ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εγγενής Ανίχνευση Λογοκλοπής Κειμένου με Ευφυείς Τεχνικές

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΑΝΔΡΙΑΝΝΑΣ ΠΟΛΥΔΟΥΡΗ

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

ΕΡΓΑΣΤΗΡΙΟ ΕΥΦΥΩΝ ΣΥΣΤΗΜΑΤΩΝ
Αθήνα, Σεπτέμβριος 2016

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Εγγενής Ανίχνευση Λογοκλοπής Κειμένου με Ευφυείς Τεχνικές

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΑΝΔΡΙΑΝΝΑΣ ΠΟΛΥΔΟΥΡΗ

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

Εγκρίθηκε από την τριμελή επιτροπή την 10η Σεπτέμβρη 2016.

.....	.....	.....
Ανδρέας-Γεώργιος Σταφυλοπάτης	Στέφανος Κόλλιας	Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Επίκουρος Καθηγητής Ε.Μ.Π

Αθήνα, Σεπτέμβριος 2016

.....

ΑΝΔΡΙΑΝΝΑ ΠΟΛΥΔΟΥΡΗ
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Η/Υ Ε.Μ.Π

Περίληψη

Στην ακαδημαϊκή κοινότητα με τον όρο *λογοκλοπή* εννοούμε την παρουσίαση δουλειάς τρίτου ως προσωπικής, ελλείψει κατάλληλης αναφοράς στην πηγή ή/και γνωστοποίησης στον συγγραφέα. Στις μέρες μας, όπου η ερευνητική δραστηριότητα αξιολογείται (και) με όρους παραγωγικότητας, ενώ, ταυτόχρονα, το διαδίκτυο προσφέρει εύκολη πρόσβαση σε αμέτρητα ερευνητικά έργα, η λογοκλοπή αποτελεί ένα φαινόμενο με ολοένα αυξανόμενη συχνότητα που πλήττει την ερευνητική ακεραιότητα και αξιοπιστία.

Η έρευνα γύρω από την ανίχνευση λογοκλοπής χωρίζεται σε δύο κατευθύνσεις: εξωγενή και εγγενή. Κατά την εξωγενή ανίχνευση θεωρείται ένα εξωτερικό σώμα αναφορών, όπου αναζητούνται οι ομοιότητες με το υπό εξέταση κείμενο. Κατά την εγγενή ανίχνευση, με μόνο το υπό εξέταση κείμενο ως πηγή πληροφορίας, αναζητούνται τα λογοκλεμμένα, σε αυτό, χωρία, με εργαλείο τη στυλιστική ανάλυση του κειμένου.

Κατά την εργασία αυτή κατασκευάστηκε ένα σύστημα εγγενούς ανίχνευσης λογοκλοπής, το οποίο αναπτύχθηκε, κυρίως, σε Java. Κύρια μέρη του συστήματος είναι: η στυλιστική ανάλυση των κειμένων, όπου χρησιμοποιήθηκαν τόσο γνωστά όσο και πρωτότυπα στυλομετρικά και σημασιολογικά χαρακτηριστικά, και ένα μοντέλο μηχανικής μάθησης για την εξαγωγή των ύποπτων χωρίων. Κατά τη στυλιστική ανάλυση χρησιμοποιήθηκε η Java βιβλιοθήκη OpenNLP της Apache. Κατά τη μηχανική μάθηση χρησιμοποιήθηκε η Python βιβλιοθήκη Scikit-Learn. Πειραματιστήκαμε με 4 διαφορετικούς αλγορίθμους εκμάθησης *(Naive Bayes, Μηχανές Διανυσμάτων Υποστήριξης, Δέντρα Απόφασης, Perceptron πολλών-στρωμάτων)*.

Ακόμη, εισήχθη, για πρώτη φορά, η ανισορροπία των δεδομένων εκμάθησης ως παράμετρος του προβλήματος. Χρησιμοποιώντας τη Github repository *Unbalanced Dataset*, η οποία προϋποθέτει την εργαλειοθήκη Scikit-Learn, πειραματιστήκαμε με 2 αλγορίθμους εξισορρόπησης *(simple SMOTE, borderline SMOTE)*.

Ως σώμα δεδομένων χρησιμοποιήθηκε αυτό του διαδικτυακού διαγωνισμού για εγγενή ανίχνευση λογοκλοπής PAN@CLEF 2011, ενώ τα αποτελέσματα συγκρίνονται (και) με αυτά των διαγωνιζόμενων συστημάτων.

Λέξεις-κλειδιά: εγγενής ανίχνευση λογοκλοπής, στυλομετρία, PAN 2011, Scikit-Learn, Apache OpenNLP, επιβλεπόμενη μάθηση, εξισορρόπηση δεδομένων εκμάθησης, SMOTE

Abstract

In the academic society the term *plagiarism* refers to the presentation of someone else's work as one's own, without proper citation and/or acknowledgment of the original author. Nowadays, that success in academic research is, as well, a matter of productivity and that worldwide web is an easily accessible, endless information source, plagiarism arises as a fast growing problem that harms research integrity and credibility.

Research for plagiarism detection involves two different approaches: extrinsic and intrinsic. In terms of extrinsic detection, a suspicious document is compared to a collection of reference documents. In terms of intrinsic detection, no reference corpus is provided and the detection of the plagiarised passages is based on the stylistic changes or inconsistencies within the document.

In this thesis, an intrinsic plagiarism detection system is constructed, which is, mostly, developed in Java programming language. Main parts of this system are the stylistic analysis of the documents - where widely known stylometrics and semantics features are used, as well as novel ones -, and a machine learning model for the extraction of the plagiarised passages. For the stylistic analysis the Java library OpenNLP of Apache is used. For the machine learning model the Python library Scikit-Learn is used. We run experiments for 4 different learning algorithms *(Naive Bayes, Support Vector Machines, Decision Trees, Multilayer Perceptron)*.

In addition, the fact of unbalanced training data is, for the first time, considered as one of the parameters of the intrinsic plagiarism detection problem. Using the Github repository *Unbalanced Dataset* - which requires the Scikit-Learn toolkit -, we experimented with 2 balancing algorithms (simple SMOTE, borderline SMOTE).

The data corpus of the PAN 2011 evaluation lab for intrinsic plagiarism detection is employed, while the results of the detection system are compared to those that took part in the evaluation lab.

*Keywords:*Intrinsic plagiarism detection (IPD), stylometry, PAN 2011, Scikit-Learn, Apache OpenNLP, supervised learning, training data balancing, SMOTE

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κ. Σταφυλοπάτη για την εμπιστοσύνη που μου έδειξε να αναλάβω αυτή τη διπλωματική. Επίσης, ευχαριστώ στον Γιώργο Σιόλα για την καθοδήγηση και την υποστήριξη καθ' όλη τη διάρκεια της εργασίας. Τέλος, ευχαριστώ τον πολύ καλό φίλο Παναγιώτη για τις πολύ γόνιμες συζητήσεις μας πάνω στο πρόβλημα αλλά και γενικά.

Περιεχόμενα

Κατάλογος πινάκων

Κεφάλαιο 1

Περιγραφή του Προβλήματος

1.1 Εισαγωγή

Plagiarism είναι ο διεθνής όρος του γενικότερου προβλήματος στο οποίο αφορά η εργασία αυτή. Στα ελληνικά ο όρος αυτός αποδίδεται ως *λογοκλοπή*. Η ίδια η λέξη plagiarism προέρχεται από τη λατινική *plagium*, που σημαίνει απαγωγή *(kidnapping)*. Στην κυριολεξία σημαίνει κλοπή, παρουσίαση υλικού που προέρχεται από κάποιον ως δουλειά άλλου [2].

Η ανίχνευση λογοκλοπής *(plagiarism detection)* ως κλάδος που αφορούσε στη φυσική γλώσσα έκανε τα πρώτα της βήματα τη δεκαετία του ’90. Πριν από την ανίχνευση λογοκλοπής σε φυσική γλώσσα, κατασκευάζονταν και ερευνώνταν, ήδη, συστήματα ανίχνευσης αντιγραφής προγραμματιστικού κώδικα αλλά και κακής χρήσης λογισμικού(π.χ. απαγορευμένη μεταπώληση λογισμικού προϊόντος), αρχής γενομένης με τη μελέτη για ανίχνευση λογοκλοπής μεταξύ φοιτητών σε προγραμματιστικό κώδικα Pascal και C τη δεκαετία του ’70 [3] [4]. Πλέον η ανίχνευση λογοκλοπής σε φυσική γλώσσα έχει εξελιχθεί σημαντικά και χρησιμοποιεί γνώσεις και εργαλεία από άλλα σχετικά πεδία, όπως ανάκτηση πληροφορίας *(Information Retrieval)*, επεξεργασία φυσικής γλώσσας *(Natural Language Processing)*, υπολογιστική γλωσσολογία *(Computational Linguistics)* κ.ά. [5].

Στην ακαδημαϊκή κοινότητα *λογοκλοπή* θεωρείται η παρουσίαση δουλειάς κάποιου (μπορεί να συμπεριλαμβάνονται προφορικό ή γραπτό κείμενο, δεδομένα, ιδέες) ως προσωπικής, ελλείψει κατάλληλης αναφοράς στην πηγή ή/και γνωστοποίησης στον συγγραφέα. Η δυσκολία σαφέστερου ορισμού του φαινομένου της λογοκλοπής εμποδίζει την δημιουργία ενός ενιαίου, διεθνώς αποδεκτού συνόλου κανόνων για την αποφυγή του αλλά και για τον καθορισμό ενός πλαισίου, πέρα από το οποίο ένα έργο θα θεωρείται προϊόν λογοκλοπής. Η δυσκολία, δε, εξαγωγής ενός σαφέστερου ορισμού έγκειται κυρίως στη ρευστότητα της έννοιας της ”κοινής γνώσης” -για την οποία δεν είναι απαραίτητες οι βιβλιογραφικές αναφορές- ανάλογα με την εξειδίκευση του θέματος, τον κλάδο κ.ά. , αλλά και στην αδυναμία καταγραφής μιας πλήρους περιπτωσιολογίας για την ”παρουσίαση δουλειάς άλλου” σε αντιστοιχία με την ”κατάλληλη αναφορά” σε αυτήν που θα πρέπει να την συνο-

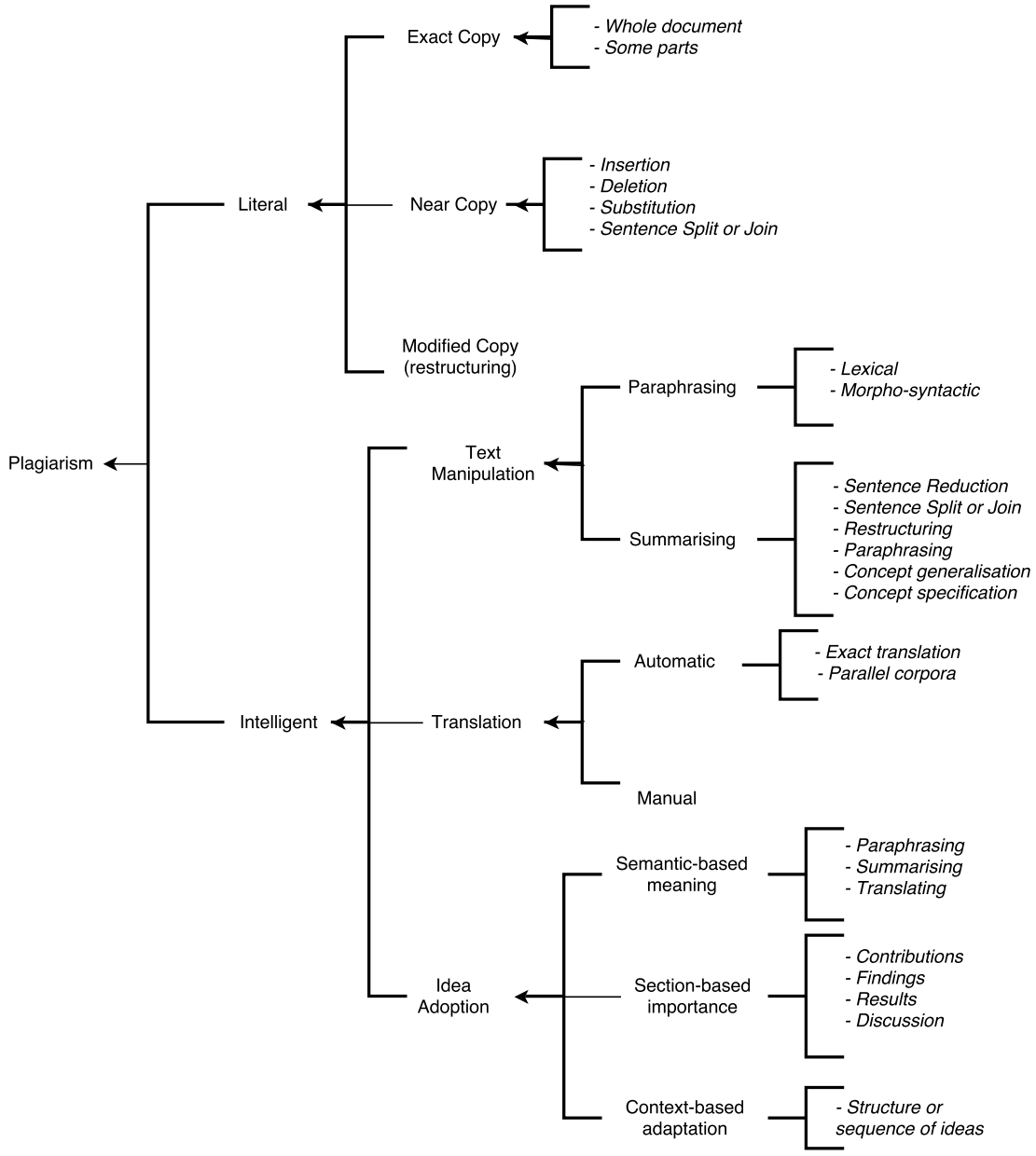
δεύει. Έτσι, ένα αμφιλεγόμενο έργο εξετάζεται πάντα ως ξεχωριστή περίπτωση. Παρόλαυτά, στις μέρες μας, όπου η ερευνητική δραστηριότητα αξιολογείται (και) με όρους παραγωγικότητας, θεωρείται ένα πρόβλημα που με ολοένα αυξανόμενη συχνότητα πλήττει βαθιά την ακαδημαϊκή και ερευνητική ακεραιότητα και αξιοπιστία, ενώ, ταυτόχρονα, το στίγμα του λογοκλόπου αποτελεί, ίσως, τον βαρύτερο αφορισμό για έναν ερευνητή.

1.2 Plagiarism types - Μία απόπειρα ταξινόμησης

Προκειμένου να λύσει κανείς ένα πρόβλημα, θα πρέπει πρώτα να έχει εξοικειωθεί καλά με τη φύση του προβλήματος. Η λογοκλοπή μπορεί να είναι εσκεμμένη ή και όχι. Μπορεί να είναι καλοδουλεμένη ή στεγνή αντιγραφή. Μπορεί να περιλαμβάνει μεταφρασμένα κομμάτια κειμένου. Δεν υπάρχει σύστημα, τουλάχιστον ακόμα, που να είναι σε θέση να ανιχνεύσει το ίδιο καλά λογοκλοπή σε κάθε έκφανσή της. Σε κάθε περίπτωση, για λόγους αποδοτικότητας αλλά και αποτελεσματικότητας, το σχεδιασμό ενός συστήματος ανίχνευσης λογοκλοπής θα πρέπει να συνοδεύει μια καθαρή στόχευση σε συγκεκριμένο τύπο λογοκλοπής ή, τουλάχιστον, σε συναφείς μεταξύ τους τύπους. Η κατηγοριοποίηση των πιθανών τύπων λογοκλοπής είναι σημαντική, μεταξύ άλλων, διότι μας διευκολύνει στο σχεδιασμό νέων μετρικών και εργαλείων, περισσότερο εξειδικευμένων και αποτελεσματικών σε κάθε τύπο ξεχωριστά.

Η ίδια η δουλειά της κατηγοριοποίησης δεν είναι εύκολη, ωστόσο έχουν γίνει αρκετές απόπειρες με διαφορετικές προσεγγίσεις. Μια ενδιαφέρουσα προσέγγιση, που στοχεύει στην κατανόηση της πηγής του προβλήματος, ακολουθείται από τους Alzahrani *et al* [5], όπου βασικό κριτήριο είναι η τακτική/ες που χρησιμοποιεί κάποιος που διαπράττει λογοκλοπή. Με αυτόν τον τρόπο αναδεικνύεται το φάσμα της δυσκολίας στο οποίο μπορεί να κινείται το πρόβλημα της ανίχνευσης. Για την εν λόγω ταξινόμηση συγκεντρώθηκαν δεδομένα από συνεντεύξεις σε ακαδημαϊκό προσωπικό με 10-20 χρόνια προϋπηρεσία διδασκαλίας, όπου οι ερωτήσεις αφορούσαν κυρίως τις τακτικές λογοκλοπής που υιοθετούσαν οι φοιτητές [5]. Σε αυτήν, η λογοκλοπή χωρίζεται σε δύο μεγάλες κατηγορίες: την ”κατά λέξη” *(literal)* και την ”έξυπνη” *(intelligent)* λογοκλοπή. Στην ”κατά λέξη λογοκλοπή” περιλαμβάνονται οι περιπτώσεις, όπου δεν καταβάλλεται ιδιαίτερη προσπάθεια για τη ”συγκάλυψη” της λογοκλοπής. Για παράδειγμα η απλή αντιγραφή-επικόλληση από το διαδίκτυο. Στην ”έξυπνη λογοκλοπή”, από την άλλη πλευρά, έχουμε τις πιο σοβαρές περιπτώσεις, όπου ο λογοκλόπος προσπαθεί να εξαπατήσει και να διαμορφώσει το λογοκλεμμένο κομμάτι έτσι ώστε να φαίνεται σαν δική του δουλειά.
Η ταξινόμηση αυτή δίνεται στο Σχήμα 1.1.

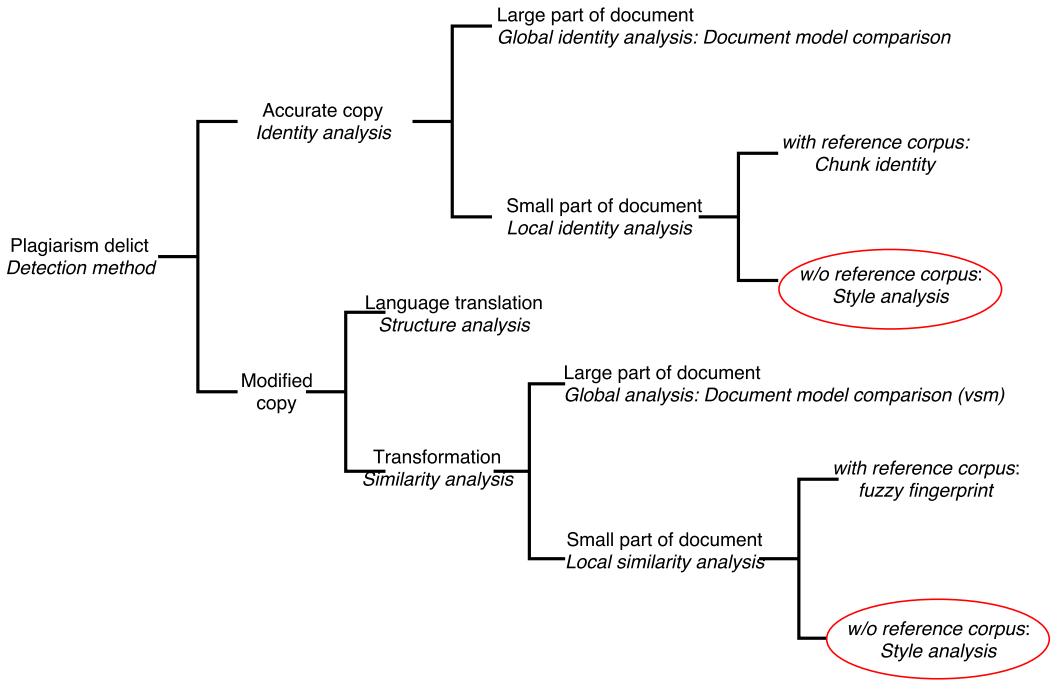
Μια άλλη προσέγγιση ακολουθείται από Z.Ceska *et al* [6] με άξονα τις κοινές αρχές των μεθόδων ανίχνευσης. Στην ίδια λογική κινείται και η πιο απλή αλλά



Σχήμα 1.1 Ταξινόμηση λογοκλοπής

σαφής ταξινόμηση που έκαναν οι Sven Meyer zu Eissen και Benno Stein [7], η οποία φαίνεται στο Σχήμα 1.2 και θα χρησιμοποιήσουμε ως σημείο αναφοράς για την κατηγοριοποίηση του προβλήματος με το οποίο ασχολείται αυτή η εργασία.

Τα κυκλωμένα τμήματα στο παραπάνω σχήμα υποδεικνύουν το ευρύτερο αντικείμενο της εργασίας αυτής. Θεωρούμε ότι τα λογοκλεμμένα χωρία, αν υπάρχουν



Σχήμα 1.2 Μία ταξινόμηση πράξεων λογοκλοπής και μεθόδων ανάλυσης.

τέτοια, στα κείμενα που εξετάζει το σύστημά μας, αποτελούν ένα μικρό κομμάτι του συνολικού κειμένου, ενώ για την ανίχνευσή τους επιστρατεύουμε μεθόδους στυλιστικής ανάλυσης.

Στα επόμενα θα αναφερθούμε με περισσότερες λεπτομέρειες στο σώμα δεδομένων που χρησιμοποιήθηκε.

1.3 Εξωγενής και Εγγενής Ανίχνευση Λογοκλοπής

Η αυτόματη ανίχνευση λογοκλοπής κειμένου χωρίζεται σε δύο μεγάλες κατηγορίες-προσεγγίσεις επίλυσης: την εξωγενή και την εγγενή ανίχνευση λογοκλοπής *(extrinsic and intrinsic plagiarism detection)*. Ας δούμε πώς περιγράφονται με φυσική γλώσσα καθένα από αυτά τα δύο προβλήματα.

Εξωγενής ανίχνευση λογοκλοπής: δεδομένων ενός ύποπτου (προς εξέταση) κειμένου και ενός βιβλιογραφικού σώματος, στόχος είναι να βρεθούν όλα τα κομμάτια του ύποπτου κειμένου που έχουν ως πηγή κάποιο από τα κείμενα της βιβλιογραφίας.
Εκ πρώτης όψεως η εξωγενής ανίχνευση μπορεί να φαίνεται σαν ένα απλό πρό-

βλημα ταιριάσματος *(string matching)* , στην πραγματικότητα, όμως, δεδομένου του όγκου της βιβλιογραφίας (κάποιες φορές χρησιμοποιείται ως πηγή, για παράδειγμα, ό,τι παρέχεται από τον παγκόσμιο ιστό), σίγουρα χρειαζόμαστε πιο "έξυπνες" μεθόδους. Εξάλλου, αυτός είναι ένας από τους βασικούς λόγους που η έρευνα στράφηκε προς την αναζήτηση εγγενών μεθόδων ανίχνευσης λογοκλοπής · ότι δηλαδή το σώμα ψηφιακών κειμένων αναφοράς τείνει να είναι τόσο μεγάλο όσος ο ίδιος ο παγκόσμιος ιστός.

Εγγενής ανίχνευση λογοκλοπής: δεδομένου ενός ύποπτου κειμένου και μόνο, στόχος είναι η ανίχνευση λογοκλεμμένων τμημάτων, εντοπίζοντας αλλαγές στα στυλιστικά χαρακτηριστικά γραφής.
Για παράδειγμα, αν ένας μαθητής είχε βάλει την αδερφή του να γράψει τον επίλογο στην έκθεση που είχε ως εργασία για το σπίτι, τότε ένα σύστημα εγγενούς ανίχνευσης θα έπρεπε να είναι σε θέση να εντοπίσει την τελευταία παράγραφο ως προϊόν λογοκλοπής.

Ας διατυπώσουμε τώρα έναν πιο αυστηρό ορισμό, του γενικότερου προβλήματος *ανίχνευσης λογοκλοπής*, και έπειτα να δούμε πώς, με βάση αυτόν, το πρόβλημα προσεγγίζεται από τις εξωγενείς και εγγενείς μεθόδους.

Ανίχνευση λογοκλοπής
Έστω $s = \langle s_{plg}, d_{plg}, s_{src}, d_{src} \rangle$ ένα στιγμιότυπο λογοκλοπής όπου s_{plg} είναι ένα χωρίο *(passage)* του κειμένου d_{plg} και, ταυτόχρονα, η λογοκλεμμένη εκδοχή ενός χωρίου s_{src} που βρίσκεται στο κείμενο d_{src}. Δεδομένου του d_{plg}, η δουλειά ενός ανιχνευτή λογοκλοπής είναι να εντοπίσει το s εξάγοντας ένα αντίστοιχο στιγμιότυπο λογοκλοπής $r = \langle r_{plg}, d_{plg}, r_{src}, d'_{src} \rangle$. Λέμε ότι το r ανιχνεύει το s αν και μόνο αν $s_{plg} \cap r_{plg} \neq \emptyset$, $s_{src} \cap r_{src} \neq \emptyset$ και $d_{src} = d'_{src}$.

Οι αλγόριθμοι στο πεδίο της *εξωγενούς ανίχνευσης* προσπαθούν να εντοπίσουν το στιγμιότυπο λογοκλοπής s, εντοπίζοντας το κείμενο d_{src} από μια συλλογή κειμένων-αρχείων D (π.χ το διαδίκτυο) και εξάγοντας τα s_{src} και s_{plg} από τα d_{src} και d_{plg} αντίστοιχα, με λεπτομερειακή σύγκριση των δύο κειμένων. Από την άλλη πλευρά, οι αλγόριθμοι στο πεδίο της *εγγενούς ανίχνευσης* προσπαθούν να ανιχνεύσουν το στιγμιότυπο s χωρίς να διαθέτουν ή να προσπαθούν να εντοπίσουν την πηγή, το κείμενο, δηλαδή, d_{src}, αλλά αναλύοντας το στυλιστικό χαρακτήρα γραφής του κειμένου d_{plg}. Η ιδέα είναι ότι σημαντικές στυλιστικές αλλαγές από το ένα εδάφιο στο άλλο, πιθανότατα υποδεικνύουν ότι το s_{plg} έχει γραφεί από άλλον συγγραφέα, σε σχέση με το υπόλοιπο d_{plg} αρχείο [8].

Φαίνεται τώρα καθαρά η διαφορετική προσέγγιση στο πρόβλημα μεταξύ των εγγενών και εξωγενών μεθόδων. Με δυο λόγια, στη μεν εξωγενή προσέγγιση εξετάζονται ύποπτες *ομοιότητες* μεταξύ κειμένων, ενώ, από τη άλλη, στην εγγενή προσέγγιση εξετάζονται ύποπτες *ανομοιομορφίες* μέσα στο ίδιο κείμενο. Σε κάθε περίπτωση όμως, η εσκεμμένη επεξεργασία του λογοκλεμμένου χωρίου αλλάζει άρδην τα δεδομένα του προβλήματος και ανεβάζει κατά πολλά επίπεδα τη δυσκολία της ανίχνευσης. Σε αυτές τις περιπτώσεις, όσο καλύτερη,ή αλλιώς "εξυπνότερη" η επε-

ξεργασία, η εξωγενής μέθοδος θα δυσκολεύεται να εντοπίσει ως όμοια τα passages s_{src} και s_{plg}, ενώ η εγγενής μέθοδος θα δυσκολεύεται, από την άλλη, να εντοπίσει ως πιθανότατα γραμμένα από διαφορετικό συγγραφέα τα λογοκλεμμένα χωρία σε σχέση με το υπόλοιπο κείμενο.

1.4 Αντικείμενο της Εργασίας

Η παρούσα εργασία εμπίπτει στον κλάδο της ανίχνευσης λογοκλοπής με εγγενείς μεθόδους ή αλλιώς, χωρίς τη χρήση ή αναζήτηση κειμένων αναφοράς. Στηριζόμαστε σε εγγενείς πληροφορίες του υπό εξέταση γραπτού κειμένου, όπως το στυλ γραφής *(writing style)* και σημασιολογικά χαρακτηριστικά *(semantics)*. Γενικότερα, αναζητούνται και επιστρατεύονται μέσα ποσοτικοποίησης των χαρακτηριστικών ενός κειμένου που υποδεικνύουν και σκιαγραφούν τη μοναδικότητα της γραφής ενός συγγραφέα. Γι' αυτό το σκοπό, το υπό εξέταση κείμενο χωρίζεται σε "φυσικά" τμήματα, τα οποία μπορεί να είναι προτάσεις, παράγραφοι ή κομμάτια μεγαλύτερου μεγέθους. Για κάθε τέτοιο τμήμα εξάγονται τα χαρακτηριστικά γραφής που έχουν επιλεγεί και αναλύονται με διάφορες μεθόδους οι τιμές που προκύπτουν, έτσι ώστε να ανακαλυφθούν τμήματα, όπου η μοναδικότητα της γραφής πιθανότατα διαρρηγνύεται.

Όπως και σε κάθε πρόβλημα ανάκτησης πληροφορίας, έτσι και εδώ έπρεπε αρχικά να βρούμε ένα κατάλληλο σώμα δεδομένων *(data corpus)*. Βρήκαμε αυτό που αναζητούσαμε στην ιστοσελίδα του PAN [9]. Στην ιστοσελίδα αυτή βρίσκει κανείς επιστημονικές δράσεις και δουλειά που αφορούν στην απάτη ψηφιακών κειμένων γενικότερα, ή, καλύτερα, την ανίχνευση αυτής. Κάθε χρόνο διοργανώνεται ένας διαγωνισμός με διάφορες θεματικές σχετιζόμενες με τον κλάδο αυτό, ενώ τα αποτελέσματα που προκύπτουν παρουσιάζονται στο συνέδριο CLEF *(Conference and Labs of the Evaluation Forum)* του αντίστοιχου έτους. Αποτελεί πάγια τακτική των διοργανωτών του διαγωνισμού και διαχειριστών της ιστοσελίδας να ανεβάζουν τα δεδομένα καθώς και τα αποτελέσματα από κάθε διαγωνισμό, στο πλαίσιο της ελεύθερης πρόσβασης και διάδοσης της γνώσης.
Στον διαγωνισμό που διεξήχθη το 2011 υπήρχε, μεταξύ άλλων, ένα πεδίο για εγγενή ανίχνευση λογοκλοπής, από όπου και προμηθευτήκαμε το σώμα δεδομένων, αποτελούμενο από ένα σύνολο 4753 κειμένων σε μορφή αρχείου *text*, ενώ το κάθε κείμενο συνοδεύεται από ένα αρχείο *xml*, όπου περιλαμβάνονται οι "λύσεις". Στο xml αρχείο, δηλαδή, περιέχονται οι "συντεταγμένες" των λογοκλεμμένων χωρίων που περιλαμβάνει το αντίστοιχο κείμενο, εάν υπάρχουν, καθώς και ο τύπος της επεξεργασίας που έχει εφαρμοστεί κάθε φορά. Μαζί με το αναγκαίο σώμα δεδομένων για την εργασία μας, ο διαγωνισμός του PAN μας προμήθευσε και με ένα σημείο αναφοράς για την αξιολόγηση της αποτελεσματικότητας του συστήματος ανίχνευσης που κατασκευάσαμε, που δεν είναι άλλο από τα ίδια τα αποτελέσματα του διαγωνισμού.

1.5 Συναφή Προβλήματα

1.5.1 Βανδαλισμός Βικιπαιδείας

Παρόμοιας φύσεως πρόβλημα είναι ο *βανδαλισμός* της διαδικτυακής, ανοιχτής εγκυκλοπαίδειας Βικιπαιδεία. Ως *βανδαλισμός* ορίζεται κάθε πρόσθεση, αφαίρεση, ή αλλαγή περιεχομένου εδαφίου στη Βικιπαιδεία με σκοπό την αλλοίωση της παρεχόμενης πληροφορίας. Τέτοιες ενέργειες μπορεί να είναι η προσθήκη άσχετου, προσβλητικού υλικού, η άσκοπα κενές σελίδες, η εισαγωγή προφανών ασυναρτησιών. Σκοπός είναι η αυτόματη αποκατάσταση των άρθρων από ενέργειες βανδαλισμού. Δηλαδή η αυτόματη ανίχνευση και απομάκρυνση των παρεμβάσεων στο άρθρο, οι οποίες αποτελούν βανδαλισμό. Σε αυτό το πρόβλημα συμπεριλαμβάνονται και παραποιήσεις ή προσθήκη άσχετου οπτικοακουστικού υλικού ή/και υπερσυνδέσμων *(hyperlinks)* αλλά αυτές οι περιπτώσεις δεν σχετίζονται με το πρόβλημά μας και δεν αναφερόμαστε σε αυτές.
Η σχέση του αυτού του προβλήματος με την εγγενή ανίχνευση λογοκλοπής είναι εμφανής· και στα δύο προβλήματα επιχειρούμε, δοσμένου ενός κειμένου, να εντοπίσουμε και να εξάγουμε τα χωρία εκείνα που δεν συνάδουν με το υπόλοιπο κείμενο, βάσει ορισμένων κριτηρίων. Αυτό που καθιστά τα δύο προβλήματα εντελώς διαφορετικά ως προς την προσέγγιση της επίλυσης και τις ίδιες τις μεθόδους επίλυσης, είναι η διαφορετική αφετηρία, τα διαφορετικά δεδομένα που παρέχονται εξαρχής σε καθεμία από τις δύο περιπτώσεις. Στην περίπτωση της εγγενούς ανίχνευσης λογοκλοπής είδαμε πως δεν έχουμε καμιά άλλη πηγή ανάκτησης πληροφορίας, παρά το ίδιο το υπό εξέταση κείμενο, το οποίο παραλαμβάνουμε ως ενιαίο έργο χωρίς υποδείξεις ή σαφώς οριοθετημένα χωρία για τα οποία θα πρέπει να αποφανθούμε αν είναι προϊόντα λογοκλοπής ή όχι. Στην περίπτωση του βανδαλισμού της βικιπαιδείας, από την άλλη, γνωρίζουμε ακριβώς σε ποιο χωρίο θα εστιάσουμε, μιας και η επεξεργασία του κειμένου γίνεται από τους χρήστες διαδικτυακά, οπότε μαζί με κάθε αποθήκευση του επεξεργασμένου άρθρου μπορούμε να γνωρίζουμε ποιο κομμάτι προστέθηκε, διεγράφη ή πέρασε από επεξεργασία. Με δυο λόγια, αναφερόμαστε πλέον σε *ύποπτα χωρία*, για τα οποία καλούμαστε να αποφανθούμε αντλώντας πληροφορίες από το υπόλοιπο κείμενο, το οποίο θεωρείται υγιές. Περάσαμε λοιπόν από το *ύποπτο κείμενο* στο *υγιές κείμενο με ύποπτα χωρία*, κάτι που κάνει φανερή την αυξημένη, συγκριτικά, δυσκολία της εγγενούς ανίχνευσης.

1.5.2 Επαλήθευση συγγραφέα

Το πρόβλημα της εγγενούς ανίχνευσης λογοκλοπής σχετίζεται στενά με αυτό της επαλήθευσης συγγραφέα *(authorship verification)*. Στο πλαίσιο αυτού του προβλήματος δίνονται ως δεδομένα κάποια δείγματα γραφής ενός συγγραφέα *A* και ζητείται να αποφασιστεί αν ένα κείμενο αμφίβολης προελεύσεως είναι γραμμένο ή όχι από τον *A*. Για μεγαλύτερη σαφήνεια, διατυπώνουμε το πρόβλημα σε μορφή απόφασης [10].

Επαλήθευση Συγγραφέα
Δεδομένο.Κείμενο *d* πιθανώς γραμμένο από το συγγραφέα *A*.
Ερώτηση.Υπάρχει τμήμα του *d* που να είναι γραμμένο από συγγραφέα $B \neq A$;

Φαίνεται λοιπόν πως αυτό το πρόβλημα αποτελεί ειδική περίπτωση της ανίχνευσης λογοκλοπής με εγγενείς μεθόδους. Στο πρώτο επιχειρείται ή "αποκωδικοποίηση", κατά κάποιο τρόπο, της ιδιαιτερότητας της γραφής ενός συγγραφέα *A* μέσω της ανάλυσης κειμένων που ανήκουν στον *A*, καθώς και η αποτύπωση αυτής της ιδιαιτερότητας σε συγκεκριμένα χαρακτηριστικά, έτσι ώστε τελικά να μπορεί να αναγνωριστεί ως δικό του ένα υπό εξέταση κείμενο ή να απορριφθεί. Πιο γενικό φαίνεται το πρόβλημα της εγγενούς ανίχνευσης, όπου χωρίς περαιτέρω πληροφορίες για τον συγγραφέα, παρεκτός το ίδιο το υπό εξέταση κείμενο, επιχειρείται η αναγνώριση της μοναδικότητας ή ομοιομορφίας της γραφής του αλλά και η ανίχνευση των τμημάτων εκείνων όπου αυτή η ομοιομορφία καταργείται. Ακόμα και στην πιο απλή εκδοχή του προβλήματος, όπου το λογοκλεμμένο χωρίο δεν έχει υποστεί επεξεργασία, τα λογοκλεμμένα χωρία κειμένου προσθέτουν δυσκολία στην ανίχνευση της μοναδικότητας της γραφής του συγγραφέα, θα μπορούσαμε να πούμε, σαν θόρυβο.

