ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΠΛΗΡΟΦΟΡΙΑΚΑ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΑ ΣΥΣΤΗΜΑΤΑ»

Αναγνώριση Ρητορικής Μίσους Στα Μέσα Κοινωνικής Δικτύωσης

(Hate Speech Recognition in Social Media)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κατρίνη Δημήτριου

Επιβλέπων : Δρ Ευστάθιος Σταματάτος

Μέλη εξεταστικής επιτροπής: Δρ. Καβαλλιεράτου Εργίνα
Δρ. Παπασαλούρος Ανδρέας

Σάμος, Φεβρουάριος 2020

Η σελίδα αυτή είναι σκόπιμα λευκή.

Στη μνήμη της μητέρας μου Δέσποινας

και στην οικογένεια μου, Αλεξία, Μιχάλη και Λένα

«Η Ιθάκη σ' έδωσε τ' ωραίο ταξίδι.
Χωρίς αυτήν δεν θα 'βγαινες στον δρόμο.
Άλλα δεν έχει να σε δώσει πια.
Κι αν πτωχική την βρεις, η Ιθάκη δεν σε γέλασε.
Έτσι σοφός που έγινες, με τόση πείρα,
ήδη θα το κατάλαβες οι Ιθάκες τι σημαίνουν»

Κ.Π. Καβάφης

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου Δρ. Σταματάτο Ευστάθιο για την ευκαιρία που μου έδωσε και την εμπιστοσύνη που μου έδειξε να ασχοληθώ με ένα τόσο ενδιαφέρον πεδίο της Επιστήμης Υπολογιστών καθώς και για την άριστη συνεργασία μας. Επίσης, οφείλω να ευχαριστήσω τον αδερφό μου Δρ. Κατρίνη Κωνσταντίνο για την ενθάρρυνση και τον πολύτιμο χρόνο που μου παρείχε, μαζί με τις γνώσεις και τις συμβουλές του καθ'όλη τη διάρκεια του μεταπτυχιακού και ιδιαίτερα κατά την εκπόνηση της εργασίας. Τέλος, θέλω να εκφράσω την εγκάρδια ευγνωμοσύνη μου στη σύζυγό μου Αλεξία για την υπομονή που επέδειξε καθώς και για την αμέριστη στήριξη και την παρότρυνση που μου προσέφερε, όπως επίσης και στα παιδιά μου Μιχάλη και Λένα που καθημερινά μέσα από τα μάτια τους μου μεταλαμπαδεύουν δύναμη και όραμα.

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

Λίστα Διαγραμμάτων

Λίστα Πινάκων

Ακρωνύμια

Machine Learning	ML
Support Vector Machine	SVM
Bag of words	BoW
Logistic Regression	LR
Random Forest	RF
Artificial Neural Networks	ANN
Term Frequency Inverse Document Frequency	Tf-idf
Part-of-Speech	PoS
Long short-term Memory	LSTM
Convulutional Neural Networks	CNN
Lexical Syntactic Feature	LSF

Περίληψη

Το φαινόμενο της ρητορικής μίσους έχει πάρει μεγάλες διαστάσεις ειδικά τα τελευταία χρόνια, όντας απότοκο της παγκόσμιας κοινωνικής και οικονομικής κρίσης, που οδήγησε στην ριζοσπαστικοποίηση ατόμων και κοινωνικών ομάδων και ανάδειξη ακραίων/λαϊκίστικων ρητορικών και συμπεριφορών. Ειδικά, στο διαδίκτυο και πιο συγκεκριμένα στα μέσα κοινωνικής δικτύωσης το φαινόμενο αυτό γιγαντώνεται, αφού το πολυπληθές κοινό τους, η άμεση διακίνηση της πληροφορίας, και το χαρακτηριστικό της ανωνυμίας που αυτά διασφαλίζουν, το καθιστούν εξαιρετικά επικίνδυνο για την διαταραχή της κοινωνικής συνοχής, για ψυχολογική επιβάρυνση των ομάδων ή και μεμονωμένων προσώπων στα οποία απευθύνεται, αλλά και σαν παράγοντα υποκίνησης πράξεων βίας.

Προκειμένου να διατηρήσουν την φερεγγυότητα και την αξιοπιστία τους, αρκετά μέσα κοινωνικής δικτύωσης προσπαθούν να απομονώσουν φαινόμενα ρητορικής μίσους θέτοντας όρους αρχικά στην πολιτική χρήσης τους, αλλά και χρησιμοποιώντας εργατικό δυναμικό για την ανίχνευση κακόβουλων μηνυμάτων/σχολίων και τη διαγραφή τους. Όσο όμως η χρήση τους αυξάνει γεωμετρικά, τέτοιες πρακτικές καθίστανται ανεπαρκείς, οπότε είναι επιτακτική η ανάπτυξη μεθόδων αυτόματης ανίχνευσης ρητορικής μίσους. Στην προσπάθεια αυτή πολύτιμος αρωγός είναι ο τομέας της Τεχνητής Νοημοσύνης διαμέσου μεθόδων Μηχανικής Μάθησης. Στον τομέα αυτό η παγκόσμια επιστημονική κοινότητα συμβάλει ώστε να αναπτυχθούν σχετικά μοντέλα και να αναδειχθούν τα βέλτιστα ως προς την ανίχνευση της ρητορικής μίσους.

