως απεικονίζονται στο Σχήμα 3.4. Στη φάση της προ-εκπαίδευσης, το μοντέλο εκπαιδεύεται σε δεδομένα χωρίς ετικέτα σε διαφορετικά προβλήματα. Στο στάδιο της προσαρμογής, το μοντέλο αρχικοποιείται με τις προ-εκπαιδευμένες παραμέτρους, οι τιμές των οποίων προσαρμόζονται χρησιμοποιώντας δεδομένα με ετικέτα.



Σχήμα 3.4: Προεκπαίδευση και προσαρμογή για το BERT (Πηγή: [Devl19])

Η είσοδος στο BERT μοντέλο αναπαριστά κάθε φορά ή μία πρόταση, ή δύο προτάσεις (ερώτηση-απάντηση) σε μια ακολουθία συμβόλων. Το πρώτο σύμβολο είναι το δεσμευμένο σύμβολο CLS (classification token). Η τελική τιμή του διανύσματος κάθε λέξης προκύπτει αθροίζοντας το διάνυσμα του συμβόλου και το διάνυσμα της θέσης, όπως φαίνεται στο Σχήμα 3.5. Αν σαν είσοδο έχουμε δύο προτάσεις, αυτές χωρίζονται με ένα ειδικό σύμβολο, το SEP [Devl19].



Σχήμα 3.5: Παράδειγμα εισόδου για το BERT (Πηγή: [Devl19])

Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκαν προεκπαιδευμένες ενσωματώσεις λέξεων για την ελληνική γλώσσα, οι οποίες προήλθαν από το Greek BERT [Kout20]. Πιο συγκεκριμένα, χρησιμοποιήθηκε το μοντέλο `bert-base-greek-uncased-v1`, το οποίο περιλαμβάνει κείμενα από την ελληνική Wikipedia [Wiki], το ελληνικό κομμάτι του European Parliament Proceedings Parallel Corpus [Koeh05] και το ελληνικό κομμάτι του OSCAR [Suar19]. Το μοντέλο έχει εκπαιδευτεί χρησιμοποιώντας τον κώδικα του BERT και είναι παρόμοιο με το αγγλικό μοντέλο `bert-base-uncased model` (12 κρυφά επίπεδα, 768 νευρώνες σε κάθε κρυφό επίπεδο, 12 έξοδοι, 110 εκ. παράμετροι). Επίσης, κατά την εκπαίδευσή του ελληνικού μοντέλου έχει ακολουθηθεί η ίδια διαδικασία με το αγγλικό (1 εκ. βήματα εκπαίδευσης, ομάδες 256 κειμένων μήκους 512 χαρακτήρων το κάθε ένα και αρχικός ρυθμός μάθησης 10^{-4}).

3.1.7 Ενσωματώσεις προτάσεων

Αντίστοιχα με τις ενσωματώσεις λέξεων, οι ενσωματώσεις προτάσεων προκύπτουν από την αντιστοίχιση των προτάσεων ενός κειμένου σε ένα διάνυσμα πραγματικών αριθμών. Ένας τρόπος να προκύψει η ενσωμάτωση για μία πρόταση θα μπορούσε να είναι να υπολογιστεί ο σταθμισμένος μέσος των ενσωματώσεων των λέξεων που την αποτελούν. Με αυτό τον τρόπο όμως δεν λαμβάνονται υπόψιν οι αλληλεπιδράσεις μεταξύ των λέξεων σε μία πρόταση ή η σειρά των λέξεων. Για αυτό το λόγο έχουν προταθεί διαφορετικοί αλγόριθμοι υπολογισμού ενσωματώσεων προτάσεων, όπως ο *sentence BERT* (SBERT) [Reim19].

Ο SBERT αποτελεί μία διαφοροποίηση του προεκπαιδευμένου BERT δικτύου, το οποίο χρησιμοποιεί αρχιτεκτονικές κατάλληλες για την εύρεση σημασιολογικά κοντινών ενσωματώσεων προτάσεων στο διανυσματικό χώρο, οι οποίες μπορούν να συγκριθούν χρησιμοποιώντας την ομοιότητα συνημιτόνου. Με αυτό τον τρόπο, μειώνεται η προσπάθεια για την εύρεση των σημασιολογικά κοντινών προτάσεων από 65 ώρες με τη χρήση του BERT, σε περίπου 5 δευτερόλεπτα με τη χρήση του SBERT, διατηρώντας την ίδια ακρίβεια.

Ο αλγόριθμος χρησιμοποιεί σύγχρονες τεχνικές της ΕΦΓ για τη δημιουργία των ενσωματώσεων προτάσεων, όπως:

- **Μηχανισμούς προσοχής** που επιτρέπουν στον αλγόριθμο να δημιουργεί τις ενσωματώσεις των προτάσεων, δίνοντας έμφαση στα πιο σημαντικά σημεία της εισόδου.

- **Μετασχηματιστές**. Η αρχιτεκτονική των *μετασχηματιστών* (transformers) [Vasw17] δημιουργήθηκε από την Google Brain και το Πανεπιστήμιου του Τορόντο το 2017. Η αρχιτεκτονική αυτή χρησιμοποιεί τον μηχανισμό προσοχής σε ένα νευρωνικό δίκτυο το οποίο μπορεί να παραλληλιστεί έτσι ώστε το μοντέλο να εκπαιδεύεται σε λιγότερο χρόνο.

- **Μοντέλο BERT** (Ενότητα 3.1.6).

- **Siamese Network**: Αποτελεί ένα νευρωνικό δίκτυο το όποιο εκπαιδεύεται για να συγκρίνει την ομοιότητα μεταξύ των εισόδων. Σε αυτή την περίπτωση, το Sentence-BERT εκπαιδεύεται να υπολογίζει τις ομοιότητες μεταξύ δύο προτάσεων εισόδου. Το κλειδί είναι ότι παράγει εσωτερικές αναπαραστάσεις των προτάσεων για να χρησιμοποιηθούν σε προβλήματα εύρεσης ομοιότητας. Αυτές οι αναπαραστάσεις δημιουργούνται χρησιμοποιώντας δύο BERT δίκτυα σε μια ειδική σύνδεση.

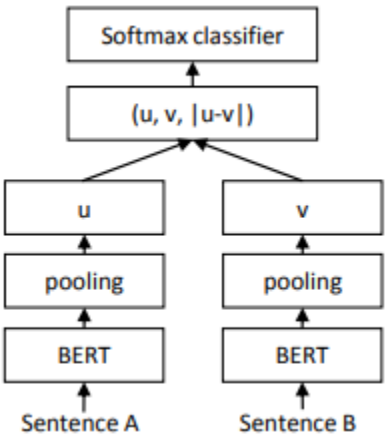
Το SBERT προσθέτει ένα επίπεδο υποδειγματοληψίας στην έξοδο του BERT για να καταλήξει σε μια συγκεκριμένου μεγέθους ενσωμάτωση πρότασης. Έχουν γίνει πειράματα με τρεις διαφορετικές τεχνικές υποδειγματοληψίας (χρήση της εξόδου του συμβόλου CLS, εύρεση της μέσης τιμής όλων των διανυσμάτων εξόδου και εύρεση του μέγιστου των διανυσμάτων εξόδου) με την πιο συχνά χρησιμοποιούμενη να είναι η μέση τιμή.

Η προσαρμογή του SBERT βασίζεται στην δημιουργία δύο δικτύων κατά την εκπαίδευση των οποίων ενημερώνονται τα βάρη, έτσι ώστε οι παραγόμενες ενσωματώσεις προτάσεων να είναι κοντά στο χώρο, όταν είναι σημασιολογικά κοντά και οι αντίστοιχες προτάσεις. Αυτό επιτυγχάνεται με την χρήση τριών αντικειμενικών συναρτήσεων και πιο συγκεκριμένα μιας συνάρτησης ταξινόμησης, μιας συνάρτησης παλινδρόμησης και μιας συνάρτησης τριπλέτας.

Στην περίπτωση της αντικειμενικής συνάρτησης ταξινόμησης συνενώνονται οι ενσωματώσεις πρότασης u, v με την διαφορά τους και στη συνέχεια γίνεται ο πολλαπλασιασμός με τα εκπαιδευμένα βάρη $W_t \in R^{3nxk}$ (Εξίσωση 3.4)

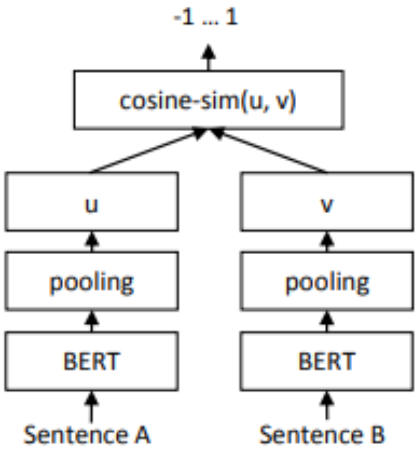
$$o = softmax(Wt(u, v, |u-v|)) \tag{3.4}$$

όπου n είναι η διάσταση των ενσωματώσεων πρότασης και k είναι ο αριθμός των ετικετών. Έτσι γίνεται βελτιστοποίηση της διασταυρούμενης εντροπίας. Αυτή η αρχιτεκτονική απεικονίζεται στο Σχήμα 3.6.



Σχήμα 3.6: Αρχιτεκτονική μοντέλου με χρήση συνάρτησης ταξινόμησης (Πηγή: [Reim19])

Στην περίπτωση της αντικειμενικής συνάρτησης παλινδρόμησης υπολογίζεται η ομοιότητα συνημιτόνου μεταξύ των δύο ενσωματώσεων πρότασης u, v και χρησιμοποιείται το μέσο τετραγωνικό σφάλμα για την αντικειμενική συνάρτηση. Η αρχιτεκτονική αυτή φαίνεται στο Σχήμα 3.7.



Σχήμα 3.7: Αρχιτεκτονική μοντέλου SBERT με χρήση συνάρτηση παλινδρόμησης (Πηγή: [Reim19])

Στην περίπτωση της αντικειμενικής συνάρτησης τριπλέτας, έχοντας μία αρχική πρόταση a, μία θετική πρόταση p και μία αρνητική πρόταση n, η συνάρτηση απώλειας προσαρμόζει το δίκτυο έτσι ώστε η διαφορά μεταξύ της a και της p είναι μικρότερη από τη διαφορά μεταξύ της πρότασης a και n (Εξίσωση 3.5)

$$\max(||sa - sp|| - ||sa - sn|| + \cdot, 0) \tag{3.5}$$

όπου sx η ενσωμάτωση πρότασης για τα a, n, p, $||\cdot||$ η μετρική για την απόσταση και περιθώριο $\cdot$. Το περιθώριο εξασφαλίζει ότι ότι το sp είναι το λιγότερο $\cdot$ πιο κοντά στο sa από το sn. Σαν μετρική χρησιμοποιείται η ευκλείδεια απόσταση και για περιθώριο το 1.

3.2 Συσταδοποίηση κειμένου

Η *συσταδοποίηση κειμένου* (text clustering) είναι μια υποπεριοχή της μη-επιβλεπόμενης μάθησης όπου το ζητούμενο είναι η τοποθέτηση μιας συλλογής κειμένων σε *συστάδες* (clusters), έτσι ώστε

τα κείμενα που ανήκουν στην ίδια συστάδα να έχουν κοινή θεματολογία. Στο περιβάλλον των συστημάτων συστάσεων ειδησεογραφικών άρθρων, η συσταδοποίηση των κειμένων μπορεί να επιφέρει σημαντικές πληροφορίες για το περιεχόμενο των κειμένων που αρέσουν στους αναγνώστες, καθώς και να βελτιώσει την ποιότητα των προτεινόμενων άρθρων για ανάγνωση.

Υπάρχουν τρία επίπεδα συσταδοποίησης κειμένων [Sudh20]:

- **Συσταδοποίηση κειμένου**: Αποτελεί τη διαδικασία ομαδοποίησης κειμένων με βάση το περιεχόμενο τους.

- **Συσταδοποίηση πρότασεων**: Αποτελεί τη διαδικασία ομαδοποίησης προτάσεων που προέρχονται από κείμενα. Ένα παράδειγμα είναι η ανάλυση των tweets.

- **Συσταδοποίηση λέξεων**: Οι συστάδες λέξεων είναι ομάδες λέξεων που αναφέρονται στο ίδιο θέμα. Ο πιο εύκολος τρόπος να δημιουργηθεί μία συστάδα λέξεων είναι η εύρεση συνώνυμων για μία συγκεκριμένη λέξη. Για παράδειγμα, το WordNet [Word10] είναι μια λεξική βάση για την αγγλική γλώσσα, η οποία ομαδοποιεί λέξεις σε σύνολα συνώνυμων λέξεων, τα οποία ονομάζονται synsets.

3.2.1 Αλγόριθμοι συσταδοποίησης κειμένων

Οι κύριοι στόχοι της συσταδοποίησης κειμένου είναι να οργανωθούν τα δεδομένα μέσω των συστάδων καθώς και να βρεθεί το ποσοστό ομοιότητας μεταξύ των κειμένων της συστάδας. Στη βιβλιογραφία, έχουν προταθεί περισσότεροι από εκατό αλγόριθμοι συσταδοποίησης και έχουν γίνει πολλές έρευνες. Οι αλγόριθμοι συσταδοποίησης μπορούν να χωριστούν στις παρακάτω κατηγορίες:

- **Διαμεριστικοί** : Οι αλγόριθμοι αυτοί ελαχιστοποιούν το κριτήριο που έχει τεθεί για την συσταδοποίηση, μετακινώντας τα στοιχεία ενός συνόλου δεδομένων μεταξύ των συστάδων, έως ότου υπάρξει μία (τοπική) βέλτιστη λύση. Σχετικό παράδειγμα αποτελεί ο αλγόριθμος των k μέσων

- **Ιεραρχικοί**: Οι αλγόριθμοι αυτοί στοχεύουν στη δημιουργία μιας ιεραρχίας από συστάδες. Χωρίζονται σε δύο κατηγορίες:

    1. **Αθροιστικοί**: Αυτή είναι μια προσέγγιση από κάτω προς τα πάνω. Αρχικά, κάθε στοιχείο ενός συνόλου δεδομένων αποτελεί μία ξεχωριστή συστάδα. Στη συνέχεια, οι συστάδες αρχίζουν και συγχωνεύονται όσο αυξάνεται η ιεραρχία.
    2. **Διαχωριστικοί**: Αυτή είναι μια προσέγγιση από πάνω προς τα κάτω. Αρχικά, τα στοιχεία ενός συνόλου δεδομένων αποτελούν μία συστάδα. Στη συνέχεια, όσο μειώνεται η ιεραρχία η συστάδα χωρίζεται δημιουργώντας νέες.

- **Γραφοθεωρητικοί**: Οι αλγόριθμοι αυτοί είναι βασισμένοι στη θεωρία των γράφων. Οι κόμβοι ενός γράφου αποτελούν τα στοιχεία ενός συνόλου δεδομένων, εδώ τα κείμενα, ενώ οι ακμές του γράφου είναι η απόσταση μεταξύ των κόμβων, δηλαδή η διαφορά μεταξύ των κειμένων.

- **Υπο-χώρων**: Οι αλγόριθμοι αυτοί βρίσκουν συστάδες σε διαφορετικούς υποχώρους (μίας ή περισσότερων διαστάσεων). Είναι μία προέκταση της κλασικής συσταδοποίησης N διαστάσεων που επιτρέπει την ταυτόχρονη ομαδοποίηση χαρακτηριστικών και παρατηρήσεων, δημιουργώντας συστάδες γραμμών και στηλών .

- **Βασισμένοι στην πυκνότητα**: Αυτοί οι αλγόριθμοι βασίζονται στην ιδέα ότι μία συστάδα στο χώρο των δεδομένων είναι μια περιοχή υψηλής πυκνότητας σημείων, η οποία διαχωρίζεται από τις άλλες συστάδες με γειτονικές περιοχές χαμηλής πυκνότητας σημείων. Τα δείγματα δεδομένων στις διαχωριστικές περιοχές χαμηλής πυκνότητας σημείων θεωρούνται συνήθως θόρυβος ή ακραίες τιμές.

- **Βασισμένοι στους περιορισμούς**: Αυτοί οι αλγόριθμοι αποτελούν μια ημι-επιβλεπόμενη προσέγγιση η οποία χρησιμοποιεί βασική γνώση με τη μορφή περιορισμών. Οι περιορισμοί εκφράζονται συνήθως σαν δυαδικές δηλώσεις και δείχνουν ότι δύο στοιχεία του συνόλου δεδομένων μπορούν ή δεν μπορούν να είναι μαζί στην ίδια συστάδα. Οι περιορισμοί μπορούν να χρησιμοποιηθούν όλοι για την δημιουργία των συστάδων, ή να χρησιμοποιηθούν περισσότερο σαν καθοδήγηση και όχι σαν απαραίτητη προϋπόθεση.

- **Νευρωνικών δικτύων**: Αυτοί οι αλγόριθμοι χρησιμοποιούν νευρωνικά δίκτυα για να ομαδοποιήσουν τα στοιχεία ενός συνόλου δεδομένων.

- **Ασαφούς συσταδοποίησης**: Με τη χρήση αυτών των αλγορίθμων τα στοιχεία ενός συνόλου δεδομένων μπορεί να ανήκουν σε περισσότερες από μία συστάδες. Οι αλγόριθμοι αυτοί χρησιμοποιούν μεθόδους ελαχίστων τετραγώνων για να βρουν την βέλτιστη θέση για κάθε στοιχείο. Αυτή η θέση μπορεί να είναι σε κάποιο χώρο πιθανότητας μεταξύ δύο ή περισσότερων συστάδων.

- **Μοντέλων κατανομής**: Αυτοί οι αλγόριθμοι σχετίζονται με την στατιστική και είναι βασισμένοι στα μοντέλα κατανομής. Οι συστάδες θα μπορούσαν να οριστούν σαν τα αντικείμενα τα οποία είναι πιο πιθανό να ανήκουν στην ίδια κατανομή.

- **Γενετικών αλγορίθμων**: Βασίζονται στους γενετικούς αλγορίθμους. Ο τρόπος λειτουργίας των γενετικών αλγορίθμων είναι εμπνευσμένος από την βιολογία. Χρησιμοποιεί την ιδέα της εξέλιξης μέσω γενετικής μετάλλαξης, φυσικής επιλογής και διασταύρωσης.

Στα πλαίσια της διπλωματικής εξετάστηκαν οι διαμεριστικοί αλγόριθμοι k μέσων και σφαιρικών k μέσων καθώς και αλγόριθμοι που βασίζονται σε νευρωνικά δίκτυα.

3.2.2 Αλγόριθμος k μέσων

Όπως προαναφέρθηκε, ο αλγόριθμος των k μέσων είναι ένας διαχωριστικός αλγόριθμος. Έστω ότι έχουμε ένα σύνολο παρατηρήσεων $(x_1, x_2, x_3, ..., x_n)$ όπου κάθε x_i είναι ένα d-διαστάσεων διάνυσμα πραγματικών αριθμών. Ο αλγόριθμος προσπαθεί να χωρίσει τις n παρατηρήσεις σε k σύνολο $S = S_1, S_2, ..., S_k$, ώστε να ελαχιστοποιείται το άθροισμα των τετραγώνων στο κάθε ένα από αυτά. Η συνάρτηση που πρέπει να ελαχιστοποιηθεί δίνεται από την Εξίσωση 3.6:

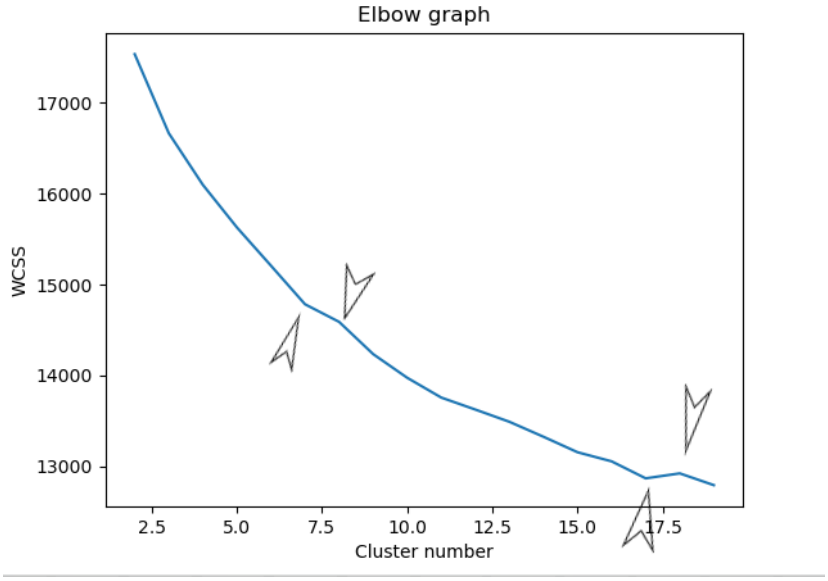
$$argmin \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 \tag{3.6}$$

όπου μ_i είναι ο μέσος όρος των στοιχείων στο S_i.