Προβλήματα κατάταξης μιας-κλάσης
Η Εγγενής Ανίχνευση Λογοκλοπής αλλά και τα συναφή προβλήματα που περιγράφηκαν παραπάνω, αποτελούν προβλήματα κατάταξης μιας-κλάσης *(one-class classification problems)*. Σε ένα πρόβλημα κατάταξης μιας-κλάσης ορίζεται μια κλάση-στόχος, για την οποία υπάρχει ένας συγκεκριμένος αριθμός παραδειγμάτων, ή αλλιώς στιγμιοτύπων, μέσα στο υπό μελέτη σώμα δεδομένων. Οτιδήποτε δεν ανήκει σε αυτή τη κλάση-στόχο, αποτελεί έκτοπο σημείο *(outlier)*, ενώ το πρόβλημα ταξινόμησης έγκειται στο διαχωρισμό των έκτοπων σημείων από τα σημεία-μέλη της κλάσης-στόχου. Συνήθως, ο αριθμός των έκτοπων σημείων είναι κατά πολύ μεγαλύτερος από την κλάση-στόχο, και, γενικά, μπορεί να συλλέγεται αυθαίρετο πλήθος από αυτά. Εκ πρώτης όψεως ένα πρόβλημα ταξινόμησης μιας-κλάσης μπορεί να φαίνεται ως πρόβλημα ταξινόμησης δύο-κλάσεων, μόνο που υπάρχει μια σημαντική διαφοροποίηση: τα μέλη της κλάσης-στόχου μπορούν να θεωρηθούν αντιπροσωπευτικά για την κλάση τους, όμως τα έκτοπα σημεία δεν μπορούν να νοηθούν ως αντιπροσωπευτικά για μια, κάποιου είδους, "κλάση-μη στόχο". Κι αυτό διότι τα έκτοπα σημεία μπορεί να έχουν διαφορετικές προελεύσεις και πιθανότατα δεν σχετίζονται με κανέναν τρόπο.
Με άλλα λόγια, η επίλυση ενός προβλήματος ταξινόμησης μιας-κλάσης απαιτεί την εκμάθηση ενός "σεναρίου" (του σεναρίου της κλάσης-στόχου) μέσω της απουσίας διαφοροποιητικών στοιχείων [10].

Κεφάλαιο 2

Συστήματα Εγγενούς Ανίχνευσης Λογοκλοπής

2.1 Μεθοδολογία ενός τυπικού συστήματος

Για την ανίχνευση λογοκλοπής με εγγενείς μεθόδους, η ιδέα είναι πως περιμένουμε τα λογοκλεμμένα κομμάτια που έχει εισάγει ο συγγραφέας από άλλες πηγές, να διαφέρουν στυλιστικά από το υπόλοιπο κείμενο που έχει γραφεί από τον ίδιο. Υπάρχει λοιπόν άμεση σύνδεση του προβλήματος με τις τεχνικές *στυλομετρίας (stylometry)*. Ως *στυλομετρία* αναφέρονται οι εφαρμογές του κλάδου της Εφαρμοσμένης Γλωσσολογίας, που έχουν στόχο τη μελέτη και ερμηνεία των κειμένων, υπό το πρίσμα του γλωσσολογικού στυλ.

Ουσιαστικά, κάθε σύστημα εγγενούς ανίχνευσης λογοκλοπής προϋποθέτει την επιλογή συγκεκριμένων στυλιστικών χαρακτηριστικών, τα οποία καλούνται να αναδείξουν το στυλ γραφής του υπό εξέταση κειμένου αλλά και τις πιθανές ανομοιομορφίες μέσα σε αυτό. Έως την εξαγωγή αυτών των χαρακτηριστικών κειμένου αλλά και την αξιοποίησή τους για την εξαγωγή των τελικών αποτελεσμάτων, πρέπει να προηγηθεί και να ακολουθήσει, αντίστοιχα, αρκετή δουλειά.
Ένα τυπικό σύστημα εγγενούς ανίχνευσης λογοκλοπής αποτελείται από τρεις φάσεις επεξεργασίας του κειμένου: την προ-επεξεργασία, την στυλιστική ανάλυση και τη μετ-επεξεργασία για την τελική εξαγωγή των, θεωρούμενων ως, λογοκλεμμένων κομματιών. Όπως φαίνεται από τις συμμετοχές στο διαγωνισμό του PAN το 2011, τα βασικά δομικά υλικά του συστήματος περιλαμβάνουν μια στρατηγική κατάτμησης κειμένου *(text chunking model)*, ένα μοντέλο εξαγωγής του στυλ γραφής *(writing style retrieval model)* και έναν αλγόριθμος ανίχνευσης των έκτοπων σημείων *(outlier detection algorithm)*. Η διαδικασία εξαγωγής των λογοκλεμμένων κομματιών μπορεί να συνοψιστεί σε μια σειρά βημάτων ως εξής [8]:

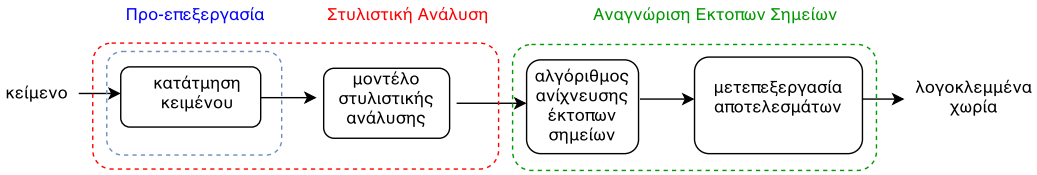
(1) το κείμενο κατατμείται,

(2) τα κομμάτια που προκύπτουν από την κατάτμηση αναπαριστώνται βάσει του μοντέλου εξαγωγής του στυλ γραφής,

(3) στυλιστικές διαφοροποιήσεις μεταξύ των κομματιών αυτών, αναγνωρίζονται βάσει του αλγορίθμου ανίχνευσης έκτοπων σημείων,

(4) ύστερα από μετεπεξεργασία, τα αναγνωρισμένα κομμάτια επιστρέφονται ως πιθανώς λογοκλεμμένα κομμάτια του υπό εξέταση κειμένου.

Στο Σχήμα 2.1 φαίνεται παραστατικά η τυπική διαδικασία εξαγωγής λογοκλεμμένων κομματιών από ένα σύστημα εγγενούς ανίχνευσης λογοκλοπής.



Σχήμα 2.1 Τυπικό σύστημα εγγενούς ανίχνευσης λογοκλοπής

Σχετικά με ό,τι επακολουθεί, επισημαίνεται πως έχουμε λάβει υπόψιν κλασικές μεθοδολογίες για τα συστήματα ανίχνευσης. Σε καμία περίπτωση δεν καλύπτεται όλο το φάσμα των προσεγγίσεων για τον σχεδιασμό και την υλοποίηση ενός τέτοιου συστήματος, κάτι που αναδεικνύεται και στο Κεφάλαιο 4 του παρόντος τόμου, όπου παρουσιάζονται τα συστήματα που συμμετείχαν στον διαγωνισμό του PAN το 2011. Πρόκειται περισσότερο για μια προσπάθεια γνωριμίας με κάποιες βασικές έννοιες και μεθοδολογίες.

2.2 Προεπεξεργασία κειμένου

Το στάδιο της προ-επεξεργασίας του κειμένου *(Text Preprocessing)* περιλαμβάνει την κατάτμηση κειμένου αλλά και την εξαγωγή κάποιων χαρακτηριστικών που, συνήθως, προαπαιτούνται για την στυλιστική ανάλυση που ακολουθεί ως επόμενο βήμα.
Όσον αφορά στα χαρακτηριστικά-προαπαιτούμενα για την στυλιστική ανάλυση, αυτά καθορίζονται από το σχεδιασμό της τελευταίας. Θα αναφέρουμε εδώ τα συνήθη προεπεξεργαστικά βήματα σε ένα σύστημα εγγενούς ανίχνευσης λογοκλοπής:

- αναγνώριση προτάσεων *(sentence detection)*
- αναγνώριση λέξεων *(token detection)*
- μετατροπή κεφαλαίων γραμμάτων σε πεζά *(de-capitalisation)*
- απομάκρυνση των αλφαριθμητικών *(alphanumerics removal)*

- αναγνώριση θεμάτων των λέξεων *(stemming)*
- απομάκρυνση ειδικών χαρακτήρων *(special characters removal)*
- απομάκρυνση των stopwords *(stopwords removal)*
- αναγνώριση των λέξεων ως μερών του λόγου *(Part-Of-Speech Tagger* ή *POS Tagger)*

Επαναλαμβάνεται πως μέρος από τα παραπάνω μπορεί να χρησιμοποιείται σε ένα σύστημα ΕΑΛ, ανάλογα με τον εκάστοτε σχεδιασμό.

Με τον όρο *stopwords* εννοούνται οι λέξεις που θεωρούνται κοινές σε μια γλώσσα, για παράδειγμα τα άρθρα ή οι λέξεις ”είναι”, ”έχει” για την ελληνική γλώσσα. Όπως είναι αναμενόμενο, δεν υπάρχει μια μόνο λίστα των stopwords για κάθε γλώσσα. Κατά μια προσέγγιση τέτοιες λέξεις δεν περιέχουν σημαντική πληροφορία γι’ αυτό και απομακρύνονται. Από μια άλλη σκοπιά, ο τρόπος χρήσης (π.χ η πυκνότητά τους μέσα στο κείμενο) μπορεί να χρησιμοποιηθεί για τη διαμόρφωση ισχυρών στυλιστικών χαρακτηριστικών. Το ίδιο συμβαίνει με τα σημεία στίξης και άλλους ειδικούς χαρακτήρες.

Με μια ματιά στην παραπάνω λίστα βλέπουμε πως πολλά από τα βήματα της προεπεξεργασίας του κειμένου, αποτελούν από μόνα τους σοβαρά αλγοριθμικά προβλήματα, τα οποία αποτελούν ανοιχτά ερευνητικά πεδία. Χαρακτηριστικά αναφέρουμε τα αναγνώριση προτάσεων, αναγνώριση των θεμάτων των λέξεων, αναγνώριση των μερών του λόγου.
Για αυτά μπορεί κανείς να χρησιμοποιήσει εργαλεία (και) ανοιχτού κώδικα που υπάρχουν στο διαδίκτυο.

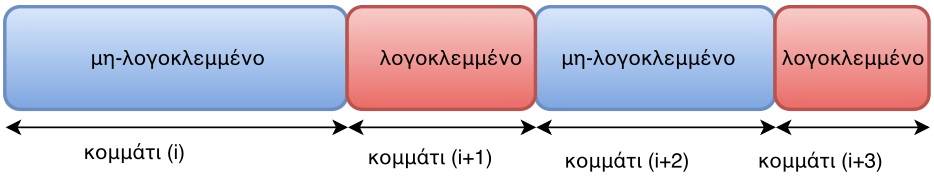
Όσον αφορά στην κατάτμηση του κειμένου, θα ασχοληθούμε με τις πλέον χρησιμοποιούμενες τεχνικές στην εγγενή ανίχνευση λογοκλοπής.
Η στρατηγική κατάτμησης στο πρόβλημα εγγενούς ανίχνευσης λογοκλοπής αποτελεί ιδιαίτερα σημαντική σχεδιαστική επιλογή και επηρεάζει άμεσα την αποτελεσματικότητα αλλά και την αποδοτικότητα του συστήματος. Κατά τη διαδικασία της κατάτμησης το κείμενο σπάει σε κομμάτια, τα οποία θα περάσουν, έπειτα, στο στάδιο της στυλιστικής ανάλυσης. Σε αυτό το στάδιο παράγεται, δηλαδή, το αντικείμενο εργασίας της στυλιστικής ανάλυσης. Αφού ποσοτικοποιηθούν τα στυλιστικά χαρακτηριστικά για καθένα από τα κομμάτια, τα τελευταία περνούν στο στάδιο της ανίχνευσης στυλιστικών ανομοιομορφιών, όπου συγκρίνονται μεταξύ τους ή/και με ολόκληρο το κείμενο. Μέσω αυτών των συγκρίσεων θα αναδειχθούν τα έκτοπα σημεία.
Είναι, λοιπόν, φανερό ότι μια αποτυχημένη στρατηγική κατάτμησης οδηγεί πιθανότατα σε μειωμένη ικανότητα αναγνώρισης των στυλιστικών ανομοιομορφιών.

Αρκεί να σκεφτούμε ένα παράδειγμα αποτυχημένης επιλογής: έστω ότι εξετάζουμε ένα κείμενο, το οποίο περιλαμβάνει λογοκλεμμένα χωρία μέσου μήκους 5 προτάσεων, και η στρατηγική κατάτμησης που έχουμε επιλέξει σπάει το κείμενο

σε τμήματα σταθερού μήκους 20 προτάσεων. Είναι φανερό πως το σύστημα μας θα ”βάζει νερά” από δυο τρύπες · αφενός, το σύστημα θα δυσκολεύεται να εντοπίσει τις στυλιστικές ανομοιομορφίες μεταξύ των κατετμημένων κομματιών, λόγω υπερίσχυσης των ”υγειών” κομματιών, που στην ευνοϊκότερη περίπτωση θα είναι 15 υγιείς έναντι 5 λογοκλεμμένων. Αφετέρου, ακόμα και στην περίπτωση μιας εξαιρετικά καλοσχεδιασμένης και ευαίσθητης στυλιστικής ανάλυσης, που θα ήταν ικανή να εντοπίσει τις ανομοιομορφίες, θα κατέληγε με πολλές ”παράπλευρες απώλειες” στην εξαγωγή των λογοκλεμμένων χωρίων.

Σε μια ιδανική κατάτμηση του κειμένου, κάθε κομμάτι είναι είτε καθαρά μη-λογοκλεμμένο τμήμα του κειμένου είτε καθαρά λογοκλεμμένο τμήμα. Αλλά όχι μόνο αυτό · στην ιδανική περίπτωση κάθε κομμάτι θα ήταν το μεγίστου δυνατού μήκους καθαρά μη-λογοκλεμμένο ή καθαρά λογοκλεμμένο τμήμα. Έτσι, στη στυλιστική ανάλυση θα περνούσαν κατά το δυνατό μεγαλύτερα κομμάτια κειμένου αλλά και ”ανόθευτα” σε κάθε περίπτωση, οπότε θα εξάγονταν αξιόπιστα στυλιστικά χαρακτηριστικά κατά το μέγιστο. Βέβαια, αν μπορούσαμε αν καταφέρουμε κάτι τέτοιο σημαίνει πως η εργασία μας θα είχε τελειώσει, ήδη, από το πρώτο βήμα.

Ας δούμε σχηματικά ποιος είναι ο στόχος, δηλαδή η ιδανική κατάτμηση ενός κειμένου.



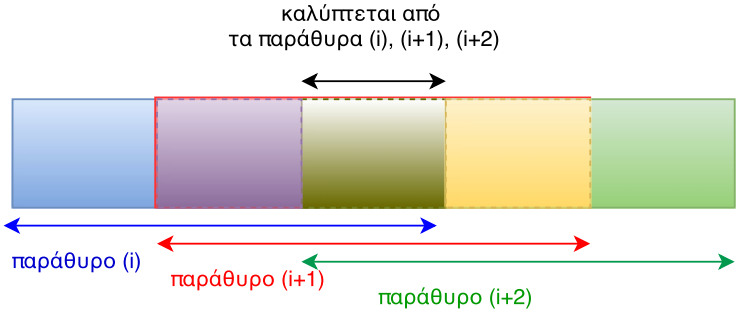
Σχήμα 2.2 Ιδανική κατάτμηση κειμένου. Με κόκκινο χρώμα τα λογοκλεμμένα χωρία.

Ακολουθούν κάποιες από τις βασικές επιλογές για την κατάτμηση κειμένου.

Τμήματα Σταθερού Μήκους
Η πιο απλή στρατηγική κατάτμησης είναι το ”σπάσιμο” του κειμένου *d* σε σταθερού μήκους τμήματα $s_1, s_2, ..., s_n$. Το μήκος των τμημάτων μπορεί να επιλεγεί σε εύρος μερικών προτάσεων έως και αρκετών παραγράφων. Το μήκος των τμημάτων πρέπει πάντα να επιλέγεται με κριτήρια το συνολικό μήκος του κειμένου, το αναμενόμενο μήκος των ενδεχόμενων λογοκλεμμένων χωρίων - αν είναι δυνατό να εκτιμηθεί αυτό - αλλά και την επεξεργαστική ισχύ που σχεδιάζεται να καταναλώνει το σύστημα.
Αυτή η πολιτική κατάτμησης είναι αρκετά απλοϊκή και με πολλά μειονεκτήματα. Η επιλογή του μήκους γίνεται αυθαίρετα, μιας και, από τη φύση του προβλήματος, δεν παρέχεται καμιά πληροφορία για τα λογοκλεμμένα χωρία και το μόνο που

μπορούμε να κάνουμε είναι αβάσιμες υποθέσεις. Από τη μια μεριά, επιλογή μεγάλου μήκους κατάτμησης οδηγεί σε αστοχία ανίχνευσης των, πιθανώς περιορισμένης έκτασης, λογοκλεμμένων χωρίων, ενώ στην αντίθετη περίπτωση δυσκολεύει το έργο της στυλιστικής ανάλυσης και ταυτόχρονα αυξάνει την απαιτούμενη επεξεργαστική ισχύ.

Μετακινούμενο παράθυρο
Η πολιτική κατάτμησης που επιλέγεται κατά κόρον στα συστήματα εγγενούς ανίχνευσης είναι το μετακινούμενο παράθυρο *(sliding window)*. Επιλέγουμε την τιμή των δύο παραμέτρων (1) *μήκος παραθύρου (μ.π.)* και (2) *βήμα (β)*. Τυπικές τιμές για το μήκος παραθύρου είναι κάποιες δεκάδες προτάσεις (ή κάποιες χιλιάδες χαρακτήρων), ενώ για βήμα επιλέγεται συνήθως τιμή μεταξύ $1/4$ και $1/3$ επί του μήκους παραθύρου. Το μετακινούμενο παράθυρο διασχίζει το κείμενο από την αρχή ως το τέλος, από τα αριστερά προς τα δεξιά. Ό,τι εμφανίζεται μέσα στο παράθυρο αποτελεί κάθε φορά ένα χωρίο, μια ακόμα μονάδα προς στυλιστική ανάλυση. Συνήθως τα παράθυρα στα οποία ”σπάει” το κείμενο είναι επικαλυπτόμενα, δηλαδή το βήμα παίρνει τιμές μικρότερες του μήκους παραθύρου *(0 < β < μ.π.)*. Σε αυτήν την περίπτωση δύο διαδοχικά παράθυρα περιλαμβάνουν κοινό κείμενο μήκους *(μ.π. − β)*. Στο επόμενο σχήμα παρασταίνεται η κατάτμηση του κειμένου με τη στρατηγική του μετακινούμενου παραθύρου, με επικαλυπτόμενα παράθυρα, για *μ.π. = 3β*.



Σχήμα 2.3 Μετακινούμενο παράθυρο, με μήκος παραθύρου 3k και βήμα k.

Πρόκειται για μια εύκολα υλοποιήσιμη και σχετικά κομψή λύση, που όμως η δυσκαμψία στο μήκος παραθύρου την καταδικάζει σε χαμηλή προσαρμοστικότητα στα εκάστοτε λογοκλεμμένα χωρία.
Η στρατηγική της κατάτμησης σε τμήματα σταθερού μήκους, δεν είναι παρά η εφαρμογή του μετακινούμενου παραθύρου με βήμα ίσο με το μήκος παραθύρου.

Όπως φαίνεται και από τα συστήματα που υλοποιήθηκαν για το διαγωνισμό του PAN, η λύση του μετακινούμενου παραθύρου για την κατάτμηση του κειμένου υιοθετείται κατά κανόνα. Πρόκειται για σχετικά ευέλικτη λύση για τις δύο αντικρουόμενες στοχεύσεις της κατάτμησης: από τη μια, σπάσιμο του κειμένου

σε όσο το δυνατόν λιγότερα και μεγαλύτερα κομμάτια, με στόχο την πιο αξιόπιστη στατιστική-στυλιστική ανάλυσή τους αλλά και τη μείωση των απαιτούμενων επεξεργαστικών πόρων · από την άλλη, κατάτμηση κειμένου σε μικρά κομμάτια, με στόχο τη ικανότητα απομόνωσης των λογοκλεμμένων από τα μη-λογοκλεμμένα χωρία για ευκολότερη ανίχνευση του πρωτότυπου συγγραφικού στυλ αλλά και των ανωμαλιών.

2.3 Στυλιστική Ανάλυση Κειμένου

Η στυλιστική ανάλυση *(Writing Style Analysis)* του κειμένου αποτελεί την καρδιά του συστήματος εγγενούς ανίχνευσης λογοκλοπής. Σε αυτό το τμήμα εξάγονται τα στυλιστικά χαρακτηριστικά των κομματιών που προέκυψαν από την κατάτμηση. Πραγματοποιείται, δηλαδή, ενός είδους ποσοτικοποίηση του στυλ για κάθε κομμάτι και, αν προβλέπεται από τη σχεδίαση του συστήματος, του συνολικού κειμένου. Έτσι, μπορούμε, σχηματικά, να πούμε, ότι κάθε κομμάτι περνάει από μια στυλιστική συνάρτηση *(style function)*, από όπου εξάγεται, ως ταυτότητα, μια τιμή ή ένα σύνολο τιμών. Η σύγκριση αυτών των ταυτοτήτων μεταξύ τους, σε επόμενο βήμα, θα αναδείξει ως έκτοπα σημεία τα πιθανώς λογοκλεμμένα χωρία. Είναι, λοιπόν, καίριας σημασίας η επιλογή στυλιστικών χαρακτηριστικών ικανών να αναδείξουν τη μοναδικότητα γραφής ενός συγγραφέα. Ταυτόχρονα, δεδομένων των τακτικών που μπορεί να χρησιμοποιηθούν με σκοπό τη συγκάλυψη της λογοκλοπής (Σχήμα 1.1) αλλά και της τεράστιας ποικιλίας σε μορφή και περιεχόμενο που μπορεί να συναντήσουμε σε ένα υπό εξέταση κείμενο, είναι φανερό, πως δεν υπάρχει ”παντοδύναμο” χαρακτηριστικό, που να εγγυάται ένα επιτυχημένο σύστημα.

Η εγγενής ανίχνευση λογοκλοπής χρησιμοποιεί εργαλεία της στυλομετρίας. Ας πάρουμε μια πρώτη γεύση για την ανίχνευση του στυλ γραφής, μέσα από τις κατηγορίες των στυλομετρικών χαρακτηριστικών που χρησιμοποιούνται συνήθως [10] :

1. στατιστικά κειμένου: σε επίπεδο χαρακτήρων
2. συντακτικά χαρακτηριστικά: σε επίπεδο προτάσεων
3. χαρακτηριστικά με χρήση των μερών του λόγου: ποσοτικοποίηση της χρήσης των διαφόρων μερών του λόγου
4. λεξιλογικά χαρακτηριστικά: ποσοτικοποίηση των σπάνιων λέξεων
5. δομικά χαρακτηριστικά - οργάνωση του κειμένου

Βλέπουμε πως, όπως επιτάσσει η κοινή λογική και η διαίσθηση, αναζητούμε τη μοναδικότητα της γραφής στο συντακτικό, το λεξιλόγιο και τον τρόπο οργάνωσης του κειμένου. Μια προσέγγιση που φαίνεται να πηγαίνει μάλλον κόντρα στη διαίσθηση, είναι αυτή της εξαγωγής στατιστικών σε επίπεδο χαρακτήρων.Η συχνότητα

χρήσης κάποιων χαρακτήρων ή κατηγορίας χαρακτήρων (π.χ σύμφωνα-φωνήεντα) ή ομάδας χαρακτήρων (π.χ συγκεκριμένες τριάδες χαρακτήρων) σε πρώτη σκέψη φαίνεται αυθαίρετη και τυχαία. Παρόλαυτά χρησιμοποιούνται και τέτοια χαρακτηριστικά και μάλιστα πετυχαίνουν αξιόλογες επιδόσεις.

Στον Πίνακα 2.1 περιλαμβάνονται τα πιο σημαντικά και δημοφιλή χαρακτηριστικά στις στυλιστικές αναλύσεις, ανά κατηγορία [10]. Τα στυλομετρικά χαρακτηριστικά που περιλαμβάνονται στον πίνακα και κατά πάσα πιθανότητα δεν είναι γνωστά εξηγούνται παρακάτω.

Πίνακας 2.1 Δημοφιλή στυλομετρικά χαρακτηριστικά

Κατηγορία	**Στυλομετρικό χαρακτηριστικό**
Λεξιλογικά χαρακτριστικά (σε επίπεδο χαρακτήρων)	Συχνότητα χαρακτήρων Συχνότητα/ποσοστό των n-gram χαρακτήρων Συχνότητα ειδικών χαρακτήρων (’(’, ’;’,’/’,κλπ.) Ποσοστό συμπίεσης
Λεξιλογικά χαρακτηριστικά (σε επίπεδο λέξεων)	Μέσο μήκος λέξης Μέσο μήκος πρότασης Μέσο πλήθος συλλαβών ανά λέξη Συχνότητα λέξεων Συχνότητα/ποσοστό n-gram λέξεων Πλήθος των άπαξ λεγομένων Δείκτης Dale-Chall Βαθμός επιπέδου Flesch Kincaid Δείκτης Gunning Fog Μέτρο του Honore R Μέτρο του Sichel S Μέτρο του Yule K Μέση κλάση συχνότητας λέξεων
Συντακτικά χαρακτηριστικά	Μέρη-του-λόγου Συχνότητα/ποσοστό n-gram μερών-του-λόγου Συχνότητα λέξεων με γραμματικό ρόλο Συχνότητα σημείων στίξης
Δομικά χαρακτηριστικά	Μέσο μήκος παραγράφου Ενδοπαραγραφοποίηση Χρήση λέξεων (απο)χαιρετισμού Χρήση υπογραφών

Ποσοστό συμπίεσης
Χρησιμοποιώντας έναν αλγόριθμο συμπίεσης δεδομένων-κειμένου εξάγουμε το μέγεθος του συμπιεσμένου κειμένου. Ως ποσοστό συμπίεσης *(compression rate)* ορίζεται το πηλίκο του μεγέθους του συμπιεσμένου κειμένου προς το μέγεθος του

ασυμπίεστου κειμένου.

$$\text{ποσοστό συμπίεσης } = \frac{\text{μέγεθος μη-συμπιεσμένου κειμένου}}{\text{μέγεθος συμπιεσμένου κειμένου}} \tag{2.1}$$

N-grams λέξεων/χαρακτήρων
Τα λεγόμενα *n-grams* χρησιμοποιούνται ευρέως σε προβλήματα εξόρυξης δεδομένων από κείμενο *(text mining)* και γλωσσικής επεξεργασίας *(*language processing). Πρόκειται για ομάδες εμφανιζόμενων λέξεων/χαρακτήρων μέσα σε ένα δεδομένο ”παράθυρο” κειμένου, στο οποίο κατά τον υπολογισμό των *n-grams* μετακινούμαστε με βήμα 1 (1 λέξη ή 1 χαρακτήρα). Ας πάρουμε για παράδειγμα την πρόταση *Μια αλεπού περπατάει στο φεγγάρι*. Για $N = 2$ τα *word n-grams* (γνωστά ως word bigrams) είναι:

- Μια αλεπού
- αλεπού περπατάει
- περπατάει στο
- στο φεγγάρι

ενώ, για την ίδια πρόταση, με $N = 3$ θα είχαμε:

- Μια αλεπού περπατάει
- αλεπού περπατάει στο
- περπατάει στο φεγγάρι

Αντίστοιχα υπολογίζονται τα *n-grams χαρακτήρων*, μετρώντας χαρακτήρες αντί για λέξεις [12].

Πλήθος των άπαξ λεγόμενων
Από την ελληνική φράση *άπαξ λεγόμενα*, πρόκειται για τις λέξεις που εμφανίζονται μόνο μια φορά σε ολόκληρο το κείμενο *hapax legomena)*.

Δείκτης Dale-Chall
Πρόκειται για μια μαθηματική φορμούλα, που δεδομένων του πλήθους των λέξεων, του πλήθους των προτάσεων και του πλήθους των ”δύσκολων λέξεων” μέσα σε ένα κείμενο, υπολογίζεται μια τιμή, η οποία αντικατοπτρίζει τη δυσκολία ανάγνωσης του κειμένου. Σε παρόμοια λογική κινούνται και τα *Βαθμός επιπέδου Flesch Kincaid* και *Δείκτης Gunning Fog*. Οι μαθηματικές τους φόρμουλες δίνονται στον Πίνακα 2.2.

Πίνακας 2.2 Στυλιστικές μετρικές δυσκολίας ανάγνωσης του κειμένου

$\text{Δείκτης Dale-Chall} = 0.1579 * \left(\frac{\text{δύσκολες λέξεις}}{\text{λέξεις}} * 100\right) + 0.0496 * \left(\frac{\text{λέξεις}}{\text{προτάσεις}}\right)$
$\text{Βαθμός Flesch Kincaid} = 206.835 - 1.015 * \left(\frac{\text{λέξεις}}{\text{προτάσεις}} * 100\right) - 84.6 * \left(\frac{\text{συλλαβές}}{\text{λέξεις}}\right)$
$\text{Δείκτης Gunning Fog} = 0.4 * \left[\left(\frac{\text{λέξεις}}{\text{προτάσεις}}\right) + 100 * \left(\frac{\text{σύνθετες λέξεις}}{\text{λέξεις}}\right)\right]$

Μέτρο του Honore (R)
Όπως και τα *Μέτρο του Sichel (S)*, *Μέτρο του Yule's (K)*, πρόκειται και πάλι για τιμή που προκύπτει από μαθηματική φόρμουλα και ποσοτικοποιεί τον πλούτο του λεξιλογίου στο κείμενο. Οι μαθηματικές τους σχέσεις δίνονται στον Πίνακα 2.3. [13]

Πίνακας 2.3 Στυλιστικές μετρικές λεξιλογικού πλούτου του κειμένου

$\text{Μέτρο του Honore (R)} = \frac{100 log_{10} N}{1 - \frac{\text{πλήθος των άπαξ λεγόμενων}}{\text{V}}}$
$\text{Μέτρο του Sichel (S)} = \frac{\text{Πλήθος των άπαξ δισλεγόμενων}}{\text{V}}$
$\text{Μέτρο του Yule (K)} = 10^4 * \left(-\frac{1}{N} + \sum_{i=1}^{V} \left(\frac{i}{N}\right)^2\right)$

όπου
V: πλήθος διαφορετικών λέξεων
V_i: πλήθος διαφορετικών λέξεων που συναντώνται *i* φορές
N: πλήθος όλων των λέξεων
άπαξ δισλεγόμενα: λέξεις που συναντώνται ακριβώς δύο φορές
άπαξ λεγόμενα: λέξεις που συναντώνται ακριβώς μία φορά

Η επιτυχία ενός συστήματος εγγενούς ανίχνευσης λογοκλοπής βασίζεται στην εύστοχη επιλογή στυλιστικών χαρακτηριστικών. Θα μπορούσαμε να παρομοιάσουμε το πρόβλημά μας, με μία με κλειστά μάτια ταξινόμηση αντικειμένων διαφορετικού υλικού, που έχουν συγκολληθεί σε μια ενιαία, συμπαγή ράβδο · μία επιτυχής κατάτμηση μπορεί να μας δώσει άρτια κομμάτια για τα οποία θα πρέπει να αποφανθούμε, μια καλή τεχνική εξαγωγής λογοκλεμμένων κομματιών μπορεί να μας παρέχει εύστοχα κριτήρια για την ταξινόμηση των κομματιών που θα κρατάμε στα χέρια μας, όμως χωρίς καλά στυλιστικά χαρακτηριστικά θα είναι σαν να στερούμαστε την αίσθηση της αφής.

2.4 Αναγνώριση Λογοκλεμμένων Κομματιών

Αφού ολοκληρωθεί η στυλιστική ανάλυση μέσω υπολογισμού των επιλεχθέντων χαρακτηριστικών θα πρέπει να μεσολαβήσουν κάποια βήματα, ώστε να είναι δυνατή η εξαγωγή των ύποπτων, για λογοκλοπή, χωρίων *(Outlier Identification)*. Κατά τα βήματα αυτά θα πρέπει να λαμβάνεται υπόψιν, ότι το σύστημα δεν σχεδιάζεται για ένα και μόνο κείμενο αλλά προορίζεται για επεξεργασία πολλών, και πολύ διαφορετικών μεταξύ τους, κειμένων. Δηλαδή, πριν από την κυρίως ειπείν αξιολόγηση των ανομοιομορφιών θα πρέπει να έχουμε προχωρήσει στην ποσοτικοποίηση αυτών των ανομοιομορφιών αλλά και να έχουμε εξασφαλίσει την ενιαία, για τα διάφορα κείμενα, κωδικοποίηση ή αναπαράστασή τους.

Ποσοτικοποίηση ανομοιομορφιών
Σε αυτό το στάδιο, δεδομένων των στυλιστικών απεικονίσεων των χωρίων του κειμένου αλλά και του συνόλου αυτού, προσπαθούμε να βρούμε έναν αποτελεσματικό τρόπο σύγκρισης που θα αναδείξει τις στυλιστικές ανομοιομορφίες. Πριν από τους συγκεκριμένους, μαθηματικούς τρόπους σύγκρισης τιμών θα πρέπει να αποφασίσουμε σχετικά με τα ίδια τα συγκρινόμενα μέρη. Οι δύο προσεγγίσεις είναι:

- σύγκριση κάθε χωρίου με το σύνολο του κειμένου
- σύγκριση καθενός χωρίου με όλα τα υπόλοιπα χωρία

Η πρώτη προσέγγιση (σύγκριση χωρίου - κειμένου) προϋποθέτει σιωπηρά, ότι το υπό εξέταση κείμενο είναι κατά το μεγαλύτερο μέρος του, πράγματι γραμμένο από τον φερόμενο ως συγγραφέα. Το στυλιστικό αποτύπωμα του συνόλου του κειμένου αποτελεί βάση εργασίας και σημείο αναφοράς, μιας και η απόκλιση των χωρίων από αυτό είναι που τα καθιστά ύποπτα λογοκλοπής ή όχι. Επομένως, στην περίπτωση που το μεγαλύτερο μέρος του κειμένου αποτελεί προϊόν λογοκλοπής, το σημείο αναφοράς καθίσταται αυτόματα αναξιόπιστο.

Η σύγκριση κάθε χωρίου-χωρίου θεωρείται, από κάποιους, ότι μπορεί να ξεπεράσει τον παραπάνω περιορισμό [1].

Αφού λοιπόν επιλεγούν τα συγκρινόμενα μέρη θα πρέπει να σχεδιάσουμε τον *τρόπο* σύγκρισης. Ουσιαστικά πρόκειται για την επιλογή μιας συνάρτησης απόστασης, που θα παίρνει ως είσοδο τα διανύσματα τιμών των δύο συγκρινόμενων

μερών και θα υπολογίζει κατά πόσο αποκλίνουν.
Ο Πίνακας 2.4 περιλαμβάνει κάποιες πολύ γνωστές και δημοφιλείς συναρτήσεις απόστασης.

Πίνακας 2.4 Συναρτήσεις απόστασης

Μέσο Τετραγωνικό Σφάλμα (MSE)

$$d_{MSE} : (\mathbf{x}, \mathbf{y}) \mapsto \frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

Απόσταση Manhattan

$$d_1 : (\mathbf{x}, \mathbf{y}) \mapsto \|\mathbf{x} - \mathbf{y}\|_1 = \sum_{i=1}^{n}(x_i - y_i)$$

Απόσταση Euclidian

$$d_2 : (\mathbf{x}, \mathbf{y}) \mapsto \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Απόσταση Minkowski

$$d_p : (\mathbf{x}, \mathbf{y}) \mapsto \|\mathbf{x} - \mathbf{y}\|_p = \left(\sum_{i=1}^{n}(x_i - y_i)^p\right)^{\frac{1}{p}}$$

Απόσταση Chebyshev

$$d_\infty : (\mathbf{x}, \mathbf{y}) \mapsto \|\mathbf{x} - \mathbf{y}\|_\infty = \lim_{p\to\infty}\left(\sum_{i=1}^{n}(x_i - y_i)^p\right)^{\frac{1}{p}} = \max_i |x_i - y_i|$$

Απόσταση Canberra

$$d_{CAD} : (\mathbf{x}, \mathbf{y}) \mapsto \sum_{i=1}^{n}\frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Απόσταση Συνημιτόνου

$$d_{cos} : (\mathbf{x}, \mathbf{y}) \mapsto 1 - \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2} = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Ενιαία κωδικοποίηση - Κανονικοποίηση των τιμών
Ένα ιδιαίτερα σημαντικό, αν και πολλές φορές παραγνωρισμένο, βήμα κατά την εξαγωγή του διανύσματος τιμών στυλιστικού αποτυπώματος, είναι η κανονικοποίηση των τιμών. Συνήθως τα συστήματα δεν κατασκευάζονται για την ανάλυση ενός μόνο ύποπτου κειμένου αλλά ένα ολόκληρο σώμα κειμένων. Όπως είναι φυσικό, τα προς εξέταση κείμενα παρουσιάζουν σημαντικές διαφορές μεταξύ τους ως προς το μέγεθος, τη δομή, τον πλούτο του λεξιλογίου κλπ.. Συνεπώς σε κάθε κείμενο αντιστοιχούν διαφορετικά όρια για το "φυσιολογικό" και του "έκτοπο". Κατά την εξαγωγή των έκτοπων σημείων - δηλαδή κατά την αναγνώριση των ύποπτων χωρίων των κειμένων - ορίζονται κάποια "κατώφλια", το ξεπέρασμα των οποίων σημαίνει *μη κανονικότητα*. Είναι φανερό πως χωρίς μια ενιαία κωδικοποίηση των στυλιστικών αποτυπωμάτων, τα κατώφλια που θα ορίσουμε είναι καταδικασμένα να αποτύχουν στα περισσότερα εκ των κειμένων. Η ίδια η επιλογή της συνάρτησης κανονικοποίησης πρέπει να αποτελεί κομμάτι του σχεδιασμού. Η συνάρτηση κανονικοποίησης μπορεί να αναδείξει ή να καταποντήσει την αποτελεσματικότητα των στυλιστικών χαρακτηριστικών που επιστρατεύουμε.
Η εμπειρία μας μπορεί να το επιβεβαιώσει αυτό: η αρχική κανονικοποίηση που επιχειρήσαμε στο σύστημά μας ήταν η απλή διαίρεση με την εκάστοτε μέγιστη τιμή, κάτι που αποδείχθηκε λάθος· η εκάστοτε μέγιστη τιμή δεν είναι καθόλου ουδέτερη κι έτσι αν και κανονικοποιημένες οι τιμές των χαρακτηριστικών, εντούτοις εξακολουθούσαν να φέρουν το στοιχείο της ιδιαιτερότητας του κάθε κειμένου, από το οποίο ακριβώς θέλαμε να απαλλαγούμε. Όταν εφαρμόσαμε άλλους τρόπους κανονικοποίησης, όπως τις συναρτήσεις *εφ(x)*, $x/(1+x)$ κ.ά., είδαμε πως η επιλογή της κατάλληλης συνάρτησης είναι αδιάσπαστο κομμάτι της στυλιστικής ανάλυσης και πρέπει πάντα να γίνεται με άξονα τα στυλιστικά χαρακτηριστικά του συστήματος για την ανάδειξή τους.