Η παρούσα διπλωματική εργασία έχει ως στόχο να αναλύσει το φαινόμενο αυτό, να αναδείξει προβλήματα που ανακύπτουν κατά την εφαρμογή τεχνικών μηχανικής μάθησης και να μελετήσει και να συγκρίνει μοντέλα μηχανικής μάθησης για την κατηγοριοποίηση μηνυμάτων ρητορικής μίσους χρησιμοποιώντας σύνολα δεδομένων από το κοινωνικό μέσο Twitter. Σύμφωνα με τα ευρήματα διαπιστώνεται ότι η Λογιστική Παλινδρόμηση αποτελεί τη βέλτιστη μέθοδο ταξινόμησης, σε σύγκριση με τους λοιπούς στατιστικούς ταξινομητές που δοκιμάστηκαν πειραματικά. Επίσης, τα πειραματικά ευρήματα αξιολογούν την επιρροή συγκεκριμένων κλάσεων γνωρισμάτων κειμένου στην επίλυση του προβλήματος, με κύρια ευρήματα την υπεροχή των n-grams λέξεων κατά την προεπεξεργασία και την μικρή συνεισφορά της ανάλυσης συναισθήματος ως γνωρίσματος.

Εστιάζοντας στην κατηγοριοποίηση μηνυμάτων της κλάσης μίσους μόνον, η παρούσα διπλωματική αναδεικνύει - τόσο σε βιβλιογραφικό επίπεδο όσο και πειραματικά - ότι περαιτέρω βήματα απαιτούνται για την πλήρη και αξιόπιστη αυτοματοποίηση της εν λόγω διεργασίας μηχανικής μάθησης (machine learning task). Τα διαθέσιμα ανοικτά σύνολα δεδομένων που τροφοδοτούν τα μοντέλα εμφανίζουν μεγάλη ανομοιογένεια ως προς τις κλάσεις. Τούτο, σε συνδυασμό με το ότι η υβριστική – προσβλητική γλώσσα συχνά δεν είναι εύκολο να διαχωριστεί από τη ρητορική μίσους, καθιστά τους παραδοσιακούς ταξινομητές που βασίζονται σε στατιστικά μοντέλα να έχουν σχετικά χαμηλή αποτελεσματικότητα ως προς την σωστή ταξινόμηση των μηνυμάτων μίσους, αν και δίνουν αρκετά καλά αποτελέσματα γενικά κατά την ταξινόμηση

(μετρικές που περιλαμβάνουν όλες τις κλάσεις). Επίσης απαιτείται πιο ενδελεχής μελέτη κατά την εξαγωγή των γνωρισμάτων με βάρος στην ανάδειξη σημασιολογικών χαρακτηριστικών.

Λέξεις Κλειδιά: *Ρητορική Μίσους, Κατηγοριοποίηση Κειμένου, Μηχανική Μάθηση, Ταξινόμηση, Τεχνητή Νοημοσύνη*

Abstract

The phenomenon of hate speech has become quite widespread in recent years, in the wake of the global social and economic crisis that has led to radicalization of individuals and social groups and the emergence of extreme/populist rhetoric and behavior. Especially on the Internet and more specifically on social media, this phenomenon is exacerbated, as their large audience, the immediate flow of information, and the anonymity they provide make it extremely dangerous for the disturbance of social cohesion, for psychological burden on the groups and/or individuals it is targeting, but also as a potential trigger for acts of violence.

In order to maintain their credibility and sustainability, many social media enterprises try to isolate hate speech by first setting terms of use in their usage policy, but also by employing manpower to detect malicious messages/comments and delete them. However, as their use increases at a great pace, such practices become inadequate, so it is imperative to develop automatic hate speech detection methods and tools for a sustainable control of such incidents. A valuable aid in this endeavor is the field of Artificial Intelligence through Machine Learning methods. In this area, the global scientific community is contributing in developing relevant models and ultimately in developing relevant tools to detect hate speech.

The present thesis aims to analyze this phenomenon, to identify problems encountered in the application of machine learning techniques, and to study and compare machine learning models for categorizing hate speech using data corpora stemming from Twitter. The findings indicate that Logistic Regression is the optimal classification method for the task, compared to other statistical classifiers that have been experimentally evaluated. Experimental work conducted herein also evaluate the efficacy of specific text classification features for the task, with the main findings suggesting the predominance of the use of word n-grams in text pre-processing and also a small contribution of sentiment analysis as a feature.

Focusing on the categorization of hate class messages only, this thesis demonstrates - both through bibliographical surveys and experimental results - that further steps are needed to fully and reliably automate this machine learning task. The available open data corpora available for use are very heterogeneous in terms of data to class distribution. This, coupled with the fact that abusive language is often not easy to separate from hate speech, makes traditional statistical model-based classifiers relatively weak in their ability to correctly classify hate messages, although they yield enough good overall results (i.e. on metrics that include all classes). A more in-depth study is also needed when engineering text classification features for the purpose of hate speech classification, with a focus on features that carry semantical meaning.

Keywords: *Hate Speech, Text Categorization, Machine Learning, Classification, Artificial Intelligence*

1 *Εισαγωγή*

1.1 Αυτόματη ανίχνευση ρητορικής μίσους με μηχανική μάθηση

Η παρούσα διπλωματική εργασία μελετάει το φαινομένου της ύπαρξης και διακίνησης μηνυμάτων ρητορικής μίσους στα μέσα κοινωνικής δικτύωσης και πως αυτά μπορούν να ανιχνευθούν και να κατηγοριοποιηθούν με τεχνικές μηχανικής μάθησης με την εφαρμογή μεθόδων επεξεργασίας φυσικής γλώσσας. Οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται κατά κόρον σε προβλήματα ταξινόμησης κειμένου, ως εκ τούτου αποτελούν ένα εργαλείο για την αυτόματη ανίχνευση ρητορικής μίσους στο διαδίκτυο. Ωστόσο δεν έχει υποδειχθεί από την επιστημονική κοινότητα μία βέλτιστη μέθοδος για την κατηγοριοποίηση μηνυμάτων μίσους. Τα σύνολα δεδομένων που είναι διαθέσιμα προς μελέτη παρουσιάζουν μεγάλη ανομοιογένεια όσον αφορά τις κλάσεις. Επίσης, είναι εξαιρετικά δύσκολο να διαχωριστεί η ρητορική μίσους από υβριστική/προσβλητική γλώσσα, με αποτέλεσμα οι παραδοσιακοί ταξινομητές να εμφανίζουν μεν εξαιρετικά καλά αποτελέσματα στις επιδόσεις τους γενικά, να αποτυγχάνουν δε να ταξινομήσουν με μεγάλη ακρίβεια σωστά τα μηνύματα που περιέχουν ρητορική μίσους. Η επιλογή των γνωρισμάτων κατά την εκπαίδευση των ταξινομητών αποτελεί ένα ακόμη πεδίο έρευνας προς την ανάπτυξη βελτιωμένων μοντέλων ανίχνευσης και ταξινόμησης ρητορικής μίσους.