Για το σκοπό της συσταδοποίησης των άρθρων, οι παρατηρήσεις $(x_1, x_2, x_3, ..., x_n)$ τέθηκαν ίσες με τον μέσο όρο του αθροίσματος των προεκπαιδευμένων BERT διανυσμάτων των λέξεων του κάθε άρθρου. Κάθε παρατήρηση είναι ένα διάνυσμα 768 διαστάσεων, όσες και οι διαστάσεις των pre-trained embeddings των λέξεων.

Για την εύρεση του βέλτιστου αριθμού συστάδων χρησιμοποιήθηκε ο *κανόνας του αγκώνα*. Πρόκειται για μια ευριστική μέθοδο, η οποία χρησιμοποιείται για τον καθορισμό του αριθμού των συστάδων σε ένα σύνολο δεδομένων. Η μέθοδος περιλαμβάνει τη δημιουργία μίας γραφικής παράστασης *επεξηγούμενης διακύμανσης*, η οποία είναι συνήθως το κλάσμα μεταξύ της διασποράς εντός της συστάδας ως προς τη συνολική διασπορά. Μετά τη δημιουργία της γραφικής παράστασης, διαλέγονται οι γωνίες («αγκώνες») που κάνει η καμπύλη. Εκεί που υπάρχουν οι γωνίες βρίσκεται και ο πιθανός βέλτιστος αριθμός των συστάδων που μπορούν να χωριστούν τα στοιχεία του συνόλου δεδομένων.

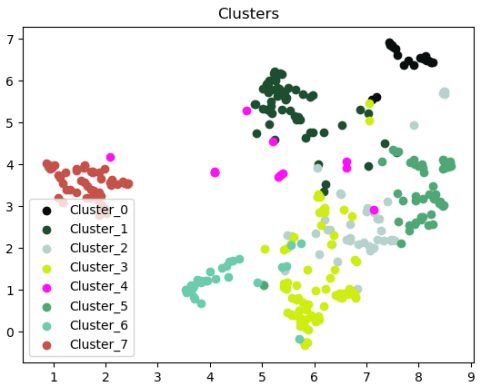
Στο Σχήμα 3.8 φαίνεται η γραφική παράσταση επεξηγούμενης διακύμανσης για μια αρχική συλλογή ειδησεογραφικών νέων που συλλέχθηκαν στο πλαίσιο της παρούσας διπλωματικής εργασίας. Τα βέλη δείχνουν τα πιθανά πλήθη σύστάδων που μπορούμε να χωρίσουμε τη συλλογή δεδομένων, έτσδι ώστε να έχουμε το βέλτιστο αποτέλεσμα.



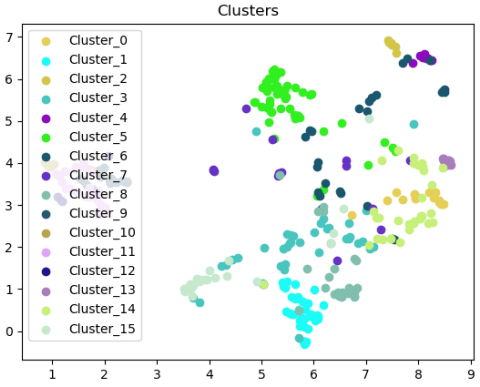
Σχήμα 3.8: Γραφική παράσταση επεξηγούμενης διακύμανσης για αρχική συλλογή ειδησεογραφικών νέων

Για πλήθος ίσο με 8 και πλήθος ίσο με 16 οι συστάδες που δημιουργούνται φαίνονται στα Σχήματα 3.9 και 3.10 αντίστοιχα. Τα διανύσματα λέξεων (διάστασης 768 στοιχείων το κάθε ένα) αναπαραστάθηκαν στο δισδιάστατο χώρο χρησιμοποιώντας την τεχνική UMAP (Uniform Manifold Approximation and Projection) για τη μείωση των διαστάσεων.

Παρατηρούμε ότι αν αυξηθεί το πλήθος των συστάδων, υπάρχουν συστάδες που μένουν ίδιες και συστάδες που διαχωρίζονται σε δύο ή περισσότερες άλλες. Και οι δύο διαχωρισμοί φαίνεται να χωρίζουν καλά τη συλλογή δεδομένων, αλλά δεν ξέρουμε ποιος είναι ο καλύτερος. Με τη χρήση του αλγόριθμου k μέσων είναι δύσκολο να προσδιοριστεί σε ποιο θέμα αναφέρεται κάθε συστάδα γιατί οι πιο συχνές λέξεις δεν είναι και αυτές που θα τις χαρακτήριζαν. Για αυτό το λόγο αναζητήθηκαν λύσεις που συνδυάζουν τη μοντελοποίηση θέματος μαζί με τις προ-εκπαιδευμένες ενσωματώσεις λέξεων.



Σχήμα 3.9: Γραφική αναπαράσταση της συλλογής δεδομένων χωρισμένης σε 8 συστάδες



Σχήμα 3.10: Γραφική αναπαράσταση της συλλογής δεδομένων χωρισμένης σε 16 συστάδες

3.2.3 Αλγόριθμος σφαιρικών k μέσων

Ο αλγόριθμος σφαιρικών k μέσων αποτελεί παραλλαγή του αλγορίθμου k μέσων που περιγράφτηκε προηγουμένως. Είναι κατάλληλος για δεδομένα κειμένων, σε αντίθεση με τον αλγόριθμο k μέσων που μπορεί να χρησιμοποιηθεί και σε διαφορετικού τύπου δεδομένα. Σκοπός του αλγορίθμου είναι να ελαχιστοποιήσει το άθροισμα της Εξίσωσης 3.7

$$\sum_{i}(1 - cos(x_i, p_{c(i)})) \tag{3.7}$$

για κάθε ανάθεση c του στοιχείου i στις συστάδες $c(i) \in 1, 2, ..., k$ και σε όλα τα πρωτότυπα $p_1, .., p_k$ στον ίδιο χώρο χαρακτηριστικών, όσο τα διανύσματα χαρακτηριστικών x_i αναπαριστούν τα στοιχεία του συνόλου δεδομένων [Horn12].

Η διαφοροποίηση του σε σύγκριση με τον αλγόριθμο k μέσων είναι το κριτήριο ελαχιστοποίησης. Στην πρώτη περίπτωση το κριτήριο είναι το άθροισμα των τετραγώνων στην κάθε συστάδα (Εξίσωση 3.6), ενώ σε αυτή την περίπτωση είναι η ομοιότητα συνημιτόνου (Εξίσωση 3.7). Με αυτή την αλλαγή, εξετάζονται και οι γωνίες μεταξύ των παρατηρήσεων και γίνεται εφικτή η κανονικοποίηση τους, για πιο δίκαιες συγκρίσεις μεταξύ τους. Συγκριτικά με τον αλγόριθμο k μέσων, ο συγκεκριμένος αλγόριθμος προκύπτει ως η προβολή των παρατηρήσεων σε μία μοναδιαία σφαίρα και χρήση της ευκλείδειας απόστασης για την ομαδοποίηση των παρατηρήσεων.

3.2.4 Αλγόριθμος σφαιρικών k μέσων και μοντελοποίηση θεμάτων

Τα *μοντέλα θέματος* (topic models) είναι ένα χρήσιμο εργαλείο ανάλυσης το οποίο μπορεί να εντοπίσει τα κρυμμένα θέματα σε ένα σύνολο άρθρων. Η συνήθης προσέγγιση είναι η χρήση πιθανοτικών μοντέλων θέματος, ωστόσο στην εργασία [Sia20] προτείνεται μία νέα μέθοδος για την εύρεση των θεμάτων. Η συγκεκριμένη τεχνική επιτρέπει τη συσταδοποίηση των άρθρων με τη χρήση προεκπαιδευμένων ενσωματώσεων λέξεων, συνδυάζοντας πληροφορίες για τα άρθρα (πχ βάρη για τις λέξεις) και αναδιάταξη των συχνά χρησιμοποιούμενων λέξεων. Η προσέγγιση αυτή έχει τα πλεονεκτήματα του μικρότερου χρόνου εκτέλεσης και της μικρότερης υπολογιστικής πολυπλοκότητας. Έτσι, στο πλαίσιο της παρούσας διπλωματικής εργασίας χρησιμοποιήθηκε η συγκεκριμένη προσέγγιση.

Τα βήματα του αλγορίθμου είναι τα ακόλουθα:

1. Προεπεξεργασία και εξαγωγή του λεξιλογίου από τα άρθρα της συλλογής δεδομένων

2. Αντιστοίχιση κάθε λέξης με το προεκπαιδευμένο διάνυσμα (ενσωμάτωσή) της

3. Συσταδοποίηση των άρθρων με τη χρήση του αλγόριθμου σφαιρικών k μέσων για τη δημιουργία k συστάδων με ή χωρίς βάρη στις λέξεις του λεξιλογίου. Η χρήση βαρών στις λέξεις είναι βασισμένη στο μοντέλο της *λανθάνουσας κατανομής Dirichlet* (latent Dirichlet allocation - LDA). Στη συγκεκριμένη περίπτωση χρησιμοποιούνται οι μετρικές TF, TF-DF και TF-IDF.

4. Εξαγωγή από κάθε συστάδα των πρώτων J λέξεων για την εκτίμηση της ποιότητας της συσταδοποίησης. Σε αλγορίθμους συσταδοποίησης που βασίζονται στο σχηματισμό κέντρων για κάθε συστάδα, οι πρώτες J λέξεις είναι είτε αυτές που είναι πιο κοντά στο κέντρο $c^{(i)}$ ή αυτές με τη μεγαλύτερη πιθανότητα με βάση τις παραμέτρους της συστάδας. Τυπικά αυτό σημαίνει ότι οι J λέξεις προκύπτουν από την Εξίσωση 3.8 παρακάτω:

$$\arg\max_{J} \sum_{j \in J} cos(c^{(i)}, x_j) \tag{3.8}$$

5. Αναδιάταξη των J λέξεων για την εκτίμηση της συστάδας. Πολλές φορές οι λέξεις που είναι πιο κοντά στο κέντρο δεν αντιπροσωπεύουν με τον καλύτερο τρόπο το νόημα της. Χωρίς την αναδιάταξη, το κριτήριο της *κανονικοποιημένης σημείου-προς-σημείου αμοιβαίας πληροφορίας* (normalized pointwise mutual information - NPMI) είναι πιο χαμηλό, παρότι οι θεματικές που προκύπτουν έχουν νόημα [Sia20].

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, στην εκτέλεση του αλγορίθμου χρησιμοποιήθηκαν προ-εκπαιδευμένες ενσωματώσεις λέξεων word2vec και fastText στην ελληνική γλώσσα, αναδιάταξη των κορυφαίων 10 λέξεων ανά συστάδα, διαφορετικός αριθμός θεμάτων καθώς και προεπεξεργασία των άρθρων και χρήση βαρών στις λέξεις του λεξιλογίου. Με βάση τις παραπάνω επιλογές, προέκυψαν τα παρακάτω συμπεράσματα:

- Όσο μεγαλύτερο ήταν το πλήθος των θεμάτων, τόσο καλύτερα ήταν τα αποτελέσματα (μικρότερο NPMI σκορ). Στον Πίνακα 3.1 με τα 7 θέματα βλέπουμε ότι το NPMI σκορ είναι μεγαλύτερο από ότι στον Πίνακα 3.2 με τα 8 θέματα.

- Οι λέξεις μετά την προεπεξεργασία μπορεί να μεταβάλλονταν, εξαιτίας των διαδικασιών της εξαγωγής του θέματος (της λέξης) καθώς και της λημματοποίησης. Για παράδειγμα, στον Πίνακα 3.2 η λέξη «κάθε» γίνεται «καθω»

- Η αναδιάταξη έδωσε λέξεις με περισσότερο νόημα για την κάθε συστάδα. Αν συγκρίνουμε τους Πίνακες 3.1 - 3.2 και 3.3 - 3.4, βλέπουμε ότι οι λέξεις των δεύτερων πινάκων έχουν περισσότερο νόημα για το κάθε θέμα από ότι οι λέξεις των πρώτων πινάκων.

- Τα λεξικά των προεκπαιδευμένων ενσωματώσεων λέξεων δεν είχαν πολλές κοινές λέξεις με τα κείμενα που αφορούσαν τον αθλητισμό, για αυτό και στους Πίνακες 3.1, 3.2, 3.3, 3.4 δεν φαίνεται ξεκάθαρα κάποιο θέμα που να αφορά τον αθλητισμό.

- Στα θέματα, ακόμα και κατόπιν της προεπεξεργασίας και της αναδιάταξης, παρέμεναν ρήματα τα οποία χρησιμοποιούνται συχνά στην ελληνική γλώσσα και δεν προσδίδουν στο νόημα της κάθε συστάδας.

- Οι ενσωματώσεις λέξεων fastText έδωσαν καλύτερα αποτελέσματα σε σύγκριση με τις αντίστοιχες του word2Vec για την ίδια συλλογή δεδομένων.

Α/Α	**Θέμα**	**NPMI σκορ**
1	πολλοί κάποιοι έκαναν έγιναν είχαν αυτοί όλοι οποίοι έχουν δώσουν	-0.92235
2	τόνισε αναφέρει πρόεδρος έδωσε αποκάλυψε έλαβε δεύτερη δήλωσε μίλησε κατέληξε	-0.97663
3	όμως προφανώς μάλλον βέβαια ταυτόχρονα αλλά πάντως επειδή αφού άλλωστε	-0.87449
4	οκτωβρίου δεκεμβρίου νοεμβρίου πέμπτη γιώργος μαρτίου γιάννης μεσημέρι τετάρτη αθήνας	-0.9705
5	ανάπτυξη πολιτικής απόφαση διαδικασίες υλοποίηση οικονομίας πολιτών σκοπό ενίσχυσης έρευνα	-1.0
6	άλλο γύρω πολλές κάποιο όπως όπου βάση αυτά εβδομάδα κέντρο	-0.9316
7	κορονοϊού κορονοϊό ασθενών κλινικές νοσοκομείο κρούσματα νοσηλεύονται εμβόλιο κλινική κρουσμάτων	-0.96755
Συνολικό NPMI σκορ		-0.94902

Πίνακας 3.1: Θέματα με τη χρήση ενσωματώσεων λέξεων word2vec χωρίς καμία επεξεργασία

Από τα παραπάνω αποτελέσματα συμπεραίνουμε ότι η χρήση της μοντελοποίησης θεμάτων, συμβάλει σημαντικά στην κατανόηση του νοήματος και του θέματος κάθε συστάδας. Η χρήση των προεκπαιδευμένων ενσωματώσεων λέξεων word2Vec και fastText δεν βελτιώνει σημαντικά τα αποτελέσματα της συσταδοποίησης, αφού οι λέξεις που παράγονται από τον αλγόριθμο για κάθε θέμα είναι κυρίως ρήματα και δεν έχουν κάποια νοηματική σχέση μεταξύ τους. Για αυτούς τους λόγους, αναζητήθηκαν λύσεις οι οποίες χρησιμοποιούν νευρωνικά μοντέλα καθώς και ενσωματώσεις λέξεων SBERT για να λαμβάνεται υπόψιν και το νόημα των λέξεων σε κάθε κείμενο.

Α/Α	Θέμα	NPMI σκορ
1	ειπε στιγμη πρωτη εκανε τελος καλο δοσει εδωσε εγινε μεγαλη	-0,51995
2	καθω τεστ πανδημια τελος δοσει ασθενει στοιχει εκατ σημειωσε pfizer	-0,84808
3	καθω χρονος συγκεκριμενος δηλωσε εκανε ανεφερε στοιχει πρωτος ανδρας προσωπικος	-0,59272
4	συμφωνα κατασταση μεσω μασκα ευρω στηριξη σχεση βαρος διαρκεια βαση	-0,8894
5	μπορω κανω γινω βρισκομαι δοσει θελω δωσω φαινομαι φτασω χρειαζομαι	-0,48728
6	καθω ομαδας υγειας χρονιας τελος θεσσαλονικης στοιχει υπουργος σημειωσε προεδρος	-0,43414
7	χωρα νοεμβριου κυβερνηση ανακοινωση αγωνα συνεντευξη πατριαρχη αποφαση γιωργο πρωην	-0,97632
8	υπαρχω υγειας ελλαδα νοεμβριου πληροφοριες εμβολιου εργασιας ελενη παιχνιδια θεμος	-0,80829
NPMI σκορ		-0.69452

Πίνακας 3.2: Θέματα με τη χρήση ενσωματώσεων λέξεων word2vec, προεπεξεργασίας των κειμένων και αναδιάταξης των τελικών λέξεων του κάθε θέματος

Α/Α	Θέμα	NPMI σκορ
1	βρει βρεθεί δώσει βγάλει πάρει καταφέρει ζητήσει θέλει φτάσει κάνει	-1,0
2	μπορούσαν ξεκινήσει αναφέρει μπορούσε αναφέρθηκε αναφέρεται ανέφερε υπάρχει κάνουν ξεκίνησε	-0,92829
3	βέβαια ακόμη ακόμα επειδή σίγουρα ωστόσο μάλλον κάποιο αρκετά πάντως	-0,8247
4	δεύτερη σημαντική τρίτη πολιτική συνολική εκείνη σημερινή μεγάλη περίοδο σχετική	-0,92317
5	συγκεκριμένο συγκεκριμένη συγκεκριμένα πραγματοποιήθηκε ταυτόχρονα χαρακτηριστικά αποτελεσματικότητα κυβέρνηση πρωθυπουργός τελευταίο	-1,0
6	είχαν έχουν είχε έχει είχαμε είχα έχουμε γνωρίζουμε κάνουμε βλέπουμε	-0,66138
7	αρκετές πολλές ανθρώπους πέντε διάστημα περίπου κάποιες άνθρωποι ημέρες αυτούς	-1,0
NPMI σκορ		-0,90536

Πίνακας 3.3: Θέματα με τη χρήση ενσωματώσεων λέξεων fastText, χωρίς καμία επεξεργασία

3.2.5 Αλγόριθμοι Νευρωνικών Δικτύων

Σε αυτή την κατηγορία εξετάστηκαν δύο διαφορετικές μεθοδολογίες και πιο συγκεκριμένα το *embedded topic model* (ETM) [Dien19] και το CombinedTM [Bian21], ένα σύστημα μοντελοποίησης θεμάτων που βασίζεται σε ενσωματώσεις λέξεων σχετικές με τα συμφραζόμενα. Το ETM είναι ένα *μοντέλο θέματος* (topic model - TM) το οποίο χρησιμοποιεί αναπαραστάσεις υπό τη μορφή ενσωματώσεων και για τις λέξεις αλλά και για τα θέματα. Περιέχει δύο έννοιες λανθάνουσας διάστασης. Πρώτα, αναπαριστά σε ενσωματώσεις το λεξιλόγιο ενός συνόλου δεδομένων σε έναν L-διαστάσεων χώρο. Στη συνέχεια, αναπαριστά κάθε κείμενο του συνόλου δεδομένων σε όρους από k θέματα.