Αξιολόγηση ανομοιομορφιών
Σε αυτό το στάδιο έχουμε για κάθε χωρίο κάποια τιμή (ή κάποιες τιμές), όπου αντικατοπτρίζεται η στυλιστική απόκλιση σχετικά με τα υπόλοιπα χωρία ή/και με το σύνολο του κειμένου. Για την τελική εξαγωγή των ύποπτων χωρίων, αυτό που ουσιαστικά χρειαζόμαστε είναι κάποιες "κόκκινες γραμμές" για τις τιμές αυτές, το ξεπέρασμα των οποίων θα σημαίνει αδικαιολόγητη στυλιστική απόκλιση.
Αυτά τα όρια-κατώφλια μπορούν να οριστούν

- αυθαίρετα, βάσει παρατηρήσεων και εμπειρίας
- με τη βοήθεια μηχανικής μάθησης

Δεδομένου ενός σώματος κειμένων με γνωστά τα λογοκλεμμένα χωρία ως υλικό εκπαίδευσης, ένας ταξινομητής *(classifier)* αρχικά θα εκπαιδευτεί και θα ρυθμίσει τις παραμέτρους του κατάλληλα, έτσι ώστε να είναι μετά σε θέση να αξιολογήσει άλλα, άγνωστα κείμενα.
Όταν το σύστημα έχει σχεδιαστεί έτσι, ώστε στο στάδιο της εξαγωγής των λογοκλεμμένων χωρίων, τα δεδομένα να είναι *μια τιμή για κάθε χωρίο*, τότε ένας καλός

ταξινομητής πρέπει να είναι σε θέση να εντοπίζει και να υιοθετεί καταλληλότερη τιμή κατωφλίου, σε σχέση με τον αυθαίρετο ορισμό της.

Η παραπάνω ρητορική για την αναγνώριση των λογοκλεμμένων χωρίων αλλάζει αν ακολουθήσουμε μια προσέγγιση πιο κοντά στα ευφυή συστήματα και αφήσουμε περισσότερο χώρο στη μηχανική μάθηση. Ένα μοντέλο μηχανικής μάθησης μπορεί να αντικαταστήσει τόσο τις συναρτήσεις απόστασης όσο και την επιλογή των κατωφλίων.

Η εργασία αυτή υιοθετεί την προσέγγιση των ευφυών τεχνικών. Η ανίχνευση λογοκλεμμένων τμημάτων κειμένου με ευφυείς τεχνικές, το πρόβλημα των Μη Ισορροπημένων Δεδομένων Εκπαίδευσης για την Εγγενή Ανίχνευση και το ξεπέρασμα του, είναι θέματα του Κεφαλαίου 3 του παρόντος τόμου.

2.5 Αξιολόγηση συστήματος εγγενούς ανίχνευσης λογοκλοπής

Σε αυτό το κομμάτι θα δούμε κάποιες μετρικές με τις οποίες μπορούν να αξιολογηθούν τα αποτελέσματα ενός συστήματος εγγενούς ανίχνευσης λογοκλοπής.
Πριν τις παρουσιάσουμε κρίνουμε απαραίτητο το να αναφερθούμε στην ιδιαιτερότητα των προβλημάτων με δεδομένα δύο κλάσεων, τα οποία δεν κατανέμονται ισορροπημένα στις κλάσεις, και στην αναγκαιότητα ειδικών μετρικών για την αξιολόγησή τους.
Στο πρόβλημα της εγγενούς ανίχνευσης λογοκλοπής καλούμαστε, σε τελική ανάλυση, να αναγνωρίσουμε λογοκλεμμένα χωρία μέσα από ένα πλήθος χωρίων κειμένου. Ας θεωρήσουμε ότι μέσα στο υποθετικό αυτό πλήθος χωρίων, τα λογοκλεμμένα έχουν ετικέτα 1, ενώ τα μη λογοκλεμμένα ετικέτα 0. Όπως θα δούμε στο Κεφάλαιο 3, και θεωρείται υπόθεση εργασίας στα πλαίσια αυτού του τόμου, τα μηδενικά τείνουν να είναι τάξεις μεγέθους περισσότερα από τους άσσους. Ταυτόχρονα συμβαίνει η ανίχνευση των άσσων να είναι εξόχως σημαντικότερη από την ανίχνευση των μηδενικών. Για το λόγο αυτό δεν μπορούμε να βασιστούμε (μόνο) σε μια απλή μετρική αξιολόγησης προβλέψεων, όπως, για παράδειγμα, η απόλυτη ακρίβεια των προβλέψεων $(\frac{\text{σωστές προβλέψεις}}{\text{σύνολο προβλέψεων}})$. Αρκεί να σκεφτούμε την επιτυχία της αφελούς ταξινόμησης όλων των χωρίων ως μηδενικά, δηλαδή μη λογοκλεμμένα · ένα τέτοιο σύστημα ταξινόμησης θα είχε επιτυχία πολύ παραπάνω από 50%.

Υπάρχουν πολλά προβλήματα τέτοιου είδους, δηλαδή δύο κλάσεων, μη ισορροπημένων σε πλήθος, δεδομένων, όπου η κλάση μειοψηφίας να έχει εξαιρετική σημασία. Τέτοιο πρόβλημα είναι, για παράδειγμα, οι προληπτικές εξετάσεις για σοβαρές ασθένειες · οι περισσότερες εκ των εξετάσεων θα βγουν αρνητικές και μια λάθος θετική πρόβλεψη δεν θα είναι καταστροφική, ενώ μια λανθασμένη αρνητική πρόβλεψη πιθανότατα θα αποβεί μοιραία για τον εξεταζόμενο.
Χρειαζόμαστε λοιπόν μετρικές που να λαμβάνουν υπόψιν αυτήν την ιδιαιτερότητα.

Είδαμε ότι οι προβλέψεις του συστήματος μπορούν να γίνουν με την πρόσδοση δυαδικών τιμών στα διάφορα χωρία κειμένου, με μια ετικέτα, δηλαδή, που θα τα σημαδεύει (υποθέσαμε ετικέτα 0 για τα μη-λογοκλεμμένα και 1 για τα λογοκλεμμένα). Λόγω της δυαδικότητας της τιμής πρόβλεψης, 4 καταστάσεις είναι πιθανές:

- Σωστά Θετικό *(True Positive ή TP)* : πρόβλεψη ως λογοκλεμμένου ενός μη-λογοκλεμμένου χωρίου
- Σωστά Αρνητικό *(True Negative ή TN)* : πρόβλεψη ως μη-λογοκλεμμένου ενός μη-λογοκλεμμένου χωρίου
- Λάθος Θετικό *(False Positive ή TP)* : πρόβλεψη ως λογοκλεμμένου ενός μη-λογοκλεμμένου χωρίου
- Λάθος Αρνητικό *(False Negative ή FN)* : πρόβλεψη ως μη-λογοκλεμμένου ενός λογοκλεμμένου χωρίου

Τα παραπάνω συνοψίζονται στον Πίνακα 2.5, όπου ως θετική θεωρείται η κλάση των λογοκλεμμένων χωρίων και ως αρνητική των μη-λογοκλεμμένων.

Πίνακας 2.5 Πίνακας σφαλμάτων

	Προβλεπόμενα θετικά	Προβλεπόμενα αρνητικά
Πραγματικά θετικά	TP *(Σωστά Θετικό)*	FN *(Λάθος Αρνητικό)*
Πραγματικά αρνητικά	FP *(Λάθος Θετικό)*	TN *(Σωστά Αρνητικό)*

Στον Πίνακα 2.6 παρουσιάζονται οι κλασικές μετρικές αξιολόγησης για το πρόβλημά ανίχνευσης λογοκλοπής [14] [15].

Πίνακας 2.6 Μετρικές Αξιολόγησης συστήματος εγγενούς ανίχνευσης λογοκλοπής

$$\textit{Precision} = \frac{TP}{TP + FP}$$

$$\textit{Recall} = \frac{TP}{TP + FN}$$

$$\textit{F-measure} = \frac{2 * Precision * Recall}{Precision + Recall}$$

Ας δούμε τις μετρικές πιο αναλυτικά.

Precision (Ακρίβεια). Πρόκειται για τον λόγο των σωστά προβλεπόμενων ως λογοκλεμμένων χωρίων προς το σύνολο των χωρίων που προβλέφθηκαν ως λογοκλεμμένων. Ουσιαστικά πρόκειται για εκτίμηση της ακρίβειας των θετικών προβλέψεων. Δίνει απάντηση στην ερώτηση *Πόσα από αυτά που προβλέπω ως λογοκλεμμένα είναι πράγματι λογοκλεμμένα;*

Recall (Ποσοστό ανάκλησης). Ο λόγος των σωστά προβλεπόμενων ως λογοκλεμμένων χωρίων προς το σύνολο των πραγματικά λογοκλεμμένων χωρίων. Δίνει απάντηση στην ερώτηση *Πόσα από τα πραγματικά λογοκλεμμένα χωρία κατάφερα να εντοπίσω;*

F-measure (Μέτρο F). Πρόκειται για τον αρμονικό μέσο των *Precision* και *Recall*.

Κεφάλαιο 3

Ευφυείς Μέθοδοι Εξαγωγής Λογοκλεμμένων Χωρίων

3.1 Εισαγωγή

Όταν μιλάμε για ευφυείς μεθόδους στο κομμάτι της εξαγωγής των λογοκλεμμένων χωρίων αναφερόμαστε σε μεθόδους και αλγορίθμους μηχανικής μάθησης που επιστρατεύουμε για να αντλήσουμε τα μέγιστα από την πληροφορία που μας παρέχει η στυλιστική ανάλυση. Είδαμε πως στα συστήματα εγγενούς ανίχνευσης λογοκλοπής συνηθίζεται η εφαρμογή μιας συνάρτησης απόστασης που συγκεντρώνει τις τιμές όλων των στυλιστικών χαρακτηριστικών σε μία, ενώ έπειτα μπορεί να ”αναλαμβάνει” ένας ταξινομητής να προβλέπει ποια χωρία είναι ή όχι λογοκλεμμένα, αφού ρυθμίσει κατάλληλα τις εσωτερικές παραμέτρους του μέσω διαδικασίας εκμάθησης *(learning)*.

Σε αυτή την εργασία δώσαμε μεγαλύτερο χώρο στις ευφυείς μεθόδους και τη μηχανική μάθηση. Έτσι, σύμφωνα με αυτή τη προσέγγιση, στα επόμενα θεωρούμε πως κάθε χωρίο παρασταίνεται με ένα διάνυσμα τιμών, και όχι μόνο με μια ενιαία τιμή όπως προκύπτει από μια συνάρτηση απόστασης. Θεωρούμε πως η απαραίτητη *σύγκριση* (χωρίο - χωρίο ή χωρίο - ολόκληρο κείμενο, ανάλογα με το σχεδιασμό) εμπεριέχεται σε κάθε τιμή του διανύσματος. Αν, για παράδειγμα, κάθε χαρακτηριστικό αποδίδεται ως μια μεμονωμένη τιμή, η ενσωμάτωση της σύγκρισης στο διάνυσμα τιμών θα ήταν η διαφορά των τιμών των αντίστοιχων χαρακτηριστικών χωρίου-χωρίου ή χωρίου - συνόλου του κειμένου.
Με αυτόν τον τρόπο μπορούμε να εκμεταλλευτούμε τις δυνατότητες που προσφέρει η μηχανική μάθηση · απόδοση βαρών στα χαρακτηριστικά ανάλογα με τη σπουδαιότητά τους, αναγνώριση συσχετισμών μεταξύ των χαρακτηριστικών, βέλτιστη ρύθμιση κατωφλίων. Επιπλέον, αποκτάται μεγαλύτερη ελευθερία στον αριθμό αλλά και στο είδος των επιλεγόμενων στυλιστικών χαρακτηριστικών, αφού αυτά αποσυμπλέκονται και μπορούν να δρουν αυτόνομα.
Στα επόμενα θα δούμε *μεθόδους επικύρωσης* για την καλύτερη δυνατή εκμετάλ-

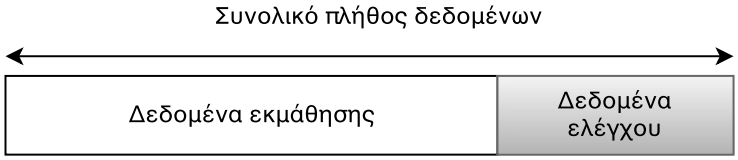
λευση του διαθέσιμου σώματος δεδομένων εκπαίδευσης, *αλγορίθμους εκμάθησης* καθώς και τρόπους να αντιμετωπίσουμε την ανισορροπία των δεδομένων εκμάθησης στο πρόβλημα.

3.2 Μέθοδοι Επικύρωσης

Στην εξόρυξη δεδομένων το αρχικό βήμα είναι και το δυσκολότερο · η ίδια η συλλογή των δεδομένων συνοδεία των ”λύσεων”. Όσον αφορά το πρόβλημα της ανίχνευσης λογοκλοπής, κομμάτι της έρευνας ασχολείται με την κατασκευή δεδομένων κειμένου που να προσιδιάζουν σε ρεαλιστικές περιπτώσεις λογοκλοπής. Τα δεδομένα μας, λοιπόν, είναι μάλλον δυσεύρετα και, γι’ αυτό, πολύτιμα.

Σε εφαρμογές μηχανικής μάθησης το διαθέσιμο σώμα των δεδομένων χωρίζεται σε δύο τμήματα:

1. δεδομένα εκμάθησης για την εκπαίδευση του μοντέλου *(training set)*
2. δεδομένα ελέγχου για την αξιολόγηση του εκπαιδευμένου μοντέλου *(test set)*



Σχήμα 3.1 Δεδομένα Εκμάθησης και Ελέγχου

Αυτή η μέθοδος όμως έχει δύο βασικά μειονεκτήματα:

- σε περιπτώσεις όπου διαθέτουμε πολύ λίγα δεδομένα μπορεί να μοιάζει πολυτέλεια το να κρατήσουμε ένα τμήμα τους για έλεγχο
- σε περίπτωση ατυχούς χωρισμού των δεδομένων σε εκμάθησης και ελέγχου, η αξιολόγηση του εκπαιδευμένου μοντέλου μπορεί να είναι παραπλανητική

Αυτοί οι περιορισμοί μπορούν να ξεπεραστούν με κάποιες μεθόδους σε βάρος, βέβαια, της υπολογιστικής πολυπλοκότητας. Οι μέθοδοι που θα παρουσιάσουμε εμπίπτουν στην γενικότερη κατηγορία *Διασταυρωμένη Επικύρωση*.

Διασταυρωμένη Επικύρωση:

- Τυχαία Υποδειγματοληψία *(Random Subsampling)*
- Κ-στρώσεων Διασταυρωμένη Επικύρωση *(K-Fold Cross validation)*
- Κράτα-ένα-εκτός Διασταυρωμένη Επικύρωση *(Leave-one-out Cross validation)*

Παρουσιάζουμε καθεμία από αυτές τις μεθόδους ξεχωριστά.

Τυχαία Υποδειγματοληψία
Η μέθοδος αυτή περιλαμβάνει τα εξής βήματα:

1. χώρισε το σώμα δεδομένων σε παραδείγματα εκαπίδευσης και ελέγχου, επιλέγοντας τυχαία έναν (σταθερό) αριθμό από παραδείγματα ελέγχου χωρίς αντικατάσταση

2. για κάθε κομμάτι εκπαίδευσε το μοντέλο *από την αρχή* με τα παραδείγματα εκπαίδευσης

3. αξιολόγησε το εκπαιδευμένο μοντέλο με τα παραδείγματα ελέγχου. Κράτα το αποτέλεσμα

4. επανάλαβε τα βήματα 1 έως 3 πολλές φορές

5. η τελική αξιολόγηση (σφάλμα) υπολογίζεται ως ο μέσος όρος των επιμέρους

Στο Σχήμα 3.2 παριστάνονται 3 επαναλήψεις διαχωρισμού των δεδομένων σε tπαραδείγματα εκπαίδευσης και ελέγχου για ένα υποθετικό σώμα δεδομένων.

Σχήμα 3.2 Τυχαία Υποδειγματοληψία

Κ-στρώσεων Διασταυρωμένη Επικύρωση
Στην Κ-στρώσεων Διασταυρωμένη Επικύρωση έχουμε την εξής μεθοδολογία:

1. χώρισε το σώμα δεδομένων σε k κομμάτια

2. για το πρώτο (επόμενο) κομμάτι στη σειρά εκπαίδευσε το μοντέλο *από την αρχή* με τα υπόλοιπα $(k-1)$ κομμάτια

3. αξιολόγησε το εκπαιδευμένο μοντέλο με το κομμάτι που απέμεινε

4. επανάλαβε τα βήματα για κάθε κομμάτι

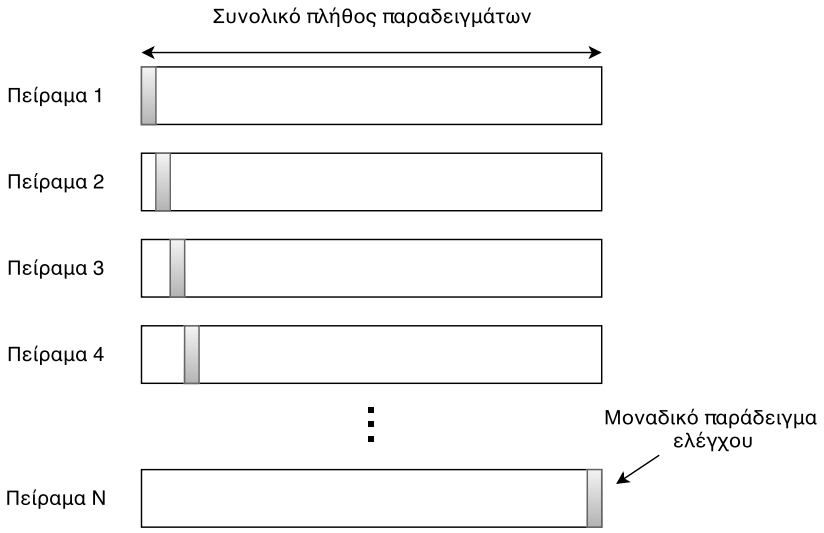
5. η τελική αξιολόγηση (σφάλμα) υπολογίζεται ως ο μέσος όρος των επιμέρους

Σχήμα 3.3 K-στρώσεων Διασταυρωμένη Επικύρωση

Στο Σχήμα 3.3 παρασταίνεται η παραπάνω διαδικασία για $k = 4$.
Με το χωρισμό του σώματος δεδομένων σε πολλά και μικρά κομμάτια επιτυγχάνεται μεγαλύτερη αξιοπιστία εκτίμησης του σφάλματος του τελικού, εκπαιδευμένου συστήματος, σε βάρος όμως της υπολογιστικής πολυπλοκότητας.

Κράτα-ένα-εκτός Διασταυρωμένη Επικύρωση
Πρόκειται για ακραία εφαρμογή της μεθόδου K-στρώσεων Διασταυρωμένη Επικύρωση, όπου για την παράμετρο k ισχύει $k = N$, όπου N τιμή του συνόλου των περιπτώσεων που περιλαμβάνονται στο σώμα δεδομένων. Χωρίζουμε, δηλαδή, το σύνολο των περιπτώσεων έτσι ώστε το σώμα των δεδομένων ελέγχου να αποτελείται κάθε φορά από 1 μόνο περίπτωση.
Και πάλι, η τελική εκτίμηση του σφάλματος προκύπτει ως μέσος όρος των επιμέρους εκπαιδεύσεων και ελέγχων. Στο Σχήμα 3.4 παρασταίνεται η μέθοδος Κράτα-ένα-εκτός Διασταυρωμένη Επικύρωση.

Τιμή της παραμέτρου k
Το ερώτημα που προκύπτει είναι το πώς επιλέγεται η κατάλληλη τιμή της παραμέτρου k. Σε κάθε περίπτωση έχουμε ένα δίλημμα μεταξύ υπολογιστικής πολυπλοκότητας, από τη μία, και της ακρίβειας της αξιολόγησης από την άλλη. Προφανώς όταν επιλέγεται μεγάλη τιμή του k έχουμε μεγάλη ακρίβεια αξιολόγησης του μοντέλου αλλά σε βάρος του χρόνου υπολογισμού. Το αντίστροφο συμβαίνει για μικρή τιμή του k.
Όπως είναι λογικό, για μεγάλα σώματα δεδομένων είναι μάλλον αδύνατο να επιλέξουμε πολύ μεγάλες τιμές του k, ενώ ταυτόχρονα έχουμε περισσότερα δεδομένα για το σώμα εκπαίδευσης κι έτσι δεν είναι πρόβλημα να κρατάμε ένα σχετικά μεγάλο ποσοστό των δεδομένων για το σώμα ελέγχου. Αντίθετα για μικρά σώματα δεδομένων έχουμε μεγαλύτερη ελευθερία επιλογής ενώ, ταυτόχρονα, μεγαλύτερη ανάγκη για οικονομία στο σώμα εκπαίδευσης. Η επιλογή της τιμής πρέπει, λοιπόν, πάντα να γίνεται με άξονα το μέγεθος του σώματος δεδομένων που έχουμε στη διάθεση μας.



Σχήμα 3.4 Κράτα-ένα-εκτός Διασταυρωμένη Επικύρωση

3.3 Μέθοδοι Εκμάθησης

3.3.1 Κατηγορίες τεχνικών εκμάθησης

Στην μηχανική μάθηση οι τεχνικές εκμάθησης χωρίζονται στις εξής 3 κατηγορίες:

- *Επιβλεπόμενη μάθηση (Supervised Learning)*
  Ο αλγόριθμος μάθησης κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους σε γνωστές-επιθυμητές εξόδους (σύνολο εκπαίδευσης), με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε πολλά προβλήματα όπως η ταξινόμηση *(classification)* και η πρόγνωση *(prediction)*.

- *Μη-επιβλεπόμενη μάθηση (Unsupervised Learning)*
  Ο αλγόριθμος μάθησης προσπαθεί να κατασκευάσει ένα μοντέλο με βάση κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων, χωρίς, όμως, να γνωρίζει τις επιθυμητές εξόδους. Έχουμε δηλαδή μη-χαρακτηρισμένα δεδομένα μάθησης *(unlabeled learning data)* και στόχος είναι η ανακάλυψη κάποιας εσωτερικής δομής που μπορεί αυτά να παρουσιάζουν. Η πιο γνωστή χρήση τέτοιων αλγορίθμων είναι στα προβλήματα συσταδοποίησης *(clustering)*.

- *Ενισχυτική μάθηση (Reinforcement Learning)*
  Ο αλγόριθμος μάθησης αναπτύσσει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον. Χρησιμοποιείται κυρίως σε προβλήματα σχεδιασμού *(planning)* όπως ο έλεγχος κίνησης ρομπότ.

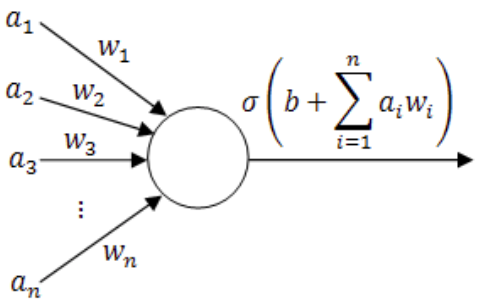
Στα πλαίσια της παρούσας εργασίας μας ενδιαφέρουν τεχνικές επιβλεπόμενης μάθησης. Ακολουθεί μια επισκόπηση αυτών των τεχνικών, παρουσιάζοντας τις βασικότερες κατηγορίες και κάποιες χαρακτηριστικές μεθόδους τους.

3.3.2 Τεχνικές Επιβλεπόμενης Μάθησης

Το Νευρωνικό Δίκτυο Perceptron

Αρκετοί δημοφιλείς αλγόριθμοι εκμάθησης βασίζονται στο νευρωνικό δίκτυο Perceptron.

Perceptron ενός-στρώματος (Single-layered Perceptron)
Το Perceptron ενός-στρώματος είναι ένας κλασικός *γραμμικός ταξινομητής (linear classifier)*, δηλαδή παίρνει τις αποφάσεις ταξινόμησης εφαρμόζοντας γραμμικό συσχετισμό των χαρακτηριστικών εισόδου.
Μπορεί να περιγραφεί περιεκτικά ως εξής:
Αν $\langle x_1, x_2..., x_n \rangle$ το διάνυσμα των χαρακτηριστικών που δίνεται ως είσοδος και $\langle w_1, w_2..., w_n$ είναι το διάνυσμα βαρών σύνδεσης, τότε το Perceptron υπολογίζει το άθροισμα $\sum_i x_i w_i$ και η έξοδος εξαρτάται από ένα ρυθμιζόμενο κατώφλι κι έτσι αν το άθροισμα είναι μεγαλύτερο της τιμής κατωφλίου τότε η έξοδος ισούται με 1, αλλιώς με 0.
Το Σχήμα 3.5 απεικονίζει ένα τέτοιο δίκτυο, όπου σ η συνάρτηση που καθορίζει το κατώφλι εξόδου και b το πιθανό bias.

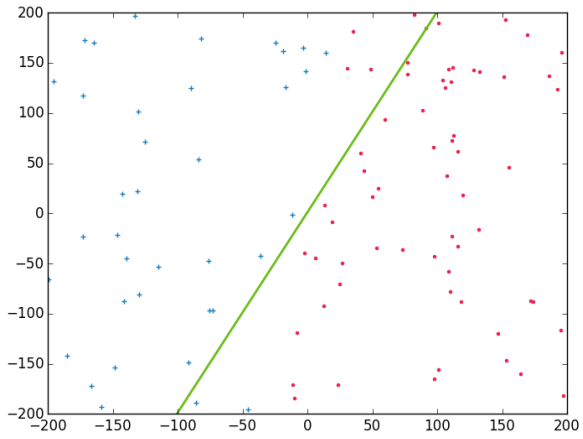


Σχήμα 3.5 Perceptron ενός-στρώματος

Η εκπαίδευση του δικτύου γίνεται με κατάλληλη ρύθμιση των βαρών σε κάθε βήμα ανάλογα με το εκάστοτε σφάλμα πρόβλεψης. Ο πιο γνωστός αλγόριθμος ρύθμισης για Perceptron ενός-στρώματος είναι ο WINNOW [16]. Σύμφωνα με αυτόν, εάν η πρόβλεψη είναι $y' = 0$ και η επιθυμητή έξοδος $y = 1$, τότε τα βάρη έχουν πολύ χαμηλές τιμές · έτσι για κάθε χαρακτηριστικό x_i για το οποίο $x_i = 1$, το βάρος ρυθμίζεται $w_i = w_i * \alpha$, όπου $\alpha > 1$ και ονομάζεται *παράμετρος προώθησης (promotion parameter)*. Αντίθετα, αν η πρόβλεψη είναι $y' = 1$ και η επιθυμητή έξοδος $y = 0$, τότε τα βάρη των χαρακτηριστικών με $x_i = 1$ είναι πολύ υψηλά και ρυθμίζονται με τον ίδιο τρόπο με μια *παράμετρο υποβιβασμού (demotion parameter)*, $0 < \beta < 1$.

Γενικά, ο WINNOW είναι ένα παράδειγμα *εκθετικού αλγορίθμου ανανέωσης (exponential update algorithm)*. Τα βάρη των ”σχετικών” χαρακτηριστικών αυξάνονται εκθετικά, ενώ τα βάρη των ”μη σχετικών” συρρικνώνονται εκθετικά.

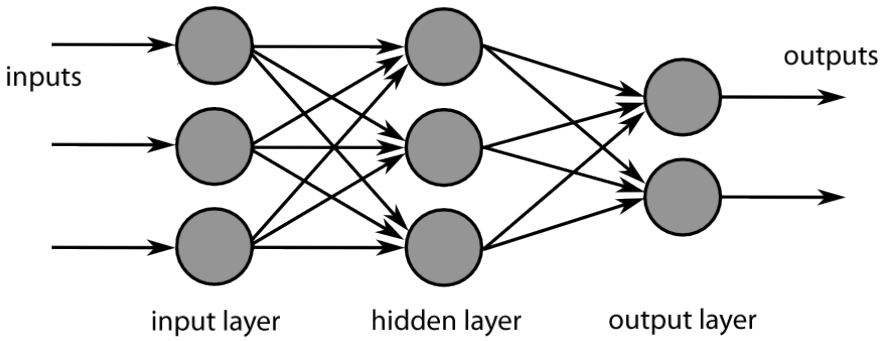
Perceptron πολλών-στρωμάτων (Multi-Layered Perceptron ή MLP)
Το απλό μοντέλο Perceptron είναι αποτελεσματικό για στιγμιότυπα *γραμμικά διαχωρίσιμα (linearly separable sets)*. Ένα σύνολο στιγμιοτύπων λέμε ότι είναι γραμμικά διαχωρίσιμο αν μια ευθεία γραμμή ή ένα επίπεδο αρκούν για να διαχωριστούν στις σωστές τους κατηγορίες.
Μια τέτοια περίπτωση στιγμιοτύπων φαίνεται στο Σχήμα 3.6, όπου τα στιγμιότυπα της μιας κλάσης παριστάνονται με μπλε χρώμα και της άλλης με κόκκινο.



Σχήμα 3.6 Γραμμικά διαχωρίσιμο σώμα δεδομένων

Όταν όμως τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα το απλό μοντέλο αποτυγχάνει. Τα perceptrons πολλών-στρωμάτων (ή Τεχνητά Νευρωνικά Δίκτυα) προσπαθούν να λύσουν αυτό το πρόβλημα [17] [18]. Ένα νευρωνικό δίκτυο πολλών-στρωμάτων αποτελείται από έναν μεγάλο αριθμό μονάδων-κόμβων που συνδέονται μεταξύ τους. Οι κόμβοι είναι τριών ειδών: i) εισόδου, οι οποίοι δέχονται την είσοδο που πρόκειται να επεξεργαστεί, ii) εξόδου, απ’ όπου προκύπτουν τα αποτελέσματα και, iii) ενδιάμεσοι ”κρυφοί” κόμβοι. Τα νευρωνικά δίκτυα με προς-τα-εμπρός-τροφοδότηση *(feed-forward neural network)* επιτρέπουν στα σήματα να μεταφέρονται μόνο προς μια κατεύθυνση, από την είσοδο προς την έξοδο. Ένα τέτοιο δίκτυο φαίνεται στο Σχήμα 3.7.

Ουσιαστικά, πρόκειται για πολλαπλά στρώματα κόμβων σε έναν κατευθυνόμενο γράφο, όπου κάθε στρώμα είναι πλήρως συνδεδεμένο με το επόμενο, δηλαδή η έξοδος ενός κόμβου στο στρώμα i δίνεται ως είσοδος σε κάθε κόμβο του στρώματος $(i + 1)$. Εκτός από τους κόμβους εισόδου, οι υπόλοιποι είναι ”νευρώνες”, δηλαδή επεξεργαστικές μονάδες, σε καθένα εκ των οποίων αντιστοιχεί μια *συνάρτηση ενεργοποίησης (activation function)*. Η γνωστότερη τεχνική εκπαίδευσης του



Σχήμα 3.7 Προς-τα-εμπρός τροφοδοτούμενο νευρωνικό δίκτυο

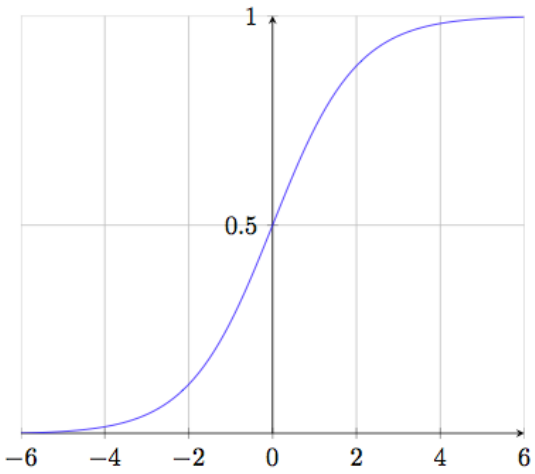
δικτύου είναι η τεχνική *backpropagation*.

Συνάρτηση ενεργοποίησης
Εάν οι συναρτήσεις ενεργοποίησης των νευρώνων είναι γραμμικές, τότε, αποδεικνύεται με γραμμική άλγεβρα, ότι για κάθε MLP υπάρχει ισοδύναμο με μόνο ένα κρυφό στρώμα νευρώνων. Αυτό που ουσιαστικά διαφοροποιεί το MLP από το απλό μοντέλο Perceptron είναι η χρήση μη γραμμικών συναρτήσεων ενεργοποίησης, συνήθως σιγμοειδών. Η επιστράτευση σιγμοειδών συναρτήσεων προέκυψε στην προσπάθεια μοντελοποίησης της ενεργοποίησης των βιολογικών νευρώνων του εγκεφάλου.
Μια χαρακτηριστική σιγμοειδής συνάρτηση που χρησιμοποιείται στα MLP δίνεται στην Σχέση 3.1, όπου S το (σταθμισμένο) άθροισμα των εισόδων του νευρώνα. Το διάγραμμα αυτής της συνάρτησης φαίνεται στο Σχήμα 3.8. Η ονομασία αυτών των συναρτήσεων πηγάζει από τη μορφή τους, που μοιάζει με συνάρτηση βήματος αλλά με αμβλυμένες τις γωνίες, έτσι ώστε να είναι διαφορίσιμες.

$$\sigma\left(S\right) = \frac{1}{1+e^{-S}} \tag{3.1}$$

Ο αλγόριθμος εκπαίδευσης Backpropagation
Όπως και στην περίπτωση του απλού Perceptron, η πληροφορία του δικτύου βρίσκεται στα βάρη των συνδέσεων μεταξύ των νευρώνων. Επομένως το πρόβλημα της εκπαίδευσης του δικτύου μπορεί να συνοψιστεί στο εξής ερώτημα: πώς εκπαιδεύουμε τα βάρη των συνδέσεων ώστε το δίκτυο να κατηγοριοποιεί με κατά το δυνατόν μεγαλύτερη ακρίβεια.
Ο αλγόριθμος backpropagation μπορεί να περιγραφεί με βήματα ως διαδικασία 6 βημάτων [19]:

1. Αρχικοποίησε τα βάρη σε τυχαίες τιμές
2. Δώσε το p-οστό στιγμιότυπο εκπαίδευσης $\mathbf{X}_p = (X_{p1}, X_{p2}, ..., X_{pN})$ με επιθυμητή έξοδο $\mathbf{Y}_p = (Y_{p1}, Y_{p2}, ..., Y_{pM})$



Σχήμα 3.8 Sigmoid function

3. Πέρασε τις τιμές εισόδου στο πρώτο στρώμα κόμβων, στρώμα 0, σε αντιστοιχία ένα-προς-ένα. Για κάθε τέτοιο κόμβο i η έξοδος είναι $Y_{0i} = X_{pi}$

4. Για κάθε νευρώνα i στα επόμενα στρώματα $j = 1, 2, ..., M$ υπολόγισε την έξοδο βάσει της συνάρτησης ενεργοποίησης f

$$Y_{ji} = f\left(\sum_{k=1}^{N_{j-1}} Y_{(j-1)k} W_{kij}\right) \tag{3.2}$$

5. Πάρε την έξοδο· για κάθε νευρώνα i στο τελευταίο στρώμα, στρώμα M, η έξοδος είναι $O_{pi} = Y_{Mi}$

6. Υπολόγισε την τιμή σφάλματος δ_{ji} για κάθε νευρώνα i για όλα τα στρώματα j σε αντίθετη σειρά, $j = M, (M-1), ..., 2, 1$, από το στρώμα εξόδου προς το στρώμα εισόδου και ρύθμισε τα βάρη αναλόγως. Για το στρώμα εξόδου το σφάλμα είναι

$$\delta_{Mi} = Y_{Mi}(1 - Y_{Mi})(T_{pi} - Y_{Mi}) \tag{3.3}$$

και για τα κρυφά στρώματα είναι

$$\delta_{ji} = Y_{ji}(1 - Y_{ji}) \sum_{k=1}^{N_{j+1}} \delta_{(j+1)k} W_{(j+1)k} \tag{3.4}$$

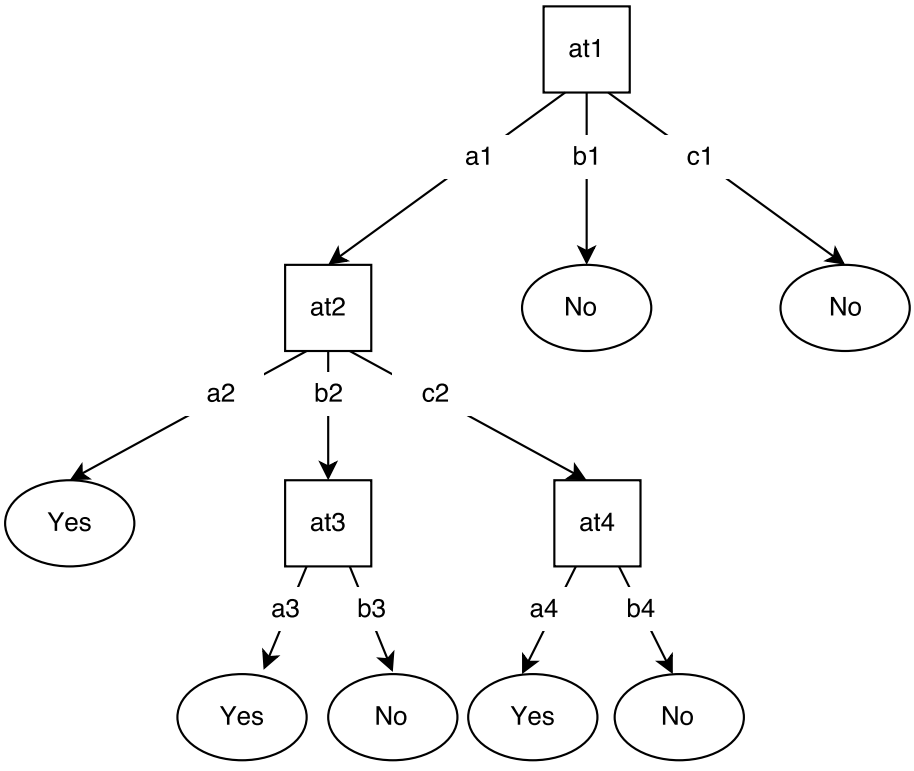
Ανανέωσε τις τιμές των βαρών σύνδεσης των νευρώνων ανάλογα με την τιμή του σφάλματος.