1.2 Αντικείμενο διπλωματικής

Στη συγκεκριμένη εργασία γίνεται μία προσπάθεια σύγκρισης κάποιων ευρέως χρησιμοποιημένων από τη διεθνή βιβλιογραφία ταξινομητών. Αφού αρχικά γίνεται μία επισκόπηση της εφαρμογής της διαδικασίας μηχανικής μάθησης στα προβλήματα επεξεργασίας κειμένων και μία όσο το δυνατόν ενδελεχέστερη επισκόπηση της σχετικής διεθνούς βιβλιογραφίας όσον αφορά την ανίχνευση ρητορικής μίσους με την εφαρμογή διαφόρων μοντέλων σε σύνολα δεδομένων που έχουν αντληθεί από το Twitter, επιχειρείται μέσα από πειράματα να υποδειχθεί το βέλτιστο μοντέλο για δύο διαφορετικά σύνολα δεδομένων και να συγκριθεί με αντίστοιχα μοντέλα που προκύπτουν από τη μελέτη της σχετικής βιβλιογραφίας. Επίσης, πραγματοποιούνται πειράματα για να διαπιστωθεί η επιρροή κάποιων γνωρισμάτων στην απόδοση του μοντέλου. Η εργασία είναι αναφέρεται στη μελέτη στατιστικών ταξινομητών που προκρίνονται από τη διεθνή βιβλιογραφία για την ανίχνευση ρητορικής μίσους στο Twitter. Αποτελεσματικότερες μέθοδοι κάνουν χρήση βαθέων νευρωνικών δικτύων, όπου επιτυγχάνεται πληρέστερη αυτοματοποίηση της κατηγοριοποίησης κειμένου, ο χρονικός όμως προϋπολογισμός της παρούσας εργασίας, καθώς και η έλλειψη ικανών πόρων δεν επέτρεψε την μελέτη τέτοιων τεχνικών , όπως και συμφωνήθηκε και με τον επιβλέποντα καθηγητή.

1.3 Ρητορική Μίσους στην Κοινωνία

Σύμφωνα με το Συμβούλιο Υπουργών της Ευρωπαϊκής Ένωσης (Σύσταση 97(20)[1]), η Ρητορική Μίσους ορίζεται σαν «κάθε μορφή έκφρασης που διαδίδει, υποκινεί, προωθεί ή δικαιολογεί το φυλετικό μίσος, την ξενοφοβία, τον αντισημιτισμό ή άλλες μορφές μίσους που στηρίζονται στη μισαλλοδοξία, όπως τη μισαλλοδοξία που εκφράζεται μέσα από επιθετικό εθνικισμό και εθνοκεντρισμό, διάκριση και εχθρικότητα κατά μειονοτήτων, μεταναστών και ανθρώπων μεταναστευτικής καταγωγής».

Τα άτομα που χρησιμοποιούν Ρητορική Μίσους (Hate Speech) υιοθετούν αρνητικά στερεότυπα πρωτίστως κατά κάποιων ομάδων ή ατόμων που μοιράζονται κοινά προσωπικά ή εθνοτικά χαρακτηριστικά, ιδέες, θρησκευτικές αντιλήψεις παραβιάζοντας έτσι αρχές όπως είναι η ισότητα και ο σεβασμός. Ειδικά στις μέρες μας που χαρακτηρίζονται από ποικίλες κοινωνικοπολιτικές εντάσεις, έντονο μεταναστευτικό κύμα, όπως επίσης και προσπάθεια διαφόρων μειονοτικών ομάδων να αποκτήσουν ισότιμα και αναγνωρισμένα δικαιώματα (π.χ. ΛΟΑΤ κοινότητα), η Ρητορική Μίσους θεωρείται ένα πολύ επικίνδυνο φαινόμενο για την κοινωνία που εκτός των διάφορων δυσλειτουργικών επιπτώσεων για την ομαλή κοινωνική ζωή που αυτή περιλαμβάνει, όπως τις διακρίσεις, την περιθωριοποίηση, την αποξένωση και τη στοχοποίηση, συχνά υποκινεί ή και στοχεύει άμεσα προς την άσκηση σωματικής βίας. Η Ευρωπαϊκή Επιτροπή κατά του Ρατσισμού και της Μισαλλοδοξίας (ECRI) υπογραμμίζει ότι η ρητορική του μίσους πρέπει να ρυθμίζεται μέσα από νομοθεσίες καθώς και άλλων μηχανισμών όπως είναι η αυτορρύθμιση, η πρόληψη και ο αντίλογος. Στην ελληνική επικράτεια και συγκεκριμένα με τον νόμο 2485/2014[2] που τροποποιεί τον νόμο 927/1979 ορίζεται και ποινικοποιείται η Ρητορική Μίσους. Ο συγκεκριμένος νόμος τιμωρεί πλέον την «υποκίνηση, πρόκληση, διέγερση ή προτροπή» σε «πράξεις ή ενέργειες» που μπορούν να προκαλέσουν «διακρίσεις, μίσος ή βία» κατά προσώπου ή ομάδας προσώπων. Εκτός από τις ομάδες που μπορούν να προσδιοριστούν με βάση τη φυλή και την εθνική καταγωγή, ο νόμος επεκτείνει την προστασία του και στις ομάδες που προσδιορίζονται με βάση το χρώμα, τη θρησκεία, τον σεξουαλικό προσανατολισμό, την ταυτότητα φύλου ή την αναπηρία, θέτοντας όμως ως προϋπόθεση της τιμωρίας το να έγινε η ανωτέρω πράξη κατά τρόπο που εκθέτει σε κίνδυνο τη δημόσια τάξη ή ενέχει απειλή για τη ζωή, την ελευθερία ή τη σωματική ακεραιότητα προσώπων. Κοντολογίς, ο νόμος δεν αρκείται στο ότι κάποιος είπε κάτι ρατσιστικό ή μισαλλόδοξο, αλλά απαιτεί για την καταδίκη να μπορούσε το λεχθέν να διακινδυνεύσει τη δημόσια τάξη, πράγμα που μένει να αποδειχθεί. Ίσως ο στόχος του νομοθέτη είναι να διαφυλάξει την έννοια της «ελεύθερης έκφρασης» που κάθε δημοκρατικό καθεστώς πρέπει να προφυλάσσει και να προωθεί. Το δικαίωμα στην ελευθερία της έκφρασης αναγνωρίζεται ως ανθρώπινο δικαίωμα σύμφωνα με το άρθρο 19 της Οικουμενικής Διακήρυξης των Δικαιωμάτων του Ανθρώπου και αναγνωρίζεται από το Διεθνές Δίκαιο για τα Ανθρώπινα Δικαιώματα στο Διεθνές Σύμφωνο για τα Ατομικά και Πολιτικά Δικαιώματα.