Στα παραδοσιακά TM, κάθε θέμα είναι μία πλήρης κατανομή πάνω στο λεξιλόγιο του συνόλου δεδομένων. Στο ETM ωστόσο, το k-οστό θέμα είναι ένα διάνυσμα $a_k \in R^L$ στο χώρο των ενσωματώσεων. Το a_k ονομάζεται *ενσωμάτωση θέματος* (topic embedding) και είναι μια κατανεμημένη αναπαράσταση του k-οστού θέματος σε ένα σημασιολογικό χώρο των λέξεων. Σε αυτή τη διαδικασία, το ETM χρησιμοποιεί την ενσωμάτωση θέματος για να δημιουργήσει μία κατανομή για κάθε θέμα πάνω στο λεξιλόγιο. Συγκεκριμένα, το ETM χρησιμοποιεί ένα λογαριθμικό γραμμικό μοντέλο το οποίο παίρνει το εσωτερικό γινόμενο μεταξύ του πίνακα των ενσωματώσεων των λέξεων και των

Α/Α	Θέμα	NPMI σκορ
1	χρονος συγκεκριμενος δηλωσε κορονοιος πρωτος εγινε παρος προσωπικος περισσοτερος μιλησε	-0,59533
2	χωρα κατασταση στιγμη πρωτη ανακοινωση θεση σχεση συνεντευξη διαδικασια μεγαλη	-0,75451
3	μπορω κανω υπαρχω γινω βρισκομαι αναφερω νοσηλευομαι θελω δωσω φαινομαι	-0,34178
4	ομαδας υγειας μεσω εμβολιου εργασιας εταιρειας ασθενει στηριξη υπουργος βαρος	-0,82795
5	μετρο ελενη θεμος θεσσαλονικη θεσσαλονικης σπιτι ανδρας νοσοκομεια μπουρλας προεδρος	-0,76271
6	συμφωνα ειπε ανεφερε σημαντικο σημειωσε συνολικα μεγαλο γεγονος επιπλεον προσθεσε	-0,54062
7	ελλαδα πανδημια κυβερνηση ερευνα ευρωπη μοναδα υγεια πολιτικη isis αναπτυξη	-0,97469
8	νοεμβριου αρχες ευρω ημερες εκατ black πεμπτη εβδομαδα news πρωι	-0,91059
NPMI σκορ		-0.71352

Πίνακας 3.4: Θέματα με τη χρήση ενσωματώσεων λέξεων fastText, προεπεξεργασίας των κειμένων και αναδιάταξης των τελικών λέξεων του κάθε θέματος

θεμάτων. Έτσι, δίνει μεγάλη πιθανότητα σε μία λέξη v στο θέμα k, μετρώντας την ομοιότητα μεταξύ τους.

Στο ETM μπορούν να χρησιμοποιηθούν προεκπαιδευμένες ενσωματώσεις ή να τα μάθει το μοντέλο κατά τη διάρκεια της εκπαίδευσης. Όταν το ETM μαθαίνει τις ενσωματώσεις κατά την εκπαίδευση, ταυτόχρονα βρίσκει τα θέματα καθώς και τον *χώρο των ενσωματώσεων* (embedding space). Αντίθετα όταν το ETM χρησιμοποιεί προεκπαιδευμένες ενσωματώσεις, μαθαίνει τα θέματα σε ένα συγκεκριμένο χώρο ενσωματώσεων. Αυτή η στρατηγική είναι χρήσιμη όταν υπάρχουν λέξεις που δεν χρησιμοποιούνται στο λεξιλόγιο. Το ETM μπορεί να υποθέσει πως οι λέξεις αυτές ταιριάζουν με τα θέματα επειδή μπορεί να υπολογίσει τις ενσωματώσεις ακόμα και λέξεων που δεν έχει «δει» [Dien19].

Το ETM χρησιμοποιήθηκε και με προεκπαιδευμένες ενσωματώσεις από το Greek BERT [Kout20], καθώς και με τις ενσωματώσεις που δημιουργούνται κατά την εκπαίδευση του μοντέλου. Τα θέματα που δημιουργήθηκαν για κάποιο αρχικό σύνολο δεδομένων ειδησεογραφικών άρθρων φαίνονται στους Πίνακες 3.5 και 3.6 για τις ενσωματώσεις του BERT και του ETM, αντίστοιχα. Το μοντέλο εκπαιδεύτηκε για 1000 εποχές και για 12 θέματα. Επίσης, δεν χρησιμοποιήθηκε κάποια προεπεξεργασία στα κείμενα. Από την εκπαίδευση του μοντέλου προέκυψαν τα παρακάτω συμπεράσματα:

- Όσο αυξάνεται το πλήθος των θεμάτων μειώνεται η ποικιλομορφία τους. Αυτό συμβαίνει, πιθανώς, γιατί στη συγκεκριμένη συλλογή δεδομένων υπήρχε καθορισμένος αριθμός θεμάτων. Έτσι όταν ξεπεράστηκε, τα θέματα άρχισαν να εκφράζουν παρεμφερείς έννοιες.

- Η αύξηση του πλήθους των θεμάτων απαιτεί και την αύξηση των εποχών εκπαίδευσης του μοντέλου. Αν δεν αυξηθούν οι εποχές, τα θέματα θα περιέχουν λέξεις γενικές χωρίς να προσδίδουν κάποιο ιδιαίτερο νόημα. Επίσης, θα μειωθεί και το ποικιλομορφία, αφού πολλά θέματα θα μοιάζουν μεταξύ τους.

- Η αύξηση του πλήθους των θεμάτων οδηγεί στη δημιουργία θεμάτων με περισσότερο νόημα και σε έναν καλύτερα κατανοητό διαχωρισμό των κειμένων.

- Όταν το μοντέλο εκπαιδεύτηκε με τις προεκπαιδευμένες ενσωματώσεις του BERT, δεν έδωσε καλά αποτελέσματα, ούτε στις λέξεις των θεμάτων ούτε στις μετρικές 3.5. Αυτό συνέβη γιατί το Greek BERT έχει εκπαιδευτεί πάνω σε δεδομένα που δεν συμπίπτουν με ειδησεογραφικά

νέα. Έτσι οι ενσωματώσεις του δεν μπορούν να αποδώσουν το νόημα των θεμάτων που έχει παράγει το μοντέλο ETM.

- Αντίθετα, όταν το μοντέλο εκπαιδεύτηκε και παρήγαγε τις δικές του ενσωματώσεις, έδωσε πολύ καλά αποτελέσματα και στις μετρικές των θεμάτων αλλά και στις λέξεις (Πίνακας 3.6).

Α/Α	**Θέμα**
1	'κατασταση', 'υγειας', 'νοσοκομειο', 'τεστ', 'κορωνοιο', 'θετικος', 'ημερες', 'νοεμβριου', 'κορονοιο'
2	'κανει', 'ελλαδα', 'νοεμβριου', 'μελισσες', 'ελενη', 'αθηνα', 'series', 'χρονια', 'αγριες'
3	'league', 'πρωην', 'εκατ', 'παιχνιδια', 'αεκ', 'παοκ', 'πρωταθλημα', 'basket', 'παικτες'
4	'κυβερνηση', 'υγειας', 'μετρα', 'εβδομαδα', 'covid', 'χωρα', 'μερα', 'αποφαση', 'συμφωνα'
5	'παοκ', 'ομαδα', 'ειπε', 'πρωτη', 'ομαδας', 'συμφωνα', 'πολλα', 'χρονο', 'αποτελεσμα'
6	'νοσοκομειο', 'ειπε', 'εβδομαδες', 'νοεμβριου', 'covid', 'χρονος', 'υγειας', 'βαλμπουενα', 'συμφωνα'
7	'συμφωνα', 'κοινωνικης', 'κορονοιου', 'ειπε', 'πανδημιας', 'ομαδα', 'δημου', 'πρωτη', 'κυκλοφορια'
8	'νοσοκομεια', 'ελλαδα', 'θεσσαλονικη', 'ειπε', 'εοδυ', 'ασθενων', 'νοσοκομειο', 'θεσσαλονικης', 'συμφωνα'
9	'εμβολιου', 'εε', 'δοσεις', 'δηλωσε', 'pfizer', 'μπουρλα', 'αρχες', 'ειπε', 'εμβολιο'
10	'of', 'χρονια', 'πολιτικη', 'κκε', 'the', 'ζωη', 'ιδρυμα', 'βημα', 'νοεμβριου'
11	'ελλαδα', 'υπουργος', 'εμβολια', 'παναγιωτοπουλος', 'δηλωσε', 'ειπε', 'τρισ', 'χωρας', 'εργασιας'
12	'παιχνιδια', 'παιδια', 'χωρα', 'βαρος', 'περιπτωση', 'isis', 'πληροφοριες', 'τρομοκρατικη', 'σχολειο'
Συνάφεια	-0.016329138304513192
Ποικιλομορφία	0.7466666666666667

Πίνακας 3.5: Θέματα με χρήση ενσωματώσεων ETM χωρίς καμία επεξεργασία

Στην περίπτωση του CombinedTM χρησιμοποιούνται προεκπαιδευμένες ενσωματώσεις SBERT για την συσταδοποίηση των άρθρων, σε συνδυασμό με τη χρήση του μοντέλου BoW για την υλοποίηση του TM. Πιο συγκεκριμένα, συνδυάζονται οι προκεπαιδευμένες αναπαραστάσεις προτάσεων με τη χρήση του SBERT και τα νευρωνικά μοντέλα θέματος. Οι προεκπαιδευμένες ενσωματώσεις BERT (Ενότητα 3.1.6) των προτάσεων καταφέρνουν να παράγουν πιο ουσιαστικά και συνεκτικά θέματα σε σύγκριση με το μοντέλο LDA ή τα ήδη υπάρχοντα νευρωνικά μοντέλα για το TM. Επίσης, το συγκεκριμένο μοντέλο αυξάνει και τη συνάφεια των θεμάτων. Η αρχιτεκτονική του απεικονίζεται στο Σχήμα 3.11.

Το μοντέλο CombinedTM βασίζεται σε δύο κύρια στοιχεία:

1. Στο νευρωνικό μοντέλο ProdLDA [Sriv17]. Το ProdLDA είναι ένα νευρωνικό TM το οποίο βασίζεται σε *διαφοροποιημένους αυτοκωδικοποιητές* (variational autoencoders - VAE). Ο κωδικοποιητής εκπαιδεύει ένα νευρωνικό δίκτυο, έτσι ώστε να αντιστοιχίζει την BoW αναπαράσταση ενός κειμένου, σε μία συνεχή λανθάνουσα αναπαράσταση. Στη συνέχεια, ο αποκωδικοποιητής ανακατασκευάζει την BoW αναπαράσταση, παράγοντας τις λέξεις από τη λανθάνουσα αναπαράσταση.

2. Στις ενσωματώσεις SBERT (Ενότητα 3.1.7)

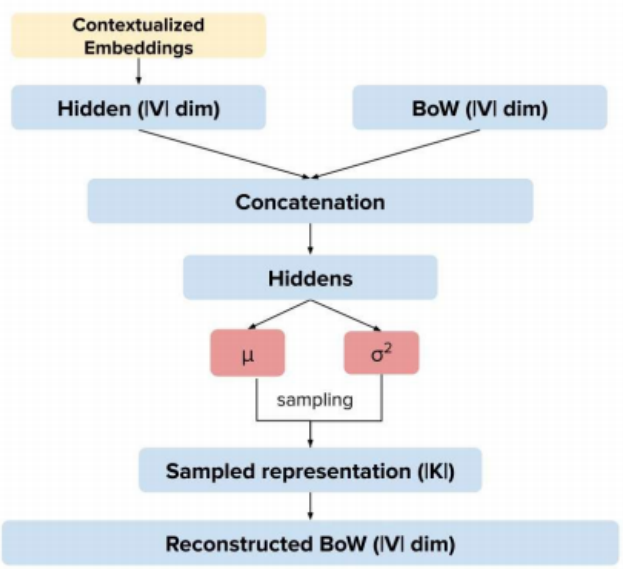
Α/Α	Θέμα
1	’επιχειρουμε’, ’λιθιασης’, ’κυμης’, ’υφων’, ’ιχνηλατηση’, ’κατασταλτικων’, ’covid’, ’αντιστρεψει’, ’αγκαλιασει’
2	’λιθιασης’, ’επιχειρουμε’, ’κυμVAEης’, ’υφων’, ’ιχνηλατηση’, ’κατασταλτικων’, ’covid’, ’αντιστρεψει’, ’διακομισθηκε’
3	’επιχειρουμε’, ’λιθιασης’, ’κυμης’, ’υφων’, ’ιχνηλατηση’, ’κατασταλτικων’, ’διακομισθηκε’, ’αντιστρεψει’, ’covid’
4	’καραντινα’, ’θετικοι’, ’μαντλιν’, ’ηλικιωμενοι’, ’αντιτρομοκρατικη’, ’evil’, ’lite’, ’επιχειρουμε’, ’mauro’
5	’λιθιασης’, ’επιχειρουμε’, ’υφων’, ’κυμης’, ’vox’, ’computerspiritfarer’, ’αγκαλιασει’, ’ιχνηλατηση’, ’συρος’
6	’λιθιασης’, ’επιχειρουμε’, ’κυμης’, ’υφων’, ’ιχνηλατηση’, ’κατασταλτικων’, ’αντιστρεψει’, ’αντιλαμβανομαστε’, ’διακομισθηκε’
7	’εμβολιου’, ’covid’, ’pfizer’, ’μπουρλα’, ’λιθιασης’, ’αγελης’, ’μασκες’, ’ιερουσαλη’, ’υφων’
8	’μεθ’, ’πανδημιας’, ’κορωνοιο’, ’κρουσματων’, ’covid’, ’κορονοιου’, ’κλινες’, ’πανδημια’, ’κορωνοιου’
9	’λιθιασης’, ’επιχειρουμε’, ’κυμης’, ’υφων’, ’ιχνηλατηση’, ’κατασταλτικων’, ’ιερουσαλη’, ’αντιστρεψει’, ’ζητουμενα’
10	’κυμης’, ’λιθιασης’, ’ιχνηλατηση’, ’αγκαλιασει’, ’επιχειρουμε’, ’αντιστρεψει’, ’βαλμπουενα’, ’συρος’, ’υφων’
11	’λιθιασης’, ’επιχειρουμε’, ’κυμης’, ’υφων’, ’ιχνηλατηση’, ’κατασταλτικων’, ’αντιστρεψει’, ’μορειρα’, ’covid’
12	’επιχειρουμε’, ’λιθιασης’, ’κυμης’, ’υφων’, ’ιχνηλατηση’, ’κατασταλτικων’, ’διακομισθηκε’, ’αντιστρεψει’, ’covid’
Συνάφεια	-0,8224182921021321
Ποικιλομορφία	0,30666666666666664

Πίνακας 3.6: Θέματα με τη χρήση των ενσωματώσεων του Greek-BERT χωρίς καμία επεξεργασία

Η μέθοδος που προτείνεται είναι ανεξάρτητη από την επιλογή του TM και των προ-εκπαιδευμένων αναπαραστάσεων, αφού το TM επεκτείνει έναν αυτοκωδικοποιητή και οι αναπαραστάσεις βασίζονται στα κείμενα του συνόλου δεδομένων. Το μοντέλο ProdLDA λοιπόν, επεκτείνεται με τις *ενσωματώσεις συμφραζομένων* (contextualized embeddings) των κειμένων από το SBERT. Οι αναπαραστάσεις των κειμένων προβάλλονται μέσω ενός κρυφού επιπέδου με διάσταση ίση με τη διάσταση του λεξιλογίου, συνδεόμενη σειριακά με τις BoW αναπαραστάσεις των κειμένων (Σχήμα 3.11).

Αυτό το μοντέλο φαίνεται να δίνει καλύτερα αποτελέσματα στις μετρικές NPMI (τ), *συνάφειας εξωτερικών ενσωματώσεων λέξεων* (α) και *ανεστραμμένης επικάλυψης εξαρτώμενης από τη θέση* (ρ), σε αρκετές συλλογές δεδομένων σε σύγκριση με άλλα μοντέλα, όπως λ.χ. το ETM. Τα αποτελέσματα φαίνονται στο Σχήμα 3.12. Για όλους τους προαναφερόμενους λόγους, συμπεριλήφθηκε στο πλαίσιο της παρούσας διπλωματικής εργασίας. Η χρήση του έδωσε για την αρχική συλλογής δεδομένων που χρησιμοποιήθηκε και στο ETM, καλύτερα αποτελέσματα στο NPMI σκορ, την συνάφεια των θεμάτων καθώς και στην ποικιλομορφία των θεμάτων. Τα αποτελέσματα φαίνονται στον Πίνακα 3.7.

Παρατηρούμε ότι εκτός από τις καλύτερες μετρικές, οι λέξεις δίνουν περισσότερο νόημα σε κάθε θέμα, από ότι οι λέξεις του ETM (Πίνακας 3.5). Για παράδειγμα, αν πάρουμε το 12^{ο} θέμα του ETM, θα δούμε ότι αναφέρεται σε άρθρα για την τρομοκρατική οργάνωση ISIS με τις λέξεις *’παιχνιδια’, ’παιδια’, ’χωρα’, ’βαρος’, ’περιπτωση’, ’isis’, ’πληροφοριες’,’τρομοκρατικη’, ’σχολειο’*. Στο ContextualTM βλέπουμε ότι το αντίστοιχο θέμα είναι το 2^{ο} με τις λέξεις *’isis’, ’βαση’, ’προστασιας’, ’πολιτη’, ’εργου’, ’κατηγοριες’,’τρομοκρατικη’, ’νεου’, ’βαρος’, ’κορυδαλλου’*. Οι μη κοινές λέξεις *’κορυδαλλου’, ’βαση’, ’προστασιας’, ’πολιτη’, ’κατηγοριες’* του CTM φαίνεται να ταιριάζουν καλύτερα στο θέμα από ότι οι λέξεις *’παιχνιδια’, ’παιδια’, ’περιπτωση’, ’πληροφοριες’, ’σχολειο’* του



Σχήμα 3.11: Διάγραμμα της αρχιτεκτονικής του μοντέλου CombinedTM (Πηγή: [Bian21])

Model	Avg τ	Avg α	Avg ρ
Results for the Wiki20K Dataset:			
Ours	**0.1823**	0.1980	**0.9950**
PLDA	0.1397	0.1799	0.9901
MLDA	0.1443	**0.2110**	0.9843
NVDM	-0.2938	0.0797	0.9604
ETM	0.0740	0.1948	0.8632
LDA	-0.0481	0.1333	0.9931
Results for the StackOverflow Dataset:			
Ours	**0.0280**	0.1563	0.9805
PLDA	-0.0394	0.1370	**0.9914**
MLDA	0.0136	0.1450	0.9822
NVDM	-0.4836	0.0985	0.8903
ETM	-0.4132	**0.1598**	0.4788
LDA	-0.3207	0.1063	0.8947

Results for the GoogleNews Dataset:			
Ours	**0.1207**	**0.1325**	**0.9965**
PLDA	0.0110	0.1218	0.9902
MLDA	0.0849	0.1219	0.9959
NVDM	-0.3767	0.1067	0.9648
ETM	-0.2770	0.1175	0.4700
LDA	-0.3250	0.0969	0.9774
Results for the Tweets2011 Dataset:			
Ours	**0.1008**	**0.1493**	0.9901
PLDA	0.0612	0.1327	0.9847
MLDA	0.0122	0.1272	**0.9956**
NVDM	-0.5105	0.0797	0.9751
ETM	-0.3613	0.1166	0.4335
LDA	-0.3227	0.1025	0.8169

Results for the 20NewsGroups Dataset:			
Ours	0.1025	0.1715	0.9917
PLDA	0.0632	0.1554	**0.9931**
MLDA	**0.1300**	0.2210	0.9808
NVDM	-0.1720	0.0839	0.9805
ETM	0.0766	**0.2539**	0.8642
LDA	0.0173	0.1627	0.9897

Σχήμα 3.12: Σύγκριση του ContextualTM με άλλα παρόμοια μοντέλα σε διαφορετικές συλλογές δεδομένων (Πηγή: [Bian21])

ETM. Για αυτούς τους λόγους, επιλέχθηκε το μοντέλο CTM για την υλοποίηση της συσταδοποίησης των άρθρων.