Για την ανανέωση των βαρών υπάρχουν πολλές τεχνικές και μαθηματικές φόρμουλες. Ως πιο δημοφιλή αναφέρουμε εδώ τη μέθοδο *Gradient Descent*, σύμφωνα με την οποία η διόρθωση της τιμής του εκάστοτε βάρους σύνδεσης εξαρτάται από την κλίση της συνάρτησης σφάλματος ως προς το βάρος αυτό [20].

Λογικές/Συμβολικές Τεχνικές

Ως χαρακτηριστική της κατηγορίας των λογικών/συμβολικών τεχνικών θα ασχοληθούμε με την εκμάθηση με *Δέντρα Απόφασης (Decision Trees)*.

Τα δέντρα απόφασης είναι δέντρα που κατατάσσουν στιγμιότυπα ταξινομώντας τα βάσει των τιμών των χαρακτηριστικών που τα προσδιορίζουν. Κάθε κόμβος σε ένα δέντρο απόφασης αναπαριστά ένα χαρακτηριστικό ενός υπό κατηγοριοποίηση στιγμιότυπου και κάθε κόμβος-παιδί προκύπτει με βάση την τιμή που μπορεί να πάρει ο κόμβος-γονέας. Τα στιγμιότυπα κατηγοριοποιούνται ξεκινώντας από τον κόμβο-ρίζα του δέντρου και ”διασχίζοντας” το δέντρο βάσει των τιμών των χαρακτηριστικών του [21].
Στο Σχήμα 3.9 φαίνεται ένα παράδειγμα δέντρου απόφασης για τα δεδομένα εκμάθησης του Πίνακα 3.1 [22].



Σχήμα 3.9 Παράδειγμα δέντρου απόφασης

Πίνακας 3.1 Παράδειγμα δεδομένων εκμάθησης

at1	at2	at3	at4	Class
a1	a2	a3	a4	Ναι
a1	a2	a3	b4	Ναι
a1	b2	a3	a4	Ναι
a1	b2	b3	b4	Όχι
a1	c2	a3	a4	Ναι
a1	c2	a3	b4	Ναι
b1	b2	b3	b4	Ναι
c1	b2	b3	b4	Ναι

Με βάση το δέντρο απόφασης του Σχήματος 3.9, το μονοπάτι ταξινόμησης ενός υποθετικού στιγμιότυπου $\langle at1 = a1, at2 = b2, at3 = a3, at4 = b4 \rangle$ θα διερχόταν από τους κόμβους at1, at2, at3, και θα κατέληγε να το ταξινομήσει ως θετικό *("Ναι")*.

Το χαρακτηριστικό που διαιρεί όσο το δυνατόν πιο ισορροπημένα τα δεδομένα εκμάθησης γίνεται ρίζα του δέντρου. Έχουν προταθεί διάφορες μέθοδοι για αυτό το σκοπό, όπως τα *information gain* [23], *gini index* [24]. Η ίδια διαδικασία εφαρμόζεται για κάθε επόμενο χωρισμό των δεδομένων, δημιουργώντας υπο-δέντρα έως ότου τα δεδομένα εκμάθησης να χωρίζονται σε υποσύνολα της ίδιας κλάσης.

Ένα δέντρο απόφασης, όπως και κάθε εκπαιδευμένο μοντέλο *h*, λέγεται ότι έχει υπερπροσαρμοστεί *(overfit)* στα δεδομένα εκμάθησης , εάν υπάρχει *h'* τέτοιο, ώστε να δίνει μεγαλύτερο σφάλμα πρόβλεψης,σε σχέση με το *h*, όταν εφαρμόζεται στα δεδομένα εκπαίδευσης και μικρότερο όταν εφαρμόζεται στα δεδομένα ελέγχου. Υπάρχουν δύο βασικές τεχνικές για την αποφυγή της υπερπροσαρμογής στα δέντρα απόφασης:

1. τερματισμός της διαδικασίας εκπαίδευσης πριν το σημείο του τέλειου "ταιριάσματος" με τα δεδομένα εκμάθησης

2. κλάδεμα του δέντρου απόφασης

Εάν δυο δέντρα είναι εξίσου αποτελεσματικά σε ένα σύνολο δεδομένων εκμάθησης, τότε επιλέγεται εκείνο με τα λιγότερα φύλλα.
Οι Breslow&Aha συγκεντρώνουν μεθόδους απλοποίησης των δέντρων απόφασης με στόχο την αύξηση της αποτελεσματικότητάς τους [25].

Βασικό πλεονέκτημα των δέντρων απόφασης είναι πως γίνονται άμεσα κατανοητά από τον άνθρωπο. Είναι εύκολο να καταλάβει κανείς γιατί ένα στιγμιότυπο προβλέφθηκε για μια συγκεκριμένη κλάση, με βάση την ιεραρχία των ελέγχων που γίνονται κατά τη διαδικασία εκτίμησής του.

Αλγόριθμοι Στατιστικής Εκμάθησης

Στις στατιστικές προσεγγίσεις έχουμε ένα στοχαστικό μοντέλο, το οποίο δεν κάνει απευθείας κατηγοριοποίηση αλλά δίνει ως έξοδο την πιθανότητα ένα στιγμιότυπο να ανήκει σε κάθε μία από τις πιθανές κλάσεις. Εδώ τα χαρακτηριστικά εισόδου αλλά και η μεταβλητή κλάσης είναι τυχαίες μεταβλητές.
Τα μπεϋζιανά δίκτυα *(Bayesian Networks)* είναι ο πιο γνωστός εκπρόσωπος των μεθόδων στατιστικής εκμάθησης. Ένα μπευζιανό δίκτυο (ή πιθανοτικός κατευθυνόμενος ακυκλικός γράφος) είναι ένα πιθανοτικό γραφικό μοντέλο, που αναπαριστά ένα σύνολο από τυχαίες μεταβλητές και τις μεταξύ τους εξαρτήσεις με τη βοήθεια ενός κατευθυνόμενου, ακυκλικού γράφου *(Directed Acyclic Graph* ή *DAG)*· στον γράφο οι κόμβοι αναπαριστούν τις τυχαίες μεταβλητές, ενώ οι συνδέσεις μεταξύ των κόμβων τις μεταξύ τους εξαρτήσεις.

Naive Bayes
Η πιο απλή πλην αρκετά δημοφιλής εφαρμογή των μπεϋζιανών μοντέλων στη μηχανική μάθηση είναι η προσέγγιση Naive Bayes. Πρόκειται για μεθόδους που βασίζονται στην εφαρμογή του θεωρήματος Baye's (3.5) με την "αφελή" υπόθεση της ανεξαρτησίας των χαρακτηριστικών-τυχαίων μεταβλητών ανά δύο(3.6).
Το θεώρημα Bayes διατυπώνεται, δεδομένης μιας μεταβλητής κλάσης y και ένα διάνυσμα χαρακτηριστικών $\langle x_1, x_2, ..., x_n \rangle$, ως εξής [26]:

$$P\left(y|x_1, ..., x_n\right) = \frac{P(y)P(x_1, ..., x_n|y)}{P(x_1, ..., x_n)} \tag{3.5}$$

Με την αφελή υπόθεση ανεξαρτησίας

$$P\left(x_i|y, x_1, ..., x_{i-1}, x_{i+1}, ..., x_n\right) = P(x_i|y) \tag{3.6}$$

το θεώρημα Bayes απλουστεύεται ως εξής:

$$P\left(y|x_1, ..., x_n\right) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, ..., x_n)} \tag{3.7}$$

Η πιθανότητα $P(x_1, ..., x_n)$ είναι σταθερή δεδομένης της εισόδου, οπότε ως κανόνας ταξινόμησης μπορεί να χρησιμοποιηθεί το απλούστερο

$$P\left(y|x_1, ..., x_n\right) \propto P(y)\prod_{i=1}^{n} P(x_i|y) \tag{3.8}$$

$$\Downarrow$$

$$\hat{y} = arg\max_{y} P(y)P(x_i|y) \tag{3.9}$$

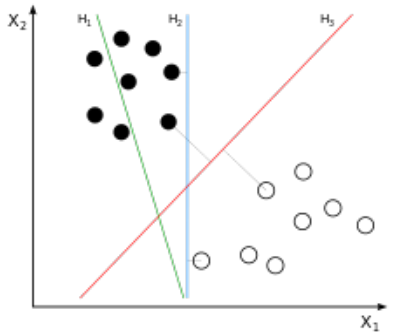
όπου $P(y)$ η σχετική συχνότητα της κλάσης y στα δεδομένα εκπαίδευσης.

Οι διάφοροι Naive Bayes ταξινομητές διαφέρουν κυρίως στις παραδοχές για την κατανομή της πιθανότητας $P(x_i|y)$. Ως παράδειγμα αναφέρουμε την Gaussian Naive Bayes προσέγγιση κατά την οποία λαμβάνεται :

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{3.10}$$

Μηχανές Διανυσμάτων Υποστήριξης
(Support Vector Machines ή SVM)

Οι μηχανές διανυσμάτων υποστήριξης παρέχουν μη-πιθανοτικές τεχνικές ταξινόμησης, τόσο για γραμμική όσο και για μη-γραμμική ταξινόμηση.
Ένα εκπαιδευμένο SVM μοντέλο αναπαριστά τα δεδομένα ως σημεία στο χώρο, με τρόπο ώστε τα στιγμιότυπα διαφορετικής κλάσης να χωρίζονται από ένα κενό, το οποίο θα πρέπει να είναι όσο το δυνατόν μεγαλύτερο.

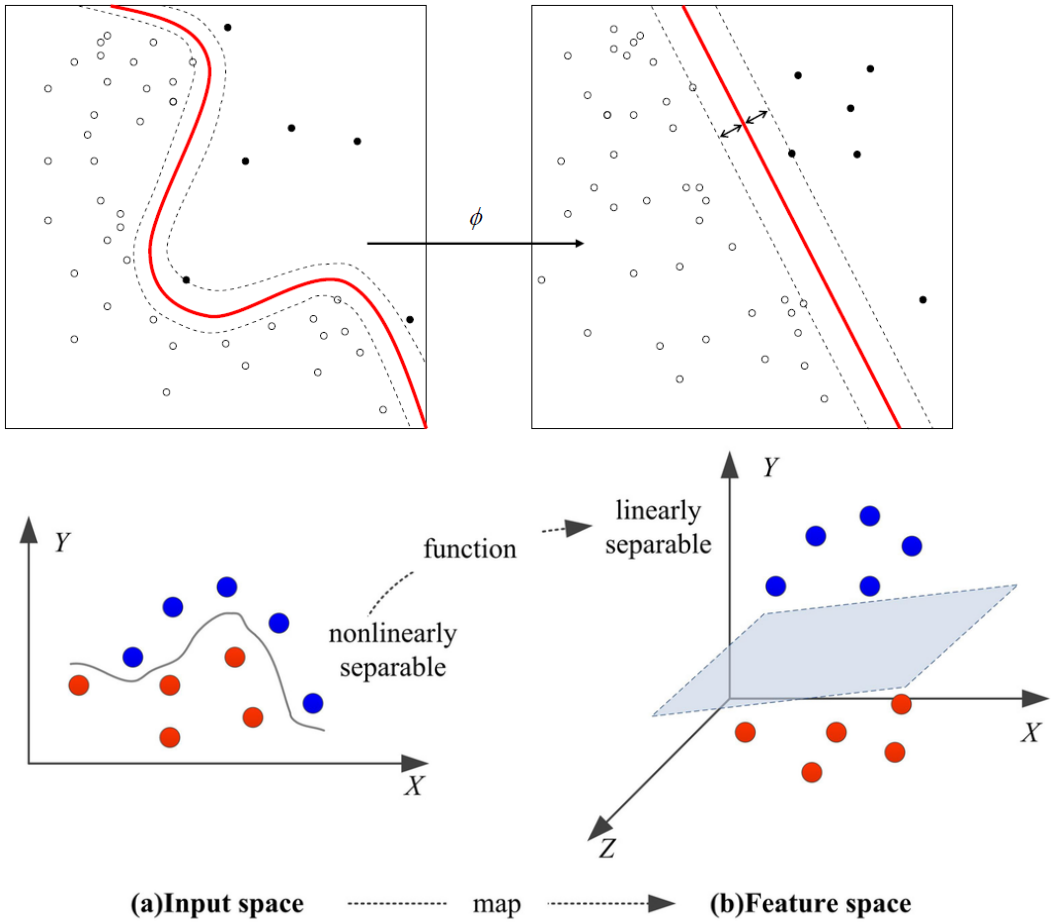


Σχήμα 3.10 SVM - Επίπεδα διαχωρισμού

Στο Σχήμα 3.10 παρουσιάζεται ένα παράδειγμα γραμμικής ταξινόμησης. Έχουμε δύο ειδών δεδομένα, τα άσπρα και μαύρα σημεία, και 3 διαφορετικές ευθείες για την ταξινόμησή τους. Βλέπουμε πως η H_1 αποτυγχάνει να διαφοροποιήσει τις κλάσεις. Η H_2 τις διαφοροποιεί αλλά δεν δίνει το μεγαλύτερο δυνατό περιθώριο. Η H_3 είναι η βέλτιστη ευθεία ταξινόμησης.
Γενικότερα, στην περίπτωση των γραμμικών SVM ταξινομητών τα στιγμιότυπα εισόδου ή, αλλιώς, σημεία δεδομένων θεωρούνται ως διανύσματα n διαστάσεων και το ζητούμενο είναι ο διαχωρισμός τους, μέσω ενός κατάλληλου υπερ-επιπέδου *(hyperplane)* $(n-1)$ διαστάσεων. Από όλα τα πιθανά υπερεπίπεδα που επιτυγχάνουν ικανοποιητική ταξινόμηση στα δεδομένα εκπαίδευσης επιλέγεται αυτό που δίνει το μεγαλύτερο περιθώριο μεταξύ των κλάσεων *(maximum margin hyperplane)*, μιας και όσο μεγαλύτερο το περιθώριο τόσο μικρότερο το γενικευμένο σφάλμα του ταξινομητή.

Εκτός από γραμμική ταξινόμηση, οι μηχανές διανυσμάτων υποστήριξης μπορούν να χρησιμοποιηθούν και για μη-γραμμική ταξινόμηση εφαρμόζοντας το λεγόμενο

κόλπο πυρήνα (kernel trick) [27] [28], κατά το οποίο απεικονίζουν την είσοδο σε χώρο μεγαλύτερων διαστάσεων.Όταν τα δεδομένα προς ταξινόμηση δεν είναι γραμμικά διαχωρίσιμα, επιχειρείται η απεικόνιση του πρωτότυπου χώρου δεδομένων σε μεγαλύτερη διάσταση, όπου ο διαχωρισμός τους θα είναι, πιθανότατα, πιο εύκολος.Δύο παραστατικά παραδείγματα αυτής της διαδικασίας απεικονίζονται στο Σχήμα 3.11.



Σχήμα 3.11 Κόλπο πυρήνα

Η απεικόνιση πραγματοποιείται με μεθόδους πυρήνα *(kernel methods)*, δηλαδή με μια, κατάλληλη για το πρόβλημα, συνάρτηση πυρήνα $k(x, y)$ *(kernel function)*. Οι μέθοδοι πυρήνα εξασφαλίζουν εύλογο υπολογιστικό φορτίο, επειδή, σε αντίθεση με άλλες τεχνικές αναγωγής σε μεγαλύτερη διάσταση, δεν χρησιμοποιούν έναν οριζόμενο από το χρήστη *χάρτη απεικόνισης* αλλά μια συνάρτηση πυρήνα · μια *συνάρτηση σύγκρισης*, δηλαδή, των δεδομένων στιγμιοτύπων-σημείων ανά δύο. Με τη συνάρτηση πυρήνα η μέθοδος ενεργεί σε έναν πολλών-διαστάσεων, ”κρυμμένο” χώρο στιγμιοτύπων, χωρίς ποτέ να χρειαστεί να υπολογίσει τις συντεταγμένες των

δεδομένων σε αυτό το χώρο, αλλά υπολογίζοντας το εσωτερικό γινόμενο των απεικονίσεων όλων των ζευγών στιγμιοτύπων στο χώρο αυτό.

Τα SVM για μηχανική μάθηση χρησιμοποιούν συνήθως θετικά ορισμένους πυρήνες *(positive definite kernels)*, που ορίζονται ως εξής:
Έστω $\boldsymbol{S}$ *ένα μη κενό σύνολο. Μια συμμετρική συνάρτηση* $\boldsymbol{K} : \boldsymbol{S} \times \boldsymbol{S} \rightarrow \mathbb{R}$ *ονομάζεται* ***θετικά ορισμένος πυρήνας*** *στο* $\boldsymbol{S}$ *εάν*

$$\sum_{i,j=1}^{n} c_i c_j \mathbf{S}(x_i x_j) \geq 0 \tag{3.11}$$

για κάθε $n \in \mathbb{N}$, $x_1, x_2, ..., x_n \in \boldsymbol{S}$, $c_1, c_2, ..., c_n \in \mathbb{R}$.

Αξίζει να σημειωθεί πως, σύμφωνα με τους Gin&Wang(2012), το πέρασμα σε μεγαλύτερες διαστάσεις αυξάνει το γενικευμένο σφάλμα του μοντέλου, ωστόσο αν δίνονται αρκετά παραδείγματα προς εκπαίδευση ο αλγόριθμος δουλεύει ικανοποιητικά.

Κάποιες κλασικές συναρτήσεις πυρήνα είναι:

- Πολυωνυμική (ομογενής): $k(\mathbf{x}_i \mathbf{x}_j) = (\mathbf{x}_i * \mathbf{x}_j)^d$
- Πολυωνυμική (μη ομογενής): $k(\mathbf{x}_i \mathbf{x}_j) = (\mathbf{x}_i * \mathbf{x}_j + 1)^d$
- Γκαουσιανή ακτινική συνάρτηση βάσης *(Gaussian radial basis function)*: $k(\mathbf{x}_i \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$ για $\gamma > 0$.

Δε θα επεκταθούμε στη μαθηματική ανάλυση και υλοποίηση του *κόλπου πυρήνα*. Ο αναγνώστης μπορεί να ανατρέξει στη σχετική βιβλιογραφία [27] [28].

3.4 Το πρόβλημα Μη-Ισορροπημένου Σώματος Δεδομένων

3.4.1 Περιγραφή του προβλήματος

Σε πολλές εφαρμογές επιβλεπόμενης μάθησης συμβαίνει οι καταχωρήσεις μιας κλάσης στο σώμα των δεδομένων να είναι πολύ λιγότερες *(κλάση μειοψηφίας)* από αυτές των υπολοίπων κλάσεων *(κλάσεις πλειοψηφίας)*. Τότε λέμε πως το σώμα δεδομένων είναι *μη-ισορροπημένο (Unbalanced Dataset)*. Σε αυτές τις περιπτώσεις οι ταξινομητές τείνουν να ταξινομούν με μεγάλη ακρίβεια την (ή τις) κλάση πλειοψηφίας, ενώ αντίθετα αποτυγχάνουν στη σωστή ταξινόμηση της κλάσης μειοψηφίας, λόγω της επίδρασης των πολυπληθών παραδειγμάτων εκμάθησης της πρώτης στα συνήθη κριτήρια εκπαίδευσης [29]. Ταυτόχρονα, συμβαίνει, στις περισσότερες από αυτές τις περιπτώσεις, η κλάση μειοψηφίας να παρουσιάζει το μεγαλύτερο ενδιαφέρον από άποψη εκπαίδευσης του μοντέλου και η αστοχία σωστής ταξινόμησής της συνεπάγεται το μεγαλύτερο κόστος [30].

Η Εγγενής ανίχνευση Λογοκλοπής είναι ένα πρόβλημα με μη-ισορροπημένο σώμα δεδομένων. Λαμβάνοντας ως (προ)υπόθεση εργασίας ότι κάθε κείμενο ως

επί το πλείστον γραμμένο από τον ίδιο συγγραφέα, περιμένουμε τα λογοκλεμμένα χωρία να είναι πολύ λιγότερα, τάξεις μεγέθους λιγότερα, των μη λογοκλεμμένων. Αυτό σημαίνει πως αν δεν τα επεξεργαστούμε με κάποιο τρόπο, ένα μοντέλο μηχανικής μάθησης θα εκπαιδευτεί ανισόρροπα, συνήθως δίνοντας πολύ μεγαλύτερη βαρύτητα στην αξιολόγηση των πολυπληθέστερων μη-έκτοπων σημείων, αφού δεν ”προλαβαίνει” να μάθει να αναγνωρίσει τα έκτοπα σημεία, ως κλάση μειοψηφίας, λόγω ανεπάρκειας δεδομένων. Είναι προφανές ότι μια τέτοια προοπτική μπορεί να καταστρέψει τα αποτελέσματα ενός άρτια, κατά τα άλλα, σχεδιασμένου συστήματος. Η σημαντικότητα της κλάσεως μειοψηφίας στο πρόβλημα εγγενούς ανίχνευσης λογοκλοπής φάνηκε και στην ενότητα *Αξιολόγηση ενός συστήματος εγγενούς συστήματος λογοκλοπής*, όπου ορίστηκαν μετρικές αξιολόγησης που δίνουν βαρύτητα στην επιτυχία ταξινόμησης των λογοκλεμμένων χωρίων.

Η έρευνα έχει δώσει κάποιες λύσεις-προτάσεις για το πρόβλημα αυτό, τις τυπικότερες εκ των οποίων θα παρουσιάσουμε στα επόμενα.

3.4.2 Αλγόριθμοι Εξισορρόπησης

Οι λύσεις που προτείνονται για το πρόβλημα των μη-ισορροπημένων δεδομένων, μπορούν να ταξινομηθούν σε 3 βασικές κατηγορίες [31]:

1. *Δειγματοληψία δεδομένων (Data sampling):* τα παραδείγματα εκπαίδευσης προσαρμόζονται κατάλληλα ώσπου να παρουσιάζουν μια πιο ισορροπημένη κατανομή κλάσεων, έτσι ώστε κατά την εκπαίδευση να αντιπροσωπεύονται επαρκώς όλες οι κλάσεις [32]

2. *Αλγοριθμικές προσαρμογές (Algorithmic modifications):* επιλέγονται ή προσαρμόζονται μέθοδοι εκπαίδευσης ώστε να συντονίζονται με το πρόβλημα των μη-ισορροπημένων δεδομένων [33]

3. *Εκπαίδευση με ευαισθησία κόστους (Cost-sensitive learning):* σε αυτή την κατηγορία περιλαμβάνονται προσεγγίσεις τόσο σε επίπεδο δεδομένων όσο και σε επίπεδο αλγορίθμων αλλά και συνδυασμού αυτών των δύο. Κατά την εκπαίδευση αποδίδεται μεγαλύτερο κόστος αστοχίας για τα παραδείγματα της κλάσης μειοψηφίας, σε σχέση με την/τις κλάση/εις πλειοψηφίας, έτσι ώστε η προσπάθεια αποφυγής αστοχιών μεγάλου κόστους για τα μεν, να αντισταθμίσει την υπερπληθώρα παραδειγμάτων για τα δε [34].

Από τις παραπάνω κατηγορίες θα εστιάσουμε στις μεθόδους *Δειγματοληψίας δεδομένων*, μιας και είναι αυτές που συμπεριλάβαμε στα πειράματα του συστήματός μας. Οι μέθοδοι δειγματοληψίας δεδομένων χωρίζονται στις τρεις ακόλουθες υποκατηγορίες [31]:

1. *υπο-δειγματοληπτικές μέθοδοι (Undersampling)*, όπου δημιουργείται ένα υποσύνολο του αρχικού σώματος δεδομένων αφαιρώντας κάποιες καταχωρήσεις (συνήθως παραδείγματα των κλάσεων πλειοψηφίας)

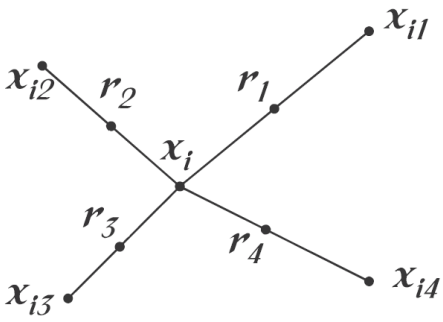
2. *υπερ-δειγματοληπτικές μέθοδοι (Oversampling)*, όπου δημιουργείται ένα υπερσύνολο του αρχικού σώματος δεδομένων, μέσω επανάληψης των υπαρχόντων παραδειγμάτων ή με τη δημιουργία νέων

3. *υβριδικές μέθοδοι (Hybrid methods)*, που συνδυάζουν τις δύο παραπάνω μεθόδους

Ας δούμε από κοντά την υποδειγματοληψία και την υπερδειγματοληψία.

Υποδειγματοληψία
Σε αυτή τη μέθοδο δημιουργείται ένα υποσύνολο του αρχικού σώματος δεδομένων εκπαίδευσης, απομακρύνοντας παραδείγματα της κλάσης πλειοψηφίας έως ότου επέλθει εξισορρόπηση. Τα δεδομένα εκπαίδευσης είναι λιγότερα κι έτσι βελτιώνεται και η υπολογιστική πολυπλοκότητα κατά την εκπαίδευση. Η πιο συχνή τεχνική είναι η απλή *τυχαία υποδειγματοληψία (Random Undersampling* ή *RUS)*. Κατά την τεχνική RUS παραδείγματα της κλάσης πλειοψηφίας επιλέγονται και αφαιρούνται τυχαία.
Το βασικό μειονέκτημα της μεθόδου υποδειγματοληψίας είναι ότι από τα δεδομένα εκπαίδευσης αφαιρείται πιθανώς σημαντική πληροφορία, κάτι που μπορεί να οδηγήσει σε αναξιόπιστη εκπαίδευση του μοντέλου. Για την αντιμετώπιση αυτού του προβλήματος αναζητούνται ”έξυπνοι” τρόποι εντοπισμού και αφαίρεσης παραδειγμάτων της κλάσης πλειοψηφίας. Τέτοια παραδείγματα είναι οι *Σύνδεσμοι Tomek (Tomek links)* [35], ο *Κανόνας των Συνεπτυγμένων Κοντινότερων Γειτόνων (Condensed Nearest Neighbor Rule)* [36] και η μέθοδος *Επιλογή Από-Μια-Μεριά (One-Sided Selection* ή *OSS)* [37]. Στην τελευταία αναζητούνται παραδείγματα που θεωρούνται είτε περιττά είτε ”θορυβώδη”.

Υπερδειγματοληψία
Κατά την υπερδειγματοληψία επιχειρείται η διόγκωση της κλάσης μειοψηφίας, έτσι ώστε να αντιπροσωπεύονται επαρκώς κατά την εκπαίδευση του classifier. Η πιο απλή εφαρμογή είναι, και εδώ, η τυχαία επανάληψη των παραδειγμάτων της κλάσης μειοψηφίας *(Random Oversampling* ή *ROS)*. Μια πιο ενδιαφέρουσα και αρκετά δημοφιλής μέθοδος είναι η *SMOTE (Synthetic Minority Oversampling Technique)* που προτάθηκε αρχικά από τους Chawla et al. [38]. Σύμφωνα με αυτήν την τεχνική η κλάση μειοψηφίας υπερδειγματοληπτείται κατασκευάζοντας ”συνθετικά” παραδείγματα και όχι απλά επαναλαμβάνοντας ήδη εμφανιζόμενα στο σώμα δεδομένων. Για κάθε καταχώρηση της κλάσης μειονότητας και τους k κοντινότερους γείτονες, δημιουργούνται ”συνθετικά” παραδείγματα, τα οποία τοποθετούνται στις γραμμές που θα ”ένωναν” το σημείο της καταχώρησης με τα σημεία των γειτόνων. Η τιμή του k εξαρτάται από το πλήθος των συνθετικών παραδειγμάτων που πρόκειται να παραχθούν. Μια απεικόνιση της μεθόδου παρουσιάζεται στο Σχήμα 3.5, όπου x_i είναι μια καταχώρηση της κλάσης μειοψηφίας, x_1, x_2, x_3, x_4 είναι οι τέσσερις κοντινότεροι γείτονες, και r_1, r_2, r_3, r_4 είναι τα συνθετικά παραδείγματα που δημιουργούνται.



Σχήμα 3.12 Δημιουργία ”συνθετικού” παραδείγματος με την τεχνική SMOTE

Το πρόβλημα με την αρχική αυτή εκδοχή της τεχνικής που περιγράφηκε είναι με τον τρόπο που δημιουργούνται τα νέα παραδείγματα · συγκεκριμένα, οι κοντινότεροι γείτονες εμπλέκονται στη διαδικασία κατασκευής νέων καταχωρήσεων χωρίς να λαμβάνεται υπόψιν η κλάση στην οποία αυτοί ανήκουν. Έτσι, το πιθανότερο είναι πως δημιουργούνται παραδείγματα που επιδεινώνουν την επικάλυψη μεταξύ των κλάσεων [39]. Προκειμένου να ξεπεραστεί αυτός ο περιορισμός έχουν προταθεί παραλλαγές-βελτιώσεις της μεθόδου. Χαρακτηριστικά αναφέρουμε τις τεχνικές *Borderline-SMOTE* [40] και *Adaptive Synthetic Sampling* [41].

Στην Borderline-SMOTE μέθοδο η λύση δίνεται με τον έλεγχο των επιλεγόμενων γειτόνων, έτσι ώστε αυτοί να ανήκουν στην κλάση μειοψηφίας.

Στην μέθοδο Adaptive Synthetic Smote επιχειρείται η κατασκευή συνθετικών παραδειγμάτων δίνοντας βαρύτητα στα παραδείγματα της κλάσης μειοψηφίας που είναι ”δύσκολο” για τον ταξινομητή να μάθει να τα αναγνωρίζει. Ο βαθμός αυτής της δυσκολίας καθορίζεται βάσει των γειτόνων της κάθε καταχώρησης: υπολογίζεται η κατανομή των γειτόνων σε παραδείγματα κλάσεως μειοψηφίας έναντι κλάσεως πλειοψηφίας. Όσο περισσότερο υπερισχύει η κλάση πλειοψηφίας τόσο περισσότερα συνθετικά παραδείγματα κατασκευάζονται. Για την κατασκευή τους λαμβάνονται υπόψιν μόνο οι συντεταγμένες των γειτόνων που ανήκουν στην κλάση μειοψηφίας, όπως και στη Borderline-SMOTE μέθοδο.

3.5 Η βιβλιοθήκη Scikit-Learn

Το Scikit-Learn είναι μια Python βιβλιοθήκη για μηχανική μάθηση ανοιχτού κώδικα [1]. Μεταξύ άλλων περιλαμβάνει αλγορίθμους εκμάθησης για προβλήματα κατηγοριοποίησης *(classification problems)*, συσταδοποίηση *(clustering)*, επιλογή μοντέλου *(model selection)* [42] [43]. Περιλαμβάνονται όλες οι μέθοδοι εκμάθησης που παρουσιάστηκαν στην αντίστοιχη ενότητα, οι οποίες είναι και οι εφαρμογές που χρησιμοποιούμε. Επιπλέον της καθαυτής Scikit-Learn βιβλιοθήκης έχει δημιουργηθεί ένα Github αρχείο με αλγορίθμους εξισορρόπησης για ανισόρροπα δε-

[1] http://scikit-learn.org/stable/

δομένα εκπαίδευσης, που συνδέονται με το Scikit-Learn [2]. Εκεί υπάρχουν, μεταξύ πολλών άλλων, οι τεχνικές εξισορρόπησης που αναφέρθηκαν στην προηγούμενη ενότητα. Στις ιστοσελίδες που επισημαίνονται διατίθεται επίσης αρκετό εκπαιδευτικό υλικό και παραδείγματα χρήσης.

[2] https://github.com/scikit-learn-contrib/imbalanced-learn

Κεφάλαιο 4

Διαγωνιζόμενα Συστήματα στον διαγωνισμό PAN'11

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα παρουσιάσουμε τις 4 συμμετοχές στο διαγωνισμό του PAN 2011 για την εγγενή ανίχνευση λογοκλοπής, μιας και είναι το σώμα δεδομένων αυτού του διαγωνισμού που χρησιμοποιούμε και, άρα, τα αποτελέσματα του συστήματός μας είναι άμεσα συγκρίσιμα με αυτών των συμμετοχών. Τα συστήματα θα παρουσιαστούν με φθίνουσα σειρά επιτυχίας, ενώ τα αποτελέσματα θα παρουσιαστούν και θα σχολιαστούν ξεχωριστά. Κύριες πηγές μας είναι οι αναφορές των ίδιων των συμμετεχόντων Oberreuter et al. [44], Kestemont et al. [1], Akiva [45], Rao et al. [46] αλλά και το survey paper από του διοργανωτές του διαγωνισμού Potthast et al. [8].

4.2 Σύστημα των *Oberreuter et al.*

Η σκέψη πίσω από την μοντελοποίηση του συστήματος των Oberreuter et al. [44] και, κυρίως, για την επιλογή και σχεδιασμό του στυλιστικού χαρακτηριστικού ήταν η εξής, όπως περιγράφεται από τους ίδιους: εάν κάποιες από τις λέξεις που χρησιμοποιούνται στο κείμενο αναδεικνύουν την μοναδικότητα του συγγραφέα, τότε μπορούμε να υποθέσουμε ότι αυτές οι λέξεις θα είναι συγκεντρωμένες στις παραγράφους (ή, γενικότερα, στα χωρία του κειμένου) που ο αναφερόμενος ως συγγραφέας πράγματι έγραψε.
Ακολουθείται λοιπόν μια λεξιλογική προσέγγιση.

Προ-επεξεργασία κειμένου
Αρχικά κάθε κείμενο προ-επεξεργάζεται απομακρύνοντας τα αλφαριθμητικά, και τους χαρακτήρες που δεν ανήκουν στην ομάδα a-z. Όλοι οι χαρακτήρες θεωρούνται πεζοί, ενώ τα stopwords *δεν* απομακρύνονται.
Για την κατάτμηση του κειμένου χρησιμοποιείται η μέθοδος του μετακινούμενου

παραθύρου με μέγεθος παραθύρου,*m*, 400 λέξεων. Αναφέρεται δε, πως οι παράμετροι της κατάτμησης προσαρμόζονται ανάλογα με το μέγεθος του υπό εξέταση αρχείου-κειμ [1]ένου, χωρίς περαιτέρω πληροφορίες επ' αυτού. Σε επίπεδο συμβολισμού θεωρούμε πως το κείμενο κατατμείται σε κομμάτια *c* που ανήκουν στο σύνολο *C*.

Στυλιστική ανάλυση & αναγνώριση ύποπτων κομματιών

Το στυλιστικό χαρακτηριστικό βασίζεται στη συχνότητα όρου *(term frequency* ή *tf)*. Πρώτα υπολογίζεται για όλο το κείμενο ένας πίνακας, **v**, που περιλαμβάνει τη συχνότητα εμφάνισης των όρων-λέξεων που απέμειναν από την προ-επεξεργασία. Έπειτα, για κάθε κομμάτι κειμένου $c \in C$, υπολογίζεται νέος πίνακας συχνοτήτων, v_c, με τις συχνότητες εμφάνισης των λέξεων στο συγκεκριμένο κομμάτι. Έπειτα το χαρακτηριστικό υπολογίζεται για κάθε κομμάτι κειμένου όπως φαίνεται στον Αλγόριθμο 1. Για το σύνολο του κειμένου υπολογίζεται μια *style function*. Πρόκειται για το άθροισμα της τιμής του λεξιλογικού-στυλιστικού χαρακτηριστικού σε όλα τα κομμάτια κειμένου, διαιρώντας το με το πλήθος των κομματιών. Πρόκειται, δηλαδή, για έναν απλό μέσο όρο. Ύστερα, με βάση την απόκλιση της τιμής του χαρακτηριστικού σε κάθε κομμάτι κειμένου από αυτόν το μέσο όρο και ορίζοντας (αυθαίρετα) ένα κατώφλι, *δ*, σημειώνονται τα κομμάτια του κειμένου ως ύποπτα λογοκλοπής ή όχι. Όλα τα παραπάνω παρουσιάζονται σε σαφή βήματα στον Αλγόριθμο 1.

Algorithm 1: Intrinsic plagiarism evaluation, Oberreuter et al. [44]

Data: C, **v**, *m*, *δ*
for $c \in C$ **do**
  $d_c \leftarrow 0$;
  build v_c using frequencies on segment *c*;
  for *word* $w \in v_c$ **do**
    $d_c \leftarrow d_c + \frac{|freq(w,\mathbf{v}) - freq(w,v_c)|}{|freq(w,\mathbf{v}) + freq(w,v_c)|}$;
  end
end
$style \leftarrow \frac{1}{|C|} \sum_{c \in C} d_c$;
for $c \in C$ **do**
  if $d_c < style - \delta$ **then**
    mark segment c as outlier and potential plagiarised passage
  end
end

Βλέπουμε πως η τακτική εντοπισμού λογοκλεμμένων κομματιών στηρίζεται στη σύγκριση των τμημάτων του κειμένου με το στυλιστικό αποτύπωμα του συνόλου του κειμένου. Υπογραμμίζεται πω για τον υπολογισμό του στυλιστικού χαρακτηριστικού για κάθε κομμάτι κειμένου λαμβάνονται υπόψιν μόνο οι λέξεις που περιέχονται στο κομμάτι. Αν το χαρακτηριστικό, d_c, λαμβάνει μικρή τιμή τότε πι-

θανότατα θα βρεθεί να αποκλίνει πλέον της τιμής κατωφλίου από το σύνολο του κειμένου. Η μικρή τιμή σημαίνει πως η συχνότητα εμφάνισης των λέξεων στο συγκεκριμένο κομμάτι κειμένου είναι περίπου ίση με αυτή σε όλο το κείμενο και άρα, πιθανότατα, πρόκειται για λέξεις απομονωμένες στο συγκεκριμένο χωρίο, κάτι που συνιστά ένδειξη λογοκλοπής. Οι συγγραφική ομάδα αναφέρει, ακόμα, ότι οι παράμετροι του συστήματος (*m*, *δ*) καθορίζονται ανά κείμενο ανάλογα με το μήκος του, χωρίς, όμως, να δίνουν περισσότερες πληροφορίες.