[1] https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680505d5b

[2] https://.ministryofjustice.gr%2Fsite%2FLinkClick.aspx%3Ffileticket%3DIk2xQr3jIkg%253D%26tabid%3D132&usg=AOvVaw2-qoudl2-3GM-GxbYB8w3T

Η στάθμιση μεταξύ της ρητορικής μίσους και της ελευθερίας της έκφρασης είναι ένα πεδίο διαρκούς και δύσκολου προβληματισμού. Αυτό που διαφοροποιεί την πρώτη από τη δεύτερη είναι, ιδίως, η υποκίνηση σε πράξεις βίας ή η άμεση στοχοποίηση ατόμων ή κοινωνικών ομάδων λόγω ενός ιδιαίτερου κοινωνικού, σωματικού ή πνευματικού χαρακτηριστικού τους.

1.4 Ρητορική Μίσους στο Διαδίκτυο

Αναντίρρητα ζούμε στην εποχή όπου το Διαδίκτυο αποτελεί τον κύριο «χώρο» έκφρασης σκέψεων, ανταλλαγής ιδεών, αφού λόγω του παγκόσμιου χαρακτήρα του, την εξαιρετικά μεγάλη, έως και άμεση, ταχύτητα μετάδοσης πληροφορίας καθώς και τη δυνατότητα για αμφίδρομη επικοινωνία καθιστούν τους χρήστες του άμεσους παρατηρητές και σχολιαστές των εξελίξεων. Το Διαδίκτυο θεωρείται ένα άκρως δημοκρατικό μέσο μαζικής επικοινωνίας, για αυτό και βρίσκει απήχηση σε όλες τις κοινωνικές ομάδες και ηλικίες, με κύρια απήχηση όμως στους νέους. Χωρίς προσπάθεια δαιμονοποίησής του, η ανωνυμία και η μαζικότητα που αυτό προσφέρει το καθιστά ένα πρόσφορο εργαλείο και μέσο διακίνησης Ρητορικής Μίσους, ιδιαίτερα στα κοινωνικά δίκτυα (social media) που αποτελούν το κύριο πεδίο ανταλλαγής ιδεών ή καταγραφής/έκφρασης απόψεων μέσω του Διαδικτύου. Επίσης η δυσκολία της ρύθμισης ελέγχου του Διαδικτύου και το γεγονός ότι οι περισσότεροι άνθρωποι τολμούν να πουν περισσότερα στο διαδίκτυο παρά στην πραγματική ζωή εκτός Διαδικτύου έχει ως αποτέλεσμα να γίνεται πιο περίπλοκος και δυσχερής ο έλεγχος του διαδικτυακού μίσους.

Για να ανταποκριθούν στο εντεινόμενο πρόβλημα της ρητορικής μίσους στο διαδίκτυο, η Ευρωπαϊκή Επιτροπή και τέσσερις μεγάλες επιχειρήσεις (Facebook, Microsoft, Twitter και YouTube) παρουσίασαν τον Μάιο του 2016 έναν «κώδικα δεοντολογίας για την καταπολέμηση της παράνομης ρητορικής μίσους στο διαδίκτυο»[3]. Κατά τη διάρκεια του 2018, οι ακόλουθες τέσσερις νέες εταιρείες αποφάσισαν να προσχωρήσουν στον κώδικα δεοντολογίας: Google+, Instagram, Snapchat, Dailymotion. Οι εταιρείες αφαιρούν το παράνομο περιεχόμενο όλο και ταχύτερα, αλλά αυτό δεν οδηγεί σε υπερβολική αφαίρεση περιεχομένου: το ποσοστό αφαίρεσης δείχνει ότι ο έλεγχος που πραγματοποιούν οι εταιρείες εξακολουθεί να σέβεται την ελευθερία της έκφρασης. Επίσης οι εταιρείες ζητούν από άλλους χρήστες να επισημαίνουν ή και να καταγγέλλουν περιεχόμενο που σχετίζεται με Ρητορική Μίσους ή προβαίνουν και σε μπλοκαρίσματα λογαριασμών

Οι περισσότερες υπάρχουσες μέθοδοι εντοπισμού Ρητορικής Μίσους επαφίενται στον ανθρώπινο παράγοντα και στην γενικότερη αντίληψη για το τι στοιχειοθετεί ρητορική μίσους, επομένως η ανίχνευσή της δεν είναι πάντα αντικειμενικά σωστή. Ειδικά σε περιπτώσεις σαρκασμού ή χρήσης απλά καθομιλουμένης υβριστικής γλώσσας (χωρίς να υπάρχει η πρόθεση να θιχτεί κάποιος) είναι πολύ εύκολο κάποιος να οδηγηθεί σε λανθασμένα συμπεράσματα.