3.3 Αναγνώριση επώνυμων οντοτήτων

Ο όρος *επώνυμη οντότητα* (named entity) αναφέρεται σε μία λέξη ή φράση η οποία μπορεί να ξεχωρίσει ένα αντικείμενο από ένα σύνολο δεδομένων με παρόμοια χαρακτηριστικά. Παραδείγματα επώνυμων οντοτήτων είναι ο *οργανισμός* (organization), το *άτομο* (person), η *τοποθεσία* (location) και άλλα ανάλογα με τον τομέα στον οποίο αναφέρεται το σύνολο των δεδομένων. Η *αναγνώριση επώνυμων οντοτήτων* (named entity recognition - NER) είναι η διαδικασία με την οποία εντοπίζονται και κατηγοριοποιούνται οι επώνυμες οντότητες των κειμένων σε προκαθορισμένες κατηγορίες. Στο Σχήμα 3.13 φαίνεται ένα παράδειγμα όπου το NER σύστημα εντοπίζει τρεις επώνυμες οντότητες στη δοθείσα πρόταση. [Li20]

Το NER αποτελεί ένα σημαντικό βήμα προεπεξεργασίας για διάφορες λειτουργίες όπως η απάντηση ερωτήσεων και η μηχανική μετάφραση. Στα πλαίσια της παρούσας διπλωματικής εργασίας, το NER χρησιμοποιήθηκε για τον εντοπισμό λέξεων κλειδιών στα κείμενα, έτσι ώστε να μπορεί να γίνει

A/A	Θέμα
1	’ιδρυμα’, ’aux’, ’εργασιας’, ’βρουτσης’, ’ευρωζωνης’, ’επιχειρησεων’, ’κλιματισμου’, ’στηριξη’, ’υποβολης’, ’προιοντα’
2	’isis’, ’βαση’, ’προστασιας’, ’πολιτη’, ’εργου’, ’κατηγοριες’, ’τρομοκρατικη’, ’νεου’, ’βαρος’, ’κορυδαλλου’
3	’εε’, ’τουρκια’, ’πλευρα’, ’μπορελ’, ’κυρωσεων’, ’αφορα’, ’ελλαδα’, ’τουρκιας’, ’ελλαδας’, ’δεκεμβριου’
4	’audi’, ’smartphone’, ’δροσω’, ’οθονη’, ’μοντελο’, ’στοχο’, ’συγχυση’, ’benz’, ’play’, ’maybach’
5	’πανδημιας’, ’μαθητες’, ’κοινωνικης’, ’κοινωνια’, ’εκπαιδευση’, ’ρεπορταζ’, ’εκδηλωση’, ’εφημεριδα’, ’μορφη’, ’συνθηκες’
6	’ασθενων’, ’θεσσαλονικης’, ’νοσοκομεια’, ’νοσηλευονται’, ’βορεια’, ’lockdown’, ’κλινες’, ’ασθενεις’, ’θεσσαλονικη’, ’συμφωνα’
7	’διαγωνιζομενη’, ’ριαλιτι’, ’μενεγακη’, ’αννα’, ’πηραν’, ’αστυνομικοι’, ’δις’, ’χαριτες’, ’εξαμαρτειν’, ’βιντιαδης’
8	’εμβολιου’, ’δοσεις’, ’δηλωσε’, ’pfizer’, ’εμβολιο’, ’κλινικες’, ’τελος’, ’ξεκινησει’, ’χωρες’, ’covid’
9	’παγκο’, ’συμβολαιο’, ’καταλανος’, ’πεπ’, ’τεχνικος’, ’μεσι’, ’φοντ’, ’μαντσεστερ’, ’γκουαρντιολα’, ’σιτι’
10	’κατασταση’, ’υγειας’, ’αρχιεπισκοπος’, ’νοσηλευεται’, ’ηπια’, ’συμπτωματα’, ’θετικος’, ’αναφερει’, ’μηνυμα’, ’ευαγγελισμο’
11	’of’, ’basket’, ’παιχνιδια’, ’stoiximan’, ’παιχνιδι’, ’πρωταθλημα’, ’us’, ’last’, ’part’, ’the’
12	’πατερα’, ’μαντλιν’, ’praia’, ’luz’, ’da’, ’χρηστος’, ’πατερας’, ’αγγλια’, ’εκπομπης’, ’εξαφανιστηκε’
NPMI σκορ	-0,13788552935440798
Συνάφεια	0,47572705704031054
Ποικιλομορφία	1,0

Πίνακας 3.7: Θέματα με τη χρήση ενσωματώσεων SBERT χωρίς καμία επεξεργασία

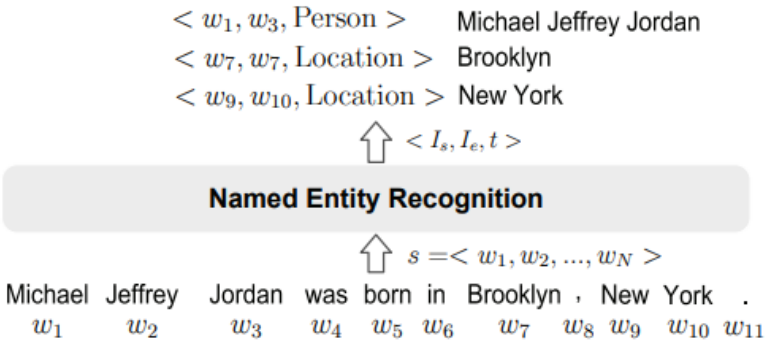
ο εντοπισμός γεγονότων στα άρθρα ενός συνόλου δεδομένων. Για το σκοπό αυτό χρησιμοποιήθηκε η βιβλιοθήκη DeepPavlov [Burt18b], η οποία περιέχει προεκπαιδευμένα NER μοντέλα σε διάφορες γλώσσες. Ένα παράδειγμα χρήσης της βιβλιοθήκης απεικονίζεται στο Σχήμα 3.14, όπου φαίνονται οι επώνυμες οντότητες που εντόπισε ο αλγόριθμος σε κάποιο δοθέν άρθρο. [Burt18a]

Παρατηρούμε ότι το NER, εντοπίζει επώνυμες οντότητες για αρκετές κατηγορίες και βοηθάει στην κατανόηση του γενικού νοήματος του άρθρου.

3.3.1 Εντοπισμός γεγονότων σε άρθρα

Ο *εντοπισμός γεγονότων* (event detection and tracking - EDT) είναι μέρος ενός συνόλου προβλημάτων που ανήκουν στην κατηγορία του *εντοπισμού θεμάτων* (topic detection and tracking - TDT). Το TDT είναι μία σημαντική περιοχή έρευνας που έχει προσελκύσει πολύ το ενδιαφέρον στο πεδίο της ανάλυσης πληροφορίας για τον εντοπισμό σημαντικών γεγονότων και την μελέτη της εξέλιξής τους με την πάροδο του χρόνου. Ο στόχος του EDT είναι να ομαδοποιήσει άρθρα τα οποία αναφέρονται σε κάποιο συγκεκριμένο γεγονός (λ.χ. ένα σεισμό στην Αθήνα) και να μελετήσει την εξέλιξη του στο χρόνο. [Mele17]

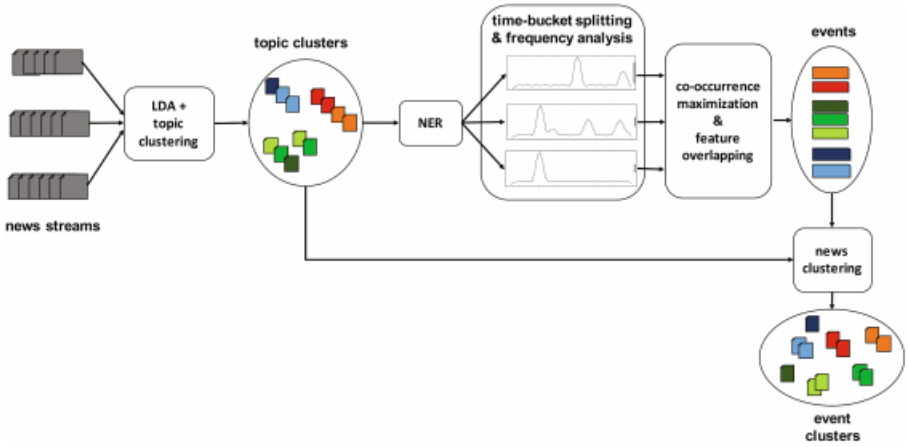
Ο εντοπισμός γεγονότων χρησιμοποιήθηκε στα πλαίσια της παρούσας διπλωματικής για την περαιτέρω συσταδοποίηση των άρθρων σε ομάδες γεγονότων, αφού σε κάθε συστάδα υπήρχαν άρθρα που αναφέρονταν σε διαφορετικά θέματα. Για τον εντοπισμό των γεγονότων χρησιμοποιήθηκε η τεχνική EDCN [Mele17] που απεικονίζεται στο Σχήμα 3.15 και η οποία τροποποιήθηκε για τους σκο-



Σχήμα 3.13: Παράδειγμα για το NER (Πηγή: [Li20])

Η κατάσταση έκτακτης ανάγκης τέθηκε σε ισχύ στις 9 Νοεμβρίου DATE και περιλαμβάνει την απαγόρευση της κυκλοφορίας κατά τη διάρκεια της νύχτας σε όλη τη χώρα και ένα μερικό λοκντάουν τα Σαββατοκύριακα στους 191 CARDINAL από τους 308 CARDINAL δήμους της Πορτογαλίας GPE. Πρόκειται να λήξει στις 23 Νοεμβρίου DATE, όμως οι βουλευτές πιθανότατα θα στηρίξουν το αίτημα του προέδρου . Με βάση το Σύνταγμα LAW, κατάσταση έκτακτης ανάγκης μπορεί να κηρυχθεί μόνο για 15 ημέρες DATE, ωστόσο είναι δυνατόν να ανανεώνεται επ ' αόριστον . « Η πρόληψη είναι ουσιαστικής σημασίας και ξεκινάει από τον καθένα από εμάς » , είπε ο Ρεμπέλο ντε Σόουζα PERSON μιλώντας σε δημοσιογράφους . Η κυβέρνηση είναι πιθανόν να ανακοινώσει νέα μέτρα το Σάββατο GPE. Η Πορτογαλία GPE, μια χώρα 10 εκατομμυρίων QUANTITY κατοίκων , έχει καταγράψει 236 . 015 CARDINAL κρούσματα και 3 . 631 CARDINAL θανάτους από την Covid - 19 LAW μέχρι σήμερα DATE, όμως οι μολύνσεις αυξάνονται και θέτουν υπό πίεση το σύστημα υγείας .

Σχήμα 3.14: Παράδειγμα για το NER για κάποιο άρθρο από το από τη συλλογή δεδομένων



Σχήμα 3.15: Τα κύρια βήματα του αλγόριθμου EDNC (Πηγή: [Mele17])

πούς της εργασίας. Τα βήματα του αλγορίθμου είναι τα εξής:

1. Συσταδοποίηση των άρθρων σε κάποιο αριθμό συστάδων.
2. Εύρεση των επώνυμων οντοτήτων κάθε άρθρου με τη χρήση του NER.
3. Ταξινόμηση των άρθρων κατά ομάδες με βάση την ώρα που δημοσιεύτηκαν και υπολογισμός των συχνοτήτων εμφάνισης των επώνυμων οντοτήτων σε κάθε ομάδα. Από αυτές τις οντότητες κρατούνται αυτές που εμφανίζονται περισσότερες φορές (λχ οι πρώτες 10%).
4. Μεγιστοποίηση συνεμφάνισης και επικάλυψη χαρακτηριστικών: σε αυτό το βήμα τα γεγονότα διαχωρίζονται και προσδιορίζεται η χρονική διάρκειά τους.

5. Τέλος, για τις συστάδες γεγονότων χρησιμοποιείται η κλασσική IR τεχνική, η οποία για κάθε άρθρο ελέγχει πόσο όμοιο είναι με το κάθε γεγονός, στη βάση της ομοιότητας συνημιτόνου.

Με βάση τον παραπάνω αλγόριθμο, υλοποιήθηκε η συσταδοποίηση των άρθρων, η οποία περιγράφεται στην Ενότητα 4.3.

Κεφάλαιο 4

Πρακτική Υλοποίηση Συστήματος Συστάσεων Ειδησεογραφικών Άρθρων

4.1 Ροές RSS

Τα τελευταία χρόνια υπάρχουν όλο και περισσότερες ιστοσελίδες που παρέχουν τα ειδησεογραφικά νέα τους με τη χρήση ροών RSS, λόγω της εύκολης πρόσβασης και παροχής άρθρων. Το RSS (Really Simple Syndication) είναι μια μορφή για την συχνή διαμοίραση ανανεωμένων πληροφοριών του διαδικτύου, όπως άρθρα και blogs, και βασίζεται στην eXtensible Markup Language (XML) [Bray98]. Μια ροή RSS, που ονομάζεται επίσης και «ροή νέων», συνήθως περιέχει τίτλους, περιλήψεις για κάποια άρθρα μιας ιστοσελίδας, καθώς και συνδέσμους για τις ιστοσελίδες των άρθρων. Με τη χρήση του RSS καθίσταται δυνατή η ενημέρωση για καινούργιο περιεχόμενο σε κάποιο ιστότοπο, χωρίς να χρειάζεται συνεχής έλεγχος.

Το παράδειγμα μιας ροής RSS φαίνεται στο Σχήμα 4.1. Παρατηρούμε ότι κάθε άρθρο εκφράζεται μέσω τις ετικέτας item και παρέχονται πληροφορίες για τον τίτλο, τον σύνδεσμο, την περιγραφή, τη μέρα και ώρα που δημοσιεύτηκε το άρθρο, σύνδεσμο για τυχόν εικόνες που περιέχει το άρθρο και τέλος το globally unique identifier (quid) του άρθρου.

```
<rss xmlns:atom="http://www.w3.org/2005/Atom"
xmlns:content="http://purl.org/rss/1.0/modules/content/"
xmlns:media="http://search.yahoo.com/mrss/" xmlns:dc="http://purl.org/dc/elements/1.1/"
version="2.0">
  <channel>
    <title>Εφημερίδα των Συντακτών</title>
    <link>https://www.efsyn.gr</link>
    <description>Νέα, Ειδήσεις και Απόψεις από την Εφημερίδα των Συντακτών στο
    διαδίκτυο</description>
    <item>
      <title>
        <![CDATA[ Καταδικάστηκε ο αδελφός του Ναβάλνι στη Ρωσία ]]>
      </title>
      <link>https://www.efsyn.gr/kosmos/eyropi/305447_katadikastike-o-adelfos-toy-nabalni-
      sti-rosia</link>
      <description>
        <![CDATA[ Ρωσικό δικαστήριο καταδίκασε τον αδελφό του φυλακισμένου ηγέτη της
        αντιπολίτευσης Αλεξέι Ναβάλνι, με την κατηγορία ότι κάλεσε σε διαδηλώσεις, κατά
        παράβαση των περιορισμών του κορονοϊού. ]]>
      </description>
      <pubDate>Fri, 06 Aug 2021 17:35:29 +0300</pubDate>
      <media:content
      url="https://www.efsyn.gr/sites/default/files/styles/default/public/2021-08/alexei-
      oleg-navalny-1.jpg?itok=52vKaAe9" medium="image"/>
      <guid>https://www.efsyn.gr/kosmos/eyropi/305447_katadikastike-o-adelfos-toy-nabalni-
      sti-rosia</guid>
    </item>
```

Σχήμα 4.1: Παράδειγμα ροής RSS από την ιστοσελίδα της «Εφημερίδας των Συντακτών»

Για τα δημιουργία του ΣΣΕΑ, αρχικά έπρεπε να δημιουργηθεί ένα σύνολο ειδησεογραφικών άρθρων. Για αυτό το σκοπό, επιλέχθηκαν ιστοσελίδες ειδησεογραφικών νέων που να έχουν δημόσια διαθέσιμες ροές. Πιο συγκεκριμένα επιλέχθηκαν οι 20 πιο δημοφιλείς ιστότοποι του ελληνικού Ίντερνετ,

όπως προκύπτεουν από τις μετρήσεις της Alexa και της Amazon, με την κατάταξη να διαμορφώνεται από τον συνδυσαμό του μέσου όρου των καθημερινών επισκεπτών και των προβολών τους [tops20]. Η ανανέωση της κατάταξης πραγματοποιήθηκε στις 12 Μαρτίου του 2020 και περιέχει τους 20 πρώτους ενημερωτικούς ιστοτόπους -εκτός από τους αθλητικούς - που βρίσκονται στο ημερήσιο Top 50 της Alexa. Ακόμα, η συγκεκριμένη συλλογή εμπλουτίστηκε και με ιστοσελίδες που επιλέχθηκαν στη βάση της συχνότητας εμφάνισής τους στο Google News [goog21]. Τέλος προστέθηκαν και ιστότοποι αθλητικών και καθαρά τεχνολογικών νέων. Έτσι, το σύνολο των ειδησεογραφικών ιστοσελίδων που χρησιμοποιήθηκε συνοψίζεται στον στον Πίνακα 4.1:

Όνομα	**URL**
Πρώτο Θέμα	`https://www.protothema.gr/`
NewsIT	`https://www.newsit.gr/`
Τα νέα online	`https://www.tanea.gr/`
Εφημερίδα των συντακτών	`https://www.efsyn.gr/`
Documento	`https://www.documentonews.gr/`
Το Βήμα	`https://www.tovima.gr/`
Techblog	`https://techblog.gr/`
Techmaniacs	`https://techmaniacs.gr/`
Gazzetta	`https://www.gazzetta.gr/`
EPT	`https://www.ert.gr/`
Star	`https://www.star.gr/`
Zappit	`https://www.zappit.gr/`
Onsports	`https://www.onsports.gr/`
Sportime	`https://www.sportime.gr/`
Metrosport	`https://www.metrosport.gr/`
Unboxholics	`https://unboxholics.com/`
Itechnews	`https://itechnews.gr/`
Eternity	`https://www.enternity.gr/rss.html`
Capital	`https://www.capital.gr/`
Enikonomia	`https://www.enikonomia.gr/`
Newsbomb	`https://www.newsbomb.gr/`

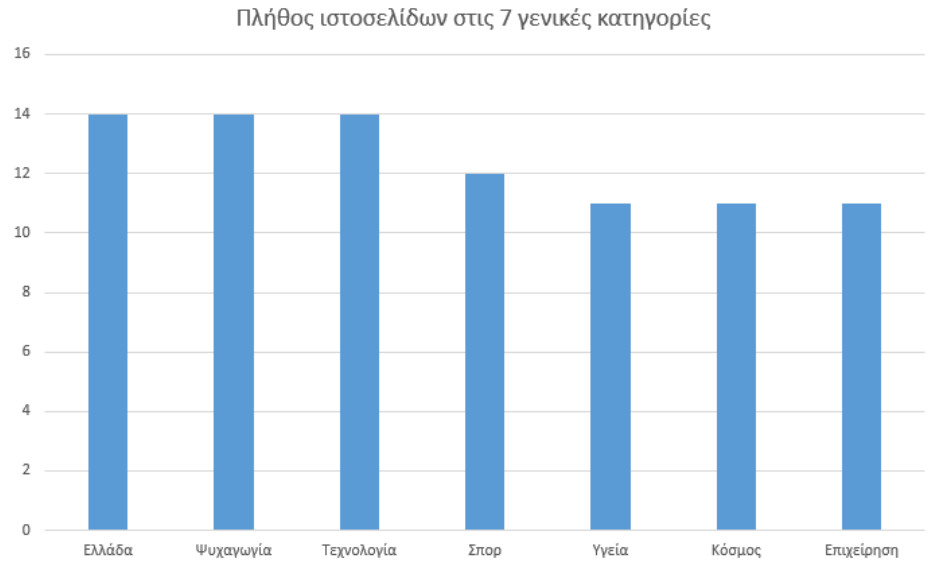
Πίνακας 4.1: Ειδησεογραφικές πηγές που χρησιμοποιήθηκαν στην παρούσα διπλωματική εργασία

Στο Σχήμα 4.2 απεικονίζονται συγκεντρωτικές στατιστικές πληροφορίες για τις πηγές που αναφέρονται στον Πίνακα 4.1. Παρατηρούμε ότι κάθε ιστότοπος μπορεί να δημοσιεύει άρθρα για διαφορετικές κατηγορίες.