4.3 Σύστημα των *Kestemont et al.*

Το σύστημα των Kestemont et al. [1] βασίζεται στα n-grams χαρακτήρων, συγκεκριμένα σε trigrams χαρακτήρων. Υπενθυμίζουμε τα trigrams χαρακτήρων *(character trigrams)* με το παράδειγμα που χρησιμοποιούν και οι ίδιοι, κατά την περιγραφή του συστήματός τους. Από τη λέξη *plagiarism* προκύπτει το εξής σύνολο από trigrams:
$\{$‘pla’, ‘lag’, ‘agi’, ‘gia’, ‘iar’, ‘ari’, ‘ris’, ‘ism’$\}$

Το καινοτόμο και πρωτότυπο στο σύστημα τους είναι πως δεν αντλούν τα trigrams χαρακτήρων από το σώμα των δεδομένων του διαγωνισμού. Εξετάζοντας ένα άλλο σώμα δεδομένων (για την ακρίβεια, αυτό του αντίστοιχου διαγωνισμού του PAN για το 2009), δημιούργησαν μια λίστα με τα πιο συχνά εμφανιζόμενα trigrams σε αυτό. Ύστερα χρησιμοποίησαν αυτή τη λίστα κατά τη στυλιστική ανάλυση των κομματιών του κειμένου, με άξονα τη συχνότητα εμφάνισής τους σε αυτά. Η επιλογή αυτή έγινε αρχικά λόγω πίστης στην αποτελεσματικότητα των n-grams χαρακτήρων σε προβλήματα στυλιστικής ανάλυσης και, τελικά, για λόγους εξοικονόμησης υπολογιστικού χρόνου, όπως θα εξηγηθεί παρακάτω. Πρωτοτύπησαν, ακόμα, και στην απόδοση στυλιστικού αποτυπώματος στα κομμάτια του κειμένου, μιας και η στυλιστική απεικόνιση καθενός γίνεται σε ένα πίνακα συσχετισμού όλων των κατατετμημένων χωρίων ανά δύο. Ας δούμε το σύστημα τους βήμα προς βήμα [1].

Προ-επεξεργασία κειμένου
Κατά την προ-επεξεργασία του κειμένου δε γίνεται καμιά ενέργεια πέραν της κατάτμησης του κειμένου. Για την τελευταία, εφαρμόζεται η μέθοδος του μετακινούμενου παραθύρου με *μ.π.* $= 5000$ *χαρακτήρες* και *β* $= 2.000$ *χαρακτήρες*.

Στυλιστική ανάλυση
Για τη στυλιστική ανάλυση δημιουργήθηκε και χρησιμοποιήθηκε μια λίστα με τα 2.500 πιο δημοφιλή trigrams χαρακτήρων του σώματος δεδομένων για το διαγωνισμό του PAN το 2010, τον αντίστοιχο διαγωνισμό, δηλαδή, δύο χρόνια πριν την εν λόγω συμμετοχή.
Έστω ότι προκύπτουν *n* χωρία από την κατάτμηση του κειμένου. Τότε αυτά τα *n* χωρία, μαζί με τη λίστα των trigrams χαρακτήρων, χρησιμοποιούνται για την κατασκευή ενός πίνακα συνδιασποράς *(covariance matrix)* διαστάσεων $n \times n$. Για ένα

κείμενο που έχει κατατμηθεί σε *n* ισομήκη παράθυρα-χωρία, $w_1, w_2, ..., w_{n-1}, w_n$, η απεικόνιση παρασταίνεται στον Πίνακα 4.1. Τα κυτία του πίνακα συνδιασποράς συμπληρώνονται με εφαρμογή της συνάρτησης απόστασης , *Δ*, που θα παρουσιάσουμε αμέσως μετά.

Πίνακας 4.1 Συμμετρικός πίνακας αποστάσεων για τη στυλιστική απεικόνιση ενός ύποπτου κειμένου, σύστημα Kestemont et al. [1]

	w_1	w_2	$\cdots$	w_{n-1}	w_n
w_1	0	$\Delta(w_1, w_2)$	$\cdots$	$\Delta(w_1, w_{n-1})$	$\Delta(w_1, w_n)$
w_2	$\Delta(w_2, w_1)$	0	$\cdots$	$\Delta(w_2, w_{n-1})$	$\Delta(w_2, w_n)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
w_{n-1}	$\Delta(w_{n-1}, w_1)$	$\Delta(w_{n-1}, w_2)$	$\cdots$	0	$\Delta(w_{n-1}, w_n)$
w_n	$\Delta(w_n, w_1)$	$\Delta(w_n, w_2)$	$\cdots$	$\Delta(w_n, w_{n-1})$	0

Η συνάρτηση απόστασης, *Δ*

Όπως φαίνεται και στον Πίνακα 4.1, οι Kestemont et al. [1] επιλέγουν να αποδώσουν αποστάσεις στα κομμάτια του κειμένου, συγκρίνοντάς τα στυλιστικά μεταξύ τους και όχι με ολόκληρο το κείμενο. Υποστηρίζουν, μάλιστα, πως η σύγκριση ενός κομματιού με όλο το κείμενο δεν μπορεί να στηριχθεί θεωρητικά, αφού οι στυλιστικοί δείκτες σχετίζονται άμεσα με το μέγεθος του υπό ανάλυση κειμένου, και φυσικά ένα κομμάτι κειμένου είναι κατά κανόνα πολύ μικρότερο σε σχέση με το σύνολο του κειμένου. Επιπλέον, επισημαίνουν πως μια τέτοια στρατηγική απόδοσης αποστάσεων στα κομμάτια, προϋποθέτει την επισφαλή υπόθεση πως το κείμενο έχει γραφτεί ως επί το πλείστον από τον ίδιο και φερόμενο ως αυθεντικό συγγραφέα, έτσι ώστε το στυλιστικό προφίλ του συνόλου του κειμένου να αναδεικνύει τις συγγραφικές επιλογές του ιδίου και να αποτελεί αξιόπιστο μέτρο σύγκρισης.

Οι ίδιοι έχουν ως βάση τη συνάρτηση απόστασης που εισήγαγε ο Stamatatos [47]. Πρόκειται για την *κανονικοποιημένη απόσταση (normalised distance* ή nd_1*)*, που χρησιμοποιείται συχνά στις στυλιστικές αναλύσεις με χρήση *character n-grams* για IPD συστήματα.

Σύμφωνα με τη σχέση του Stamatatos [47], για τον υπολογισμό της συνάρτησης απόστασης nd_1 για δύο παράθυρα w_x και w_y δημιουργείται η λίστα με όλα τα *n-grams* του παραθύρου w_x (αλλά όχι απαραίτητα του w_y). Αυτή η λίστα αποτελεί το προφίλ *(profile)* του w_x, $P(w_x)$. Με $|P(w_x)|$ συμβολίζεται το μέγεθος του $P(w_x)$, ή αλλιώς, το συνολικό πλήθος από *n-grams* στο w_x. Ο μαθηματικός τύπος υπολογισμού της λεγόμενης απόστασης μεταξύ των δύο παραθύρων δίνεται στη σχέση (4.1), όπου $f_{w_x}(g)$ η συχνότητα του n-gram g στο w_x:

$$nd_1(w_x, w_y) = \sum_{g \in P(w_x)} \frac{\left(\frac{2\left(f_{w_x}(g) - f_{w_y}(g)\right)}{f_{w_x}(g) + f_{w_y}(g)} \right)^2}{4|P(x)|} \tag{4.1}$$

Στη σχέση 4.1 η παρονομαστής $4|P(x)|$ εξασφαλίζει το κανονικοποιημένο σύνολο τιμών, με τις ακραίες τιμές να υποδηλώνουν πλήρης ομοιότητα (στην περίπτωση $nd_1 = 0$) και μηδενική ομοιότητα (στην περίπτωση $nd_1 = 1$).
Από την παραπάνω περιγραφή υπολογισμού της nd_1 βλέπουμε πως πρόκειται για μια υπολογιστικά ιδιαίτερα απαιτητική διαδικασία, αφού για κάθε παράθυρο υπολογίζουμε διαφορετική λίστα n-grams, η οποία θα αποτελέσει τη βάση για τη σύγκριση. Επιπλέον, πρόκειται για μη συμμετρική σχέση, αφού κατά τον υπολογισμό $nd_1(w_x, w_{y\neq x})$ λαμβάνονται υπόψιν τα n-grams που περιλαμβάνονται $P(w_x)$, οπότε ισχύει ότι $nd_1(w_x, w_{y\neq x}) \neq nd_1(w_y, w_{x\neq y})$. Ο ίδιος ο Stamatatos πρότεινε αυτή τη συνάρτηση απόστασης για σύγκριση χωρίου-συνολικού κειμένου [47], κι έτσι οι απαιτήσεις σε υπολογιστικό χρόνο είναι βιώσιμες.

Για να ξεπεράσουν το εμπόδιο της υψηλής χρονικής πολυπλοκότητας, κατασκεύασαν και χρησιμοποίησαν μια λίστα 2.500 trigrams ανεξάρτητων του σώματος δεδομένων στο οποίο θα εξεταζόταν το σύστημά τους. Όπως αναφέρθηκε πρόκειται για τα 2.500 δημοφιλέστερα trigrams στο σώμα δεδομένων του αντίστοιχου διαγωνισμού δύο χρόνια πριν τη δική τους συμμετοχή. Συγκρίνοντας λοιπόν τα παράθυρα με βάση αυτή την σταθερή λίστα από trigrams, αφενός δε χρειάζεται να υπολογίζουν νέα λίστα για κάθε παράθυρο, αφετέρου συνάρτηση απόστασης καθίσταται συμμετρική.
Έτσι, η σχέση 4.1, αν L η λίστα των trigrams, γίνεται:

$$\Delta(w_x, w_y) = \sum_{g\in L} \frac{\left(\frac{2\left(f_{w_x}(g) - f_{w_y}(g)\right)}{f_{w_x}(g) + f_{w_y}(g)}\right)^2}{4|L|} \tag{4.2}$$

Αναγνώριση ύποπτων χωρίων
Λόγω της επιλογής της στυλιστικής απεικόνισης με πίνακα μορφής του Πίνακα 4.1, στην περίπτωση μεγάλου κειμένου και κατάτμησής του σε σχετικά μικρά χωρία, ο πίνακας θα λαμβάνει μεγάλες διαστάσεις. Για το λόγο αυτό εφαρμόστηκε μια τεχνική αναγνώρισης έκτοπων σημείων σε μεγάλες διαστάσεις, όπως προτάθηκε από τους Filzmoser, Maronna και Werner [48], χρησιμοποιώντας το λογισμικό εργαλείο R. Η μέθοδος αυτή είναι υπολογιστικά αποδοτική, αφού μειώνει το μέγεθος των δεδομένων εφαρμόζοντας μια τεχνική της *Principal Components Analysis* (PCA), μία τεχνική που εφαρμόζεται στη στυλομετρία για μείωση διαστάσεων [49], ενώ μετά τη μείωση των διαστάσεων, για την αναγνώριση των έκτοπων σημείων, εφαρμόζεται η *Mahalanobis distance*. Ο ορισμός της απόστασης Mahalanobis δίνεται στην εξίσωση 4.3
Η απόσταση *Mahalanobis* μιας παρατήρησης $\mathbf{x} = (x_1, x_2, ..., x_n)$ από ένα σύνολο παρατηρήσεων με μέση τιμή $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_n)$ και πίνακα συνδιασποράς S είναι ίση με

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T S^{-1} (\mathbf{x} - \boldsymbol{\mu})} \tag{4.3}$$

4.4 Σύστημα του *Navot Akiva*

Το σύστημα του Akiva [45] ακολουθεί, επίσης, μια λεξιλογική προσέγγιση, εστιάζοντας στις σπανιότερα εμφανιζόμενες λέξεις του κειμένου, τις οποίες χρησιμοποιεί ως στυλιστικά σημάδια για να ξεχωρίσει τα ύποπτα χωρία του κειμένου.

Προ-επεξεργασία κειμένου & Στυλιστική ανάλυση
Η προ-επεξεργασία περιλαμβάνει την κατάτμηση του κειμένου , ενώ συνεχίζει με την ομαδοποίηση των κατατετμημένων κομματιών *(clustering)*, απ' όπου ξεκινάει και η στυλιστική ανάλυση.

Κατάτμηση κειμένου
Το κείμενο σπάει σε ισομήκη κομμάτια των 1000 χαρακτήρων, μη επικαλυπτόμενα. Έπειτα αναγνωρίζονται οι 100 πιο σπάνιες λέξεις του κειμένου, υπό την προϋπόθεση ότι εμφανίζονται σε ποσοστό τουλάχιστον 5% του συνόλου των κομματιών. Η ιδέα είναι να επιλεγούν 100 σπάνιες λέξεις αλλά όχι τόσο σπάνιες, έτσι ώστε να μην καθίστανται άχρηστες ως στυλιστικά κριτήρια. Κάθε κομμάτι απεικονίζεται από έναν πίνακα-στήλη μήκους 100 και περιλαμβάνει την πληροφορία για παρουσία ή απουσία καθεμίας από τις 100 λέξεις στο συγκεκριμένο κομμάτι.
Ύστερα τα κομμάτια συγκρίνονται μεταξύ τους ανά δύο εφαρμόζοντας την απόσταση συνημιτόνου 2.4.

Ομαδοποίηση κομματιών κειμένου
Σε δεύτερο βήμα χρησιμοποιείται μια τεχνική φασματικής ομαδοποίησης *(spectral clustering)*, που λέγεται *n-cut* [51], για την κατηγοριοποίηση των κομματιών σε ομάδες *(clusters)*.

Έπειτα από την ομαδοποίηση, εκτελείται μια δεύτερη διαδικασία ανάλυσης, αυτή τη φορά αναλύοντας τις ομάδες αλλά και τα χωρία των ομάδων. Για το σκοπό αυτό εξάγονται χαρακτηριστικά όπως

- σχετικό και απόλυτο μέγεθος ομάδας
- ομοιότητα ενός χωρίου με την ομάδα στην οποία ανήκει
- ομοιότητα ενός χωρίου με τις άλλες ομάδες
- ομοιότητα ενός χωρίου με το σύνολο του κειμένου

Όπως παρατηρεί κι ο Akiva [45], αναμένεται τα χωρία που απέχουν λίγο από το "κέντρο" της ομάδας τους, ενώ ταυτόχρονα απέχουν πολύ από το "κέντρο " του συνολικού κειμένου, να είναι πιθανότερο να περιλαμβάνουν λογοκλοπή.

Αναγνώριση ύποπτων κομματιών κειμένου
Αφού γίνει η ομαδοποίηση και η στυλιστική ανάλυση όπως περιγράφηκαν παραπάνω, επιχειρείται η αναγνώριση των χωρίων που πράγματι περιλαμβάνουν λογοκλοπή. Για το σκοπό αυτό ο Akiva [45] εκμεταλλεύεται τα δεδομένα εκπαίδευσης που παρέχονται. Κάθε κομμάτι των δεδομένων εκπαίδευσης απεικονίζεται από

ένα διάνυσμα με τις τιμές των χαρακτηριστικών που αναφέρθηκαν παραπάνω, ενώ του αποδίδεται και η κατάλληλη ετικέτα που το σημαδεύει ως *λογοκλεμμένο* ή *μη-λογοκλεμμένο*, ανάλογα με τις "λύσεις". Ύστερα, από αυτήν την ανάλυση, τα χαρακτηρισμένα δεδομένα εκπαίδευσης χρησιμοποιούνται σε μέθοδο επιβλεπόμενης μάθησης με *Δέντρα Απόφασης*, κάνοντας χρήση του λογισμικού WEKA [52].

4.5 Σύστημα των *Rao et al.*

Το σύστημα των Rao et al. [46] επιστρατεύει ένα πλήθος στυλιστικών χαρακτηριστικών, ενώ για τον εντοπισμό έκτοπων σημείων επιλέγεται η σύγκριση της στυλιστικής απεικόνισης των χωρίων με αυτή ολόκληρου του κειμένου.

Προ-επεξεργασία κειμένου
Για την κατάτμηση του κειμένου εφαρμόζεται η τεχνική του μετακινούμενου παραθύρου, με τιμές παραμέτρων *μ.π.* $= 2000$ *χαρακτήρες* και $\beta = 200$ *χαρακτήρες*. Δεν γίνεται λόγος για ενέργειες προ-επεξεργασίας του κειμένου, όμως από τα στυλιστικά χαρακτηριστικά που χρησιμοποιούνται, συμπεραίνουμε πως κάποια επεξεργασία είναι απαραίτητη (π.χ. αναγνώριση των θεμάτων των λέξεων).

Στυλιστική ανάλυση & Αναγνώριση ύποπτων χωρίων
Για τη στυλιστική απεικόνιση των κομματιών του κειμένου κατασκευάζεται ένα διάνυσμα που περιλαμβάνει τιμές ενός πλήθους χαρακτηριστικών:

- χαρακτηριστικό σχετιζόμενο με συχνότητα n-gram χαρακτήρων
  (frequent character n-gram based feature)
- συχνότητα διαφόρων αντωνυμιών
  (frequency of different pronouns)
- κλειστές κλάσεις λέξεων
  (closed class words)
- καταλήξεις θεμάτων λέξεων
  (stem suffixes)
- σημεία στίξης
  (punctuation marks)
- μέσο μήκος πρότασης
  (average length of statement)
- συχνότητα εμφάνισης κάποιων δεικτικών λέξεων ομιλίας
  (frequency of discourse markers)

Δεικτικές λέξεις ομιλίας ονομάζονται οι λέξεις που χρησιμοποιούνται σε γραπτό ή προφορικό λόγο και δεν έχουν ιδιαίτερη σημασιολογική αξία. Μπορεί να χρησιμοποιούνται για να γεμίσουν το λόγο, ενώ πολλές φορές προκύπτουν αυθόρμητα,

ως συνήθεια, από το συγγραφέα. Στα ελληνικά τέτοιες λέξεις είναι οι *βασικά, λοιπόν, έτσι*, ενώ στα αγγλικά οι *well, actually, then*.

Αφού λοιπόν τα κομμάτια του κειμένου αντιστοιχιστούν το καθένα με το διάνυσμα τιμών των χαρακτηριστικών, το τελευταίο περνάει από μια συνάρτηση απόστασης που δίνει την τιμή στυλιστικής απόκλισης του κομματιού από το σύνολο του κειμένου. Χρησιμοποιήθηκε η συνάρτηση απόστασης που προτάθηκε από τον Stamatatos [53] [47]:

$$d_1(A,B) = \sum_{g \in P(A)} \left(\frac{2\left(f_A(g) - f_B(g)\right)}{f_A(g) + f_B(g)} \right)^2 \tag{4.4}$$

όπου A, B είναι κανονικοποιημένα διανύσματα για τα χωρία και ολόκληρο το κείμενο, αντίστοιχα, και g οι διαφορετικές διαστάσεις των διανυσμάτων, που αντιστοιχούν στα στυλιστικά χαρακτηριστικά. Ως κατώφλι ορίστηκε η τιμή 2.0, δηλαδή για όποιο χωρίο προκύπτει $d_1 > 2.0$, τότε αυτό θεωρείται σημειώνεται ως λογοκλεμμένο.
Η κανονικοποιημένη απόσταση *(nd_1)* πού είδαμε στη σχέση 4.1 και έχει προταθεί, και πάλι, από τον Stamatatos [47] και αποτελεί ουσιαστικά την κανονικοποιημένη μορφή της σχέσης 4.4.

4.6 Παρουσίαση & Σχολιασμός Αποτελεσμάτων

Ο Πίνακας 4.2 συγκεντρώνει τα αποτελέσματα των τεσσάρων συμμετοχών στο διαγωνισμό [8]. Στα αποτελέσματα συμπεριλαμβάνεται, ακόμα, το σύστημα του νικητή του διαγωνισμού IPD, PAN 2009 [47], για το ίδιο σώμα δεδομένων. Για την αξιολόγηση υπολογίστηκαν οι μετρικές *precision, recall, F1-Score, granularity, plagdet*, οι οποίες ανταποκρίνονται στους ορισμούς που δώσαμε στο αντίστοιχο κεφάλαιο της αξιολόγησης ενός συστήματος εγγενούς λογοκλοπής(2.5).

Πίνακας 4.2 Αποτελέσματα διαγωνισμού στον PAN 2011 και του συστήματος-νικητή στον PAN 2009

	plagdet	*precision*	*recall*	*granularity*
Oberreuter et al. [44]	**0.33**	0.31	0.34	1.0
Kestemont et al. [1]	**0.17**	0.11	0.43	1.03
Akiva [45]	**0.08**	0.07	0.13	1.05
Rao et al. [46]	**0.07**	0.08	0.14	1.48
***Stamatatos* [47]**	***0.19***	*0.14*	*0.19*	*1.24*

Σημειώνεται πως τα αποτελέσματα του συστήματος που κέρδισε τον διαγωνισμό του 2009 στον Πίνακα 4.2, προέκυψαν για το σώμα δεδομένων του PAN 2011.

Ένας σημαντικός παράγοντας κατά τη σχεδίαση ενός συστήματος εγγενούς ανίχνευσης λογοκλοπής είναι το μέγεθος των υπό ανάλυση κειμένων. Επηρεάζονται άμεσα αφενός η επιλογή των παραμέτρων κατάτμησης του κειμένου, αφετέρου δε, η επιλογή της στυλιστικής ανάλυσης. Ακόμα, είναι προφανές ότι όσο μικρότερο το λογοκλεμμένο χωρίο τόσο δυσκολότερος καθίσταται ο εντοπισμός του. Αξίζει, λοιπόν, να δούμε την επίδοση των συστημάτων στις 3 κατηγορίες μεγέθους των κειμένων και των λογοκλεμμένων τμημάτων του σώματος δεδομένων. Η κατηγοριοποίηση έχει ως εξής [8]:

Πίνακας 4.3 Κατηγοριοποίηση σώματος δεδομένων PAN'11 ανάλογα με μέγεθος κειμένου, μέγεθος λογοκλεμμένου χωρίου

	Μέγεθος κειμένου	*Μέγεθος περίπτωσης λογοκλοπής*
μικρό	1- 10 σελ.	<150 λέξεις
μεσαίο	10 - 100 σελ.	150 - 1150 λέξεις
μεγάλο	100 - 1000 σελ.	>1150 λέξεις

Ο Πίνακας 4.4 περιλαμβάνει τα αποτελέσματα των συστημάτων για τις παραπάνω κατηγορίες [8]. Συμπεριλαμβάνονται και τα αποτελέσματα του συστήματος Stamatatos [47].

Πίνακας 4.4 Αποτελέσματα του PAN'11 και του συστήματος-νικητή στο PAN'09 στις υποκατηγορίες του σώματος δεδομένων

	plagdet					*precision*					*recall*					*granularity*				
	[13]	[15]	[14]	[16]	*[17]*	[13]	[15]	[14]	[16]	*[17]*	[13]	[15]	[14]	[16]	*[17]*	[13]	[15]	[14]	[16]	*[17]*
Μέγεθος λογοκλοπής																				
μικρό	**.03**	.01	.02	.01	*.01*	**.02**	.00	.01	.00	*.01*	**.16**	0.25	.09	.05	*.19*	**1.00**	1.00	1.00	1.01	*1.00*
μεσαίο	**.26**	.08	.04	.05	*.13*	**.19**	.05	.02	.03	*.08*	.45	.51	.12	.13	***.57***	**1.00**	1.00	1.01	1.05	*1.01*
μεγάλο	**.36**	.19	.11	.08	*.14*	**.26**	.12	.08	.11	*.12*	.57	.74	.25	.21	***.63***	**1.00**	1.12	1.15	2.16	*1.77*
Μέγεθος κειμένου																				
μικρό	**.38**	.20	.10	.06	*.21*	**.37**	.13	.07	.08	*.34*	**.38**	.55	.18	.11	*.16*	**1.00**	1.06	1.05	1.81	*1.00*
μεσαίο	**.40**	.28	.13	.10	*.28*	**.44**	.21	.07	.17	*.23*	.37	.47	.16	.11	***.48***	**1.00**	1.03	1.06	1.43	*1.17*
μεγάλο	**.28**	.17	.04	.12	*.18*	**.32**	.13	.11	.16	*.13*	.25	.24	.03	.10	***.53***	**1.00**	1.00	1.00	1.07	*1.33*

Όπως αναμέναμε, τα αποτελέσματα στα μικρά χωρία λογοκλοπής είναι σχεδόν μηδενικά.

Ο σχολιασμός που ακολουθεί έχει ως σημείο αναφοράς τα αποτελέσματα που αφορούν στο σύνολο του σώματος δεδομένων, δηλαδή τον Πίνακα 4.2.

Βλέπουμε πως ο νικητής του διαγωνισμού σημειώνει σχεδόν διπλάσιο σκορ στην μετρική *plagdet* από το δεύτερο καλύτερο σύστημα. Η διαφορά αυτή οφεί-

λεται στην υψηλή τιμή του *precision*, όπου κατάφερε σχεδόν τριπλάσιο σκορ από τον δεύτερο στη σειρά. Το σύστημα των Kestemont et al. [1] παρουσιάζει το καλύτερο recall αλλά χάνει εξαιτίας της χαμηλής τιμής στο precision· συμπεραίνεται πως το σύστημά τους έκανε πολλές *λάθος θετικές* προβλέψεις. Το σύστημα των Rao et.al [46] χάνει *και* λόγω υψηλής τιμής *granularity*, κάτι που θα μπορούσε εύκολα να ξεπεραστεί συνενώνοντας τα επικαλυπτόμενα κομμάτια μετά τις προβλέψεις.

Το σύστημα των *Oberreuter et al.* [44] θεωρήθηκε ιδιαίτερα επιτυχημένο, από άποψη τιμών αποτελεσμάτων. Έτσι, με μια πρώτη ματιά, τα αποτελέσματα φαίνονται ελπιδοφόρα για την εξέλιξη της έρευνας, όμως από πολλούς έχουν θεωρηθεί πλασματικά. Το σύστημα των νικητών του διαγωνισμού, βασίζεται σε ένα λεξιλογικό χαρακτηριστικό. Συγκεκριμένα, υπολογίζει την (στυλιστική) απόσταση ενός χωρίου από ολόκληρο το κείμενο με άξονα τη συχνότητα εμφάνισης των λέξεων μέσα σε αυτό σε σχέση με την αντίστοιχη συχνότητα στο κείμενο. Ουσιαστικά, το σύστημα βασίζεται στη μοναδικότητα των λέξεων στα κατατετμημένα χωρία σε σχέση σε με ολόκληρο το κείμενο, ώστε να τα χαρακτηρίσει ως λογοκλεμμένα. Το σώμα δεδομένων του διαγωνισμού κατασκευάστηκε με την προσθήκη τυχαίων αποσπασμάτων σε κείμενα που αρχικά αποτελούσαν ένα άρτιο και ”καθαρό” από λογοκλοπή σύνολο. Η τυχαιότητα της επιλογής και προσθήκης των διαφόρων αποσπασμάτων από ποικίλες πηγές είχε ως αποτέλεσμα, τα λογοκλεμμένα χωρία που περιλαμβάνονται στα κείμενα να είναι διαφορετικής θεματολογίας από ό,τι το υπόλοιπο κείμενο. Αυτό θεωρήθηκε ως αδυναμία του σώματος δεδομένων και μη ρεαλιστική εκδοχή λογοκλοπής [8] [54].

Το σύστημα των *Kestemont et al.* [1] κατάφερε αξιοπρεπή εμφάνιση. Πιστεύουμε πως επιλογή για στυλιστική σύγκριση των χωρίων μεταξύ τους και όχι με το σύνολο του κειμένου είναι ενδιαφέρουσα. Από την άλλη μεριά, η κατασκευή μιας λίστας από trigrams χαρακτήρων αντλούμενα από εξωτερική πηγή, μας φαίνεται μάλλον ατυχής. Όταν η βάση της στυλιστικής, λεξιλογικής, σημασιολογικής ανάλυσης ενός κειμένου δεν έχει ως σημείο αναφοράς το ίδιο το κείμενο, το πιθανότερο, κατά τη γνώμη μας, είναι να αποτύχει. Διαφορετικά, θα είχαμε βρει το παντοδύναμο χαρακτηριστικό που είναι πανάκεια για κάθε ύποπτο κείμενο ή σώμα δεδομένων. Το σύστημα των Kestemont et al. [1] μπορεί να συγκριθεί με αυτό του Stamatatos(2009) [47], αφού και οι δύο επιστρατεύουν τα trigrams χαρακτήρων για τη στυλιστική τους ανάλυση. Η κύριες διαφορές είναι πρώτον, ότι ο Stamatatos βρίσκει όλα τα trigrams που περιλαμβάνονται σε κάθε χωρίο και δεύτερον, ότι συγκρίνει την απόσταση κάθε χωρίο με όλο το κείμενο. Βλέπουμε ότι το σύστημα του Stamatatos είναι πιο αποτελεσματικό, όμως δεν μπορούμε να γνωρίζουμε πού ακριβώς οφείλεται αυτό. Κατά τη γνώμη μας, θα είχε ενδιαφέρον να εξεταστεί το σύστημα των Kestemont et al. [1] χωρίς την σταθερή λίστα trigrams, αλλά με δημιουργία, για παράδειγμα, μιας λίστας με τα πιο δημοφιλή trigrams σε κάθε κατατετμημένο χωρίο του κειμένου.
Ακόμα, κατά την αποσύμπλεξη των επικαλυπτόμενων παραθύρων κατάτμησης επιλέχθηκε ένα πακέτο χαρακτήρων κειμένου να θεωρούνται λογοκλεμμένα εάν του-

λάχιστον ένα παράθυρο κατάτμησης που το καλύπτει έχει προβλεφθεί ως λογοκλεμμένο. Βλέπουμε πως αυτή η επιλογή, που είναι εις βάρος της ακρίβειας πρόβλεψης, ενώ ταυτόχρονα ωθεί την ανάκληση, ζημίωσε το σύστημα με μια μη ανταγωνιστική τιμή ακρίβειας.

Ο *Akiva* [45] επιλέγει να ακολουθήσει την προσέγγιση της ομαδοποίησης των χωρίων, η οποία είναι περισσότερο δημοφιλής σε συστήματα εξωγενούς ανίχνευσης. Η κατάτμηση του κειμένου αποτελείται από δύο στάδια, σε μια (φιλόδοξη) προσπάθεια να βελτιώσει το σπάσιμο του κειμένου και να μπορέσει με την ομαδοποίηση να πάρει μεγαλύτερα κομμάτια-ομάδες ”καθαρού” κειμένου, όσο το δυνατόν πλησιέστερα στο πλήρως λογοκλεμμένο ή αυθεντικό. Ενώ το κίνητρο ανταποκρίνεται σε ένα υπαρκτό και πολύ σοβαρό πρόβλημα των συστημάτων εγγενούς ανίχνευσης λογοκλοπής, η μέθοδος ομαδοποίησης δε μας φαίνεται ιδιαίτερα επιτυχημένη. Το στυλιστικό κριτήριο που επιλέγεται, το οποίο βασίζεται στη συχνότητα εμφάνισης των πιο σπάνιων λέξεων, δεν φαίνεται ισχυρό.

Οι *Rao et al.* [46] επιστράτευσαν ένα πλήθος στυλιστικών χαρακτηριστικών. Δεν έχουμε καθαρή εικόνα για τις συγκεκριμένες μεθόδους εξαγωγής των χαρακτηριστικών αυτών. Πιστεύουμε πως το βασικό πρόβλημα του συστήματος εντοπίζεται στη μέθοδο ανίχνευσης των έκτοπων σημείων, και όχι στη στυλιστική ανάλυση. Καταρχάς για συνάρτηση απόστασης χρησιμοποιείται η μη κανονικοποιημένη συνάρτηση απόστασης που προτάθηκε από τον Stamatatos 4.4. Θεωρούμε πως η κανονικοποίηση των τιμών των στυλιστικών αποτυπωμάτων είναι σημαντικότατη και αποτελεί προϋπόθεση για την επιλογή κατάλληλου κριτηρίου διαχωρισμού των έκτοπων σημείων, μέσα σε ένα σώμα κειμένων ποικίλου μεγέθους. Συνεχίζοντας, χωρίς να δίνουν ιδιαίτερη σημασία σε αυτό το κομμάτι του συστήματος, ορίζουν οι ίδιοι, αυθαίρετα, ένα κατώφλι πέρα από το οποίο ένα χωρίο θα σημαδεύεται ως λογοκλεμμένο. Αυτό είναι άλλο ένα μεγάλο λάθος, από τη στιγμή που μπορούν να χρησιμοποιηθούν αλγόριθμοι εκμάθησης για την εύρεση βέλτιστης τιμής κατωφλίου.

Συνοψίζοντας, θεωρούμε πως το πρώτο σύστημα αξίζει να εξεταστεί και σε άλλο σώμα δεδομένων, ώστε να ελεγχθεί η ευρωστία του. Όσον αφορά στα υπόλοιπα τρία συστήματα, πιστεύουμε πως έγιναν κάποια σχεδιαστικά λάθη - αλλού περισσότερο σοβαρά και εμφανή, αλλού λιγότερο - η αποκατάσταση των οποίων θα μπορούσε να οδηγήσει σε αξιόλογη βελτίωση των αποτελεσμάτων.

Ως γενικότερο σχόλιο, θεωρούμε πως η τακτική της συμπύκνωσης των στυλιστικών χαρακτηριστικών σε μια μοναδική τιμή μέσω της συνάρτησης απόστασης είναι περιοριστική. Αντ’ αυτού προτείνουμε τη σύγκριση χωρίου-κειμένου κάθε στυλιστικού χαρακτηριστικού ξεχωριστά και την εφαρμογή μοντέλου μηχανικής μάθησης με είσοδο το διάνυσμα των στυλιστικών χαρακτηριστικών, έτσι ώστε να είναι δυνατή η ανακάλυψη αφανών συσχετισμών μεταξύ των χαρακτηριστικών.

Κεφάλαιο 5

Κατασκευή Συστήματος Εγγενούς Ανίχνευσης Λογοκλοπής

5.1 Προεπεξεργασία Κειμένου

Ως προ-επεξεργασία θεωρούμε την καθεαυτήν επεξεργασία του κειμένου, απ' όπου εξάγουμε τα δεδομένα που θα χρειαστούν για τα στυλιστικά χαρακτηριστικά, αλλά και την κατάτμηση του κειμένου.

Κατάτμηση του κειμένου
Εφαρμόσαμε τη μέθοδο του *μετακινούμενου παραθύρου*. Συνήθως το μέγεθος και το βήμα παραθύρου επιλέγονται σε επίπεδο χαρακτήρων. Χαρακτηριστικές τιμές είναι, για παράδειγμα, *μ.π. = 2000 χαρακτήρες*, *β = 500 χαρακτήρες*. Παρολαυτά θεωρήσαμε πιο ορθόδοξη την επιλογή σε επίπεδο προτάσεων, μιας και τα λογοκλεμμένα κομμάτια αποτελούν πακέτα προτάσεων που παρεμβάλλονται μεταξύ σημείων στίξης του "κανονικού" κειμένου. Εφαρμόσαμε τρεις διαφορετικές παραμετροποιήσεις, οι δύο με σταθερά και η μία με μεταβλητά τα μήκος και βήμα παραθύρου, ως εξής :

1. *μ.π.* $= 15$ προτάσεις, $\beta = 5$ προτάσεις
2. *μ.π.* $= 30$ προτάσεις, $\beta = 10$ προτάσεις
3. *μ.π.* $= 3k$ προτάσεις και $\beta = k$, όπου το k παίρνει τιμές ανάλογα με το μέγεθος του κειμένου ως εξής:
   - $k = 3$ για *μικρό* μέγεθος κειμένου ($\leq 30kB$)
   - $k = 5$ για *μεσαίο* μέγεθος κειμένου ($\leq 300kB$)
   - $k = 10$ για *μεγάλο* μέγεθος κειμένου ($> 300kB$)

Επεξεργασία κειμένου προ της στυλιστικής ανάλυσης
Η προεπεξεργαστικά βήματα που μας χρειάζονται για την μετέπειτα στυλιστική ανάλυση είναι τα εξής:

1. μετατροπή κεφαλαίων σε πεζά
2. αναγνώριση προτάσεων
3. αναγνώριση λέξεων
4. απομάκρυνση αλφαριθμητικών και ειδικών χαρακτήρων
5. αναγνώριση των λέξεων ως μερών-του-λόγου
6. αναγνώριση των θεμάτων των λέξεων

Τα *αναγνώριση προτάσεων, αναγνώριση των λέξεων ως μερών-του-λόγου, αναγνώριση των θεμάτων των λέξεων* αποτελούν από μόνα τους ανοιχτά προβλήματα στον τομέα της επεξεργασίας φυσικής γλώσσας. Για αυτά, καθώς και για το *token detection*, χρησιμοποιήσαμε το εργαλείο *OpenNLP* της Apache [55]. Πρόκειται για μία Java βιβλιοθήκη, ελεύθερη προς χρήση και ανοιχτού κώδικα, που βασίζεται σε μεθόδους μηχανικής μάθησης για την επίλυση προβλημάτων επεξεργασίας φυσικής γλώσσας. Η βιβλιοθήκη μπορεί να χρησιμοποιηθεί μέσω γραμμής εντολών ή να ενσωματωθεί στο Java project. Εμείς εργαστήκαμε ενσωματώνοντάς την στο project.
Το OpenNLP δουλεύει αξιόπιστα, με μικρά ποσοστά σφάλματος, όχι όμως τέλεια. Κατά συνέπεια γίνονται κάποια λάθη κατά την προεπεξεργασία, τα οποία όμως δεν αποτελούν σοβαρό κίνδυνο και δεν επηρεάζουν σημαντικά την αποτελεσματικότητα του συστήματος. Τέτοια λάθη μπορεί να είναι η εσφαλμένη αναγνώριση δύο προτάσεων ως μία, ή η λάθος εύρεση θέματος λέξης σε περίπλοκες περιπτώσεις. Λόγω της περιορισμένης έκτασής τους, δε λάβαμε υπόψιν αυτά τα σφάλματα κατά την αξιολόγηση του συστήματος.