[3] https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en

1.5 Ρητορική Μίσους στα Μέσα Κοινωνικής Δικτύωσης (Twitter)

Μεγάλο ενδιαφέρον στη διάδοση ρητορικής μίσους συγκεντρώνει το Twitter, το οποίο είναι ένα *microblogging* κοινωνικό δίκτυο. Το microblogging διαφέρει από το παραδοσιακό blogging, ως προς την έκταση των κειμένων, τα οποία αποτελούνται συνήθως από μικρές προτάσεις. Ο λόγος για την τόσο μεγάλη απήχηση αυτού του μέσου είναι η αμεσότητα και η απλότητα των μηνυμάτων. Το κύριο χαρακτηριστικό του είναι ότι η έκταση των μηνυμάτων περιορίζεται στους 280 χαρακτήρες (αρχικά το όριο ήταν 140, το οποίο έχει διατηρηθεί για κάποιες γλώσσες).

Αυτό ωθεί τους χρήστες να γράφουν πιο περιεκτικά μηνύματα. Επίσης είναι δυνατό να γίνουν συζητήσεις ανάμεσα σε πολλούς χρήστες, οι οποίες είναι ορατές σε όσους τους ακολουθούν. Επιπροσθέτως, ο χρήστης που γράφει ένα μήνυμα έχει τη δυνατότητα να το απευθύνει σε έναν άλλο χρήστη ή ομάδα χρηστών (mention). Ένα άλλο πολύ σημαντικό χαρακτηριστικό αφορά την εύκολη αναμετάδοση των μηνυμάτων (retweet). Εάν ένας χρήστης διαβάσει ένα μήνυμα το οποίο θέλει να διαδώσει, μπορεί να το κάνει πολύ εύκολα αναμεταδίδοντας το σε όσους τον ακολουθούν. Ομοίως, μερικοί από τους ακολούθους του αρχικού χρήστη, ενδεχομένως να αποφασίσουν να το αναμεταδώσουν στους δικούς τους ακολούθους. Όπως είναι προφανές, αυτό μπορεί να οδηγήσει σε μία "αλυσιδωτή αντίδραση", με αποτέλεσμα σχόλια που χαρακτηρίζονται από ρητορική μίσους να διαδίδονται πάρα πολύ γρήγορα σε μεγάλη μάζα χρηστών.

Συνοψίζοντας, η ρητορική μίσους στο διαδίκτυο είναι μία πραγματικότητα. Η ευχέρεια διάδοσης της πληροφορίας, το εύρος διάδοσης, σε πολλές περιπτώσεις ανωνυμία αλλά και εκμετάλλευση του για ιδιοτελείς σκοπούς, κάνουν αυτή την πραγματικότητα ακόμα πιο πολύπλοκη. Βάσει νομοθεσίας, τα κριτήρια χαρακτηρισμού έκφρασης ως ρητορικής μίσους είναι υποκειμενικά, απαιτούν ιδιαίτερη προσοχή και ακρίβεια για να μην καταπατούν τις αρχές ελευθερίας της έκφρασης αλλά και για να διατηρήσουν τη φύση του διαδικτύου ως ενός ελεύθερου και δημοκρατικού μέσου.

Δεδομένης της κλίμακας πληροφορίας[4], είναι αυτονόητο ότι οποιαδήποτε μη-αυτοματοποιημένη λύση ανίχνευσης ρητορικής μίσους καθίσταται ιδιαιτέρως ακριβή. Η πρόοδος που έχει συντελεστεί τελευταία στον κλάδο της τεχνητής νοημοσύνης και μηχανικής μάθησης, οι ολοένα αναπτυσσόμενες τεχνικές και εφαρμογές επεξεργασίας φυσικής γλώσσας/ανάλυσης κειμένου, καθιστούν τη χρήση τέτοιων τεχνικών ως μίας (ίσως και της μοναδικής) πολλά υποσχόμενης λύσης αυτοματοποιημένης ανίχνευσης ρητορικής μίσους στο διαδίκτυο. Η εστίαση της μελέτης σε κείμενα κοινωνικής δικτύωσης και συγκεκριμένα στο μέσο Tweeter εξυπηρετεί τους εξής σκοπούς: α) αξιολόγηση της αποδοτικότητας της λύσης σε σύνολα δεδομένων που παρουσιάζουν το μέγιστο βαθμό δυσκολίας (λόγω των χαρακτηριστικών που περιγράφηκαν παραπάνω) και β) λόγω της διάδοσης χρήσης και κλίμακας παραγωγής πληροφορίας/επιρροής του μέσου.