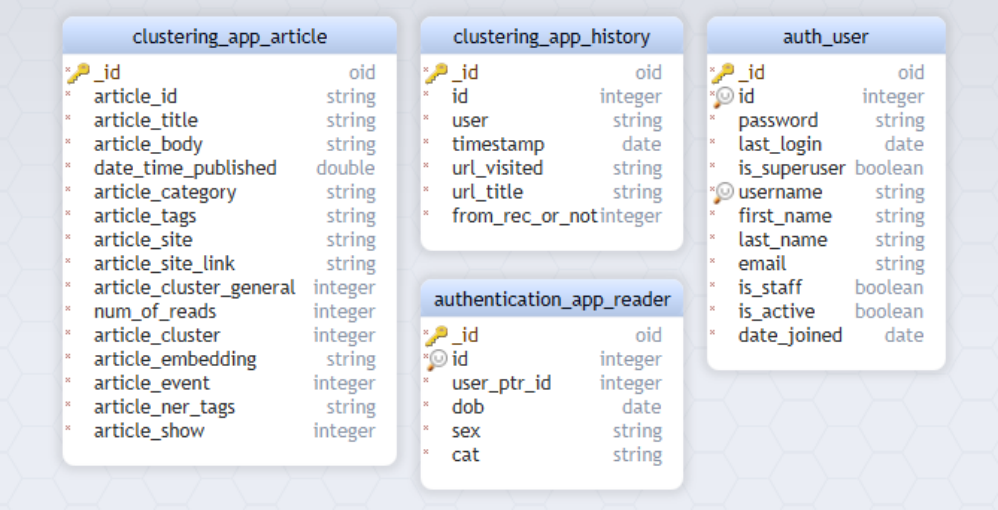
4.2 Σχεδιασμός Βάσης Δεδομένων

Τα δεδομένα που συλλέγονται από τις ροές επιλέχθηκε να αποθηκευτούν σε μια NoSQL βάση δεδομένων, την MongoDB [Chod13]. Η συγκεκριμένη επιλογή έγινε γιατί η αποθήκευση και η αναζήτηση ημι-δομημένων δεδομένων φαίνεται να είναι πιο αποτελεσματική σε μία NoSQL βάση δεδομένων. Αυτό συμβαίνει γιατί μία σχεσιακή βάση δεδομένων δεν είναι εύκολο να διαχειριστεί δεδομένα που συνεχώς αυξάνονται. [Gaut15]. Η MongoDB είναι μία βάση δεδομένων *προσανατολισμένη στο έγγραφο* (document-oriented), *ανοιχτού κώδικα* (open-source), η οποία αποθηκεύει τα δεδομένα σε μία μορφή που μοιάζει με JSON. Το κάθε έγγραφο είναι ένα σύνολο από ζεύγη κλειδιών-τιμής και έχει δυναμικό σχήμα. Πολλά έγγραφα ομαδοποιούνται για να δημιουργήσουν μια εγγραφή. Το έγγραφο και η συλλογή είναι το αντίστοιχο της εγγραφής και του πίνακα σε μια σχεσιακή βάση δεδομένων, αντίστοιχα.

Για την υλοποίηση του συστήματος συστάσεων χρειάστηκαν τέσσερεις συλλογές, μια για την αποθήκευση των στοιχείων των χρηστών, μια για την αποθήκευση των δεδομένων των άρθρων, μια



Σχήμα 4.2: Γραφική παράσταση του πλήθους των ιστοσελίδων που ανήκουν σε κάθε μία από τις 7 γενικές κατηγορίες (Ελλάδα, Κόσμος, Τεχνολογία, Σπορ, Υγεία, Επιχείρηση και Ψυχαγωγία)



Σχήμα 4.3: Συλλογές της βάσης δεδομένων που δημιουργήθηκε για το σύστημα συστάσεων.

για την αποθήκευση του ιστορικού των χρηστών και τέλος μια για τη σύνδεση των στοιχείων των χρηστών με την εφαρμογή του ΣΣΕΑ (Σχήμα 4.3):

1. **authentication_app_reader**: Κάθε έγγραφο αυτής της συλλογής αποθηκεύει τα στοιχεία που δίνουν οι χρήστες κατά την εγγραφή τους. Αυτά είναι το *id*, το *user_ptr_id* (το id είναι ίδιο με το id της συλλογής *auth_user*), την ημερομηνία γέννησης, το φύλο και τις κατηγορίες άρθρων που αρέσουν στο χρήστη να διαβάζει (Ελλάδα, Κόσμος, Υγεία, Τεχνολογία, Σπορ, Ψυχαγωγία, Επιχείρηση).

2. **clustering_app_article**: Κάθε έγγραφο αυτής της συλλογής αποθηκεύει τα δεδομένα κάθε άρθρου, καθώς και δεδομένα που παρέχονται από την αλληλεπίδραση των χρηστών με την ιστοσελίδα. Αυτά είναι τα εξής: id, τίτλος, κείμενο, ημέρα και ώρα δημοσίευσης, κατηγορία στην οποία ανήκει, ετικέτες που έχει, ο σύνδεσμος του άρθρου, ο σύνδεσμος της ιστοσελίδας στην οποία δημοσιεύτηκε, η γενική συστάδα στην οποία ανήκει, πόσες φορές έγινε ανάγνωση του από τους χρήστες, η ενσωμάτωσή του, το γεγονός στο οποίο ανήκει, τις ετικέτες που προκύ-

πτουν από το NER και το αν θα πρέπει να φαίνονται στην ιστοσελίδα του συστήματος συστάσεων ή όχι (*article_show*).

3. **clustering_app_history**: Κάθε έγγραφο αυτής της συλλογής αποθηκεύει πληροφορίες για το ποια άρθρα έχει διαβάσει ο χρήστης, αν είναι από τα προτεινόμενα ή όχι, καθώς και ποια είναι τα προτεινόμενα άρθρα για αυτόν κάθε φορά που ανανεώνει την ιστοσελίδα.

4. **auth_user**: Η ιστοσελίδα του ΣΣΕΑ δημιουργήθηκε με τη βοήθεια του framework Django (Ενότητα 4.5). Το Django περιέχει κάποιες έτοιμες συλλογές για την αποθήκευση των αναγκαίων δεδομένων των χρηστών. Έτσι, σε αυτή τη συλλογή περιέχονται τα βασικά δεδομένα των χρηστών όπως το id τους, το όνομα, το όνομα χρήστη, τον κωδικό, το email, την ημέρα που εγγράφηκαν, τα δικαιώματα που έχουν στη βάση κ.α.

4.3 Αλγόριθμος συσταδοποίησης κειμένου

Με βάση όσα παρουσιάστηκαν στο Κεφάλαιο 3, επιλέχθηκαν οι προεκπαιδευμένες ενσωματώσεις SBERT για την αναπαράσταση των κειμένων, το μοντέλο CTM για την συσταδοποίηση των άρθρων και το μοντέλο DeepPavlov για τον εντοπισμό των επώνυμων οντοτήτων. Για την εύρεση των γεγονότων στα άρθρα ακολουθήθηκαν τα εξής βήματα:

1. Υπολογισμός της συνεμφάνισης των επώνυμων οντοτήτων των άρθρων κάθε συστάδας. Η συνεμφάνιση των λέξεων εκφράζει το πόσο συχνά δύο λέξεις του λεξιλογίου του συνόλου δεδομένων εμφανίζονται μαζί. Για τον υπολογισμό της χρησιμοποιήθηκε αρχικά η μετρική της *σημείου-προς-σημείο αμοιβαίας πληροφορίας* (pointwise mutual information - PMI). Η μετρική υπολογίζει τη σχετική απόσταση μεταξύ της παρατηρούμενης και της αναμενόμενης συνεμφάνισης των λέξεων, θεωρώντας ότι αυτές οι δύο είναι ανεξάρτητες [Preo16] (Εξίσωση 4.1).

$$PMI(X,Y) = \alpha \cdot \frac{P(x,y)}{P(x) \cdot P(y)} \tag{4.1}$$

   όπου α ο παράγοντας κανονικοποίησης και PMI

$$\in -\infty \leq \text{pmi}(x;y) \leq \min\left[-\log p(x), -\log p(y)\right] \tag{4.2}$$

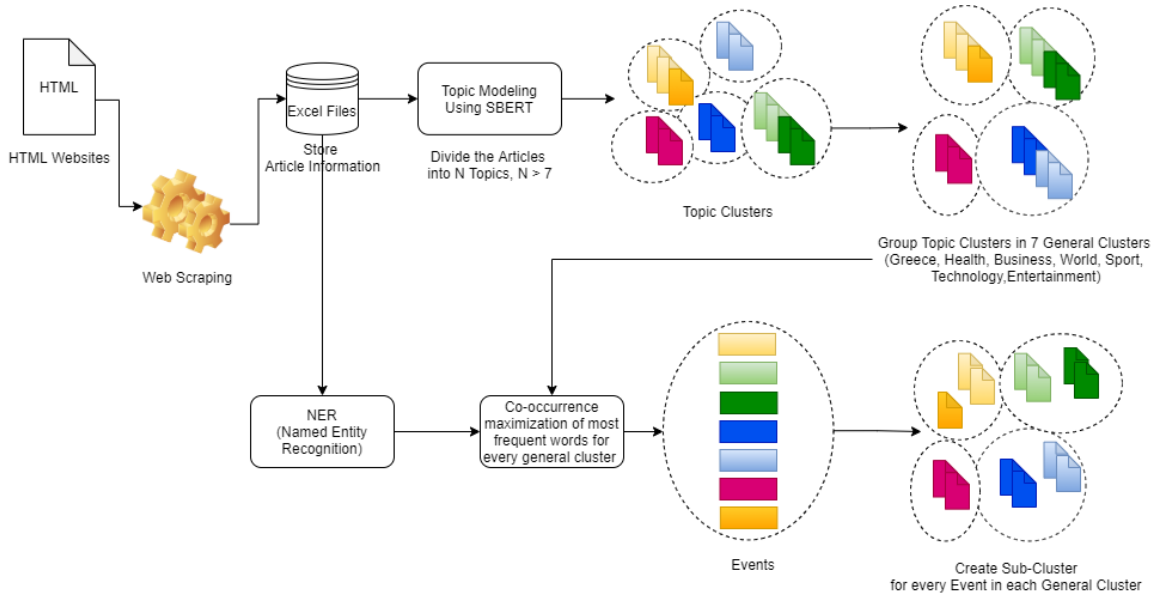
   Επειδή το PMI σκορ δεν είναι σταθερό και δυσχεραίνει την εύρεση των γεγονότων, επιλέχθηκε η μετρική NPMI (Εξίσωση 4.3)

$$NPMI(X,Y) = \frac{1}{P(x) \cdot P(y)} \cdot \frac{P(x,y)}{P(x) \cdot P(y)} \tag{4.3}$$

   όπου NPMI $\in [-1,1]$. Συγκρίνοντας τις Εξισώσεις 4.1 και 4.3, παρατηρούμε ότι η 4.3, προκύπτει από την 4.1 θέτοντας το α ίσο με $\frac{1}{P(x)\cdot P(y)}$. Ο όρος αυτός οριοθετεί το NPMI σκορ στο διάστημα $[-1,1]$. Έτσι, για κάθε συνδυασμό των επώνυμων οντοτήτων υπολογίστηκε το NPMI σκορ τους από την Εξίσωση 4.3.

2. Τα ζευγάρια λέξεων με NPMI σκορ μεγαλύτερο από a ορίστηκαν ως τα γεγονότα της συστάδας. Στη περίπτωση που εξετάσαμε δεν χρειάστηκε να χωρίσουμε τα άρθρα σε χρονικές περιόδους αφού τα άρθρα που συσταδοποιούνται κάθε φορά αναφέρονται σε μία συγκεκριμένη ημέρα. Επίσης, ως παράμετρος a ορίστηκε το 0 διότι όσο πιο κοντά είναι στο 1 το NPMI σκορ, τόσο περισσότερο οι λέξεις εντοπίζονται μαζί στα κείμενα της συλλογής. Επίσης, το μηδέν δείχνει ότι οι λέξεις είναι ανεξάρτητες, άρα αν το NPMI σκορ είναι μεγαλύτερο του μηδενός οι λέξεις έχουν συσχέτιση μεταξύ τους.

Η διαδικασία της συσταδοποίησης των άρθρων φαίνεται στο Σχήμα 4.4. Αναλυτικά τα βήματα του αλγορίθμου είναι τα εξής:



Σχήμα 4.4: Βήματα της διαδικασίας συσταδοποίησης των κειμένων της συλλογής

1. Κατέβασμα των άρθρων από τις ροές RSS και αποθήκευση των απαραίτητων πληροφοριών τους (τίτλος, κείμενο, κ.α).

2. Χρήση του αλγορίθμου CTM, για την ομαδοποίηση των άρθρων σε 25 κατηγορίες. Ο αριθμός 25 επιλέχθηκε γιατί ο αριθμός των άρθρων κάθε φορά είναι περίπου 400. Άρα υποθέσαμε ότι σε κάθε θέμα μπορεί να ανήκουν περίπου 16 άρθρα. Επίσης, στα 25 θέματα πήραμε καλύτερες λέξεις ανά θέμα, καθώς και καλύτερες μετρικές για τη συσταδοποίηση.

3. Χωρισμός των 25 θεματικών συστάδων σε 7 γενικές συστάδες. Επιλέχθηκαν 7 συστάδες για τις κατηγορίες Ελλάδα, Κόσμος, Υγεία, Τεχνολογία, Επιχείρηση, Ψυχαγωγία και Σπορ. Η αντιστοίχηση των 25 συστάδων στις 7 έγινε στη βάση των λέξεων των θεμάτων. Δεν επιλέχθηκε να γίνει από την αρχή ο χωρισμός των άρθρων σε 7 συστάδες γιατί δεν γινόταν καλός διαχωρισμός, αφού τα άρθρα αναφερόντουσαν σε διαφορετικά γεγονότα τα οποία δεν είχαν κάποια σύνδεση. Έτσι, τα θέματα που προέκυπταν δεν ανήκαν μόνο σε μία από τις 7 κατηγορίες αλλά τουλάχιστον σε δύο διαφορετικές.

4. Καθορισμός των επώνυμων οντοτήτων μέσω του NER για κάθε άρθρο.

5. Υπολογισμός των γεγονότων για κάθε συστάδα με την εύρεση του της συνεμφάνισης των επώνυμων οντοτήτων.

6. Περαιτέρω συσταδοποίηση των άρθρων κάθε συστάδας ανάλογα με το γεγονός στο οποίο αναφέρονται. Για κάθε γεγονός, ελέγχεται αν κάποιο άρθρο αναφέρεται σε αυτό. Αυτό γίνεται ελέγχοντας αν οι λέξεις που εκφράζουν το γεγονός περιέχονται στις επώνυμες οντότητες του κάθε άρθρου.

Ένα παράδειγμα συσταδοποίησης των άρθρων φαίνεται στα Σχήματα 4.5 και 4.6 για τις συστάδες Υγεία και Κόσμος. Βλέπουμε ότι τα άρθρα έχουν χωριστεί στις 2 κατηγορίες, και στη συνέχεια σε υποκατηγορίες οι οποίες είναι τα γεγονότα. Στο Σχήμα 4.6 βλέπουμε ότι το γεγονός είναι η επικοινωνία του Μπορέλ-Τσαβούσογλου για τις πυρκαγιές στην Αττάλεια και πως και τα τρία άρθρα που ανήκουν στη γενική συστάδα Κόσμος αναφέρονται σε αυτό το γεγονός για αυτό και ομαδοποιήθηκαν.

Υγεια

ΕΟΔΥ: Οι εννιά περιφέρειες που μετρούν εκατοντάδες κρούσματα
Documento, πριν από 48 ώρες και 7 λεπτά, ΕΟΔΥ_ΚΟΡΟΝΟΪΟΣ_ΚΡΟΥΣΜΑΤΑ

- Κορωνοϊός: 258 νέα κρούσματα στη Θεσσαλονίκη και 35 στη Χαλκιδική
  Metrosport, πριν από 45 ώρες και 33 λεπτά, ΝΕΑ_Θεσσαλονίκη_Χαλκιδική_Κορωνοϊός
- Κρούσματα σήμερα: 948 νέες μολύνσεις στην Αττική - Πτώση καταγράφει η Θεσσαλονίκη με 258
  Newsbomb, πριν από 48 ώρες και 59 λεπτά,
- Κορονοϊός: Σε 39 περιοχές διψήφια κρούσματα – Πρωταγωνιστεί πάλι η Αττική, ανησυχεί η Κρήτη
  NewsIT, πριν από 48 ώρες και 56 λεπτά, ΑΤΤΙΚΗ_ΚΟΡΟΝΟΪΟΣ_ΚΡΟΥΣΜΑΤΑ
- Κατανομή κρουσμάτων: Πρωταγωνιστεί πάλι η Αττική – Ανησυχία στην Κρήτη
  Star, πριν από 47 ώρες και 56 λεπτά, κορωνοϊός κρούσματα_γεωγραφική κατανομή κρουσμάτων_κορωνοϊός
- Κορονοϊός – Πού εντοπίζονται τα 2.760 κρούσματα
  Τα νέα online, πριν από 48 ώρες και 55 λεπτά,

Σχήμα 4.5: Παράδειγμα εύρεσης γεγονότων για τα άρθρα στη συστάδα Υγεία

Κοσμος

Επικοινωνία Μπορέλ - Τσαβούσογλου για τις πυρκαγιές στην Αττάλεια
Newsbomb, πριν από 45 ώρες και 46 λεπτά,

- Τηλεφωνική επικοινωνία Ζοζέπ Μπορέλ με Μεβλούτ Τσαβούσογλου για τις πυρκαγιές στην Αττάλεια
  NewsIT, πριν από 46 ώρες και 20 λεπτά, ΖΟΖΕΠ ΜΠΟΡΕΛ_ΜΕΒΛΟΥΤ ΤΣΑΒΟΥΣΟΓΛΟΥ
- ΕΕ: Επικοινωνία Ζοζέπ Μπορέλ με τον Μεβλούτ Τσαβούσογλου για τις πυρκαγιές στην Αττάλεια
  Πρώτο Θέμα, πριν από 46 ώρες και 27 λεπτά, Τουρκία_Αττάλεια

Βρετανία: Ο Μπόρις Τζόνσον και η σύζυγός του Κάρι περιμένουν το δεύτερο παιδί τους
Newsbomb, πριν από 46 ώρες και 1 λεπτά,

- Βρετανία: Το έκτο του παιδί αναμένει ο Μπόρις Τζόνσον - Η τρίτη σύζυγός του Κάρι ανακοίνωσε το χαρμόσυνο γεγονός
  Πρώτο Θέμα, πριν από 47 ώρες και 42 λεπτά, Βρετανία_Μπόρις Τζόνσον

Σχήμα 4.6: Παράδειγμα εύρεσης γεγονότων για τα άρθρα τ στη συστάδα Κόσμος

4.4 Αλγόριθμος συστήματος συστάσεων

Ο αλγόριθμος για την εύρεση των προτάσεων για κάθε χρήστη βασίστηκε στη CB τεχνική. Επιλέχθηκε αυτή η τεχνική για πολλούς λόγους. Αρχικά, έναντι της CF τεχνικής, η CB δίνει καλύτερα αποτελέσματα και δεν χρειάζεται δεδομένα για άλλους χρήστες για να βρει προτάσεις για κάποιο χρήστη. Για αυτό το λόγο δεν χρησιμοποιήθηκε και κάποια υβριδική τεχνική που να συνδυάζει τις CF και CB. Ακόμα, η χρήση αλγορίθμων βαθειάς μάθησης ήταν αρκετά δύσκολη λόγω της έλλειψης δεδομένων. Οι DL αλγόριθμοι χρειαζόταν αρκετά δεδομένα για να μπορέσουν να εκπαιδευτούν τα μοντέλα, έτσι ώστε να μπορούν να κάνουν προβλέψεις για τους χρήστες του ΣΣΕΑ. Ειδικότερα στα ελληνικά δεν υπάρχει κάποια συλλογή δεδομένων με ιστορικό χρηστών από ΣΣΕΑ και έτσι απορρίφθηκε και αυτή η λύση, παρόλο που όπως φάνηκε από την ανάλυση στο Κεφάλαιο 2 είναι αυτή που

δίνει τα καλύτερα αποτελέσματα.