Για την απομάκρυνση των stopwords χρησιμοποιήσαμε την ευρέως χρησιμοποιούμενη λίστα των 429 λέξεων [56].

5.2 Στυλιστική Ανάλυση

Για τη στυλιστική ανάλυση εξάγονται 11 χαρακτηριστικά. Για τον εντοπισμό ανομοιομορφιών επιλέξαμε τη σύγκριση κάθε χωρίου με το σύνολο του κειμένου. Τα χαρακτηριστικά είναι τα εξής και θα αναλυθούν στη συνέχεια:

1. μέσο μήκος πρότασης *(average sentence length)*
2. μέσο πλήθος συλλαβών ανά λέξη *(average syllable count of each token)*
3. δείκτης δυσκολίας ανάγνωσης Flesch-Kinkaid *(Flesch-Kinkaid grade)*
4. συχνότητα της λέξης *of (frequency of word ”of”)*
5. ποσοστό συμπίεσης ρημάτων *(verbs’ compression rate)*
6. ποσοστό συμπίεσης επιρρημάτων *(adverbs’ compression rate)*

7. ποσοστό συμπίεσης επιθέτων *(adjectives' compression rate)*

8. μέσος όρος των θετικών διαφορών των κλάσεων συχνοτήτων λέξεων κειμένου-χωρίου *(Σημασιολογικό χαρακτηριστικό No.1 ή wfc_1)*

9. τυπική απόκλιση των θετικών διαφορών των κλάσεων συχνοτήτων λέξεων κειμένου-χωρίου. *(Σημασιολογικό χαρακτηριστικό No.2 ή wfc_2)*

10. ποσοστό λέξεων που εμφανίζουν θετική διαφορά κλάσης συχνότητας μεταξύ κειμένου-χωρίου πάνω από το μέσο όρο *(Σημασιολογικό χαρακτηριστικό No.3 ή wfc_3)*

11. μέσος όρος των αρνητικών διαφορών των κλάσεων συχνοτήτων λέξεων κειμένου-χωρίου για τις συχνά εμφανιζόμενες, σε όλο το κείμενο, λέξεις *(Σημασιολογικό χαρακτηριστικό No.4 ή wfc_4)*

Αποτύπωση ανομοιομορφίας και Συναρτήσεις απόστασης
Όταν το στυλιστικό αποτύπωμα αποδίδεται ως ένα διάνυσμα τιμών, συνηθίζεται να επιλέγεται μια συνάρτηση απόστασης, η οποία παίρνει ως είσοδο τα διανύσματα των δύο συγκρινόμενων μερών - δύο χωρία ή ένα χωρίο και το σύνολο του κειμένου - και δίνει την μεταξύ τους απόσταση ως μία τιμή, που εκφράζει και την μεταξύ τους ανομοιομορφία. Εμείς, ακριβώς για το λόγο ότι χρησιμοποιούμε μηχανική μάθηση για την αναγνώριση των έκτοπων σημείων, επιλέξαμε η τελική τιμή κάθε στυλιστικού χαρακτηριστικού χωρίου να φέρει μέσα της την σύγκριση με το σύνολο του κειμένου κι έτσι να κρατάμε όλο το διάνυσμα των χαρακτηριστικών ως στυλιστικό αποτύπωμα ενός χωρίου. Με αυτόν τον τρόπο αφήνουμε περισσότερο χώρο δράσης στις τεχνικές εκμάθησης του επόμενου σταδίου, ώστε να αποδώσουν τα κατάλληλα βάρη ανάλογα με τη σημαντικότητα κάθε χαρακτηριστικού αλλά και να "ανακαλύψουν" τους μεταξύ τους συσχετισμούς.
Μιας και τα χαρακτηριστικά $1-7$ αντιστοιχούν το καθένα σε μια πραγματική τιμή, η σύγκριση ενός χωρίου με το σύνολο του κειμένου δε θα μπορούσε να γίνει, παρά με τη μόνη συνάρτηση απόστασης για δύο πραγματικές τιμές, δηλαδή την διαφορά τους. Όσον αφορά στα λεξιλογικά-σημασιολογικά χαρακτηριστικά $8-11$, αυτά αποτυπώνουν εκ κατασκευής την ανομοιομορφία με το σύνολο του κειμένου και, επομένως, δε χρειάζεται περαιτέρω σύγκριση.

Κανονικοποίηση χαρακτηριστικών
Όσα χαρακτηριστικά, στυλιστικά και σημασιολογικά, δεν λαμβάνουν εκ κατασκευής τιμές στο εύρος [-1,1], κανονικοποιούνται με εφαρμογή της συνάρτησης 5.1 :

$$\text{normalise}(x) = \frac{x}{1+|x|} \tag{5.1}$$

Ανάλυση των στυλιστικών χαρακτηριστικών
Η διαδικασία υπολογισμού των χαρακτηριστικών $1-4$ είναι προφανής. Θα παρουσιάσουμε την διαδικασία υπολογισμού των χαρακτηριστικών $5-7$. Έπειτα θα

εστιάσουμε στα σημασιολογικά χαρακτηριστικά της στυλιστικής ανάλυσης.

*Ποσοστό συμπίεσης κλάσης λέξεων.*Η ιδέα μας για τα χαρακτηριστικά συμπίεσης προήλθε από τη δουλειά των Seaward και Stan [57], οι οποίοι πρότειναν στυλιστικές μετρικές σχετιζόμενες με την πολυπλοκότητα Kolmogorov *(Kolmogorov Complexity)*. Η βασική τους ιδέα, και την οποία υιοθετήσαμε, είναι πως κάθε τμήμα ενός κειμένου παρουσιάζει μια κατανομή για κάθε κλάση λέξεων. Για παράδειγμα, παίρνοντας την κλάση *ουσιαστικό*, ένα κομμάτι κειμένου θα παρουσιάζει μια κατανομή για τις λέξεις που είναι ή δεν είναι ουσιαστικά, την οποία μπορούμε να φανταστούμε ως μια ακολουθία από 0 και 1. Το ίδιο μπορεί να γίνει και με άλλες κλάσεις λέξεων. Ας πάρουμε για παράδειγμα τις μικρές και μεγάλου μήκους λέξεις · έστω ότι αναπαριστούμε με 0 τις μικρές και με 1 τις μεγάλες, και πως δύο προτάσεις δίνουν τις εξής δυαδικές ακολουθίες:

$$0100001111010000100011110100000001$$
$$0000000011111000000011100000011111$$

Οι δύο προτάσεις έχουν ίδιο αριθμό μεγάλων - μικρών λέξεων αλλά αυτές κατανέμονται τελείως διαφορετικά, με την πρώτη να παρουσιάζει περισσότερο τυχαία και πολύπλοκη κατανομή [57].
Οι Seaward και Stan [57] προτείνουν την εξαγωγή τέτοιων δυαδικών ακολουθιών για διάφορες κλάσεις λέξεων και ύστερα τη συμπίεση τους με κάποιον μη απωλεστικό αλγόριθμο συμπίεσης *(lossless compression algorithm)* και, τέλος, τον υπολογισμό της στυλιστικής μετρικής ως ποσοστού συμπίεσης με τη βοήθεια κάποιας σχέσης παρόμοιας με την 2.1. Στη δική μας εφαρμογή αυτής της ιδέας εργαστήκαμε με τα ρήματα, τα επιρρήματα και τα επίθετα. Απαραίτητη για την εξαγωγή αυτών των χαρακτηριστικών ήταν η πρόσδοση ετικετών μερών-του-λόγου στις λέξεις του κειμένου, η οποία έγινε με τη βοήθεια του POS-Tagger των βιβλιοθηκών OpenNLP της Apache. Συνεχίζοντας, για κάθε τέτοια ακολουθία εφαρμόστηκε η απλή κωδικοποίηση run-length *run-length encoding*. Ένα παράδειγμα δίνεται παρακάτω, όπου την δυαδική ακολουθία ακολουθεί η κωδικοποίησή της:

δυαδική ακολουθία: 0011111000011110
κωδικοποίηση run-length: **2***0***5***1***4***0***4***1***1***0***

Η στυλιστική ανάλυση ολοκληρώνεται υπολογίζοντας το λόγο 5.2

$$\text{Ποσοστό συμπίεσης } = \frac{\text{Μήκος κωδικοποίησηςρRun-length}}{\text{Μήκος δυαδικής ακολουθίας}} \tag{5.2}$$

ενώ, η τελική τιμή που καταχωρείται στο διάνυσμα τιμών του χωρίου προκύπτει από τη σύγκρισή με το σύνολο του κειμένου, όπως εξηγήσαμε παραπάνω.

Λεξιλογικά-Σημασιολογικά χαρακτηριστικά. Τα χαρακτηριστικά $8 - 11$ έχουν ως πυρήνα την έννοια της *κλάσης συχνότητας λέξης (word frequency class)*. Πρόκειται, ουσιαστικά, για δείκτη της σπανιότητας μιας λέξης μέσα στο κείμενο. Τα

σημασιολογικά χαρακτηριστικά και, άρα, η κλάση συχνότητας λέξης έπαιξαν καθοριστικό ρόλο στην αποτελεσματικότητα του συστήματός μας. Ο υπολογισμός της "κλάσης" για κάθε λέξη περιγράφεται σε φυσική γλώσσα , ακολουθώντας την περιγραφή των Stein & Eissen [7].

Κλάση συχνότητας λέξης
Έστω C ένα σώμα δεδομένων και $|C|$ το πλήθος των λέξεων μέσα στο C. Ακόμα, έστω $f(w)$ η συχνότητα εμφάνισης μιας λέξης $w \in C$. Η κλάση συχνότητας, $c(w)$, για τη λέξη $w \in C$, ορίζεται ως το πηλίκο

$$c(w) = log_2 \left(\frac{f(w^*)}{f(w)} \right) \tag{5.3}$$

όπου w^* η πιο συχνά χρησιμοποιούμενη λέξη στο C.
Η κλάση συχνότητας λέξης, λοιπόν, αναδεικνύει τη σπανιότητα μιας λέξης μέσα στο κείμενο, μέσω της σύγκρισης με τις υπόλοιπες λέξεις του κειμένου. Από την σχέση (5.1) προκύπτει ότι λαμβάνει τιμές θετικές. Την τιμή 0 κατέχει η πιο συχνά χρησιμοποιούμενη λέξη, ενώ υψηλές τιμές δηλώνουν σπανιότητα εμφάνισης της λέξης μέσα στο κείμενο.

Συνεχίζουμε με την παρουσίαση και τον σχολιασμό των σημασιολογικών - λεξιλογικών χαρακτηριστικών.Σημειώνουμε πως τα προ-επεξεργαστικά βήματα 4 και 6 που αναφέρονται παραπάνω *(απομάκρυνση των stopwords, αναγνώριση θεμάτων των λέξεων)* μας χρειάζονται και εφαρμόζονται μόνο για την εξαγωγή των σημασιολογικών χαρακτηριστικών.
Στην παρουσίαση των αλγορίθμων που δίνουν τα σημασιολογικά χαρακτηριστικά, χρησιμοποιείται ο εξής συμβολισμός:

- WT *(Whole Text)* $=$ ολόκληρο το κείμενο
- P *(Passage)* $=$ "παράθυρο" από την κατάτμηση του κειμένου
- w *(Word)* $=$ λέξη (που έχει υποστεί stemming)
- $c(WT, w)$ *(class of w in WT)* $=$ word frequency class της λέξης *w* σε ολόκληρο το κείμενο
- $c(P, w)$ *(class of w in P)* $=$ word frequency class της λέξης *w* στο passage P

Για την εξαγωγή των σημασιολογικών-λεξιλογικών χαρακτηριστικών, αντικείμενο εργασίας ήταν τα θέματα των λέξεων του κειμένου, μιας και το θέμα μιας λέξης είναι αυτό που καθορίζει τη σημασιολογία του. Στο εξής όταν μιλάμε για λέξη εννοούμε το θέμα της λέξης.
Αρχική ιδέα ήταν ο εντοπισμός σημασιολογικών διαφορών μεταξύ των ύποπτων χωρίων του κειμένου από τη μία, και του συνόλου του κειμένου από την άλλη. Η σκέψη είναι πως τα στατιστικά που εξάγονται για το σύνολο του κειμένου θα

απεικονίζουν τις λεξιλογικές επιλογές του συγγραφέα, ακόμα κι αν είναι, ενδεχομένως, ”νοθευμένα” λόγω της παρουσίας κάποιων λογοκλεμμένων κομματιών. Από την άλλη, χωρία του κειμένου, που προκύπτουν από την κατάτμησή του, τα οποία αποτελούνται σε σημαντικό ποσοστό από λογοκλεμμένα κομμάτια, θα αποκλίνουν στατιστικά από το σύνολο του κειμένου, αφού θα αναδεικνύουν τις λεξιλογικές επιλογές των αυθεντικών συγγραφέων τους.
Για την πραγματοποίηση μιας τέτοιας ανάλυσης θα θέλαμε να γνωρίζουμε πράγματα όπως τις λέξεις που ”λάμπουν” περισσότερο μέσα σε ένα χωρίο αλλά και αυτές που φαίνεται πως ”λάμπουν” περισσότερο για όλο το κείμενο. Όταν οι σημαντικές λέξεις στο χωρίο παρουσιάζονται ως ήσσονος σημασίας σε ολόκληρο το κείμενο και το αντίστροφο, έχουμε ένδειξη λογοκλοπής.
Για την ανάδειξη τέτοιου είδους στατιστικών αφετηρία μας είναι το word frequency class των λέξεων αφενός σε όλο το κείμενο και, αφετέρου, σε κάθε passage. Η τιμές των κλάσεων που αποδίδονται στις λέξεις είναι άμεσα συνδεδεμένες με το υπό ανάλυση χωρίο, λόγω της κανονικοποίησης με τη συχνότητα της πιο δημοφιλούς λέξης στο εν λόγω χωρίο (βλ. σχέση (5.1)). Έτσι, στην ανάλυσή μας κάθε χωρίο, αλλά και όλο το κείμενο, είναι σαν μια μικρή κοινότητα με τους δικούς της ”αστέρες” και ”περιθωριακούς”.

Ακόμη, τα λεξιλογικά-σημασιολογικά χαρακτηριστικά στους αλγορίθμους που περιγράφουν την διαδικασία εξαγωγής τους, συμβολίζονται ως $wfc_i, i = 1, 2, 3, 4$, κι αυτό διότι βασίζονται στην κλάση συχνότητας λέξης ***(word frequency class)***.

Σημασιολογικό χαρακτηριστικό No.1
Πρόκειται για τον κανονικοποιημένο μέσο όρο των θετικών διαφορών των κλάσεων συχνοτήτων λέξεων μεταξύ κειμένου-χωρίου. Αφού έχουμε εξάγει τις κλάσεις συχνότητας λέξεων για όλο το κείμενο αλλά και για τα κομμάτια (των 15 ή 30 προτάσεων) που προκύπτουν από την κατάτμηση του κειμένου, εργαζόμαστε ως εξής: υπολογίζουμε τον μέσο όρο των θετικών διαφορών $c(WT, w) - c(P, w)$, μόνο για τις λέξεις που εμφανίζονται τουλάχιστον μια φορά στο passage. Μεγάλη τιμή σημαίνει υψηλή κλάση $c(WT, w)$ και χαμηλή $c(P, w)$, που σημαίνει, με τη σειρά του, σπανιότητα στο σύνολο του κειμένου αλλά ταυτόχρονα υψηλή σχετική συχνότητα στο χωρίο και συνεπώς, σημαντική διαφοροποίηση. Σημειώνεται πως μας ενδιαφέρουν εδώ μόνο οι θετικές διαφορές, διότι σκοπός μας σε αυτό το χαρακτηριστικό είναι αρχικά να εστιάσουμε στις πιο ”λαμπερές” λέξεις σε κάθε χωρίο, που ταυτόχρονα είναι σχετικά περιθωριακές από σκοπιά συνολικού κειμένου.
Ο αλγόριθμος υπολογισμού του χαρακτηριστικού για ένα χωρίο P δίνεται στον Αλγόριθμο 2.

Σημασιολογικό χαρακτηριστικό No.2
Πρόκειται για την τυπική απόκλιση των τιμών του σημασιολογικού χαρακτηριστικού No.1. Θέλουμε με αυτό, να δούμε τις διακυμάνσεις κατά τη σύγκριση της σημαντικότητας των ίδιων λέξεων από τη μία στο χωρίο και, από την άλλη, σε ολό-

Algorithm 2: Semantic Feature 1 for a passage, P

Data: WT, P, c()
Result: wfc_1
while *w in P* **do**
  if $c(WT, w) - c(P, w) > 0$ **then**
    $sum = sum + (c(WT, w) - c(P, w))$;
    $count = count + 1$;
  end
end
$wfc_1 = sum/count$;

κληρο το κείμενο. Η διακύμανση μπορεί να βοηθήσει τον εντοπισμό λογοκλεμμένων κομματιών σε περιπτώσεις κειμένων, όπου χρησιμοποιούνται κατά κόρον κοινές λέξεις, προσδίδοντας μια επιφανειακή ομοιομορφία στο κείμενο , ακόμα κι αν αυτό περιλαμβάνει λογοκλεμμένα κομμάτια. Λόγω μαζικότητας των κοινών λέξεων οι μέσοι όροι δεν μπορούν να αναδείξουν ανομοιομορφία στη σημασιολογία. Δίνουμε τον Αλγόριθμο 3 που δεν είναι παρά μια εφαρμογή του μαθηματικού τύπου της τυπικής απόκλισης.

Algorithm 3: Semantic Feature 2 for a passage, P

Data: WT, P, c()
Result: wfc_2
while *w in P* **do**
  if $c(WT, w) - c(P, w) > 0$ **then**
    $sum = sum + c^2(P, w)$;
    $count = count + 1$;
  end
end
$wfc_2 = squareRoot(sum/count)$;

Σημασιολογικό χαρακτηριστικό No.3
Πρόκειται για το ποσοστό των λέξεων που εμφανίζουν θετική διαφορά κλάσης συχνότητας μεταξύ κειμένου-χωρίου πάνω από το μέσο όρο, δηλαδή πάνω από την τιμή του πρώτου σημασιολογικού χαρακτηριστικού. Αποτελεί συμπληρωματικό χαρακτηριστικό της τυπικής απόκλισης των θετικών διαφορών στις κλάσεις συχνότητας λέξης, όπου όσο μεγαλύτερο το ποσοστό των λέξεων που ξεφεύγουν τόσο μεγαλύτερη η ένδειξη για λογοκλοπή. Τα βήματα υπολογισμού δίνονται στον Αλγόριθμο 4.

Σημασιολογικό χαρακτηριστικό No.4
Πρόκειται για το μέσο όρο των αρνητικών διαφορών των κλάσεων συχνοτήτων λέ-

Algorithm 4: Semantic Feature 3 for a passage, P

Data: WT, P, c(), wfc1(P), countPositive
Result: wfc_3
while *w in P* **do**
  if $c(WT, w) - c(P, w) > wfc1(P)$ **then**
    $count = count + 1$;
  end
end
$wfc_3 = count/countPositive$;

ξεων κειμένου-χωρίου για τις συχνά εμφανιζόμενες, σε όλο το κείμενο, λέξεις. Παρόμοιας λογικής με το πρώτο σημασιολογικό-λεξιλογικό χαρακτηριστικό, εδώ επικεντρωνόμαστε όχι στις σημαντικές λέξεις ενός χωρίου, αλλά στις λέξεις-θέματα που συχνά στο σύνολο του κειμένου. Ο μέσος όρος υπολογίζεται και πάλι από το άθροισμα των (αρνητικών) διαφορών $c(WT, w) - c(P, w)$, αλλά μόνο για λέξεις που θεωρούνται σημαντικές στο σύνολο το κειμένου, ορίζοντας ένα ανώτατο κατώφλι κλάσεως για αυτό το σκοπό. Το κατώφλι ορίστηκε βάσει των παρατηρήσεων σε διάφορα κείμενα του σώματος δεδομένων σχετικά με τις τιμές των κλάσεων.

Σημειώνεται πως λόγω της σχέσης 5.2, όταν μια λέξη του κειμένου δεν υπάρχει στο υπό ανάλυση χωρίο, η τιμή της κλάσης της πάει στο άπειρο. Για το χαρακτηριστικό αυτό όμως μας ενδιαφέρουν και λέξεις που πιθανώς δεν εμφανίζονται ποτέ στο χωρίο που επεξεργαζόμαστε. Για την διαχείριση αυτών των περιπτώσεων ορίσαμε ως τιμή κλάσης για τις λέξεις αυτές μια αρκετά μεγάλη τιμή.
Τα βήματα υπολογισμού του χαρακτηριστικού δίνονται στον Αλγόριθμο 5.

Algorithm 5: Semantic Feature 4 for a passage, P

Data: WT, P, c(), ϑ
Result: wfc_4
while *w in P* **do**
  if $c(WT, w) - c(P, w) < 0$ **then**
    if $c(WT, w) < \vartheta$ **then**
      $sum = sum + (-1) * (c(WT, w) - c(P, w))$;
      $count = count + 1$;
    end
  end
  $wfc_4 = sum/count$;
end
$wfc_4 = sum/count$;

5.3 Εξαγωγή Λογοκλεμμένων Χωρίων

5.3.1 Εισαγωγή

Τελικός σκοπός του συστήματός μας είναι η εξαγωγή των χωρίων που θεωρούνται λογοκλεμμένα. Στο διαγωνισμό του PAN'11 η αξιολόγηση των προβλέψεων γινόταν σε επίπεδο χαρακτήρων. Για παράδειγμα, υπολογιζόταν το πλήθος των, πιθανώς, επικαλυπτόμενων χαρακτήρων μεταξύ του (ή των) πράγματι λογοκλεμμένου χωρίου και του προβλεπόμενου ως λογοκλεμμένο χωρίο.
Εμείς θεωρούμε πως η ανά-πρόταση ανάλυση είναι πιο ορθόδοξη για το πρόβλημά μας. Έτσι, για είσοδο ενός κειμένου, αυτό που δίνει το σύστημα μας ως τελική πρόβλεψη είναι ένα διάνυσμα-στήλη από μηδενικά και άσσους · κάθε δυαδικός αριθμός στο διάνυσμα αντιστοιχεί σε μια πρόταση του κειμένου και τη χαρακτηρίζει ως λογοκλεμμένη ή μη. Συγκρίνοντας με το διάνυσμα-στήλη που περιέχει τις πραγματικές ετικέτες για κάθε πρόταση *(ground-truth labels)* υπολογίζονται οι διάφορες μετρικές και τα στατιστικά επιτυχίας του συστήματος.
Από την άλλη μεριά, η (στυλιστική) ανάλυση κάθε κειμένου έχει βασιστεί στο χωρισμό του με την τεχνική του μετακινούμενου παραθύρου. Από τη φάση της στυλιστικής ανάλυσης λαμβάνουμε το στυλιστικό αποτύπωμα κάθε κειμένου ως μια λίστα διανυσμάτων τιμών · κάθε καταχώρηση στη λίστα αντιστοιχεί σε ένα στιγμιότυπο του μετακινούμενου παραθύρου στο κείμενο. Εφαρμόζουμε λοιπόν μια "ανά παράθυρο" στυλιστική ανάλυση, και μάλιστα χρησιμοποιούμε επικαλυπτόμενα παράθυρα, όπως αναλύθηκε στην ενότητα 5.1.
Ακόμα, τα αποτελέσματα της στυλιστικής ανάλυσης προορίζονται να αποτελέσουν εν μέρει τα δεδομένα εκπαίδευσης και εν μέρει τα δεδομένα ελέγχου ενός συστήματος επιβλεπόμενης μηχανικής μάθησης. Αυτό σημαίνει πως το στυλιστικό αποτύπωμα των διαδοχικών παραθύρων κειμένου θα πρέπει να συνοδεύεται από μια ετικέτα που τα κατατάσσει στην κατάλληλη κλάση (λογοκλεμμένα ή μη). Στην πολιτική που εφαρμόσαμε η ετικέτα αυτή αποφασίζεται ως εξής: εάν σε ένα παράθυρο το λογοκλεμμένο τμήμα αποτελεί το 40% ή πλέον του συνόλου των χαρακτήρων, τότε το παράθυρο αυτό χαρακτηρίζεται ως λογοκλεμμένο.
Από τη λίστα στυλιστικών αποτυπωμάτων των "παραθύρων" των κειμένων έως το τελικό διάνυσμα-στήλη με τις ανά-πρόταση προβλέψεις παρεμβάλλεται το σύστημα μηχανικής μάθησης. Στόχος είναι η ακριβέστερη δυνατή πρόβλεψη και ταξινόμηση των "παραθύρων" - των κειμένων που αποτελούν τα δεδομένα ελέγχου - ως λογοκλεμμένα ή μη.

Η τακτική χωρισμού του σώματος δεδομένων σε δεδομένα εκπαίδευσης και ελέγχου παρουσιάζεται στα επόμενα, στην ενότητα *Μέθοδος Επικύρωσης*, οι αλγόριθμοι εκπαίδευσης που εφαρμόστηκαν, στην ενότητα *Μέθοδοι Εκμάθησης*, όπου συμπεριλαμβάνονται και οι αλγόριθμοι εξισορρόπησης των δεδομένων εκμάθησης που δοκιμάστηκαν.

Από το σύστημα μηχανικής μάθησης, λοιπόν, λαμβάνουμε τις ανά-παράθυρο

προβλέψεις των κειμένων που αποτελούν τα δεδομένα ελέγχου. Από αυτό το σημείο απομένει να χρησιμοποιήσουμε τις πληροφορίες που μας δίνουν αυτές οι προβλέψεις, ώστε να αποφανθούμε για καθεμία εκ των προτάσεων που συναποτελούν τα κείμενα. Αυτά τα βήματα παρουσιάζονται στην ενότητα *Αποσύμπλεξη επικαλυπτόμενων παραθύρων & εξαγωγή λογοκλοπής ανά-πρόταση*.
Τέλος, οι μετρικές που χρησιμοποιούνται για την αξιολόγηση του συνόλου του συστήματος παρουσιάζονται στην ενότητα *Μετρικές Αξιολόγησης*.

5.3.2 Μέθοδος Επικύρωσης

Το σώμα δεδομένων που χρησιμοποιούμε αποτελείται από 4753 κείμενα. Η μέθοδος επικύρωσης που επιλέχθηκε είναι η k-στρώσεων διασταυρωμένη επικύρωση *(k-fold cross-validation)*, με $k = 5$. Έχοντας περίπου 5000 κείμενα στη διάθεσή μας, δεδομένων των απαιτούμενων χρόνων επεξεργασίας, ο χωρισμός σε 5 κομμάτια, με κάθε κομμάτι να περιλαμβάνει περίπου 1000 κείμενα ως δεδομένα ελέγχου και τα υπόλοιπα ως δεδομένα εκπαίδευσης, αποτελεί μια λύση που δίνει αξιόπιστα αποτελέσματα από πλευρά μηχανικής μάθησης - περίπου 80% των δεδομένων για εκπαίδευση και 20% για έλεγχο- αλλά και είναι βιώσιμη υπολογιστικά. Οι 5 χωρισμοί (splits) του σώματος δεδομένων, σε δεδομένα εκπαίδευσης και ελέγχου φαίνονται στον Πίνακα 5.1.

Πίνακας 5.1 5-στρώσεων διασταυρωμένη επικύρωση στο σώμα δεδομένων του PAN'11

	Texts Included				
	split1	*split2*	*split3*	*split4*	*split5*
train set	$\{1001, ..., 4753\}$	$\{1, ..., 1000\}$ $\cap$ $\{2001, ..., 4753\}$	$\{1, ..., 2000\}$ $\cap$ $\{3001, ..., 4753\}$	$\{1, ..., 3000\}$ $\cap$ $\{4001, ..., 4753\}$	$\{1, ..., 3802\}$
test set	$\{1, ..., 1000\}$	$\{1001, ..., 2000\}$	$\{2001, ..., 3000\}$	$\{3001, ..., 4000\}$	$\{3803, ..., 4753\}$

Όπως προβλέπεται από τη μέθοδο k-στρώσεων διασταυρωμένη επικύρωση, για κάθε έναν από τους παραπάνω χωρισμούς του σώματος δεδομένων εκπαιδεύουμε το μοντέλο μηχανικής μάθησης, ύστερα παίρνουμε τις προβλέψεις που δίνει το εκπαιδευμένο μοντέλο για τα δεδομένα ελέγχου και με κατάλληλη επεξεργασία υπολογίζουμε τα αποτελέσματα που μας ενδιαφέρουν. Όταν αυτή η διαδικασία ολοκληρωθεί για όλους τους χωρισμούς του σώματος δεδομένων, υπολογίζουμε τα τελικά αποτελέσματα ως τον μέσο όρο των επιμέρους αποτελεσμάτων. Με αυτόν τον τρόπο, διασφαλίζεται πως τα τελικά αποτελέσματα είναι αντιπροσωπευτικά του συνόλου του σώματος δεδομένων.

5.3.3 Μέθοδοι Εκμάθησης

Για την εκπαίδευση του συστήματος μας πειραματιστήκαμε με τεσσάρων ειδών αλγορίθμους εκπαίδευσης:

1. Δέντρα Απόφασης
2. Perceptron πολλών στρωμάτων
3. Naive Bayes
4. Μηχανές Διανυσμάτων Υποστήριξης

Τα Δέντρα Απόφασης έδωσαν τα καλύτερα αποτελέσματα γι' αυτό και πειραματιστήκαμε περισσότερο με αυτά. Οι Μηχανές Διανυσμάτων Υποστήριξης είχαν τεράστιες απαιτήσεις σε υπολογιστικό χρόνο, κάτι που στάθηκε μεγάλο εμπόδιο για περαιτέρω πειραματισμούς.

5.3.4 Αποσύμπλεξη επικαλυπτόμενων παραθύρων & εξαγωγή λογοκλοπής ανά-πρόταση

Είπαμε πως από το σύστημα μηχανικής μάθησης λαμβάνουμε τις ετικέτες-προβλέψεις, για το αν ένα τμήμα-παράθυρο του κειμένου αποτελεί προϊόν λογοκλοπής ή όχι. Πριν δώσουμε τις ετικέτες πρόβλεψης για καθεμία πρόταση του κειμένου, πρέπει να αποφασίσουμε πώς θα διαχειριστούμε την επικάλυψη των διαδοχικών παραθύρων της κατάτμησης του κειμένου. Σύμφωνα με τη στρατηγική κατάτμησης που εφαρμόσαμε, και στις τρεις διαφορετικές περιπτώσεις, το μετακινούμενο παράθυρο έχει μήκος $3k$ και βήμα k προτάσεων. Αυτό σημαίνει πως, εξαιρώντας την αρχή και το τέλος του κειμένου, κάθε δέσμη k προτάσεων θα καλύπτεται από 3 διαφορετικά, διαδοχικά παράθυρα (Σχήμα 2.3). Οι τελικές ετικέτες πρόβλεψης, σε επίπεδο προτάσεων, αποφασίζονται ανά δέσμη k προτάσεων ανάλογα με τις προβλέψεις των παραθύρων που τις συμπεριλαμβάνουν. Έτσι, καθεμία πρόταση μιας τέτοιας δέσμης χαρακτηρίζεται ως λογοκλεμμένη αν τουλάχιστον n από τα επικαλυπτόμενα, σε αυτήν, παράθυρα έχουν χαρακτηριστεί ως τέτοια, όπου $n = \{1, 2, 3\}$. Στα αποτελέσματά μας συγκρίνονται οι τρεις αυτές περιπτώσεις.

5.3.5 Μετρικές αξιολόγησης

Για την αξιολόγηση του συστήματός μας υπολογίστηκαν οι μετρικές Precision, Recall, F-measure, οι μαθηματικοί τύποι των οποίων δίνονται στον Πίνακα 2.6. Επιλέχθηκαν αυτές οι μετρικές, αφενός διότι εστιάζουν στην επιτυχία και ακρίβεια πρόβλεψης της κλάσεως μειοψηφίας, και, αφετέρου, διότι είναι αυτές που χρησιμοποιήθηκαν και στο διαγωνισμό του PAN'11, οπότε προσφέρονται για άμεση σύγκριση με τα διαγωνιζόμενα συστήματα.

Ενδεικτικά παρουσιάζονται, στα επόμενα, και κάποια αποτελέσματα των υπολοίπων μετρικών που συμπεριλαμβάνονται στον Πίνακα 2.6.

Κεφάλαιο 6

Αποτελέσματα Συστήματος

Αρχικά τα αποτελέσματα παρουσιάζονται για κάθε ταξινομητή ξεχωριστά, όπως υπολογίστηκαν με τη μέθοδο *5-στρώσεων διασταυρωμένη επικύρωση*. Για κάποιους από τους 4 ταξινομητές έχουν γίνει πειράματα με όλες τις παραμέτρους του συστήματος, ενώ για κάποιους όχι, ανάλογα με την αποτελεσματικότητα του καθενός αλλά και την χρονική τους πολυπλοκότητα, η οποία στάθηκε σοβαρό εμπόδιο στην περίπτωση των Μηχανών Διανυσμάτων Υποστήριξης.
Τα αποτελέσματα αποτελούνται από 3 υποκατηγορίες ανάλογα με τα στυλιστικά χαρακτηριστικά που επιστρατεύουμε:

- όλα τα χαρακτηριστικά που περιγράφηκαν στην ενότητα 5.2
- μόνο τα στυλομετρικά χαρακτηριστικά
- μόνο τα σημασιολογικά-λεξιλογικά χαρακτηριστικά

Κατά τη διαδικασία εξαγωγής των αποτελεσμάτων με τη μέθοδο επικύρωσης 5-στρώσεων διαπιστώθηκε, στους περισσότερους ταξινομητές, σημαντική διακύμανση της αποτελεσματικότητας για τους διαφορετικούς χωρισμούς του σώματος δεδομένων σε σύνολα εκμάθησης και ελέγχου. Η διακύμανση αυτή είναι ενδεικτική της ευστάθειας του συστήματος. Για το λόγο αυτό θα παρουσιαστούν και κάποια από τα επιμέρους αποτελέσματα της διαδικασίας 5-στρώσεων διασταυρωμένη επικύρωση, που θα μας βοηθήσουν να αποφανθούμε και για τον καταλληλότερο ταξινομητή.

Τα Δέντρα Απόφασης έδωσαν τα δεύτερα καλύτερα αποτελέσματα, ακολουθώντας τις Μηχανές Διανυσμάτων Υποστήριξης. Λόγω των μεγάλων απαιτήσεων σε χρόνο επεξεργασίας των τελευταίων, οι περισσότεροι πειραματισμοί του συστήματος έγιναν με Δέντρα Απόφασης.

Ως μέρος των αποτελεσμάτων δίνουμε μια λίστα των χαρακτηριστικών του συστήματος μας, ταξινομημένα σύμφωνα με το F-measure που έδωσε το καθένα από αυτά στα πειράματά μας. Για λόγους σύγκρισης, υλοποιήθηκε, και ”μετρήθηκε”, το χαρακτηριστικό του συστήματος των Oberreuter et al. [44] σύμφωνα με τον Αλγόριθμο 1, αλλά κρατώντας τη δική μας προ-επεξεργασία κειμένου.

Σημαντική θέση στα αποτελέσματα λαμβάνει η εξισορρόπηση των δεδομένων εκπαίδευσης · κάθε ταξινομητής δοκιμάστηκε με και χωρίς εφαρμογή εξισορρόπησης.

Για όλα τα αποτελέσματα χρησιμοποιούμε τις μετρικές precision, recall και F-measure, οι οποίες είναι οι πλέον χρησιμοποιούμενες για την εγγενή ανίχνευση λογοκλοπής και που μας επιτρέπουν άμεση σύγκριση με τα αποτελέσματα των διαγωνιζόμενων συστημάτων στον PAN'11. Δεν συμπεριλαμβάνουμε την τιμή της granularity, μιας και, εκ κατασκευής, δεν υπάρχει επικάλυψη μεταξύ των προτάσεων για τις οποίες δίνεται τιμή πρόβλεψης και , επομένως, ισχύει πάντοτε ότι $granularity = 1$.

6.1 Δέντρα Απόφασης

Αρχικά παρουσιάζουμε τα αποτελέσματα που προέκυψαν με τη μέθοδο 5-στρώσεων διασταυρωμένη, για 3 ομαδοποιήσεις των στυλιστικών χαρακτηριστικών

$i)$ όλα τα χαρακτηριστικά που παρουσιάστηκαν στην ενότητα 5.2 $(1-11)$,

$ii)$μόνο τα στυλομετρικά χαρακτηριστικά $(1-7)$,

$iii)$ μόνο τα σημασιολογικά χαρακτηριστικά $(8-11)$.

Για τα αποτελέσματα του Πίνακα 6.1 κάθε σετ n προτάσεων, όπου n το εκάστοτε βήμα μετακίνησης του παραθύρου, κρίνεται ως λογοκλεμμένο εάν τουλάχιστον 1 από τα 3 διαδοχικά παράθυρα που το εμπεριέχουν έχει προβλεφθεί ως λογοκλέμμενο, ενώ για τον Πίνακα 6.2, τουλάχιστον 2 από τα 3.