Στο υπόλοιπο του κεφαλαίου, γίνεται μία συνοπτική εισαγωγή στους όρους και τις τεχνικές μηχανικής μάθησης, καθώς και μία συνοπτική περιγραφή της Εξόρυξης από Κείμενο (Text Mining)

[4]Το 2019, ο αριθμός ημερησίων αναρτήσεων (tweets) ανήλθε στις 500 εκατομμύρια αναρτήσεις (πηγή: https://www.omnicoreagency.com/Twitter-statistics/)

1.6 Μηχανική Μάθηση

Η Μηχανική Μάθηση (Machine Learning) αποτελεί υποπεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη (Artificial Intelligence). Το 1959, ο A. Samuel όρισε τη μηχανική μάθηση ως "*Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί*". Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων και τεχνικών που μπορούν να μαθαίνουν από τα δεδομένα, αναζητώντας πρότυπα (patterns) και σχέσεις με απώτερο σκοπό να μπορεί να μοντελοποιήσει τα δεδομένα και να κάνει προβλέψεις πάνω σε αυτά .[5] Ο T. Mitchell δίνει έναν πιο τεχνικό ορισμό για τη Μηχανική Μάθηση και συγκεκριμένα: «*Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από εμπειρία E ως προς μία κλάση εργασιών T και ένα μέτρο επίδοσης P, αν η επίδοσή του σε εργασίες της κλάσης T, όπως αποτιμάται από το μέτρο P, βελτιώνεται με την εμπειρία E*»[6].

Κατά τη γενικότερη διαδικασία της Μηχανικής μάθησης, ένα σύστημα τροφοδοτείται με εμπειρικά δεδομένα, που καλούνται *παραδείγματα εκπαίδευσης* (*training data*) και εν συνεχεία με την εφαρμογή ενός αλγορίθμου μηχανικής μάθησης, το σύστημα επιχειρείται να εκπαιδευτεί πάνω στα δεδομένα αυτά, ώστε να βρεθούν πιθανές γενικεύσεις ή εξειδικεύσεις. Κατά τη φάση αυτή, που ονομάζεται *εκπαίδευση* (*training*), το σύστημα *μαθαίνει από την εμπειρία* και μοντελοποιεί τα δεδομένα. Στη συνέχεια, τροφοδοτείται με *παραδείγματα δοκιμής* (*testing data*), τα οποία δεν έχει συναντήσει κατά την εκπαίδευσή του και με βάση τις προβλέψεις στα δεδομένα αυτά, αξιολογείται η επίδοσή του.

1.7 Κατηγορίες Μηχανικής Μάθησης

Οι εργασίες της μηχανικής μάθησης ταξινομούνται στις εξής κατηγορίες, ανάλογα με την «ανατροφοδότηση» που είναι διαθέσιμη σε ένα σύστημα εκμάθησης.

Επιβλεπόμενη Μάθηση (Supervised Learning) είναι η διαδικασία όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους σε γνωστές επιθυμητές εξόδους (ζεύγη εισόδων/εξόδων), με στόχο τη μάθηση ενός γενικότερου κανόνα που θα χρησιμοποιηθεί για την αντιστοίχιση νέων εισόδων (που δεν γνωρίζουμε τις έξοδούς τους) με τα αποτελέσματα. Η μέθοδος αυτή χρησιμοποιείται σε προβλήματα:

- Ταξινόμησης (Classification)
- Πρόγνωσης (Prediction)
- Διερμηνείας (Interpretation)

Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning), όπου ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων (υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους (unlabeled data). Δεν υπάρχει δυνατότητα αξιολόγησης της επίδοσης του αλγορίθμου, αφού δεν είναι γνωστή η επιθυμητή απόκριση. Συνήθως σκοπός αυτού του είδους μηχανικής

[5] https://el.wikipedia.org/wiki/Μηχανική_Μάθηση

[6]Mitchell, T. (1997). *Machine Learning*, McGraw Hill, *Machine Learning*, McGraw Hill, p.2

μάθησης είναι η ανακάλυψη κρυφών δομών, μοτίβων ή συσχετίσεων των δεδομένων. Χρησιμοποιείται σε προβλήματα:

- Ανάλυσης Συσχετισμών (Association Analysis)
- Ομαδοποίησης (Clustering)

Ενισχυτική Μάθηση (Reinforcement Learning), όπου ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον προκειμένου να επιτευχθεί ένας στόχος (π.χ. όπως η οδήγηση ενός αυτόνομου οχήματος), χωρίς όμως κάποιος εξωτερικός να «πληροφορεί» άμεσα το σύστημα αν έχει φτάσει κοντά στο στόχο του. Το σύστημα λαμβάνει ανάδραση μέσα από μια σειρά από επιβραβεύσεις ή ποινές, καθώς εξερευνά το χώρο του προβλήματος. Χρησιμοποιείται κυρίως στον έλεγχο κίνησης ρομπότ, σε ηλεκτρονικά παιχνίδια με αντίπαλο ή και σε διαδικασίες βελτιστοποίησης εργασιών σε εργοστασιακούς χώρους.

Μια εναλλακτική κατηγοριοποίηση των εργασιών της μηχανικής μάθησης προκύπτει βάσει του τύπου με τον τύπο της εξόδου του εκάστοτε συστήματος – και κατ' επέκταση τη φύση του προβλήματος που αποσκοπεί να λυθεί μέσω μηχανικής μάθησης – όπου συναντάμε τις εξής κατηγορίες:

1. Ταξινόμηση (classification): η είσοδος διαχωρίζεται σε δύο (binary) ή παραπάνω κλάσεις (multi-class), και ο αλγόριθμος – στη συγκεκριμένη περίπτωση, ταξινομητής- προσπαθεί να κατασκευάσει ένα μοντέλο που να ταξινομεί τα άγνωστα δεδομένα σε δύο ή περισσότερες κλάσεις, με το ελάχιστο δυνατό σφάλμα πρόβλεψης.

2. Παλινδρόμηση (regression): η έξοδος είναι συνεχόμενη μεταβλητή αντί για διακριτή.

3. Ομαδοποίηση (clustering): ένα σύνολο δεδομένων χωρίζεται σε ομάδες παρόμοιων χαρακτηριστικών, οι οποίες δεν είναι γνωστές εκ των προτέρων.