Με τη χρήση της CB τεχνικής, το πρόβλημα της ψυχρής εκκίνησης αντιμετωπίστηκε με την χρήση τεχνικών απόσπασης προτιμήσεων. Κατά την εγγραφή των χρηστών στο ΣΣΕΑ, τους ζητείται υποχρεωτικά να επιλέξουν από μία έως τρεις κατηγορίες για τις οποίες τους αρέσει να διαβάζουν άρθρα. Αν ο χρήστης συνδέεται στον ιστότοπο του ΣΣΕΑ για πρώτη φορά, τότε στο πεδίο των προτάσεων θα του εμφανιστούν τα πιο πρόσφατα νέα για τις κατηγορίες που έχει επιλέξει στην εγγραφή του. Αν υπάρχει ιστορικό τότε οι προτάσεις θα δημιουργηθούν από την CB τεχνική.

Η συγκεκριμένη CB τεχνική που επιλέχθηκε [Sama19] χρησιμοποιεί τον μέσο όρων των ενσωματώσεων των οντοτήτων των κειμένων για να αναπαραστήσει τα κείμενα σε ένα διανυσματικό χώρο. Ακόμα, για τους χρήστες κρατάει στο ιστορικό τους τις οντότητες από κάθε άρθρο που έχουν διαβάσει, τη συχνότητα που εμφανίζεται η οντότητα στα κείμενα που διαβάζουν, ποια ήταν η τελευταία φορά που διάβασαν άρθρο με αυτή την οντότητα και πόσες φορές έχει διαβαστεί άρθρο με αυτή την οντότητα από όλους τους χρήστες. Με βάση αυτά τα έμμεσα δεδομένα, από την Εξίσωση 4.4 υπολογίζεται μια βαθμολογία, η οποία πολλαπλασιάζεται με τις ενσωματώσεις των οντοτήτων (Εξίσωση 4.5) για την αναπαράσταση του κάθε χρήστη στον ίδιο διανυσματικό χώρο με τα άρθρα.

$$s(k,u) = w_1 \cdot \frac{rc(k)}{\max_{m \in P(u)} rc(m)} \cdot 2^{-\frac{t_p - k_t}{h}} + w_2 \cdot \frac{c(k)}{\max_{m \in P(u)} c(m)} \tag{4.4}$$

$$\vec{u} = \sum_{k \in P(u)} s(k,u) \cdot embedding(k) \tag{4.5}$$

όπου $w_1 + w_2 = 1$, rc τα πρόσφατα κλικ στην κάθε οντότητα τις τελευταίες k μέρες, c τα κλικ που έγιναν από το χρήστη σε κάθε οντότητα, t_p η τρέχουσα χρονική στιγμή, k_t η χρονική στιγμή που έγινε το τελευταίο κλικ στην οντότητα, h ένα διάστημα υποδιπλασιασμού μετρημένο σε δευτερόλεπτα και P είναι το προφιλ των ενδιαφερόντων του χρήστη. Η βαθμολογία προκύπτει ως ο μέσος όρος του ενδιαφέροντος του χρήστη για κάθε οντότητα και μειώνεται με την πάροδο του χρόνου αφού, όπως φάνηκε και στην Ενότητα 2.2.3, οι προτιμήσεις των χρηστών επηρεάζονται από το πόσο πρόσφατο είναι κάποιο άρθρο και άρα και οι οντότητες που αναφέρει. Τέλος υπολογίζεται η ομοιότητα μεταξύ των χρηστών και των άρθρων για να κατασκευαστεί η λίστα των προτάσεων.

Ο αλγόριθμος αυτός, προσαρμόστηκε στην σχεδίαση του ΣΣΕΑ, λαμβάνοντας υπόψιν και τον τρόπο αναπαράστασης των άρθρων σε γεγονότα μέσω της συσταδοποίησης. Έτσι, τα έμμεσα δεδομένα για τον κάθε χρήστη είναι η συχνότητα με την οποία διάβασε νέα από κάποιο συγκεκριμένο γεγονός, πότε διάβασε άρθρο από το γεγονός αυτό τελευταία φορά και πόσα άρθρα έχουν διαβαστεί από τη συστάδα που ανήκει το γεγονός. Έτσι, η Εξίσωση 4.4 διαμορφώνεται ως εξής (Εξίσωση 4.6)

$$s(e,u) = w_1 \cdot \frac{c(e)}{\max_{m \in P(u)} c(m)} \cdot 2^{-\frac{t_p - k_t}{h}} + w_2 \cdot \frac{cc(cluster_e)}{\max_{cluster_m \in P(u)} cc(cluster_m)} \tag{4.6}$$

όπου $P(u)$ είναι το ιστορικό του χρήστη, $c(e)$ ο αριθμός των κλικ που έγιναν από το χρήστη στο γεγονός e και $cc(cluster_e)$ είναι τα κλικ που έγιναν από το χρήστη στη συστάδα που ανήκει το γεγονός e. Τα υπόλοιπα είναι ίδια με την Εξίσωση 4.4. Με βάση αυτή τη βαθμολογία και τις ενσωματώσεις των άρθρων του ιστορικού του χρήστη, το διάνυσμα για την αναπαράσταση του στο διανυσματικό χώρο δίνεται από την Εξίσωση 4.7. Για κάθε τιμή της βαθμολογίας για κάποιο γεγονός e, αυτή πολλαπλασιάζεται με το άθροισμα των ενσωματώσεων των άρθρων a που έχει διαβάσει ο χρήστης από αυτό το γεγονός. Έτσι, το διάνυσμα $\vec{u}$ υπολογίζεται ως το άθροισμα αυτού του γινομένου για κάθε γεγονός e που έχει ο χρήστης στο ιστορικό του.

$$\vec{u} = \sum_{e \in P(u)} s(e,u) \cdot \sum_{a \in e} embedding(a) \tag{4.7}$$

Ο πρώτος όρος της Εξίσωσης 4.6, δείχνει το ενδιαφέρον του χρήστη για κάποιο γεγονός. Το ενδιαφέρον αυτό μειώνεται με τον χρόνο μέσω ενός παράγοντα απόσβεσης, ενώ ο δεύτερος όρος ποσοτικοποιεί το ενδιαφέρον του χρήστη γενικότερα για τη συστάδα στην οποία ανήκει το γεγονός. Με

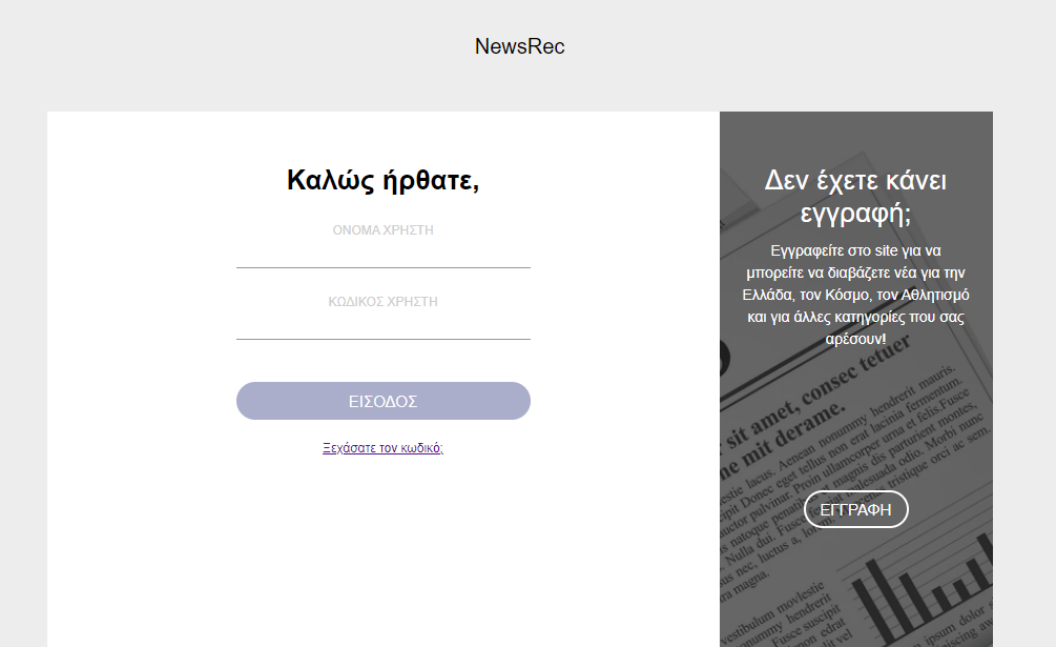
τη βοήθεια του πρώτου όρου βλέπουμε αν ο χρήστης δείχνει μεγάλο ενδιαφέρον σε συγκεκριμένα γεγονότα και πόσο επηρεάζεται το ενδιαφέρον του με το χρόνο, ενώ με τον δεύτερο όρο βλέπουμε αν ενδιαφέρεται γενικά για τη συγκεκριμένη συστάδα ή αν έτυχε να τον ενδιαφέρει κάποιο γεγονός κάποια χρονική περιόδο.

Για την εύρεση των άρθρων που θα προταθούν στον κάθε χρήστη χρησιμοποιείται η ομοιότητα συνημιτόνου ανάμεσα στα δύο διανύσματα $\vec{a}$ και $\vec{u}$ όπου $\vec{a} = embedding(a), a \in History(u)$. Τα άρθρα που προτείνονται πρέπει να είναι καινούργια άρθρα και να μην τα έχει ξαναδιαβάσει ο χρήστης.

4.5 Σχεδιασμός της ιστοσελίδας του συστήματος συστάσεων ειδησεογραφικών άρθρων

Η δημιουργία του συστήματος συστάσεων ειδησεογραφικών άρθρων έγινε με τη βοήθεια του Django [Djan]. Το Django είναι ένα περιβάλλον για τη γρήγορη δημιουργία εφαρμογών διαδικτύου το οποίο: παρέχει αρκετά έξτρα πακέτα για τη δημιουργία των εφαρμογών όπως ταυτοποίηση χρηστών και διαχείριση περιεχομένου, παρέχει ασφάλεια πχ για την εγγραφή των χρηστών, ενώ είναι επεκτάσιμο και ευέλικτο. Για τη δημιουργία του συστήματος συστάσεων, χρησιμοποιήθηκε το Djongo [djon21] το οποίο είναι μια επέκταση του Django για την χρήση της MongoDB σαν βάση δεδομένων.

Η υλοποίηση της ιστοσελίδας έγινε με τη χρήση των γλωσσών HTML, CSS, SCSS, Javascript και Bootstrap χρησιμοποιώντας τα πρότυπα Ramayana [1] για την ιστοσελίδα και το Login/Registration Form Transition [2] για την εγγραφή των χρηστών. Εικόνες από το τελικό template της εγγραφής/εισόδου της εφαρμογής καθώς και της κεντρικής σελίδας αντίστοιχα, φαίνονται στα Σχήματα 4.7 και 4.8

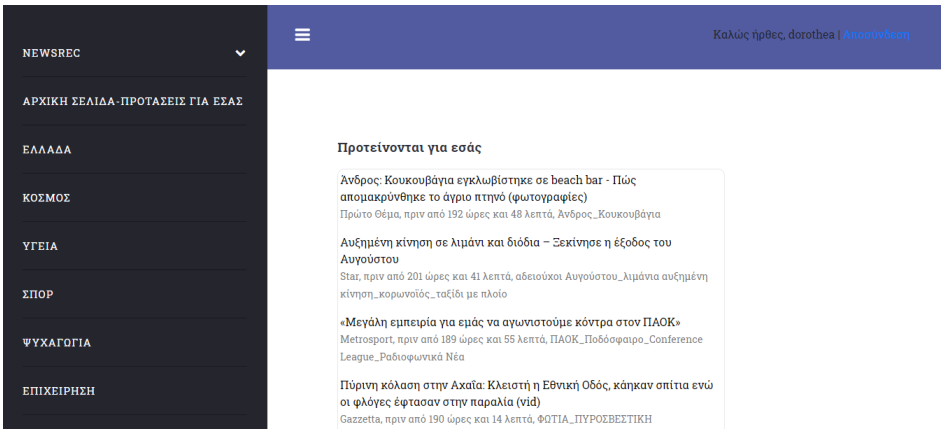


Σχήμα 4.7: Σελίδα εισόδου της εφαρμογής.

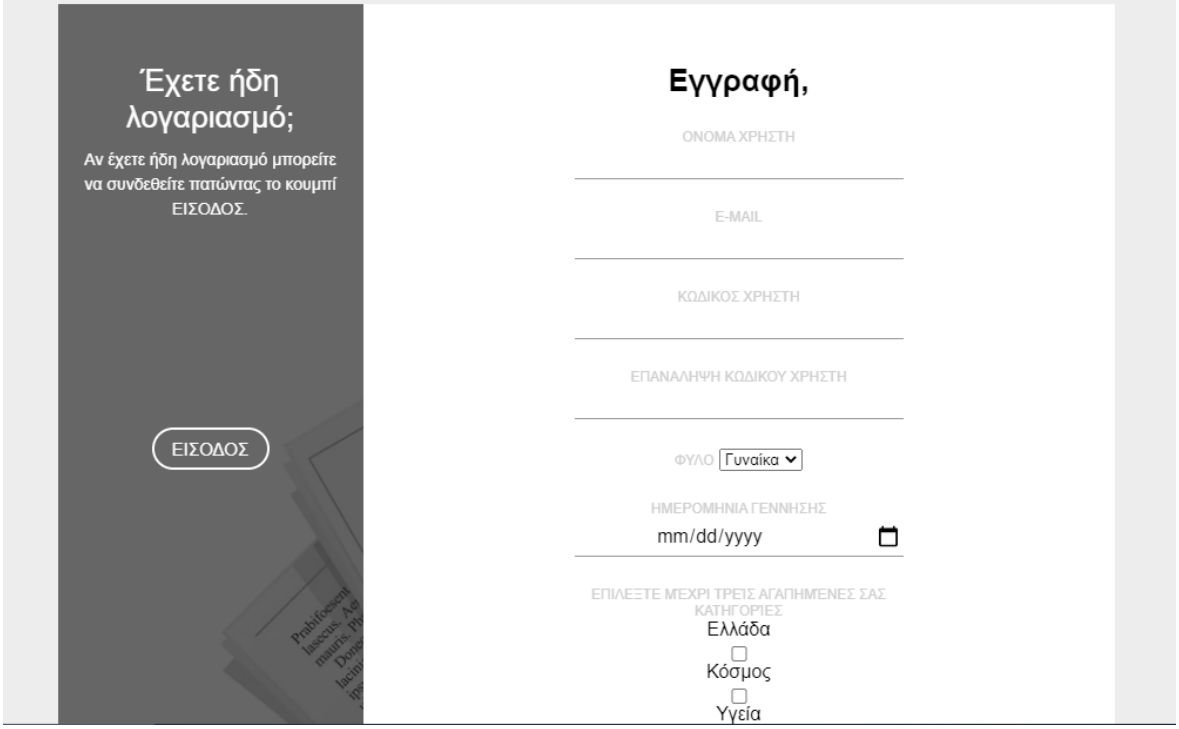
Κάθε χρήστης που θέλει να χρησιμοποιήσει την εφαρμογή αρχικά θα πρέπει να κάνει εγγραφή και να συμπληρώσει την φόρμα που φαίνεται στο Σχήμα 4.9. Στη συνέχεια κατευθύνεται στη σελίδα που φαίνεται στο Σχήμα 4.7 για την είσοδο και τέλος στην κεντρική σελίδα που φαίνεται στο Σχήμα 4.8.

Από τη στιγμή που κάποιος χρήστης συνδέεται στην εφαρμογή μπορεί να δει τα προτεινόμενα άρθρα για εκείνον με βάση το ιστορικό του, να διαβάσει νέα για όποια κατηγορία τον ενδιαφέρει

[1] https://templatemo.com/tm-529-ramayana
[2] https://codepen.io/suez/pen/RpNXOR



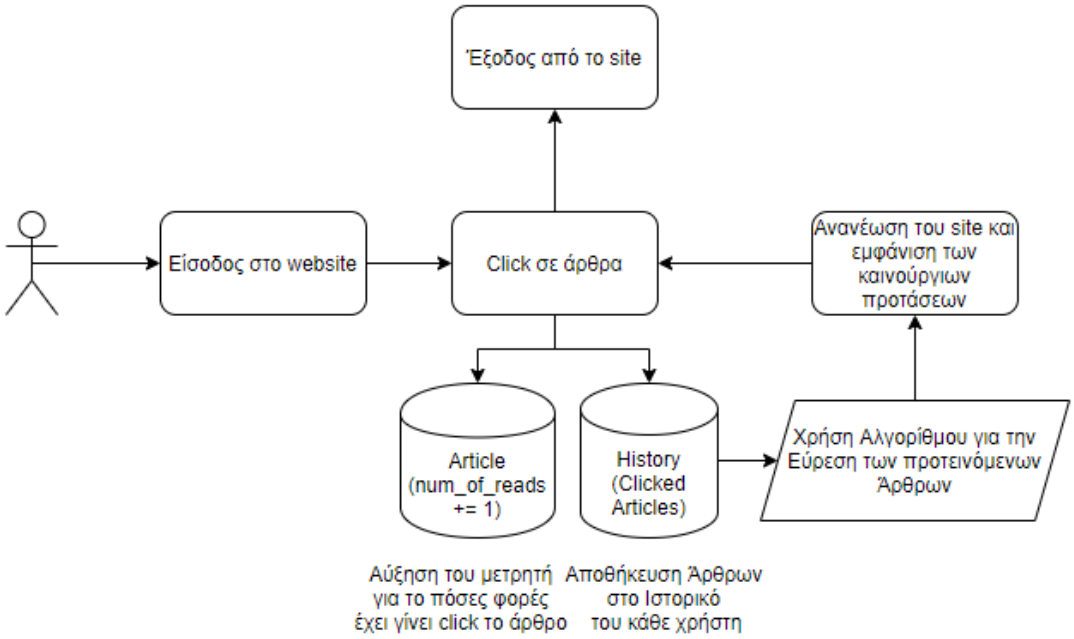
Σχήμα 4.8: Κεντρική σελίδα της εφαρμογής.



Σχήμα 4.9: Σελίδα εγγραφής των χρήστων.

και όταν έχει ενημερωθεί για τα νέα να αποσυνδεθεί από την εφαρμογή. Κάθε φορά που είναι στην κεντρική σελίδα και κάνει ανανέωση, εμφανίζονται καινούργια άρθρα αν έχει ανανεωθεί το ιστορικό του. Η διαδικασία αυτή φαίνεται στο Σχήμα 4.10.