Παρατηρώντας κάθε πίνακα ξεχωριστά συμπεραίνουμε αρχικά πώς τα μετακινούμενα παράθυρα σταθερού μήκους και βήματος είναι ελαφρώς πιο αποδοτικά από το παράθυρο με προσαρμοζόμενες τιμές ανάλογα με το μέγεθος του κειμένου. Αυτό το αποτέλεσμα πάει κόντρα στη διαίσθηση. Θα περιμέναμε το προσαρμοζόμενο παράθυρο να παρουσιάζει μεγαλύτερη ευελιξία στον εντοπισμό των λογοκλεμμένων χωρίων, τα οποία αναμένονται να έχουν μήκος ανάλογο του μεγέθους του κειμένου. Αυτά τα αποτελέσματα δεν αρκούν, βέβαια, για να καταρρίψουν αυτήν την υπόθεση. Βλέπουμε πως το παράθυρο σταθερού μήκους 30 προτάσεων, μήκος που θεωρούσαμε μεγάλο, είναι ιδιαίτερα ανταγωνιστικό. Αυτό σημαίνει πως, πιθανώς, το μικρό παράθυρο μήκους 9 προτάσεων και βήματος 3, είναι ακατάλληλο ακόμα και για τα μικρά κείμενα. Ίσως το συγκεκριμένο σώμα δεδομένων να χρειαζόταν μια κλίμακα προσαρμοζόμενου παραθύρου μεγαλύτερων τιμών.
Συγκρίνοντας τους Πίνακες 6.1, 6.2 μεταξύ τους, συμπεραίνουμε πως η επιλογή για πρόβλεψη μιας δέσμης προτάσεων ως plagiarised, αν τουλάχιστον 2 από τα 3 παράθυρα που την εμπεριέχουν θεωρούνται plagiarised υπερτερεί. Αυτή η επιλογή φαίνεται να επιτυγχάνει τον καλύτερο συμβιβασμό μεταξύ των precision και recall. Αυτό δεν ισχύει μόνο στην περίπτωση που τα σημασιολογικά χαρακτηριστικά λειτουργούν μόνα τους. Βλέπουμε πως στην επιλογή *2 από τα 3* η ακρίβεια πρόβλεψης *(precision)* ενισχύεται σε βάρος της ανάκλησης *(recall)*. Αξίζει να σημειωθεί, ότι σε πειραματισμό που κάναμε με την επιλογή *3 από τα 3*, η ακρίβεια ενισχυόταν

Πίνακας 6.1 Αποτελέσματα 5-στρώσεων διασταυρωμένης επικύρωσης, με Δέντρα Απόφασης. Μια δέσμη προτάσεων χαρακτηρίζεται λογοκλεμμένη αν τουλάχιστον *1 από τα 3* παράθυρα που την εμπεριέχουν προβλέπονται ως λογοκλεμμένα

Χαρακτηριστικά συστήματος	Μετακινούμενο παράθυρο	Precision	Recall	F-measure
Όλα	$\{\mu.\pi. = 15, \beta = 5\}$	0.167	0.497	0.250
	$\{\mu.\pi. = 30, \beta = 10\}$	0.168	0.522	0.255
	$\{\mu.\pi. = 3k, \beta = k\}$, $k = \{3, 5, 10\}$	0.165	0.454	0.243
Μόνο στυλομετρικά	$\{\mu.\pi. = 15, \beta = 5\}$	0.115	0.363	0.175
	$\{\mu.\pi. = 30, \beta = 10\}$	0.117	0.387	0.179
	$\{\mu.\pi. = 3k, \beta = k\}$, $k = \{3, 5, 10\}$	0.115	0.355	0.174
Μόνο σημασιολογικά	$\{\mu.\pi. = 15, \beta = 5\}$	0.136	0.411	0.204
	$\{\mu.\pi. = 30, \beta = 10\}$	0.137	0.442	0.209
	$\{\mu.\pi. = 3k, \beta = k\}$, $k = \{3, 5, 10\}$	0.132	0.369	0.195

Πίνακας 6.2 Αποτελέσματα 5-στρώσεων διασταυρωμένης επικύρωσης, με Δέντρα Απόφασης. Μια δέσμη προτάσεων χαρακτηρίζεται λογοκλεμμένη αν τουλάχιστον *2 από τα 3* παράθυρα που την εμπεριέχουν προβλέπονται ως λογοκλεμμένα

Χαρακτηριστικά συστήματος	Μετακινούμενο παράθυρο	Precision	Recall	F-measure
Όλα	$\{\mu.\pi. = 15, \beta = 5\}$	0.395	0.231	0.292
	$\{\mu.\pi. = 30, \beta = 10\}$	0.396	0.230	0.291
	$\{\mu.\pi. = 3k, \beta = k\}$, $k = \{3, 5, 10\}$	0.387	0.182	0.248
Μόνο στυλομετρικά	$\{\mu.\pi. = 15, \beta = 5\}$	0.295	0.149	0.198
	$\{\mu.\pi. = 30, \beta = 10\}$	0.294	0.134	0.184
	$\{\mu.\pi. = 3k, \beta = k\}$, $k = \{3, 5, 10\}$	0.293	0.114	0.165
Μόνο σημασιολογικά	$\{\mu.\pi. = 15, \beta = 5\}$	0.327	0.142	0.198
	$\{\mu.\pi. = 30, \beta = 10\}$	0.312	0.156	0.208
	$\{\mu.\pi. = 3k, \beta = k\}$, $k = \{3, 5, 10\}$	0.307	0.122	0.174

ακόμη περισσότερο φτάνοντας ακόμα και το 85%, καταβαραθρώνοντας όμως την ανάκληση σε σχεδόν μηδενικές τιμές, και δίνοντας, τελικά, απογοητευτικά αποτελέσματα στο F-measure.

Πριν προχωρήσουμε στο σχολιασμό για την αποτελεσματικότητα και ανταγωνιστικότητα του συστήματός μας, θα παρουσιάσουμε, στον Πίνακα 6.3, τα επιμέρους αποτελέσματα για τους 5 χωρισμούς του σώματος δεδομένων σε δεδομένα εκπαίδευσης και ελέγχου.

Πίνακας 6.3 Αποτελέσματα για τους 5 χωρισμούς του κειμένου, με Δέντρα Απόφασης. Μετακινούμενο παράθυρο με *μ.π.* = 15, *β* = 5 προτάσεις.

Χαρακτηριστικά συστήματος	***F-measure***				
	split1	***split2***	***split3***	***split4***	***split5***
Όλα	0.209	0.205	0.252	0.333	0.363
Μόνο στυλομετρικά	0.105	0.095	0.119	0.262	0.284
Μόνο σημασιολογικά	0.141	0.132	0.159	0.264	0.286

Από τον παραπάνω πίνακα είναι εμφανής η ανομοιομορφία μέσα στο σώμα δεδομένων, η οποία παρουσιάζει, μάλιστα, μια κάποια κανονικότητα · βλέπουμε πως η αποτελεσματικότητα αυξάνεται, όσο τα σώμα των κειμένων ελέγχου μετακινείται από τα πρώτα προς τα τελευταία κείμενα του σώματος δεδομένων. Συγκεκριμένα, η διαφορά από το χωρισμό split1 και split5, όπου το σώμα ελέγχου αποτελείται από τα 1000 πρώτα και τελευταία κείμενα αντίστοιχα, φτάνει περίπου το 15%, τιμή που είναι πολύ μεγάλη για τα δεδομένα του προβλήματος. Υπενθυμίζουμε πως η τιμή 30% στο F-measure είναι ιδιαίτερα υψηλό για τα δεδομένα του προβλήματος, ανεξάρτητα από το συγκεκριμένο διαγωνισμό. Μάλιστα, τα αποτελέσματα της συμμετοχής των Oberreuter et al. [44] έχουν κριθεί ως παραπλανητικά, λόγω του χαρακτηριστικού που επικεντρώνεται στη μοναδικότητα των λέξεων σε συνδυασμό με την ιδιαιτερότητα κατασκευής του συγκεκριμένου σώματος δεδομένων, όπου δεν υπάρχει επικάλυψη θεμάτων μεταξύ των λογοκλεμμένων και μη χωρίων ενός κειμένου. Προκειμένου να αποφύγουμε αυτήν την παγίδα με τα δικά μας σημασιολογικά χαρακτηριστικά, τα πειράματά μας περιλαμβάνουν διαχωρισμό των χαρακτηριστικών σε στυλομετρικά και σημασιολογικά. Τα αποτελέσματα που λαμβάνουμε είναι ενθαρρυντικά: βλέπουμε πως τα στυλομετρικά χαρακτηριστικά ανταγωνίζονται τα σημασιολογικά. Μάλιστα το σύστημά μας αναδεικνύεται ανώτερο απέναντι σε συστήματα που συμμετείχαν στο διαγωνισμό, ακόμα και με μόνο τη δράση των στυλομετρικών χαρακτηριστικών. Στην ευνοϊκότερη περίπτωση χωρισμού, το split5, βλέπουμε, μάλιστα πως ανταγωνίζεται και το νικητήριο σύστημα του διαγωνισμού, με το αμφιλεγόμενης ικανότητας γενίκευσης χαρακτηριστικό.
Παρολαυτά πιστεύουμε πως τα σημασιολογικά χαρακτηριστικά που επιστρατεύουμε προσφέρουν στο σύστημα ανεξάρτητα από την ιδιαιτερότητα του συγκεκριμένου σώματος δεδομένων. Θεωρούμε ακόμα σημαντική την συνύπαρξη σημα-

σιολογικών και στυλομετρικών χαρακτηριστικών, επειδή με αυτό τον τρόπο συλλέγουμε διαφορετικού είδους πληροφορίες για το κείμενο, τις οποίες μπορούμε να αξιοποιήσουμε με τη βοήθεια μηχανικής μάθησης, για να ανακαλύψουμε μεταξύ τους συσχετισμούς που βοηθούν στις προβλέψεις. Το τελευταίο γίνεται εμφανές και στους πειραματισμούς μας.

6.1.1 Εξισορρόπηση δεδομένων εκπαίδευσης

Όπως εξηγήθηκε στα προηγούμενα κεφάλαια το πρόβλημα έχει την ιδιαιτερότητα της ανισορροπίας των δεδομένων στις δύο κλάσεις ταξινόμησης - λογοκλεμμένο χωρίο ή μη. Ενδεικτικά, στον επόμενο πίνακα δίνουμε τα απόλυτα μεγέθη, καθώς και την απλή μετρική της ακριβείας σε σύγκριση με την μετρική F-measure, για τα αποτελέσματα σε χωρισμό split5 του σώματος δεδομένων.

Πίνακας 6.4 Ανισορροπία δεδομένων στο σώμα δεδομένων του PAN'11. *Κείμενα 3803-4753*

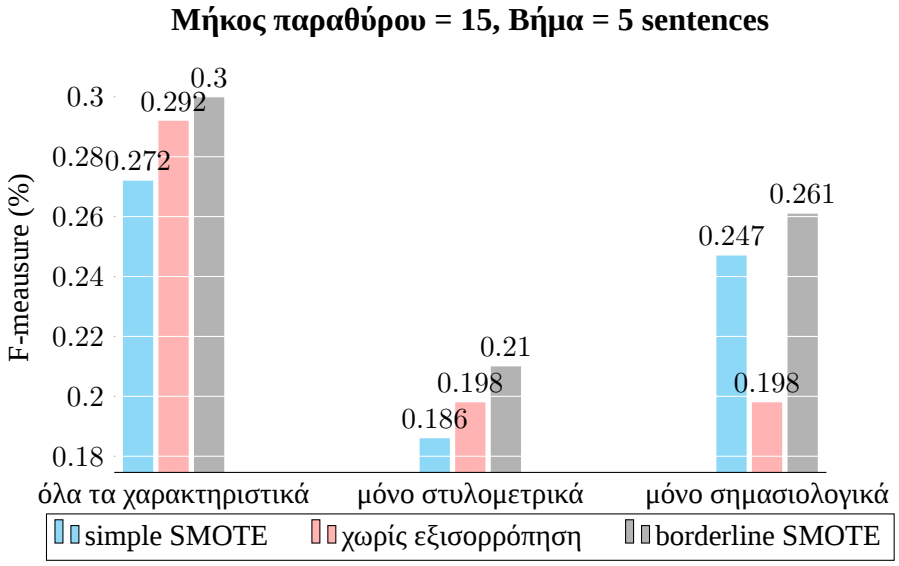
	λογοκλεμμένα		*μη-λογοκλεμμένα*	
	TP	FN	TN	FP
πλήθος προτάσεων	18161	44540	1111024	19149
Accuracy	0.946			
F-measure	0.363			

Από τον Πίνακα 6.4 γίνεται φανερό πως τα μη-λογοκλεμμένα χωρία είναι κατά τάξεις μεγέθους περισσότερα από τα λογοκλεμμένα. Βλέπουμε ακόμα τεράστια διαφορά μεταξύ της τιμής της απλής ακρίβειας, που είναι και η πιο συνηθισμένη μετρική αποτελεσματικότητας, και της μετρικής F-measure, η οποία λαμβάνει υπόψιν της το αυξημένο ενδιαφέρον μας για την κλάση μειοψηφίας, δηλαδή τα λογοκλεμμένα χωρία (εν προκειμένω προτάσεις). Αναδεικνύεται λοιπόν και η εγγενής δυσκολία του προβλήματος Εγγενούς Ανίχνευσης Λογοκλοπής, αφού ένα σύστημα που, σύμφωνα με τον κλασικό υπολογισμό της ακρίβειας πρόβλεψης $\frac{\text{σωστές προβλέψεις}}{\text{σύνολο προβλέψεων}}$, έχει αξιοπιστία πρόβλεψης 94%,στη μετρική που μας ενδιαφέρει το ποσοστό πέφτει στο 36%.
Για την εξισορρόπηση των δεδομένων εκπαίδευσης δοκιμάστηκαν οι εξής δύο αλγόριθμοι που αναλύθηκαν στην ενότητα 3.4.2:

i) simple SMOTE

ii) borderline SMOTE

Ο simple SMOTE συγκρίνεται με τον borderline SMOTE αλγόριθμο, καθώς και με τα αποτελέσματα για μη-εφαρμογή εξισορρόπησης στα δεδομένα εκπαίδευσης, στο διάγραμμα που ακολουθεί. Οι τιμές που παρουσιάζονται υπολογίστηκαν, και πάλι, με τη μέθοδο 5-στρώσεων διασταυρωμένη επικύρωση, για μετακινούμενο παράθυρο μήκους 15 προτάσεων και βήματος 5 προτάσεων.



Από το παραπάνω διάγραμμα συμπεραίνουμε πως μια άστοχη απόπειρα εξισορρόπησης των δεδομένων εκπαίδευσης μπορεί ακόμα και να δυσχεράνει την αποτελεσματικότητα του συστήματος. Βλέπουμε πως με την εφαρμογή του αλγορίθμου *simple SMOTE* το σύστημα δίνει χειρότερα αποτελέσματα στην περίπτωση των στυλομετρικών χαρακτηριστικών. Παρόλο που έχουμε σημαντική πρόοδο στην περίπτωση των σημασιολογικών χαρακτηριστικών, στο σύνολο των χαρακτηριστικών η αποτελεσματικότητα ζημιώνεται. Από την άλλη μεριά, κατά την εξισορρόπηση με τον αλγόριθμο *borderline SMOTE* - ο οποίος, υπενθυμίζεται ότι, φροντίζει να παράγει συνθετικά παραδείγματα εκπαίδευσης μόνο από γείτονες που ανήκουν στην κλάση μειοψηφίας - το σύστημα ευνοείται. θα πρέπει να σημειωθεί ότι η εφαρμογή εξισορρόπησης των δεδομένων εκπαίδευσης προσθέτει υπολογιστικό βάρος στο σύστημα, κάτι που πολλές φορές δεν είναι εύκολο να αγνοηθεί. Η πολυπλοκότητα του αλγορίθμου εξισορρόπησης μπορεί να αποτελέσει σημαντικό κριτήριο για την επιλογή του ή μη. Από τους δικούς μας πειραματισμούς αναφέρουμε, ενδεικτικά, τις περιπτώσεις των αλγορίθμων simple SMOTE, borderline SMOTE και ADASYN. Οι δύο πρώτοι αλγόριθμοι δεν επιβάρυναν το σύστημα πάνω από $\sim$1min για την εξισορρόπηση σε 4000 κείμενα, ενώ ο τελευταίος για τον ίδιο όγκο χρειάστηκε $\sim$90min.

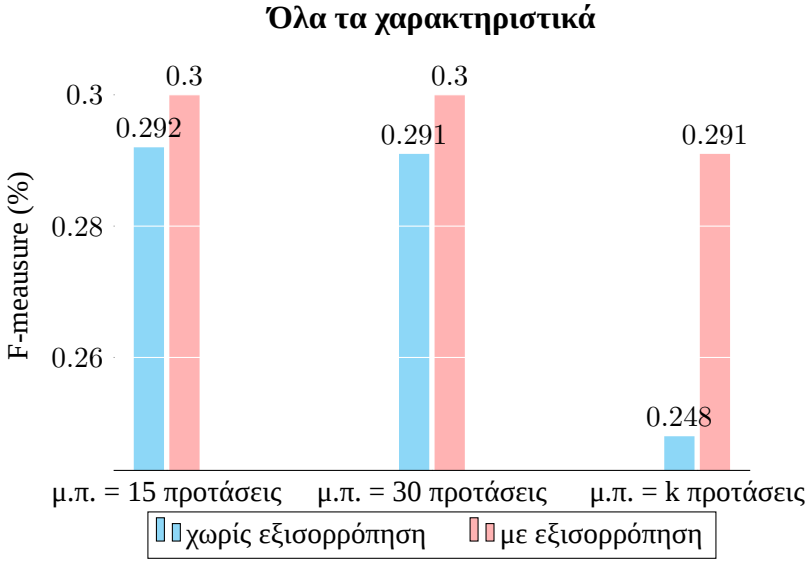
Η ενδιαφέρουσα περίπτωση εξισορρόπησης των δεδομένων εκπαίδευσης με τον αλγόριθμο *borderline SMOTE* αξίζει περαιτέρω ανάλυσης.
Ο Πίνακας 6.5 περιλαμβάνει τα αποτελέσματα του συστήματος με εξισορροπημένα δεδομένα εκπαίδευσης με τον αλγόριθμο SMOTE borderline και Δέντρα Απόφασης, ενώ για την εξαγωγή των λογοκλεμμένων προτάσεων, ένα σετ προτάσεων θεωρείται λογοκλεμμένο εάν τουλάχιστον 2 από τα 3 επικαλυπτόμενα, σε αυτό, παράθυρα έχουν προβλεφτεί ως λογοκλεμμένα.

Συγκρίνοντας τα αποτελέσματα με και χωρίς εξισορρόπηση των δεδομένων εκπαίδευσης - δηλαδή τους Πίνακες 6.5 και 6.2 αντίστοιχα - βλέπουμε πως η εξι-

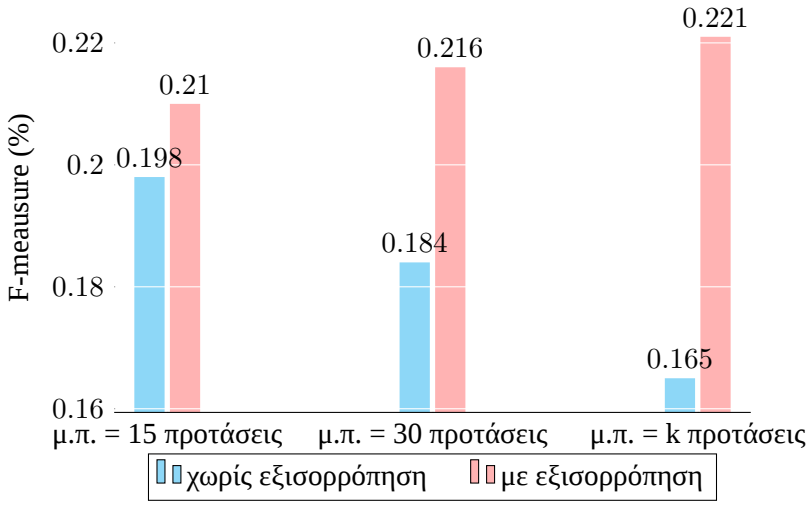
Πίνακας 6.5 Αποτελέσματα 5-στρώσεων διασταυρωμένης επικύρωσης, με Δέντρα Απόφασης και εξισορρόπηση δεδομένων εκπαίδευσης με *SMOTE borderline*.

Χαρακτηριστικά συστήματος	Μετακινούμενο παράθυρο	Precision	Recall	F-measure
Όλα	$\{\mu.\pi. = 15, \beta = 5\}$	0.246	0.386	0.300
	$\{\mu.\pi. = 30, \beta = 10\}$	0.249	0.389	0.304
	$\{\mu.\pi. = 3k, \beta = k\}$, $k = \{3, 5, 10\}$	0.252	0.344	0.291
Μόνο στυλομετρικά	$\{\mu.\pi. = 15, \beta = 5\}$	0.190	0.234	0.210
	$\{\mu.\pi. = 30, \beta = 10\}$	0.193	0.245	0.216
	$\{\mu.\pi. = 3k, \beta = k\}$, $k = \{3, 5, 10\}$	0.210	0.232	0.221
Μόνο σημασιολογικά	$\{\mu.\pi. = 15, \beta = 5\}$	0.239	0.288	0.261
	$\{\mu.\pi. = 30, \beta = 10\}$	0.237	0.302	0.266
	$\{\mu.\pi. = 3k, \beta = k\}$, $k = \{3, 5, 10\}$	0.232	0.237	0.234

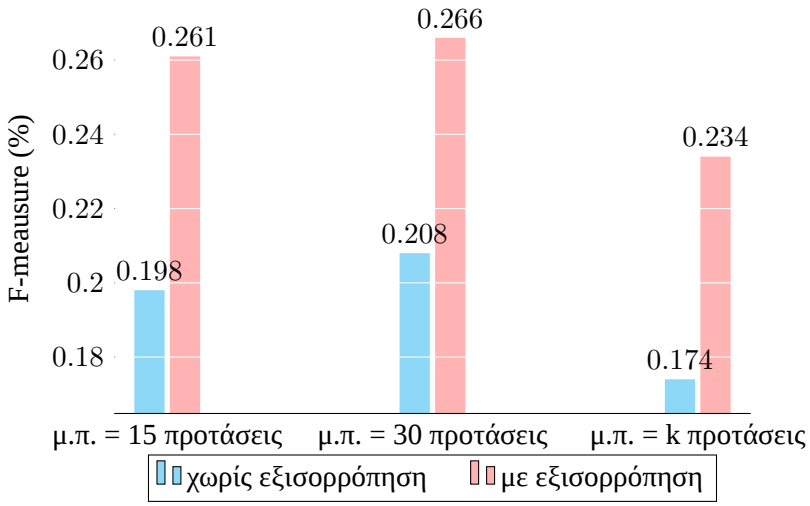
σορρόπηση δίνει σαφή ώθηση στο σύστημα. Το πλεονέκτημα παρασταίνεται στα επόμενα τρία διαγράμματα, ένα για κάθε κατηγορία χαρακτηριστικών.



Μόνο στυλομετρικά χαρακτηριστικά



Μόνο σημασιολογικά χαρακτηριστικά



Στον Πίνακα 6.6 έχουμε ταξινομήσει τα χαρακτηριστικά του συστήματός μας, όπως αυτά παρουσιάστηκαν στο αντίστοιχο κεφάλαιο, καθώς και το χαρακτηριστικό του συστήματος των Oberreuter et al. [44], βάσει της αποτελεσματικότητα τους. Η αποτελεσματικότητα καθενός από τα χαρακτηριστικά έγινε ως εξής: για έναν χωρισμό του σώματος δεδομένων (χωρισμός split5, συγκεκριμένα) σε δεδομένα εκπαίδευσης και ελέγχου, κρατήσαμε, διαδοχικά, κάθε ένα από αυτά τα χαρακτηριστικά ως το μοναδικό για τη στυλιστική ανάλυση των κειμένων. Χρησιμοποιήθηκε μετακινούμενο παράθυρο μ.π. $= 15$, β $= 5$ προτάσεις και εξισορρόπηση δεδομένων με SMOTE borderline.
Πρόκειται για μια ταξινόμηση που αναδεικνύει, μεν, την σχετική ισχύ των χαρακτηριστικών αλλά δεν πρέπει να παραγνωρίζεται το γεγονός πως η αποτελεσματικότητα ενός συστήματος πολλές φορές δεν κρίνεται από τη δύναμη ενός μεμονωμένου στυλιστικού χαρακτηριστικού αλλά στους συσχετισμούς ενός πλήθους χαρακτηριστικών.

Πίνακας 6.6 Ταξινόμηση στυλιστικών χαρακτηριστικών βάσει F-measure

Κατηγορία	**Χαρακτηριστικό**	**F-measure**
Στυλομετρικά χαρακτηριστικά	μέσο πλήθος συλλαβών ανά λέξη	0.185
	ποσοστό συμπίεσης ρημάτων	0.159
	ποσοστό συμπίεσης επιρρημάτων	0.133
	ποσοστό συμπίεσης επιθέτων	0.130
	συχνότητα της λέξης ”of”	0.126
	Βαθμός Flesch-Kinkaid	0.117
	μέσο μήκος πρότασης	0.117
Σημασιολογικά χαρακτηριστικά	wfc_4	0.196
	χαρακτηριστικό oberreuter et al. [44]	0.178
	wfc_1	0.153
	wfc_3	0.134
	wfc_2	0.116

Το χαρακτηριστικό του συστήματος των Oberreuter et al. [44] υλοποιήθηκε και συμπεριλαμβάνεται στον παραπάνω πίνακα για λόγους σύγκρισης. Υπογραμμίζεται ότι κρατήσαμε την προ-επεξεργασία όπως αυτή σχεδιάστηκε για το σύστημά μας. Βλέπουμε ότι σε επίπεδο σημασιολογίας το πρώτο σημασιολογικό-λεξιλογικό χαρακτηριστικό μας είναι, τουλάχιστον, εξίσου δυνατό όσο και αυτό των Oberreuter et al. [44]
Εντυπωσιακά είναι τα αποτελέσματα των δύο πρώτων στυλομετρικών χαρακτηριστικών - μέσο πλήθος συλλαβών ανά λέξη, ποσοστό συμπίεσης ρημάτων. Το ποσοστό συμπίεσης ρημάτων χρησιμοποιείται εδώ πρώτη φορά, και το οποίο εξάγαμε με έναν απλό αλγόριθμο, με αφορμή άλλους ιδιαίτερα πολύπλοκους αλγορίθμους.

6.2 Perceptron πολλών στρωμάτων

Το νευρωνικό δίκτυο Perceptron πολλών-στρωμάτων δεν ανταποκρινόταν χωρίς εξισορρόπηση των δεδομένων εκπαίδευσης. Συγκεκριμένα, τα παραδείγματα της κλάσεως μειοψηφίας ήταν πολύ λίγα για να ανταγωνιστούν την υπερπληθώρα παραδειγμάτων της κλάσεως πλειοψηφίας, με αποτέλεσμα το δίκτυο να αδυνατεί να εκπαιδευτεί για πρόβλεψη των πρώτων. Κατά τον έλεγχο, οι προβλέψεις γίνονταν υπέρ την κλάσεως πλειοψηφίας για το 100% των παραδειγμάτων ελέγχου. Εφαρμόζοντας τον αλγόριθμο SMOTE borderline για εξισορρόπηση των δεδομένων εκπαίδευσης, το σύστημα κατάφερε να ανταποκριθεί. Πειραματιστήκαμε με αυτή τη μέθοδο εκπαίδευσης και για τα τρία είδη μετακινούμενου παραθύρου . Τα αποτελέσματα από τη μέθοδο 5-στρώσεων διασταυρωμένη επικύρωση παρουσιάζονται στον Πίνακα 6.7. Σημειώνεται ότι σε όλα τα αποτελέσματα που παρουσιάζονται εδώ για το δίκτυο Perceptron, ένα σετ προτάσεων θεωρείται προϊόν λογοκλοπής, εάν τουλάχιστον 2 από τα 3 παράθυρα που το καλύπτουν θεωρούνται.

Πίνακας 6.7 Αποτελέσματα 5-στρώσεων διασταυρωμένης επικύρωσης, με Perceptron πολλών στρωμάτων και εξισορρόπηση δεδομένων εκπαίδευσης με *SMOTE borderline*.

Χαρακτηριστικά συστήματος	Μετακινούμενο παράθυρο	Precision	Recall	F-measure
Όλα	$\{\mu.\pi. = 15, \beta = 5\}$	0.193	0.528	0.283
	$\{\mu.\pi. = 30, \beta = 10\}$	0.223	0.434	0.295
	$\{\mu.\pi. = 3k, \beta = k\}$, $k = \{3, 5, 10\}$	0.082	0.887	0.151
Μόνο στυλομετρικά	$\{\mu.\pi. = 15, \beta = 5\}$	0.155	0.166	0.160
	$\{\mu.\pi. = 30, \beta = 10\}$	0.118	0.542	0.157
	$\{\mu.\pi. = 3k, \beta = k\}$, $k = \{3, 5, 10\}$	0.092	0.581	0.159
Μόνο σημασιολογικά	$\{\mu.\pi. = 15, \beta = 5\}$	0.143	0.601	0.231
	$\{\mu.\pi. = 30, \beta = 10\}$	0.198	0.487	0.281
	$\{\mu.\pi. = 3k, \beta = k\}$, $k = \{3, 5, 10\}$	0.103	0.746	0.182

Παρατηρούμε πως, συγκρίνοντας με τα Δέντρα Απόφασης, τα αποτελέσματα είναι ελαφρώς χειρότερα, με τη ψαλίδα να ανοίγει στην περίπτωση των στυλομετρικών χαρακτηριστικών. Παρολαυτά, λαμβάνοντας υπόψιν τα διαγωνιζόμενα συστήματα στον PAN'11 και τα ποσοστά επιτυχίας στον κλάδο, γενικότερα, το σύστημα φαίνεται να είναι ανταγωνιστικό, κάτι που θα ήταν αδύνατο χωρίς την εξισορρόπηση των δεδομένων εκμάθησης.

Το Perceptron πολλών-στρωμάτων παρουσίασε, και αυτό, διακύμανση κατά

την εξαγωγή των αποτελεσμάτων στους 5 χωρισμούς του σώματος δεδομένων. Στον Πίνακα 6.8 δίνονται τα αποτελέσματα για τους 5 χωρισμούς του σώματος δεδομένων.

Πίνακας 6.8 Αποτελέσματα για τους 5 χωρισμούς του κειμένου, με Perceptron πολλών-στρωμάτων και εξισορρόπηση δεδομένων εκπαίδευσης με *SMOTE borderline*. Μετακινούμενο παράθυρο με *μ.π.* = 15, *β* = 5 προτάσεις.

Χαρακτηριστικά συστήματος	***F-measure***				
	split1	***split2***	***split3***	***split4***	***split5***
Όλα	0.267	0.186	0.252	0.263	0.303
Μόνο στυλομετρικά	0.033	0.117	0.199	0.143	0.160
Μόνο σημασιολογικά	0.252	0.155	0.155	0.231	0.238

6.3 Naive Bayes

Η εκπαίδευση με Naive Bayes έδωσε τα ίδια, χείριστα, αποτελέσματα με το Perceptron πολλών-στρωμάτων για μη-εξισορρόπηση των δεδομένων εκπαίδευσης. Όλα τα παραδείγματα ελέγχου προβλέπονταν ως μη-λογοκλεμμένα, λόγω της ανισόρροπης εκπαίδευσης. Σε αντίθεση με το Perceptron πολλών-στρωμάτων, αυτή η μέθοδος εκπαίδευσης δεν απέδωσε με εφαρμογή της εξισορρόπησης δεδομένων. Τα αποτελέσματα για μετακινούμενο παράθυρο μήκους 15 προτάσεων και βήματος 5 προτάσεων, σε 5-στρώσεων διασταυρωμένη επικύρωση παρουσιάζονται στον Πίνακα 6.9.

Πίνακας 6.9 Αποτελέσματα 5-στρώσεων διασταυρωμένης επικύρωσης, με Naive Bayes ακι εξισορρόπηση δεδομένων εκπαίδευσης με SMOTE borderline. Μετακινούμενο παράθυρο με *μ.π.* = 15, *β* = 5 προτάσεις.

Χαρακτηριστικά συστήματος	**Precision**	**Recall**	**F-measure**
Όλα	0.073	0.512	0.128
Μόνο στυλομετρικά	0.073	0.513	0.128
Μόνο σημασιολογικά	0.055	0.999	0.105

Βλέπουμε πως η ακρίβεια στα λογοκλεμμένα κομμάτια κειμένου βυθίζεται σε τιμές κάτω του 10%. Παρατηρούμε ανικανότητα της εκπαίδευσης να αντλήσει και να εκμεταλλευτεί την πληροφορία που κρύβεται στα σημασιολογικά χαρακτηριστικά · η ανάκληση των λογοκλεμμένων χωρίων παίρνει ουσιαστικά ποσοστό

100% με ταυτόχρονη ισοπέδωση του ποσοστού της ακρίβειας, που σημαίνει ότι δεν υφίσταται διαφοροποίηση μεταξύ των λογοκλεμμένων και μη χωρίων για το εκπαιδευμένο μοντέλο. Έτσι, τα αποτελέσματα για τα στυλομετρικά χαρακτηριστικά από τη μία, και το σύνολο των χαρακτηριστικών, από την άλλη, σχεδόν ταυτίζονται. Αυτό φανερώνει και την αδυναμία του Naive Bayes σε σύγκριση με τα Δέντρα Απόφασης, να ανακαλύψει συσχετισμούς μεταξύ διαφορετικών χαρακτηριστικών. Αυτός ο αλγόριθμος εκπαίδευσης αποδείχτηκε ο πλέον ακατάλληλος για το σύστημά μας και γι' αυτό το λόγο δεν έγιναν περαιτέρω πειραματισμοί με τις παραμέτρους του συστήματος.

6.4 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης αποδείχτηκαν η πιο ισχυρή μέθοδος εκπαίδευσης τόσο σε ποσοστά επιτυχίας προβλέψεων όσο και σε ευστάθεια. Έχουν όμως το μειονέκτημα των υψηλών απαιτήσεων σε χρόνο επεξεργασίας, ενώ απαιτούν συγκριτικά μεγάλες ποσότητες μνήμης κατά την εκπαίδευση. Συγκεκριμένα, για έναν χωρισμό του σώματος δεδομένων οι υπόλοιποι ταξινομητές χρειάζονταν περίπου 1min, ενώ το SVM περίπου 40h στη μέση περίπτωση. Αυτό ήταν και το βασικό εμπόδιο στο να πειραματιστούμε εκτενώς με αυτό το εργαλείο. Τα αποτελέσματα για τις Μηχανές Διανυσμάτων Υποστήριξης αφορούν μόνο στο μετακινούμενο παράθυρο με παραμέτρους *μ.π. = 15* και *β = 5 προτάσεις*. Αυτά τα αποτελέσματα αρκούν για να αναδείξουν την υπεροχή τους έναντι των υπόλοιπων ταξινομητών.

Ο Πίνακας 6.10 περιλαμβάνει τα αποτελέσματα από την 5-στρώσεων διασταυρωμένη επικύρωση, για τις τρεις κατηγορίες των χαρακτηριστικών.

Πίνακας 6.10 Αποτελέσματα 5-στρώσεων διασταυρωμένης επικύρωσης, με SVM και εξισορρόπηση δεδομένων εκπαίδευσης με SMOTE borderline. Μετακινούμενο παράθυρο με *μ.π.* = 15, *β* = 5 προτάσεις.

Χαρακτηριστικά συστήματος	Precision	Recall	F-measure
Όλα	0.246	0.506	0.331
Μόνο στυλομετρικά	0.201	0.288	0.237
Μόνο σημασιολογικά	0.179	0.516	0.267

Στον Πίνακα 6.11 φαίνονται τα αποτελέσματα για τη μετρική F-measure για κάθε έναν από τους 5 χωρισμούς του σώματος δεδομένων σε δεδομένα εκπαίδευσης και ελέγχου.

Παρατηρούμε πως υπάρχει πολύ μικρή διακύμανση των τιμών για τους 5 χωρισμούς του σώματος δεδομένων, γεγονός που καθιστά το σύστημα ευσταθές. Αυτό

Πίνακας 6.11 Αποτελέσματα για τους 5 χωρισμούς του κειμένου, με SVM και εξισορρόπηση δεδομένων εκπαίδευσης με *SMOTE borderline*. Μετακινούμενο παράθυρο με *μ.π.* = 15, *β* = 5 προτάσεις.

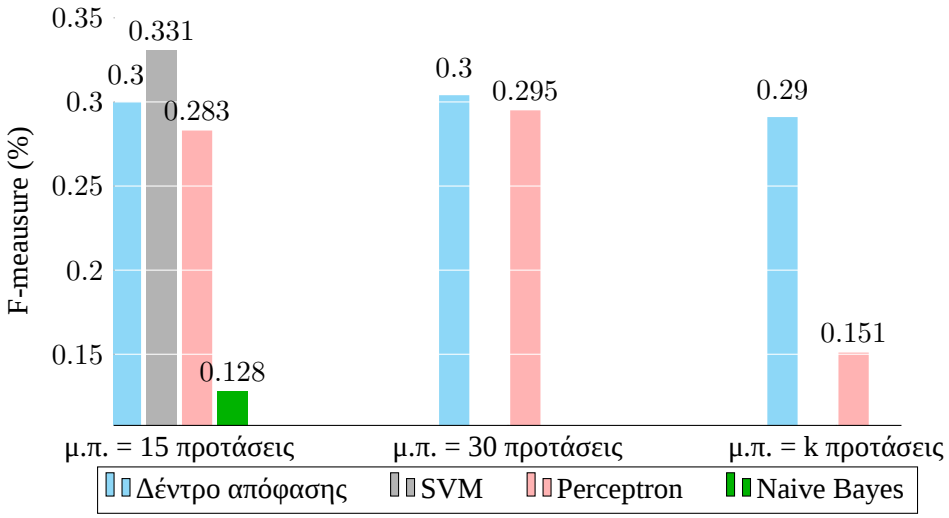
Χαρακτηριστικά συστήματος	*F-measure*				
	split1	***split2***	***split3***	***split4***	***split5***
Όλα	0.325	0.322	0.351	0.335	0.320
Μόνο στυλομετρικά	0.228	0.222	0.261	0.243	0.227
Μόνο σημασιολογικά	0.267	0.263	0.282	0.267	0.250

σημαίνει μεγαλύτερη αξιοπιστία αποτελεσμάτων κατά την εφαρμογή του μοντέλου πρόβλεψης σε άγνωστο σώμα παραδειγμάτων. Αναμένουμε, δηλαδή, οι προβλέψεις να είναι επιτυχείς σε ποσοστό περίπου 32% στο άγνωστο σώμα, σε συμφωνία με τη μέθοδο επικύρωσης.