1.8 Text Mining

Έχοντας αποκτήσει μία πρώτη εικόνα για την Μηχανική Μάθηση, ας αναφερθούμε σε ένα επιστημονικό πεδίο που οι συγκεκριμένες τεχνικές σε συνδυασμό με άλλες έχουν αποκτήσει αυξημένη εφαρμογή και δεν είναι άλλο από την Εξόρυξη Κειμένου (Text Mining).

Κατά την εξόρυξη κειμένου καλούμαστε να εξάγουμε με αυτόματο τρόπο νέα, έγκυρη και χρήσιμη πληροφορία ή γνώση από διαφορετικά σύνολα δεδομένων. Η πληροφορία αυτή μπορεί να είναι συσχετίσεις, patters ή τάσεις που δεν είναι αρχικά αναγνωρίσιμες στην αρχική μορφή των κειμένων. Κατά την εξόρυξη κειμένου, μη δομημένα ή ημί-δομημένα κείμενα, μέσα από μεθόδους επεξεργασίας φυσικής γλώσσας μετασχηματίζονται έτσι ώστε να παράξουν δομημένα δεδομένα που μπορούν να χρησιμοποιηθούν για απευθείας ανάλυση τους ή να αποτελέσουν εισόδους σε αλγορίθμους μηχανικής μάθησης για την εξαγωγή συμπερασμάτων. Στην παρούσα εργασία αδόμητα κείμενα υπό την μορφή μηνυμάτων από το twitter γραμμένα σε φυσική γλώσσα, καλούνται να μετασχηματιστούν σε τέτοια μορφή, έτσι ώστε να «διαβαστούν» και να «κατανοηθούν» από αλγορίθμους μηχανικής μάθησης, με στόχο να ταξινομηθούν με γνώμονα αν περιέχουν ή όχι ρητορική μίσους. Η **Κατηγοριοποίηση Κειμένων** (Text Classification) είναι συνακόλουθα ένα πεδίο εφαρμογής της εξόρυξης κειμένου και στοιχειοθετεί την διαδικασία κατάταξης εγγράφων ή κειμένων σε προκαθορισμένες κατηγορίες. Σκοπός της τεχνικής αυτής είναι ο προσδιορισμός των

κύριων θεμάτων μέσα από μια συλλογή εγγράφων ή κειμένων. Άλλα πεδία εφαρμογής του Text Mining είναι η **Ομαδοποίηση Κειμένων** (Text Clustering), όπου μία συλλογή εγγράφων διαιρείται σε ομάδες, αναλόγως με ένα ή περισσότερα κοινά χαρακτηριστικά των εγγράφων αυτών, η **Περιληπτική Παρουσίαση Πληροφορίας** (Text Summarization), όπου μειώνεται το ποσό κειμένου σε ένα έγγραφο, διατηρώντας όμως τις βασικές έννοιες του περιεχομένου του, ο **Γλωσσικός Προσδιορισμός** (Language Identification), όπου μπορεί να ανιχνευθεί η γλώσσα/ες στην οποία ένα κείμενο γράφεται , όπως επίσης και η απόδοση ενός κειμένου σε συγγραφέα με βάση υφολογικά κυρίως χαρακτηριστικά και τέλος η **Ανάκτηση Κειμένου** (Text Retrieval), όπου χρησιμοποιείται για την έρευνα εσωτερικών συλλογών από έγγραφα ή και για αναζήτηση σχετικών εγγράφων στο Ίντερνετ. Επίσης η εξόρυξη κειμένου μπορεί να χρησιμοποιηθεί για την εξαγωγή χαρακτηριστικών γνωρισμάτων από τα κείμενα - εφαρμογή που λαμβάνει χώρα στην παρούσα εργασία- για την Απεικόνιση κειμένων, για την κατασκευή Οντολογιών, όπως επίσης και για την Διασύνδεση Εννοιών.

Ένα πραγματικό μοντέλο Text Mining θα πρέπει να πληροί τις παρακάτω προϋποθέσεις [7] :

- να είναι λειτουργικό σε τεράστιες συλλογές κειμένων, γραμμένες σε φυσική γλώσσα.
- να βασίζεται στη χρήση αλγορίθμων.
- να εξάγει μονάδες πληροφοριών όπως για παράδειγμα υποδείγματα (patterns).
- να ανακαλύπτει νέα γνώση.

1.9 Δομή της διπλωματικής

Η παρούσα εργασία είναι δομημένη ως εξής:

Το Κεφάλαιο 2 ασχολείται με την Ταξινόμηση Κειμένων. Αναλύεται συγκεκριμένα το πρόβλημα της κατηγοριοποίησης κειμένου με αναφορά στις μεθόδους προεπεξεργασίας, στην εξαγωγή γνωρισμάτων (feature engineering), καθώς και στα μοντέλα που χρησιμοποιούνται κατά την αναπαράσταση των κειμένων. Τέλος, γίνεται μία συνοπτική περιγραφή των συνήθων αλγορίθμων που χρησιμοποιούνται στα σχετικά προβλήματα καθώς και στις μετρικές αξιολόγησής τους

Στο Κεφάλαιο 3 γίνεται μια ενδελεχής επισκόπηση της βιβλιογραφίας σχετικά με την αυτόματη ανίχνευση ρητορικής μίσους σε σύνολα δεδομένων του Twitter, αποτυπώνονται τα αποτελέσματα των ερευνητών και επιχειρείται μία συγκριτική συζήτηση ευρημάτων της βιβλιογραφίας.

Στο Κεφάλαιο 4 περιγράφεται αναλυτικά η πειραματική μέθοδος που ακολουθήσαμε κατά τη διεξαγωγή των πειραμάτων μας.