Τα άρθρα στην ιστοσελίδα ανανεώνονται ανεβάζοντας στην βάση δεδομένων ένα αρχείο με τα καινούργια άρθρα. Η διαδικασία αυτή δεν γίνεται αυτόματα διότι θα έπρεπε στον εξυπηρετητή να ανέβει και το μοντέλο CTM που έχει προκύψει από τη διαδικασία της εκπαίδευσης. Το μοντέλο χρειάζεται για να μπορούν τα καινούργια άρθρα να ταξινομηθούν στις συστάδες που έχουν ήδη δημιουργηθεί, χωρίς να χρειάζεται να εκπαιδευτεί το μοντέλο από την αρχή, ώστε να ανανεώνεται αυτόματα το περιεχόμενο της ιστοσελίδας. Όμως, το CTM χρειάζεται επιταχυντές υλικού (λ.χ. GPU) για να λειτουργήσει που δεν ήταν διαθέσιμες στον εξυπηρετητή που φιλοξενεί την ιστοσελίδα του συστήματος συστάσεων. Έτσι, η εκπαίδευση του μοντέλου καθώς και η εύρεση των γεγονότων πραγματοποιούνται εκτός του εξυπηρετητή (offline) και τα καινούργια άρθρα ανεβαίνουν στην ιστοσελίδα μη αυτόματα. Αν το μοντέλο ήταν ανεβασμένο στην ιστοσελίδα, θα χρειαζόταν να εκπαιδευτεί κάθε φορά που τα νέα αλλάζουν πολύ, έτσι ώστε να μπορεί να ταξινομεί σωστά τα καινούργια άρθρα για κάποιες μέρες



Σχήμα 4.10: Χάρτης περιήγησης στο site.

χωρίς να χρειάζεται εκπαίδευση.

Κεφάλαιο 5

Συμπεράσματα και Μελλοντικές Κατευθύνσεις

5.1 Συμπεράσματα

Στα πλαίσια της παρούσας διπλωματικής εργασίας πραγματοποιήθηκε η δημιουργία ενός ΣΣΕΑ. Η δημιουργία του ΣΣΕΑ, είχε χωριστεί σε τρεις μεγάλες κατηγορίες, την συσταδοποίηση των άρθρων, τη δημιουργία ενός αλγορίθμου προτάσεων άρθρων και τη δημιουργία μίας λειτουργικής ιστοσελίδας. Η υλοποίηση αυτών των τριών στόχων επιτεύχθηκε σε μεγάλο βαθμό και έδωσε αρκετά σημαντικά αποτελέσματα.

Γενικό συμπέρασμα είναι ότι για να εκπαιδευτούν τα μοντέλα μηχανικής μάθησης για κάποιο ΣΣΕΑ και να δώσουν καλά αποτελέσματα, χρειάζονται μεγάλο όγκο δεδομένων εισόδου χωρίς «θόρυβο». Κατά τη συσταδοποίηση των άρθρων, όσο περισσότερα ήταν τα άρθρα τόσο καλύτερες ήταν οι αντίστοιχες μετρικές. Επίσης, όσον αφορά τον «θόρυβο», στα περισσότερα κείμενα στο τέλος, υπήρχαν λέξεις όπως «πηγή» ή «ενημερωθείτε για». Όταν αυτές οι λέξεις δεν είχαν αφαιρεθεί, επηρέαζαν την μοντελοποίηση θεμάτων γιατί είχαν μεγάλη συχνότητα. Οι λέξεις αυτές δηλαδή, φαινόταν ότι χαρακτήριζαν κάποια συστάδα και επηρέαζαν αρνητικά το ΤΜ. Τέλος, για την εκπαίδευση των μοντέλων DL, για την εύρεση των προτάσεων, θα πρέπει να υπάρχουν αρκετά δεδομένα στο ιστορικό των χρηστών καθώς και μεγάλος αριθμός εγγεγραμμένων αναγνωστών στο ΣΣΕΑ.

Για το κομμάτι των ενσωματώσεων λέξεων ή προτάσεων, η μεταφορά μάθησης αποτελεί μία εξαιρετική λύση για δημιουργία τους. Δυστυχώς, οι ελληνικές προεκπαιδευμένες ενσωματώσεις λέξεων δεν έδωσαν καλά αποτελέσματα στα ειδησεογραφικά κείμενα, όπως φαίνεται και στην Ενότητα 3.2.4. Επίσης, το ελληνικό BERT, δεν είναι εκπαιδευμένο σε συλλογές δεδομένων κοντά στα είδη κειμένων που μελετήθηκαν, οπότε ούτε αυτό έδωσε καλά αποτελέσματα, παρόλο που το BERT θεωρείται κορυφαίο μοντέλο ΕΦΓ. Παρόλα αυτά, η χρήση προ-εκπαιδευμένων πολυγλωσσικών μοντέλων έδωσε αρκετά καλά αποτελέσματα και στη συσταδοποίηση, αλλά και στον εντοπισμό των οντοτήτων των κειμένων μέσω του NER. Τα προεκπαιδευμένα πολυγλωσσικά μοντέλα, χρησιμοποιήθηκαν για τη δημιουργία ενσωματώσεων προτάσεων για τα άρθρα, τα οποία χρησιμοποιήθηκαν στη συσταδοποίηση και στον αλγόριθμο εύρεσης προτάσεων.

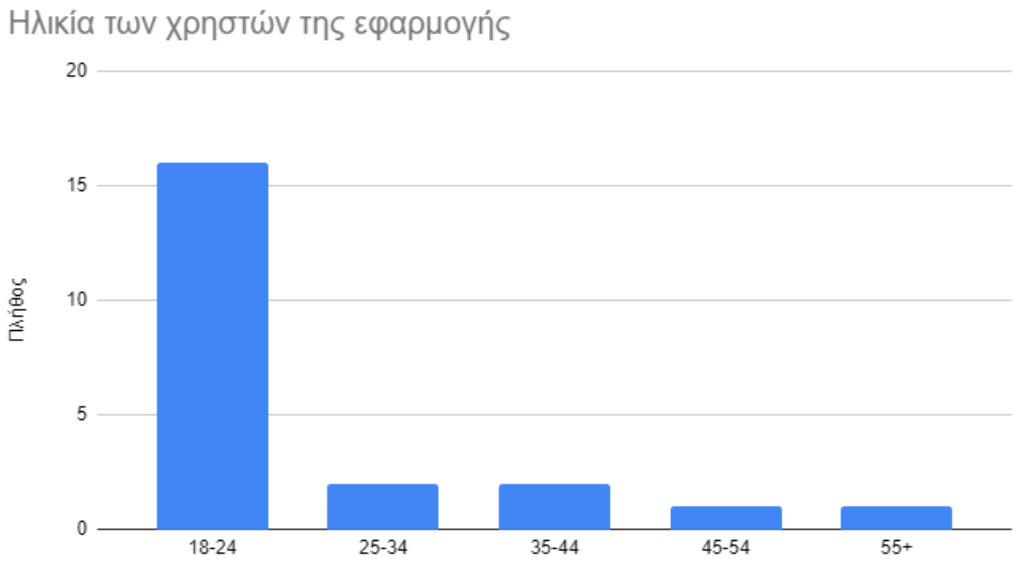
Όσον αφορά τους αλγόριθμους συσταδοποίησης, οι κλασικοί αλγόριθμοι όπως ο (σφαιρικός) k μέσων, ενώ φαίνονται να κάνουν καλή ομαδοποίηση, δεν είναι εύκολο να καθοριστεί κάθε συστάδα άρθρων σε ποια κατηγορία αναφέρεται. Έτσι, η μοντελοποίηση θεμάτων αποτελεί την καλύτερη λύση για την περίπτωση που εξετάστηκε στη διπλωματική. Επίσης, το ΤΜ με τη χρήση νευρωνικών μοντέλων έδωσε καλύτερα αποτελέσματα από τα συμβατικά μοντέλα που εξετάστηκαν.

Τέλος, ο CB αλγόριθμος που υλοποιήθηκε για την εύρεση προτάσεων για τους χρήστες του ΣΣΕΑ, φαίνεται να παράγει καλές προτάσεις προς τους χρήστες. Όσο περισσότερα άρθρα διαβάζει κάποιος χρήστης, τόσο καλύτερα άρθρα του προτείνει το ΣΣ. Αυτό συμβαίνει γιατί η ενσωμάτωση που αναπαριστά τον κάθε χρήστη προσαρμόζεται στα άρθρα που υπάρχουν κάθε φορά στο ιστορικό του.

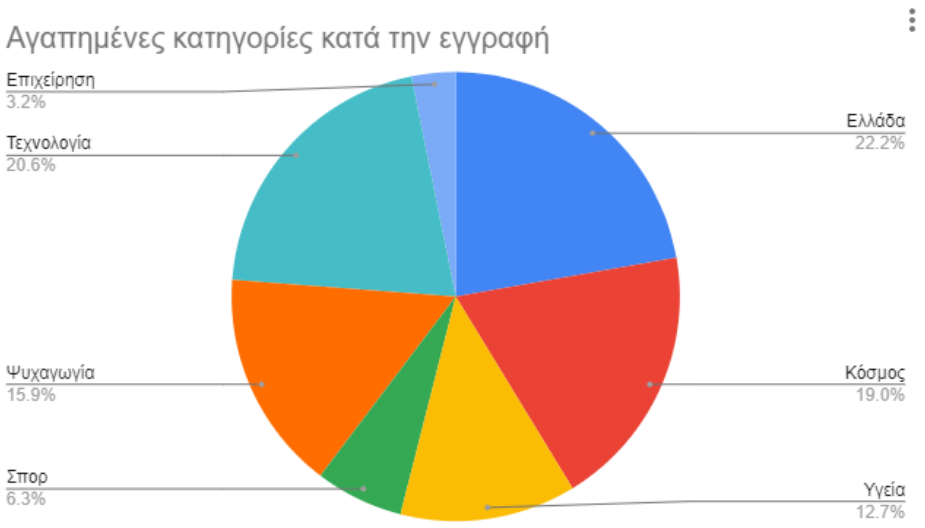
Συνολικά, φαίνεται ότι το ΣΣΕΑ που αναπτύχθηκε προτείνει άρθρα στους χρήστες σχετικά με τα ενδιαφέροντα τους και κάνει μια καλή ομαδοποίηση των άρθρων, αρχικά στις επτά γενικές κατηγορίες (Ελλάδα, Κόσμος, Τεχνολογία. Υγεία, Επιχείρηση, Σπορ, Ψυχαγωγία) και στη συνέχεια σε γεγονότα. Τα στατιστικά που προέκυψαν από τη χρήση του ΣΣΕΑ παρουσιάζονται στην αμέσως επόμενη Ενότητα.

5.2 Στατιστικά χρήσης της ιστοσελίδας

Στην ιστοσελίδα που δημιουργήθηκε για το ΣΣΕΑ, εγγράφηκαν 22 χρήστες. Τα στατιστικά για την ηλικία και για τα ενδιαφέροντα των χρηστών, φαίνονται στα Σχήματα 5.1 και 5.2 αντίστοιχα. Όπως φαίνεται από το Σχήμα 5.1, το μεγαλύτερο μέρος των χρηστών ανήκει στις ηλικίες 18 μέχρι 24, ενώ στις υπόλοιπες ηλικίες ανήκουν ένας ή δύο χρήστες. Για αυτούς τους χρήστες, η αγαπημένη τους κατηγορία είναι η Ελλάδα με ποσοστό 22, 2%, και στη συνέχεια η Τεχνολογία με ποσοστό 20, 6% όπως φαίνεται στο Σχήμα 5.2.



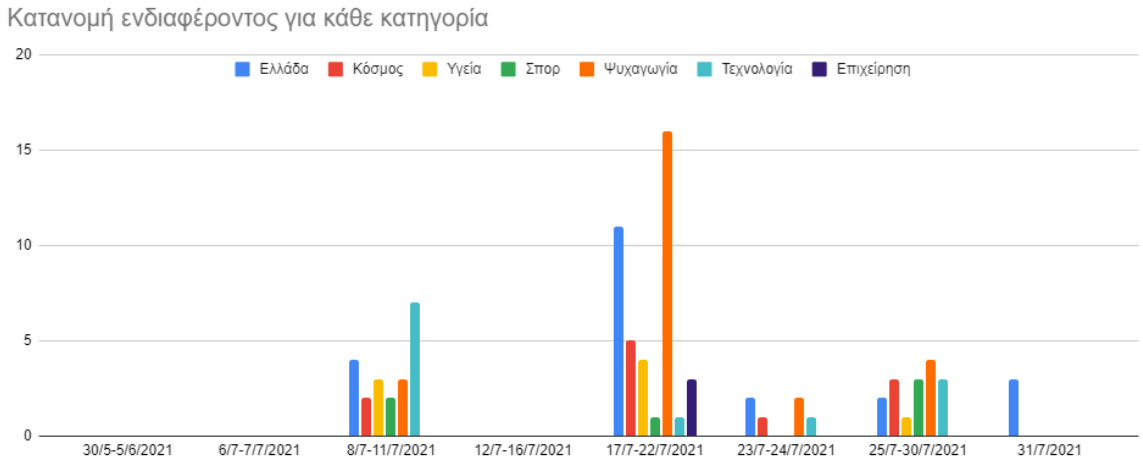
Σχήμα 5.1: Κατανομή ηλικίας των χρηστών του ΣΣΕΑ.



Σχήμα 5.2: Αγαπημένες κατηγορίες που επέλεξαν οι χρήστες κατά την εγγραφή τους στο ΣΣΕΑ.

Στην ιστοσελίδα ανέβηκαν συνολικά 2.354 διαφορετικά άρθρα στο διάστημα 6 Ιουλίου έως 31 Ιουλίου και στις 30 Μαΐου του 2021. Η κατανομή ενδιαφέροντος για τους χρήστες του ΣΣΕΑ φαίνεται στο Σχήμα 5.3. Παρατηρούμε ότι κατά μέσο όρο τα περισσότερα άρθρα που αναγνώστηκαν ανήκουν στην κατηγορία Ψυχαγωγία, παρόλο που η κατηγορία αυτή είναι τέταρτη στις αγαπημένες κατηγορίες των χρηστών με ποσοστό 15, 9%. Αυτό μας δείχνει ότι τα ενδιαφέροντα των χρηστών ως προς τα

άρθρα αλλάζουν σύμφωνα με την επικαιρότητα και δεν σχετίζονται πάντα με τις μακροχρόνιες επιθυμίες τους. Συγκεκριμένα, βλέπουμε ότι σε κάθε χρονική περίοδο έχουμε και διαφορετική κατηγορία στην οποία ανήκουν τα περισσότερα άρθρα που διαβάστηκαν. Αυτό μπορεί να συμβαίνει για δύο λόγους. Πρώτον, οι χρήστες που μπήκαν αυτά τα χρονικά διαστήματα, ενδιαφέρονται περισσότερο για αυτές τις κατηγορίες. Δεύτερον, τα γεγονότα της επικαιρότητας σε αυτές τις κατηγορίες ήταν πιο δημοφιλή από τα γεγονότα των άλλων κατηγοριών. Για παράδειγμα, στο διάστημα 8 Ιουλίου με 11 Ιουλίου, πρώτη κατηγορία είναι η Τεχνολογία και δεύτερη η Ελλάδα, ενώ στο διάστημα 17 Ιουλίου με 22 Ιουλίου είναι πρώτη η Ψυχαγωγία και δεύτερη η Ελλάδα.

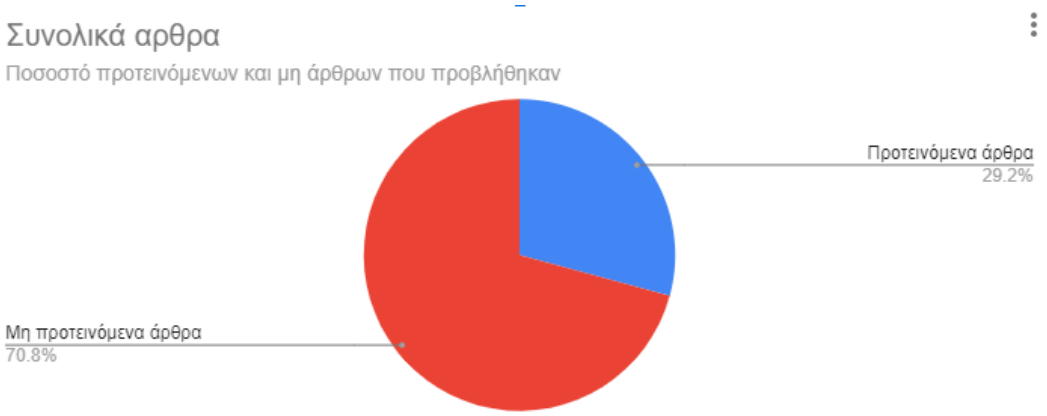


Σχήμα 5.3: Κατανομή ενδιαφέροντος για τις ημερομηνίες που ανανεώθηκαν τα άρθρα.

Από τους 22 χρήστες που εγγράφηκαν στο ΣΣΕΑ, μόνο οι 14 μπήκαν στην εφαρμογή πάνω από μία φορά και πάτησαν άρθρα, έτσι ώστε να τους προτείνει το σύστημα προσωποποιημένες προτάσεις. Συνολικά έγιναν 175 προτάσεις στους χρήστες, ενώ προβλήθηκαν συνολικά 113 άρθρα. Στο Σχήμα 5.4 βλέπουμε ότι το ποσοστό των άρθρων που προβλήθηκαν από τα προτεινόμενα είναι 18, 9%. Άρα, από τις 175 προτάσεις άρθρων που έγιναν στους χρήστες, οι 33 προτάσεις ήταν εύστοχες και τους άρεσαν. Αυτό σημαίνει ότι περίπου μία στις πέντε προτάσεις που έκανε το σύστημα ενδιέφερε κάποιο χρήστη. Στο Σχήμα 5.5 φαίνεται ότι το ποσοστό των προτεινόμενων από τα συνολικά άρθρα που αναγνώστηκαν από τους χρήστες είναι 29, 2%. Συγκεκριμένα, όπως αναφέρθηκε, τα 33 άρθρα από τα 113 που προβλήθηκαν ήταν από την λίστα των προτάσεων. Αυτό σημαίνει ότι περίπου 1 στα 4 άρθρα που διαβάστηκαν ήταν από τη λίστα των προτάσεων.

Σχήμα 5.4: Ποσοστό άρθρων που προβλήθηκαν από τα προτεινόμενα άρθρα στους χρήστες.

Συμπερασματικά, η χρήση ενός CB αλγόριθμου, ήταν η καλύτερη επιλογή για την υλοποίηση του ΣΣΕΑ, δεδομένης της έλλειψης αντίστοιχων δεδομένων στα ελληνικά για την υλοποίηση ενός DL



Σχήμα 5.5: Ποσοστό των προτεινόμενων άρθρων από τα συνολικά άρθρα που προβλήθηκαν από τους χρήστες.

συστήματος προτάσεων. Αυτό φαίνεται από το γεγονός ότι τα ποσοστά των προτεινόμενων άρθρων που αναγνώστηκαν είναι αρκετά καλά. Συγκεκριμένα, σύμφωνα με το ιστορικό των χρηστών, κατά μέσο όρο, ένα στα πέντε άρθρα που διαβάζουν οι χρήστες είναι από τη λίστα των προτάσεων. Με τον αλγόριθμο που υλοποιήθηκε λοιπόν, η λίστα των προτάσεων προσαρμοζόταν από τα πρώτα άρθρα που διάβασε ο εκάστοτε χρήστης, και οι προτάσεις ήταν προσωποποιημένες για αυτόν.