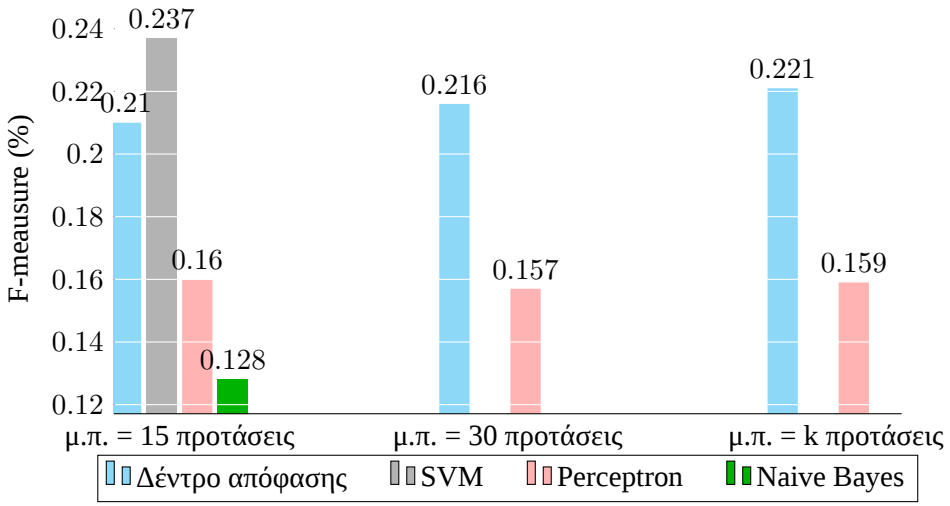
6.5 Σύγκριση των μεθόδων εκπαίδευσης

Θα συγκρίνουμε τα 4 μοντέλα εκπαίδευσης συγκεντρώνοντας με παραστατικό τρόπο τα αποτελέσματά τους. Αρχικά θα παρουσιάσουμε τη γενική εικόνα, συγκρίνοντας τους μέσους όρους όπως προκύπτουν από τη μέθοδο επικύρωσης σε διαγράμματα ραβδογράμματος. Σε δεύτερο στάδιο θα ρίξουμε μια βαθύτερη ματιά, συγκρίνοντας τόσο τις τιμές και την διακύμανσή τους στους 5 χωρισμούς του σώματος δεδομένων.

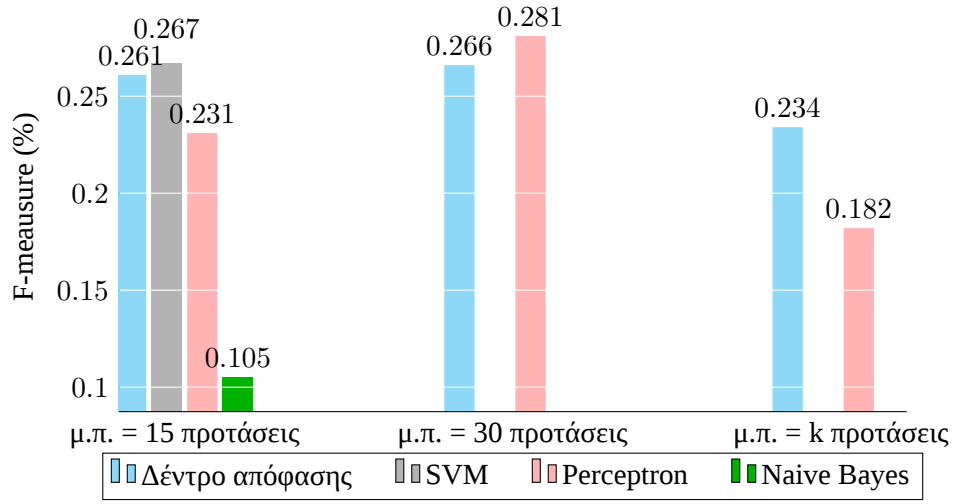
Όλα τα χαρακτηριστικά – Εξισορρόπηση δεδομένων με *SMOTE borderline*



Στυλομετρικά χαρακτηριστικά– Εξισορρόπηση δεδομένων με *SMOTE borderline*

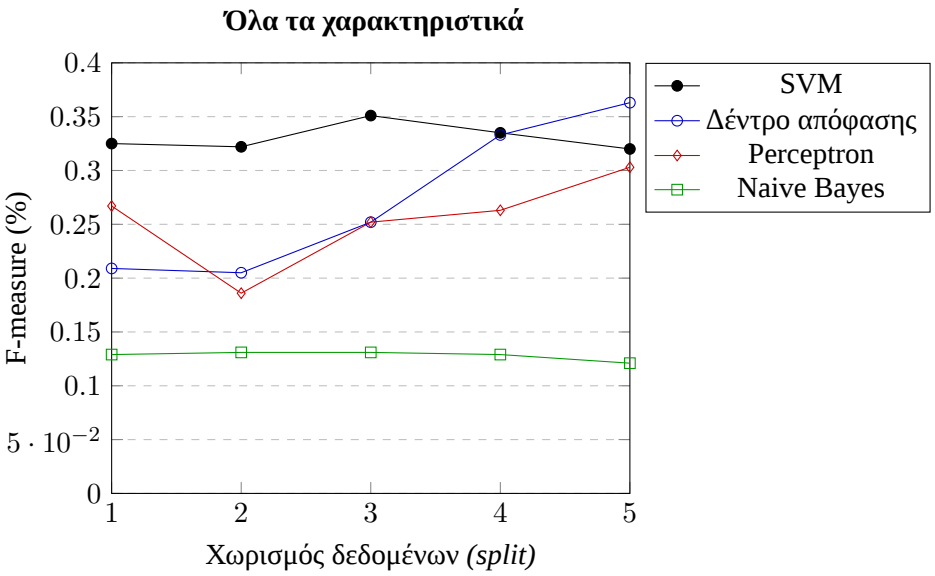


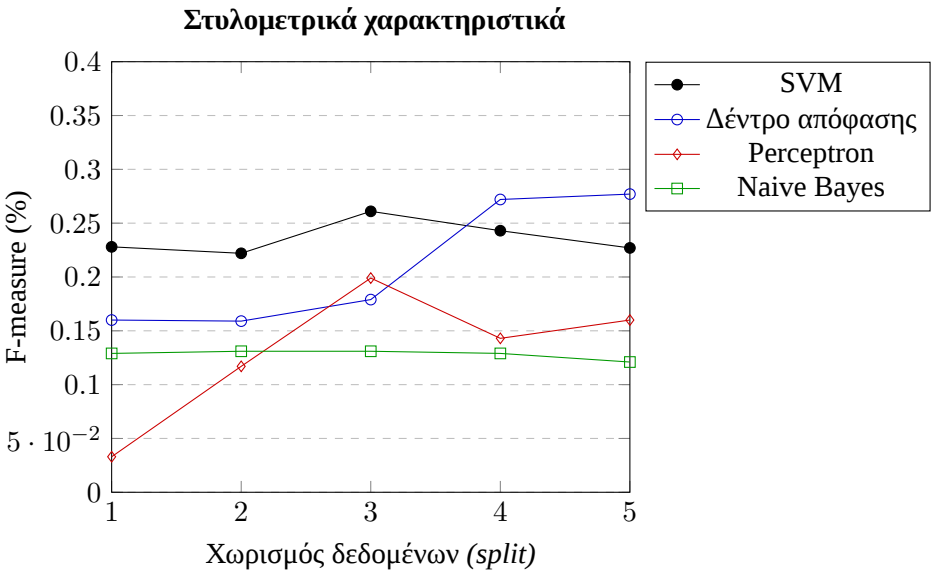
Σημασιολογικά χαρακτηριστικά– Εξισορρόπηση δεδομένων με *SMOTE borderline*

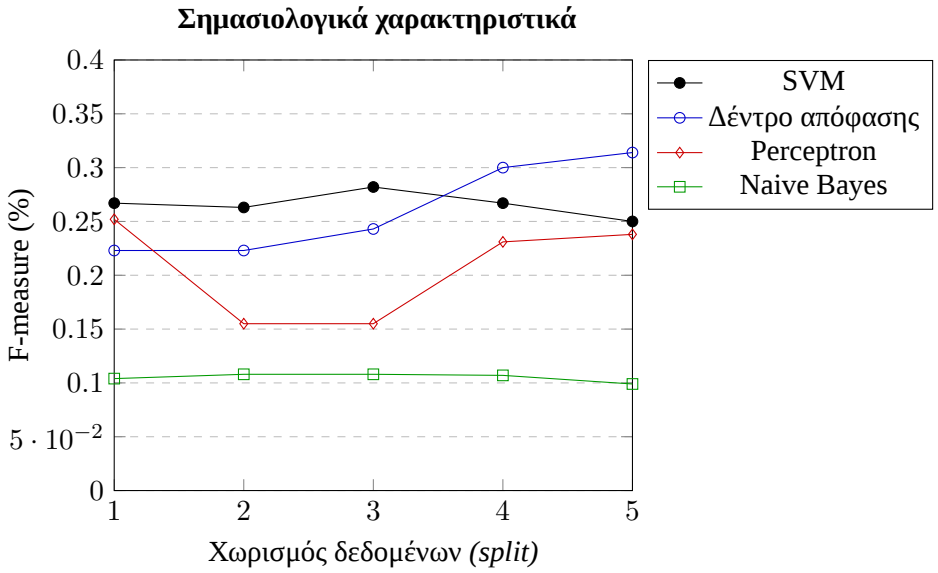


Από τα ραβδογράμματα είναι εμφανής η υπεροχή των Μηχανών Διανυσμάτων Υποστήριξης, ενώ ακολουθούν τα Δέντρα Απόφασης με ανταγωνιστικά αποτελέσματα. Το νευρωνικό δίκτυο Perceptron φαίνεται επίσης ότι προσπαθεί να ανταγωνιστεί τα Δέντρα Απόφασης, αν και με τα αποτελέσματα στην περίπτωση των στυλομετρικών χαρακτηριστικών τοποθετούνται κατώτερα στην ιεραρχία. Τέλος, η χειρότερη είναι η περίπτωση της Naive Bayes μεθόδου, που και στις 3 κατηγορίες χαρακτηριστικών δίνει αποτελέσματα σε πολύ χαμηλές τιμές.

Στις γραφικές παραστάσεις που ακολουθούν φαίνονται τα αποτελέσματα των 4 μεθόδων εκπαίδευσης για τους 5 χωρισμούς του σώματος δεδομένων ξεχωριστά, για παραμετροποίηση του μετακινούμενου παραθύρου μ.π.=15 προτάσεις, β = 5 προτάσεις.







Τα διαγράμματα αυτά μας επιτρέπουν να δούμε ένα επίπεδο βαθύτερα στο σύστημα και στην αξιοπιστία της κάθε μεθόδου εκπαίδευσης. Παρατηρούμε πως ο μέσος όρος των τιμών των επιμέρους αποτελεσμάτων στους 5 χωρισμούς του σώματος δεδομένων, μπορεί να είναι παραπλανητικός · μία ανταγωνιστική τιμή στο μέσο όρο μπορεί να έχει προκύψει από τεράστιες διακυμάνσεις των τιμών κατά τη διαδικασία επικύρωσης. Ο συνδυασμός ενός πολύ χαμηλού και ενός υψηλού ποσοστού επιτυχίας πρόβλεψης, μπορεί να καταλήγει σε έναν λογικό και ανταγωνιστικό μέσο όρο, όμως φανερώνει χαμηλή αξιοπιστία. Προορισμός κάθε μοντέλου μηχανικής μάθησης είναι η χρήση του σε πραγματικά δεδομένα, για τα οποία συνήθως δεν γνωρίζουμε τις σωστές απαντήσεις. Για ένα σύστημα με μεγάλες διακυμάνσεις σε ένα ενιαίο σώμα δεδομένων, κάθε υπόθεση για το ποσοστό επιτυχίας προβλέψεων σε ένα εντελώς άγνωστο σώμα ελέγχου θα είναι εντελώς αναξιόπιστη · δεν μπορούμε να γνωρίζουμε αν το συγκεκριμένο σώμα ελέγχου αποτελεί μια ευνοϊκή περίπτωση για το μοντέλο ή όχι. Αντίθετα, όταν ένα μοντέλο έχει επιδείξει σταθερότητα κατά τα επιμέρους βήματα της μεθόδου επικύρωσης, έχουμε έναν πολύ καλό λόγο να πιστεύουμε ότι η απόδοσή του δε θα αλλάξει δραματικά στο νέο σώμα ελέγχου.

Βλέπουμε, λοιπόν, πως αστάθεια παρουσιάζει το δίκτυο Perceptron, ειδικά στην περίπτωση των στυλομετρικών χαρακτηριστικών, ενώ και τα Δέντρα Απόφασης δεν δίνουν πολύ σταθερά αποτελέσματα, όμως παρουσιάζουν μια κανονικότητα στη διακύμανση ως προς τους 5 χωρισμούς του σώματος δεδομένων. Συγκεκριμένα, φαίνεται πως αποδίδουν καλύτερα για τα κείμενα που βρίσκονται προς το τέλος του σώματος δεδομένων.

Οι Μηχανές Διανυσμάτων Υποστήριξης παρουσιάζουν ευσταθή συμπεριφορά, γεγονός που επισφραγίζει την υπεροχή και την καταλληλότητά τους για το σύστημά μας.

Η μέθοδος Naive Bayes δίνει, επίσης, σταθερές, πλην όμως πολύ χαμηλές τιμές.

6.6 Ανταγωνιστικότητα συστήματος - Σύγκριση με τα αποτελέσματα του PAN'11

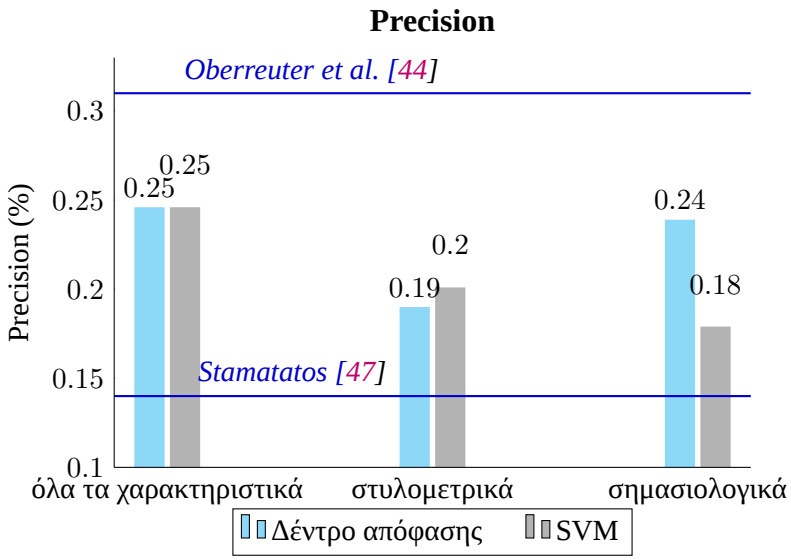
Θα συγκρίνουμε το σύστημα για τα δύο καλύτερα μοντέλα εκπαίδευσης - Μηχανές Διανυσμάτων Υποστήριξης και Δέντρα Απόφασης, με εξισορρόπηση των δεδομένων εκπαίδευσης με τον αλγόριθμο SMOTE borderline - με τα συστήματα των δύο καλύτερων συμμετοχών στο διαγωνισμό του PAN'11, καθώς και με το σύστημα που κέρδισε το διαγωνισμό του PAN το 2009. Τα αποτελέσματα είναι άμεσα συγκρίσιμα, αφού αφορούν το ίδιο σώμα δεδομένων, με τη διαφορά ότι τα αποτελέσματα των συστημάτων του διαγωνισμού αφορούν ένα συγκεκριμένο χωρισμό του σώματος δεδομένων, που, όμως, δεν τον γνωρίζουμε. Το γεγονός αυτό ενδέχεται να επηρεάζει ελαφρώς τη σύγκριση του μοντέλου εκπαίδευσης με τα Δέντρα Απόφασης λόγω της διακύμανσης των αποτελεσμάτων κατά τους πειραματισμούς κατά μήκος του σώματος δεδομένων. Εν πάσει περιπτώσει, ακόμα και στην περίπτωση των Δέντρων Απόφασης η αναντιστοιχία δεν μπορεί να επηρεάσει σημαντικά τα συμπεράσματα. Στα επόμενα θεωρούμε πως δεν υπάρχει καμιά αναντιστοιχία στα αποτελέσματα.

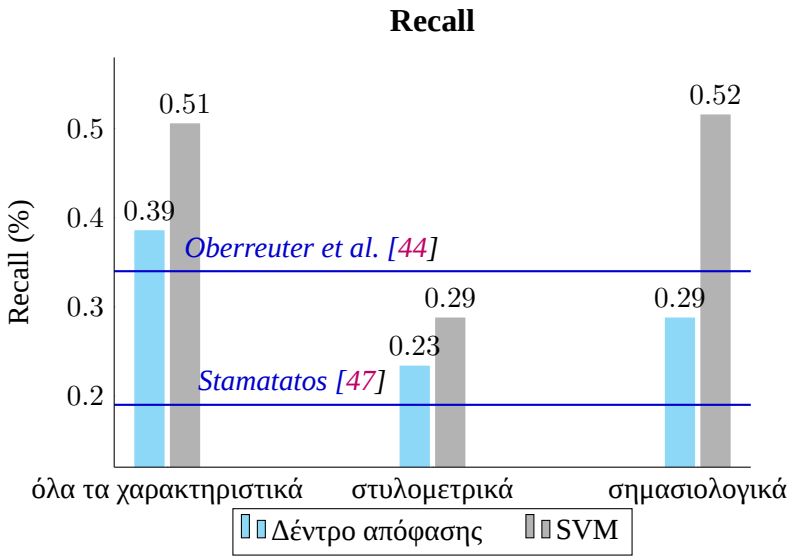
Τα αναλυτικά αποτελέσματα του διαγωνισμού έχουν δοθεί στον Πίνακα 4.2.

Υπενθυμίζουμε ότι τα αποτελέσματα του συστήματος των Oberreuter et al. [44] θεωρήθηκαν πολύ υψηλά για τα δεδομένα του προβλήματος. Η γενική αίσθηση ήταν πως πρόκειται για παραπλανητικές τιμές, λόγω της καθαρά σημασιολογικής προσέγγισης σε ένα σώμα δεδομένων, όπου τα λογοκλεμμένα χωρία έχουν επιλεγεί από διάφορες πηγές και εισαχθεί στα κείμενα τυχαία , με τη σχετικότητα του θέματος να *μην* αποτελεί κριτήριο. Έτσι, θεωρήθηκε πως το σύστημα δεν θα μπορούσε να αποδώσει σε πραγματικά δεδομένα, ή σε ένα πιο καλοστημένο σώμα δεδομένων. Προκειμένου να αποφύγουμε παρόμοιες επικρίσεις, ένα σκέλος των αποτελεσμάτων μας αποτελούν τα καθαρά στυλομετρικά χαρακτηριστικά. Ωστόσο, υποστηρίζουμε πως η ποικιλία των χαρακτηριστικών σε συνδυασμό με το κατάλληλο μοντέλο εκμάθησης, μπορεί να ανακαλύψει λανθάνοντες συσχετισμούς και να αυξήσει θεαματικά την αποτελεσματικότητα του συστήματος.

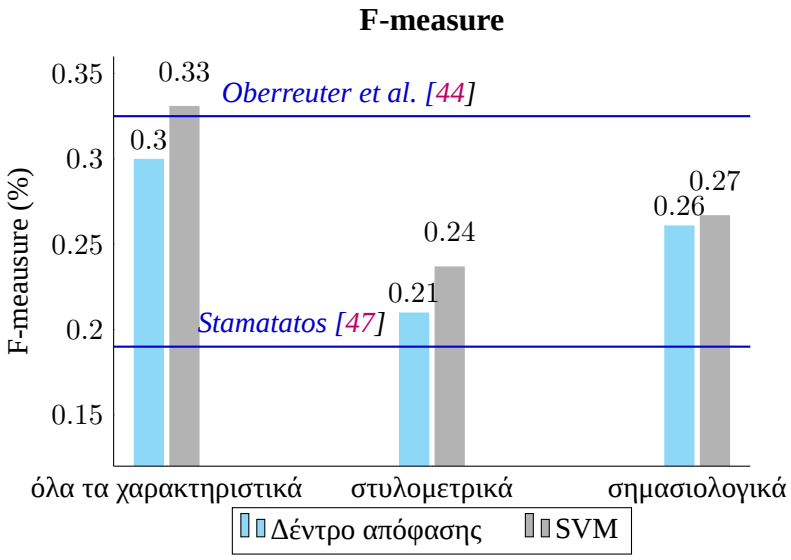
Ως δεύτερο κριτή έχουμε το σύστημα του Stamatatos [47], το οποίο ξεπέρασε τα υπόλοιπα τρία συστήματα του διαγωνισμού του 2011 και θεωρείται πιο "αντικειμενικό".

Ακολουθούν 3 ραβδογράμματα για τις τρεις μετρικές αξιολόγησης, Precision, Recall και F-measure, με την τελευταία να αποτελεί και το τελικό κριτήριο για την σύγκριση των συστημάτων.
Όσον αφορά στο Δέντρο Απόφασης, οι τιμές που παρουσιάζονται δεν είναι οι καλύτερες όλων των πειραμάτων που έγιναν για κάθε ξεχωριστή περίπτωση. Για λόγους συνέπειας, δείχνουμε τα αποτελέσματα για μετακινούμενο παράθυρο μ.π. = 15 προτάσεις, β = 5 προτάσεις.


Precision
Oberreuter et al. [44]
Stamatatos [47]
Precision (%)
0.3
0.25
0.2
0.15
0.1
0.25
0.25
0.19
0.2
0.24
0.18
όλα τα χαρακτηριστικά
στυλομετρικά
σημασιολογικά
Δέντρο απόφασης
SVM


Recall
Oberreuter et al. [44]
Stamatatos [47]
Recall (%)
0.5
0.4
0.3
0.2
0.39
0.51
0.23
0.29
0.29
0.52
όλα τα χαρακτηριστικά
στυλομετρικά
σημασιολογικά
Δέντρο απόφασης
SVM

.



Βλέπουμε αρχικά ότι το σύστημά μας ξεπερνάει αυτό του Stamatatos [47], τόσο με SVM όσο και με Δέντρα Απόφασης, σε όλες τις μετρικές και για τις 3 κατηγορίες χαρακτηριστικών, και μάλιστα με πολύ μεγάλη διαφορά για την εκμάθηση με SVM. Πολύ σημαντικά είναι τα αποτελέσματα στην κατηγορία των στυλομετρικών χαρακτηριστικών, επειδή είναι εντελώς ”ουδέτερα” ως προς το σώμα δεδομένων, δηλαδή δεν εκμεταλλεύονται την αδυναμία του των διαφορετικών θεμάτων. Σε αυτή την κατηγορία ουσιαστικά ανταγωνιζόμαστε το σύστημα του Stamatatos [47] και καταφέρνουμε να το ξεπεράσουμε με τρόπο που δεν αφήνει περιθώρια αμφισβήτησης της υπεροχής του συστήματός μας.

Σε σύγκριση με το σύστημα των Oberreuter et al. [44], βλέπουμε πως το σύστημα με όλα τα χαρακτηριστικά και εκμάθηση με SVM καταφέρνει για λίγο μεν, αλλά πάντως να το ξεπεράσει. Και στην περίπτωση των Δέντρων Απόφασης το σύστημα καταφέρνει να είναι ανταγωνιστικό.

Σε κάθε περίπτωση βλέπουμε πως το σύστημα παρουσιάζει αυξημένες τιμές στην ανάκληση των λογοκλεμμένων χωρίων και χαμηλότερες στην ακρίβεια. Αυτό είναι ένα πεδίο όπου μπορεί να βελτιωθεί το σύστημα. Θυμίζουμε πως κατά την προετοιμασία των δεδομένων εκπαίδευσης ένα χωρίο χαρακτηριζόταν ”λογοκλεμμένο” ήταν τέτοιο κατά το 40% της έκτασής του. Η αύξηση αυτού του ποσοστού, π.χ. στο 50%, θα δρούσε σε βάρος της ανάκλησης αλλά υπέρ της ακρίβειας πρόβλεψης.

Κεφάλαιο 7

Επίλογος

7.1 Συμπεράσματα

Στυλιστικά χαρακτηριστικά
Από τον Πίνακα 6.6 φαίνεται πως κάποια από τα χαρακτηριστικά που σχεδιάσαμε και εφαρμόσαμε έχουν πολύ καλές προοπτικές. Συγκεκριμένα, ιδιαιτέρως καλές επιδόσεις πετυχαίνουν τα

- ποσοστό συμπίεσης ρημάτων
- μέσος όρος των αρνητικών διαφορών των κλάσεων συχνοτήτων λέξεων κειμένου-χωρίου για τις συχνά εμφανιζόμενες, σε όλο το κείμενο, λέξεις (wfc_4)
- μέσος όρος των θετικών διαφορών των κλάσεων συχνοτήτων λέξεων κειμένου-χωρίου (wfc_1)

Μέθοδος εξαγωγής λογοκλεμμένων χωρίων
Σύμφωνα με την κλασική τακτική για τον εντοπισμό των στυλιστικών ανομοιομορφιών, επιλέγεται μια συνάρτηση απόστασης για τη σύγκριση χωρίου-κειμένου (ή χωρίου- χωρίου), από την οποία εξάγεται μια τιμή - η απόσταση των συγκρινόμενων μερών βάσει του στυλιστικού διανύσματος τιμών.

Σε αυτή την εργασία ακολουθήσαμε μια καινούρια προσέγγιση που δίνει περισσότερη ελευθερία στη μηχανική μάθηση. Αντί να χρησιμοποιήσουμε μια συνάρτηση απόστασης που θα συγκρίνει τις τιμές των διάφορων στυλιστικών χαρακτηριστικών συλλήβδην και θα τις ενσταλάζει σε μια τιμή, επιλέξαμε να κρατήσουμε το διάνυσμα τιμών ως στυλιστικό αποτύπωμα για κάθε χωρίο, αφού ενσωματώσουμε τη σύγκριση με το σύνολο του κειμένου σε καθένα από τα χαρακτηριστικά του διανύσματος. Με αυτόν τον τρόπο αφενός αποφεύγουμε την ισοπέδωση της ιδιαιτερότητας κάθε ξεχωριστού χαρακτηριστικού μέσω της ενιαίας αξιολόγησης της ”μικρής” ή ”μεγάλης” απόστασης, ενώ ταυτόχρονα αφήνουμε στο μοντέλο εκμάθησης να ανακαλύψει πιθανούς συσχετισμούς μεταξύ των χαρακτηριστικών.

Εξισορρόπηση των δεδομένων εκπαίδευσης
Σε αυτήν την εργασία θίχτηκε, για πρώτη φορά, το πρόβλημα των μη ισορροπημένων δεδομένων στην εγγενή ανίχνευση λογοκλοπής. Εφαρμόσαμε γνωστές μεθόδους εξισορρόπησης και αποδείχτηκε πως βοηθούν σημαντικά στην αποτελεσματικότητα του συστήματος. Εκτός αυτού, χάρη στην εξισορρόπηση των δεδομένων εκπαίδευσης μπορέσαμε να εφαρμόσουμε κάποια μοντέλα εκμάθησης, που δεν ανταποκρίνονταν χωρίς αυτή και αποδείχτηκαν ιδιαίτερα αποτελεσματικά (Μηχανές Διανυσμάτων Υποστήριξης).

Precision, Recall
Σε όλα μας τα αποτελέσματα παρατηρούνται τιμές ανάκλησης αρκετά υψηλότερες από αυτές της ακρίβειας. Πιστεύουμε πως αυτό οφείλεται στις σχεδιαστικές επιλογές για τα δεδομένα εκπαίδευσης. Συγκεκριμένα επιλέξαμε ένα χωρίο να θεωρείται λογοκλεμμένο αν τουλάχιστον 40% της έκτασής του είναι λογοκλεμμένο. Η αύξηση του ποσοστού αυτού πιστεύουμε ότι θα κατέληγε να εξισορροπήσει τις τιμές των Precision, Recall να βελτιώσει τα τελικά αποτελέσματα.

Κατάτμηση κειμένου
Υψηλότερα αποτελέσματα έδωσε το μετακινούμενο παράθυρο για σταθερό μήκος παραθύρου (μ.π. = 15, β = 5 προτάσεις και μ.π. = 30, β = 10 προτάσεις) σε σχέση με το μεταβλητό μήκος παραθύρου. Πιστεύουμε πως αυτό οφείλεται στις χαμηλές τιμές τις κλίμακας που εφαρμόσαμε · αφού το σταθερό μήκος παραθύρου 30 προτάσεων έδωσε τόσο υψηλά αποτελέσματα, η χαμηλή κλίμακα 9 προτάσεων ήταν μάλλον άστοχη και ζημίωσε ένα μέρος της ανάλυσης.

7.2 Προτάσεις για μελλοντική έρευνα

Πιστεύουμε πως σημαντικό κομμάτι της έρευνας στην εγγενή ανίχνευση λογοκλοπής πρέπει να είναι η κατάτμηση του κειμένου με τεχνικές που επιτρέπουν την προσαρμοστικότητα στο μέγεθος του λογοκλεμμένου χωρίου. Παρόλο που στα πειράματα αυτής της εργασίας το μετακινούμενο παράθυρο μεταβλητού μήκους δεν έδωσε τα καλύτερα αποτελέσματα, είναι ένα πρώτο βήμα προς αυτήν την κατεύθυνση που θα πρέπει να διερευνηθεί εκτενέστερα.

Βιβλιογραφία

[1] M. Kestemont, K. Luyckx, and W. Daelemans, "Intrinsic plagiarism detection using character trigram distance scores," *Notebook for PAN at CLEF 2011*.

[2] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3558294/.

[3] K. J. Ottenstein, "An algorithmic approach to the detection and prevention of plagiarism," 1976.

[4] S. Grier, "A tool that detects plagiarism in pascal programs," *SIGCSE '81 Proceedings of the twelfth SIGCSE technical symposium on Computer science education*, pp. 15–20, 1981.

[5] N. S. Salha M. Alzahrani and A. Abraham, "Understanding plagiarism linguistic patterns, textual features, and detection methods," *IEEE Transactions on Systems, Man. and Cybernetics*, vol. 42, no. 2, pp. 133–149, March 2012.

[6] Z. Ceska, M. Toman, and K. Jezek, "Multilingual plagiarism detection," ser. Lecture Notes in Computer Sciennce (including subseries Lecture Notes in Artificial Intelligence Lecture Notes in Bioinformatics), vol. 5253 LNAI, 2008, pp. 83–92.

[7] S. M. zu Eissen and B. Stein, "Intrinsic plagiarism detection," in *Proceedings of the 28th European Conference on IR Research*. London, UK: Springer, 2006, pp. 565–569.

[8] M. Potthast, A. Eiselt, A. Barron-Cedeno, B. Stein, and P. Rosso, "Overview of the 3rd international competition on plagiarism detection," 2011.

[9] http://pan.webis.de.

[10] B. Stein, N. Lipka, and P. Prettenhofer, "Intrinsic plagiarism analysis," *Language Resources and Evaluation*, January 2010.

[11] O. Ferret, "How to thematically segment texts using lexical cohesion?"

[12] http://www.text-analytics101.com/2014/11/what-are-n grams.html.

[13] N. Cheng, R. Chandramouli, and K. Subbalakshmi, "Author gender identification from text," *Digital Investigation*, vol. 8, pp. 78–88, April 2011.

[14] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced."

[15] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[16] N. Littlestone and M. Warmuth, "The weighted majority algorithm," in *Information and Computation 108*, 1994, pp. 262–273.

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1," D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds. Cambridge, MA, USA: MIT Press, 1986, ch. Learning Internal Representations by Error Propagation, pp. 318–362.

[18] G. P. Zhang, "Neural networks for classification: A survey," *Trans. Sys. Man Cyber Part C*, vol. 30, no. 4, pp. 451–462, Nov. 2000.

[19] R. Beale and T. Jackson, *Neural Computing: An Introduction*. Philadelphia: Hilger, 1991.

[20] S. Bubeck, "Convex optimization: Algorithms and complexity," *Found. Trends Mach. Learn.*, vol. 8, no. 3-4, pp. 231–357, Nov. 2015.

[21] Murthy, "Automatic construction of decision trees from data: A multi-disciplinary survey," in *Data Mining and Knowledge Discovery 2*, 1998, pp. 345–389.

[22] S. B. Kotsiantis, "Supervised machine learning: A review of classification," in *Informatica 31*, 2007, pp. 249–268.

[23] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Wadsforth International Group, 1984.

[24] E. Hunt, J. Martin, and P. Stone, *Experiments in Induction*. New York: Academic Press, 1966.

[25] L. Breslow and D. Aha, "Simplifying decision trees: A survey." in *Knowledge Engineering Review 12*, 1997, pp. 1–40.

[26] H. Zhang, "The Optimality of Naive Bayes." in *FLAIRS Conference*, V. Barr and Z. Markov, Eds. AAAI Press, 2004.

[27] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "Theoritical foundations of the potential function method in pattern recognition learning," in *Automation and Remote Control*, 1964, pp. 821–837.

[28] B. E. Boser, I. M. Guyon, and V. Vapnik, “A training algortihm for optimal margin classifiers,” in *Proceedings of the fifth annual workshop on computational learning theory - COLT ’92*, 1992, p. 144.

[29] V. Ganganwar, “An overview of classification algorithms for imbalanced datasets,” *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, April 2012.

[30] C. Elkan, “The foundations of cost-sensitive learning,” in *Proceedings of the 17th IEEE International Joint Conference on Artificial Intelligence (IJCAI’01)*, 2001, pp. 973–978.

[31] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics,” *Information Sciences*, vol. 250, pp. 113–141, 2013.

[32] G. Batista, R. Prati, and M. Monard, “A study of the behaviour of several methods for balancing machine learning training data,” in *SIGKDD Explorations 6*, vol. 1, 2012, pp. 20–29.

[33] B. Zadrozny and C. Elkan, “Learning and making decisions when costs and probabilities are both unknown,” in *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining (KDD’01)*, 2001, pp. 204–213.

[34] B. Zadrozny, J. Langford, and N. Abe, “Cost-sensitive learning by cost-proportionate example weighting,” in *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM’03)*, 2003, pp. 435–442.

[35] T. Ivan, “An experiment with the edited nearest-neighbor rule,” in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 6. IEEE, 1976, pp. 448–452.

[36] P. Hart, “The condensed nearest neighbor rule,” in *IEEE Transactions on Information Theory*, vol. 14. IEEE, 2003, pp. 515–516.

[37] M. Kubat and S. Matwin, “Addressing the curse of imbalanced training sets: One-sided selection,” in *In Proceedings of the Fourteenth International Conference on Machine Learning*. Morgan Kaufmann, 1997, pp. 179–186.

[38] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, “Smote: Synthetic minority oversampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[39] B. Wang and N. Japkowicz, “Imbalanced data set learning with synthetic examples,” in *Proceedings of the IRIS Machine Learning Workshop*, 2004.

[40] H. Han, W. Wang, and B. Mao, “Borderline-smote: a new oversampling method in imbalanced data sets learning,” in *Proceeding if the 2005 International Conference on Intelligent Computing (ICIC’05), Lecture Notes in computer Science*, vol. 3644, 2005, pp. 878–887.

[41] H. He, Y. Bai, E. Garcia, and S. Li, “Adasyn: adaptive synthetic sampling approach for imbalanced learning,” in *Proceedings for the 2008 IEEE International Joint Conference on Neural Networks (IJCNN’08)*, 2008, pp. 1322–1328.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[43] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, “API design for machine learning software: experiences from the scikit-learn project,” in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[44] G. Oberreuter, G. L’Huillier, S. A. Rios, and J. D. Velasquez, “Approaches for intrinsic and external plagiarism detection,” *Notebook for PAN at CLEF 2011*.

[45] N. Akiva, “Using clustering to identify outlier chunks of text,” *Notebook for PAN at CLEF 2011*.

[46] S. Rao, P. Gupta, K. Singhal, and P. Majumder, “External and intrinsic plagiarism detection: Vsm and discourse markers based approach,” *Notebook for PAN at CLEF 2011*.

[47] E. Stamatatos, “Intrinsic plagiarism detection using character n-gram profiles,” in *Proceedings of the 3rd International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*, 2009.

[48] P. Filzmoser, R. Maronna, and M. Werner, “Outlier identification in high dimensions,” *Computational Statistics and Data Analysis*, 2008.

[49] J. Binongo and W. Smith, “The application of principal components analysis to stylometry,” *Literary and Linguistic Computing*, 1999.

[50] https://en.wikipedia.org/wiki/Mahalanobis distance.

[51] I. S. Dhillon, Y. Guan, and B. Kulis, “Kernel kmeans:spectral clustering and normalized cuts.” in *Proceedings ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2004, pp. 551–556.

[52] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, “The weka data mining software: An update,” *SIGKDD Explorations*, 2009.

[53] E. Stamatatos, “Author identification using imbalanced and limited training texts,” in *Proceedings of the 4th International Workshop on Text-based Information Retrieval*, 2007, pp. 237–241.

[54] S. M. Alzahrani, N. Salim, and A. Abraham, “Understanding plagiarism linguistic patterns, textual features, and detection methods,” *International Journal for Educational Integrity*, vol. 9, no. 1, pp. 50–71, June 2013.

[55] https://opennlp.apache.org/.

[56] http://www.lextek.com/manuals/onix/stopwords1.html.

[57] S. Leanne and S. Matwin, “Intrinsic plagiarism detection using complexity analysis,” 2009.