5.3 Μελλοντικές κατευθύνσεις

Υπάρχουν αρκετές μελλοντικές κατευθύνσεις για την επέκταση των τεχνικών που χρησιμοποιήθηκαν στην παρούσα διπλωματική εργασία, καθώς και για την βελτίωση του ΣΣΕΑ. Κάποιες από τις κατευθύνσεις επικεντρώνονται στη διαδικασία της συσταδοποίησης και κάποιες στον αλγόριθμο προτάσεων και στο ΣΣΕΑ. Όσον αφορά τη συσταδοποίηση οι μελλοντικές κατευθύνσεις είναι οι εξής:

- Δημιουργία και χρήση βάσεων γνώσεων για την προσθήκη επιπλέον πληροφοριών στα κείμενα για τις οντότητές τους. Η προσθήκη αυτή θα οδηγήσει σε καλύτερα αποτελέσματα στη συσταδοποίση. Ακόμα, η χρήση βάσεων γνώσης μπορεί να χρησιμοποιηθεί και για την αντιστοίχηση τις κάθε συστάδας σε μία από τις 7 γενικές κατηγορίες κατηγορίες (Ελλάδα, Υγεία, Κόσμος, Τεχνολογία, Τεχνολογία, Σπορ, Ψυχαγωγία), που τώρα γίνεται χειροκίνητα.

- Χρήση και εκπαίδευση ενός μοντέλου AE για τη μείωση των διαστάσεων των ενσωματώσεων των άρθρων, έτσι ώστε να γίνεται πιο γρήγορα η εκπαίδευση του ETM μοντέλου και να μειωθεί ο όγκος των δεδομένων στη βάση του ΣΣΕΑ.

- Βελτιστοποίηση των πολυγλωσσικών μοντέλων, για να προσαρμοστούν τα βάρη τους στα ειδησεογραφικά άρθρα.

- Εκπαίδευση ενός CTM μοντέλου με άρθρα που έχουν δημοσιευτεί σε διάστημα μεγαλύτερο της μίας ημέρας και χρήση του κομματιού της πρόβλεψης του μοντέλου για τη συσταδοποίηση καινούργιων άρθρων που παίρνει η ιστοσελίδα από τις ροές RSS. Αυτή τη στιγμή το μοντέλο εκπαιδεύεται κάθε φορά που προστίθενται καινούργια άρθρα στην ιστοσελίδα. Η εκπαίδευση δεν χρειάζεται να γίνεται κάθε μέρα, αλλά το χρονικό διάστημα μέχρι την επόμενη εκπαίδευση καθώς και τα πόσα άρθρα θα παίρνει ως είσοδο το μοντέλο για να κάνει σωστή συσταδοποίηση, χρειάζονται μελέτη.

Για το κομμάτι του αλγόριθμου προτάσεων καθώς και για το ΣΣΕΑ γενικότερα, οι μελλοντικές κατευθύνσεις είναι οι παρακάτω:

- Εφαρμογή DL μοντέλων για την εύρεση προτάσεων για τους χρήστες του ΣΣΕΑ. Η χρήση αυτών των αλγορίθμων δεν δοκιμάστηκε λόγω της έλλειψης κατάλληλης συλλογής δεδομένων στα ελληνικά. Με τα αρχικά δεδομένα που συλλέχθηκαν από την ιστοσελίδα μπορεί να γίνει μια αρχική σύγκριση των DL μεθόδων για τα ΣΣΕΑ που παρουσιάστηκαν στην Ενότητα 2.3.5.

- Για την αύξηση της αλληλεπίδρασης των χρηστών με την ιστοσελίδα του ΣΣΕΑ, θα μπορούσε να δοκιμαστεί και η μέθοδος των ενημερωτικών δελτίων. Για παράδειγμα, θα μπορούσε η καλύτερη πρόταση του ΣΣ να τους έρχεται σαν ειδοποίηση στο email τους.

- Συλλογή περισσότερων έμμεσων δεδομένων για τους χρήστες και χρήση τους στα DL μοντέλα.

- Εύρεση τρόπων μείωσης του χρόνου ανταπόκρισης της ιστοσελίδας. Όσο ο όγκος των άρθρων στη βάση αυξάνεται, η ανταπόκριση του ιστοτόπου γίνεται και πιο αργή. Έτσι, θα πρέπει να βρεθούν τρόποι να μειωθεί ο όγκος των δεδομένων που χρειάζονται για την πρόταση σωστών άρθρων, έτσι ώστε να μπορούν να διαγράφονται κάποια παλαιότερα άρθρα από τη βάση.

Βιβλιογραφία

[ae] Schematic structure of an autoencoder with 3 fully connected hidden layers. The code (z, or h for reference in the text) is the most internal layer. Available from `https://en.wikipedia.org/wiki/Autoencoder`.

[Auer07] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak and Zachary Ives, "DBpedia: A Nucleus for a Web of Open Data", in *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, p. 722–735, Berlin, Heidelberg, 2007, Springer-Verlag.

[Bara18] Remigiusz Baran, Andrzej Dziech and Andrzej Zeja, "A capable multimedia content discovery platform based on visual content analysis and intelligent data enrichment", *Multimedia Tools and Applications*, vol. 77, 06 2018.

[Bian21] Federico Bianchi, Silvia Terragni and Dirk Hovy, "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence", in *ACL*, 2021.

[Boja16] Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov, "Enriching Word Vectors with Subword Information", *arXiv preprint arXiv:1607.04606*, 2016.

[Bray98] Paoli J. Bray T. and Sperberg-McQueen C.M., "Extensible markup language (XML) 1.0", Available via the World Wide Web at `http://www.w3.org/TR/1998/REC-xml-19980210`, 1998.

[Burt18a] M. Burtsev, A. Seliverstov, R. Airapetyan, M. Arkhipov, Dilyara Baymurzina, E. Botvinovsky, Nickolay Bushkov, Olga Gureenkova, A. Kamenev, Vasily Konovalov, Yuri Kuratov, Denis Kuznetsov, A. Litinsky, A. Lymar, M. Petrov, Leonid Pugachev, A. Sorokin and M. Vikhreva, "DeepPavlov: An Open Source Library for Conversational AI", 2018.

[Burt18b] Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva and Marat Zaynutdinov, "DeepPavlov: Open-Source Library for Dialogue Systems", in *Proceedings of ACL 2018, System Demonstrations*, pp. 122–127, Melbourne, Australia, July 2018, Association for Computational Linguistics.

[Cant11] Iván Cantador, Pablo Castells and Alejandro Bellogín, "An Enhanced Semantic Layer for Hybrid Recommender Systems: Application to News Recommendation", *Int. J. Semant. Web Inf. Syst.*, vol. 7, no. 1, p. 44–78, January 2011.

[Chen11] Hung-Hsuan Chen, Liang Gou, Xiaolong Zhang and Clyde Lee Giles, "CollabSeer: A Search Engine for Collaboration Discovery", in *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL '11, p. 231–240, New York, NY, USA, 2011, Association for Computing Machinery.

[Chen15] Hung-Hsuan Chen, Alexander G. Ororbia II au2 and C. Lee Giles, "ExpertSeer: a Keyphrase Based Expert Recommender for Digital Libraries", 2015.

[Chen19] Hung-Hsuan Chen and Pu Chen, "Differentiating Regularization Weights – A Simple Mechanism to Alleviate Cold Start in Recommender Systems", *ACM Trans. Knowl. Discov. Data*, vol. 13, no. 1, January 2019.

[Chod13] K. Chodorow, "MongoDB: The Definitive Guide, O'Reilly Media, Inc.", 2013.

[cnn] A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets - Scientific Figure on ResearchGate. Available from: `https://www.researchgate.net/figure/Schematic-diagram-of-a-basic-convolutional-neural-network-CNN-architecture-26_fig1_336805909` [accessed 30 Sep, 2021].

[Desa14] Maunendra Sankar Desarkar and Neha Shinde, "Diversification in news recommendation for privacy concerned users", in *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 135–141, 2014.

[Devl19] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv 1810.04805, 2019.

[Dien19] Adji B Dieng, Francisco J R Ruiz and David M Blei, "Topic modeling in embedding spaces", *arXiv preprint arXiv:1907.04907*, 2019.

[Djan] Django Software Foundation, "Django".

[djon21] "Multi Database Enabled Backends - Djongo", `https://www.djongomapper.com/` [accessed 30 Sep, 2021], 2021.

[Dou16] Yingtong Dou, Hao Yang and Xiaolong Deng, "A Survey of Collaborative Filtering Algorithms for Social Recommender Systems", in *2016 12th International Conference on Semantics, Knowledge and Grids (SKG)*, pp. 40–46, 2016.

[Ekst14] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen and Joseph A. Konstan, "User Perception of Differences in Recommender Algorithms", in *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, p. 161–168, New York, NY, USA, 2014, Association for Computing Machinery.

[Elah14] Mehdi Elahi, Francesco Ricci and Neil Rubens, "Active Learning in Collaborative Filtering Recommender Systems", in Martin Hepp and Yigal Hoffner, editors, *E-Commerce and Web Technologies*, pp. 113–124, Cham, 2014, Springer International Publishing.

[Elah16] Mehdi Elahi, Francesco Ricci and Neil Rubens, "A survey of active learning in collaborative filtering recommender systems", *Computer Science Review*, vol. 20, pp. 29–50, 2016.

[euro21] "Eurobarometer: Internet users' preferences for accessing content online", `https://digital-strategy.ec.europa.eu/en/library/eurobarometer-internet-users-preferences-accessing-content-online` [Last Update: 9 March 2021], 2021.

[Felf07] Alexander Felfernig, Klaus Isak, Kalman Szabo and Peter Zachar, "The VITA Financial Services Sales Support Environment", in *Proceedings AAAI/IAAI 2007*, pp. 1692–1699, AAAI Press, 2007. AAAI-07 : AAAI Conference on Artificial Intelligence ; Conference date: 22-07-2007 Through 26-07-2007.

[Feng20] Chong Feng, Muzammil Khan, Arif Ur Rahman and Arshad Ahmad, "News Recommendation Systems - Accomplishments, Challenges & Future Directions", *IEEE Access*, vol. 8, pp. 16702–16725, 01 2020.

[Gaut15] Anjali Gautam, Tulika, Radhika Dhingra and Punam Bedi, "Use of NoSQL Database for Handling Semi Structured Data: An Empirical Study of News RSS Feeds", in N. R. Shetty, N.H. Prasad and N. Nalini, editors, *Emerging Research in Computing, Information, Communication and Applications*, pp. 253–263, New Delhi, 2015, Springer India.

[Gome16] Carlos A. Gomez-Uribe and Neil Hunt, "The Netflix Recommender System: Algorithms, Business Value, and Innovation", *ACM Trans. Manage. Inf. Syst.*, vol. 6, no. 4, December 2016.

[goog21] "Google News", `https://news.google.com/` [accessed 30 Sep, 2021], 2021.

[gru] Gated Recurrent Unit By fdeloche - Own work, CC BY-SA 4.0 `https://commons.wikimedia.org/w/index.php?curid=60466441` via Wikimedia Commons [accessed 30 Sep, 2021].

[Guo20] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong and Qing He, "A Survey on Knowledge Graph-Based Recommender Systems", 02 2020.

[Gupt13] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang and Reza Zadeh, "WTF: The Who to Follow Service at Twitter", in *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, p. 505–514, New York, NY, USA, 2013, Association for Computing Machinery.

[He17] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu and Tat-Seng Chua, "Neural Collaborative Filtering", 2017.

[Horn12] Kurt Hornik, Ingo Feinerer, Martin Kober and Christian Buchta, "Spherical k-Means Clustering", *Journal of Statistical Software, Articles*, vol. 50, no. 10, pp. 1–22, 2012.

[ipso20] "Google Media Literacy", `https://www.ipsos.com/sites/default/files/ct/news/documents/2020-10/google-media-literacy-tables-emea.pdf` [accessed 30 Sep, 2021], 2020.

[Kami16] Marius Kaminskas and Derek Bridge, "Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems", *ACM Trans. Interact. Intell. Syst.*, vol. 7, no. 1, December 2016.

[Kari18] Mozhgan Karimi, D. Jannach and Michael Jugovac, "News recommender systems - Survey and roads ahead", *Inf. Process. Manag.*, vol. 54, pp. 1203–1227, 2018.

[Koeh05] Philipp Koehn et al., "Europarl: A parallel corpus for statistical machine translation", in *MT summit*, vol. 5, pp. 79–86, Citeseer, 2005.

[Kout20] John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis and Ion Androutsopoulos, "GREEK-BERT: The Greeks visiting Sesame Street", *11th Hellenic Conference on Artificial Intelligence DOI: 10.1145/3411408.3411440*, Sep 2020.

[Lath10] Neal Lathia, Stephen Hailes, Licia Capra and Xavier Amatriain, "Temporal Diversity in Recommender Systems", in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, p. 210–217, New York, NY, USA, 2010, Association for Computing Machinery.

[Li14] Lei Li, Li Zheng, Fan Yang and Tao Li, "Modeling and Broadening Temporal User Interest in Personalized News Recommendation", *Expert Syst. Appl.*, vol. 41, no. 7, p. 3168–3177, June 2014.

[Li20] Jing Li, Aixin Sun, Jianglei Han and Chenliang Li, "A Survey on Deep Learning for Named Entity Recognition", arXiv 1812.09449, 2020.

[lstm] Long Short Term Memory By fdeloche - Own work, CC BY-SA 4.0 https://commons.wikimedia.org/w/index.php?curid=60149410 via Wikimedia Commons [accessed 30 Sep, 2021].

[Mele17] Ida Mele and Fabio Crestani, "Event Detection for Heterogeneous News Streams", in Flavius Frasincar, Ashwin Ittoo, Le Minh Nguyen and Elisabeth Métais, editors, *Natural Language Processing and Information Systems*, pp. 110–123, Cham, 2017, Springer International Publishing.

[Miko13] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space", *eprint arXiv:1301.3781*, 2013.

[mlp] An intermediate significant bit (ISB) watermarking technique using neural networks - Scientific Figure on ResearchGate. Available from https://www.researchgate.net/figure/The-Block-diagram-of-a-three-hidden-layer-multilayer-perceptron-MLP_fig14_303286711 [accessed 30 Sep, 2021].

[movi] "MovieLens", https://movielens.org [accessed 30 Sep, 2021].

[Newm20] Nic Newman, Richard Fletcher, Anne Schulz, Simge Andı and Rasmus Kleis Nielsen, "Reuters Institute Digital News Report 2020", *Reuters Institute for the Study of Journalism*, 2020.

[Newm21] Nic Newman, Richard Fletcher, Anne Schulz, Simge Andi, Craig T Robertson and Rasmus Kleis Nielsen, "Reuters Institute Digital News Report 2021", *Reuters Institute for the Study of Journalism*, 2021.

[Preo16] Daniel Preoţiuc-Pietro, P. K. Srijith, Mark Hepple and Trevor Cohn, "Studying the Temporal Dynamics of Word Co-occurrences: An Application to Event Detection", in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4380–4387, Portorož, Slovenia, May 2016, European Language Resources Association (ELRA).

[Raza21] Shaina Raza and Chen Ding, "News Recommender System: A review of recent progress, challenges, and opportunities", 2021.

[Reim19] Nils Reimers and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", 2019.

[Ricc10] Francesco Ricci, Lior Rokach and Bracha Shapira, "Recommender Systems Handbook", 10 2010.

[rl] A Machine Learning Approach for Power Allocation in HetNets Considering QoS - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Reinforcement-Learning-Agent-and-Environment_fig2_323867253 [accessed 30 Sep, 2021].

[Rube15] Neil Rubens, Mehdi Elahi, Masashi Sugiyama and Dain Kaplan, *Active Learning in Recommender Systems*, pp. 809–846, Springer US, Boston, MA, 2015.

[Sama19] Chris Samarinas and Stefanos Zafeiriou, "Personalized high quality news recommendations using word embeddings and text classification models", 06 2019.

[Samm10] Claude Sammut and Geoffrey I. Webb, editors, *TF–IDF*, pp. 986–987, Springer US, Boston, MA, 2010.

[Sara17] K. G. Saranya and G. Sudha Sadasivam, "Personalized News Article Recommendation with Novelty Using Collaborative Filtering Based Rough Set Theory", *Mob. Netw. Appl.*, vol. 22, no. 4, p. 719–729, August 2017.

[Sia20] Suzanna Sia, Ayush Dalmia and Sabrina J. Mielke, "Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!", *CoRR arXiv:2004.14914*, vol. abs/2004.14914, 2020.

[Sriv17] Akash Srivastava and Charles Sutton, "Autoencoding Variational Inference For Topic Models", arXiv 1703.01488, 2017.

[Su09] Xiaoyuan Su and Taghi M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques", *Adv. in Artif. Intell.*, vol. 2009, January 2009.

[Suar19] Pedro Javier Ortiz Suarez, Benoit Sagot and Laurent Romary, "Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures", Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pp. 9 – 16, Mannheim, 2019, Leibniz-Institut für Deutsche Sprache.

[Sudh20] Sudha-Nadchal and arvindpdmn, "Text Clustering", `https://devopedia.org/text-clustering`, 2020.

[tops20] "Τα 20 κορυφαία ενημερωτικά site του ελληνικού internet", `https://www.e-tetradio.gr/Article/22316/ta-20-koryfaia-enhmerwtika-site-toy-ellhnikoy-internet` [accessed 30 Sep, 2021], 2020.

[Vasw17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin, "Attention Is All You Need", 2017.

[Vian16] Paula Viana and Márcio Soares, "A Hybrid Recommendation System for News in a Mobile Environment", in *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*, WIMS '16, New York, NY, USA, 2016, Association for Computing Machinery.

[Wang18] Hongwei Wang, Fuzheng Zhang, Xing Xie and Minyi Guo, "DKN: Deep Knowledge-Aware Network for News Recommendation", 2018.

[Wiki] Wikipedia contributors, "Wikipedia, The Free Encyclopedia", `https://wikipedia.org` [accessed 30 Sep, 2021].

[Word10] "Princeton University "About WordNet."", `https://wordnet.princeton.edu/`. Princeton University, 2010.

[Zhan10] Yin Zhang, Rong Jin and Zhi-Hua Zhou, "Understanding bag-of-words model: A statistical framework", *International Journal of Machine Learning and Cybernetics*, vol. 1, pp. 43–52, 12 2010.

[Zhen13] Li Zheng, Lei Li, Wenxing Hong and Tao Li, "PENETRATE: Personalized news recommendation using ensemble hierarchical clustering", *Expert Systems with Applications*, vol. 40, p. 2127–2136, 05 2013.

[Zhen18] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie and Zhenhui Li, "DRN: A Deep Reinforcement Learning Framework for News Recommendation", in *Proceedings of the 2018 World Wide Web Conference*, WWW '18, p. 167–176, Republic and Canton of Geneva, CHE, 2018, International World Wide Web Conferences Steering Committee.

Παράρτημα Α

Ευρετήριο Όρων και Συντμήσεων

Α.1 Ελληνικοί Όροι

ΕΦΓ: Επεξεργασία Φυσικής Γλώσσας
ΣΣ: Συστήματα Συστάσεων
ΣΣΕΑ: Συστήματα Συστάσεων Ειδησεογραφικών Άρθρων

Α.2 Αγγλικοί Όροι

AE: Autoencoder
BOW:Bag-of-words
CB: Content Based
CBOW:Continuous bag-of-words
CF: Collaborative Filtering
CLS: Classification token
CNN: Convolutional Neural Network
DL: Deep Learning
EDT: Event Detection and Tracking
ETM: Embedding Topic Modeling
EMEA: Europe, Middle East and Asia
GRU: Gate Recurrent Unit
HIN: Heterogenous Information Network
ILS: Intra-List Similarity
IR: Information Retrieval
KB: Knowledge Based
KGE: Knowledge Graph Embeddings
LDA: Latent Dirichlet Allocation
LM: Language Model
LSTM: Long-Short-Term Memory
MLM: Masked Language Model
MLP:Multi-Layer Perceptron
NA: News Aggregator
NLP: Natural Language Processing
NPMI: Normalized Pointwise Mutual Information
PCA: Principal Component Analysis
PMI: Pointwise Mutual Information
ReLU: Rectifier Linear Unit
RL: Reinforcement Learning
RNN: Recurrent Neural Networks
RS: Recommender Systems
TDT: Topic Detection and Tracking

TF: Term Frequency
TF-DF: Term Frequency-Document Frequency
TF-IDF: Term Frequency-Inverse Document Frequency
TM: Topic Model
UMAP: Uniform Manifold Approximation and Projection
VAE: Variational Autoencoder