Στο Κεφάλαιο 5 παρουσιάζονται τα πειραματικά αποτελέσματα της διπλωματικής, γίνεται συζήτηση των αποτελεσμάτων και σύγκρισή με αντίστοιχα αποτελέσματα που προκύπτουν από τη διεθνή βιβλιογραφία.

[7] Sharp M., "Text Mining". Seminar in Information Studies, Prof, Tefko Saracevic. 11 December 2001

Στο Κεφάλαιο 6 αναπτύσσουμε τα συμπεράσματα που προκύπτουν από τη διεξαγωγή των πειραμάτων, τη μελέτη της σχετικής βιβλιογραφίας και γενικά της ενασχόλησης με το πρόβλημα στα πλαίσια της παρούσας εργασίας.

2 *Ταξινόμηση Κειμένων*

2.1 Κατηγοριοποίηση Κειμένου

Η κατηγοριοποίηση κειμένου (text classification ή text categorization) είναι ο υπο-κλάδος της μηχανικής μάθησης που αποσκοπεί στην αυτόματη ανάθεση ενός συνόλου κειμένων, γραμμένων σε φυσική γλώσσα, σε καθορισμένες κατηγορίες ανάλογα με το περιεχόμενό τους. Πιο συγκεκριμένα, η κατηγοριοποίηση κειμένου είναι η διεργασία κατά την οποία αποδίδεται μία λογική (Boolean) τιμή στα ζεύγη $(d_j,c_j) \epsilon D \times C$, όπου με D υποδηλώνεται ένα σύνολο εγγράφων και με C ένα σύνολο προκαθορισμένων κατηγοριών (κλάσεων), έτσι ώστε μία (ή και περισσότερε) κατηγορίες να αποδίδεται (ή μη) σε κάθε έγγραφο. Ένα έγγραφο μπορεί να κατηγοριοποιηθεί σε περισσότερες κατηγορίες, οπότε μιλάμε για ταξινόμηση πολλαπλών ετικετών (multi-label classification) ή σε μία μόνο κατηγορία, ταξινόμηση μονής ετικέτας (single-label classification). Μία ειδική περίπτωση της τελευταίας είναι η δυαδική ταξινόμηση (binary classification), όπου κάθε έγγραφο πρέπει να ταξινομηθεί είτε σε μία κλάση ή στην συμπληρωματική της. Στο σημείο αυτό πρέπει να επισημάνουμε ότι η δυαδική ταξινόμηση αποτελεί μία εξειδίκευση της ταξινόμησης πολλαπλών ετικετών, ως εκ τούτου ένας αλγόριθμος ταξινόμησης που εφαρμόζεται για δυαδική ταξινόμηση μπορεί να εφαρμοστεί και σε ταξινόμηση πολλαπλών ετικετών με κατάλληλο μετασχηματισμό του προβλήματος (διαιρώντας το δηλαδή σε πολλά προβλήματα δυαδικής ταξινόμησης). Το αντίστροφο όμως δεν ισχύει: ένας αλγόριθμος ταξινόμησης πολλαπλών ετικετών δεν μπορεί να χρησιμοποιηθεί σε ταξινομήσεις μονής ετικέτας ή δυαδικές ταξινομήσεις.

Τα τελευταία χρόνια η ραγδαία πρόοδος που σημειώνεται στον τομέα της Μηχανικής Μάθησης, συνετέλεσε στην ανάπτυξη έξυπνων ταξινομητών, οι οποίοι κατηγοριοποιούν ηλεκτρονικά κείμενα αυτόματα, μέσω της εκμάθησης των χαρακτηριστικών των κατηγοριών αυτών, χρησιμοποιώντας ένα ήδη ταξινομημένο σύνολο δεδομένων. Η κατηγοριοποίηση κειμένου έχει πολλές εφαρμογές όπως στην Οργάνωση Εγγράφων, στην αναγνώριση και φιλτράρισμα spam emails, στην Αυτόματη Ευρετηρίαση Συστημάτων ανάκτησης Πληροφορίας, στην Αναγνώριση Γλώσσας, στη βελτίωση των αποτελεσμάτων των μηχανών αναζήτησης, στην εξαγωγή άποψης ή συναισθήματος από κείμενα και περαιτέρω ανάλυσή της (Sentiment Analysis), όπως επίσης και σε προβλήματα Επεξεργασίας Φυσικής Γλώσσας (NLP).

Τα βήματα που ακολουθούνται συνήθως στη Κατηγοριοποίηση Κειμένου είναι τα ακόλουθα

- Συλλογή Κειμένων (σχηματισμός σε Σύνολο Δεδομένων)
- Προεπεξεργασία κειμένων
- Αναπαράσταση κειμένων
- Εξαγωγή επιθυμητών χαρακτηριστικών από το κείμενο
- Επιλογή μοντέλου εκπαίδευσης

- Εκπαίδευση των ταξινομητών
- Αξιολόγηση των ταξινομητών

2.1.1 Προεπεξεργασία Κειμένου

Κύρια διαδικασία της κατηγοριοποίησης κειμένων είναι η αναπαράσταση των κειμένων, έτσι ώστε η φυσική γλώσσα, η οποία τα στοιχειοθετεί, να μπορέσει να γίνει κωδικοποιηθεί από τους αλγορίθμους ταξινόμησης. Πριν όμως λάβει χώρα η συγκεκριμένη διεργασία, το σύνολο δεδομένων πρέπει να υποστεί μια προεπεξεργασία, έτσι ώστε αφενός να αποκτήσει ομοιομορφία, αφετέρου να εξαλειφθεί η πιθανότητα ο ταξινομητής να τροφοδοτηθεί με περιττή ή και «άχρηστη» επιπλέον πληροφορία , έτσι ώστε να βελτιστοποιηθεί η ταχύτητά και η απόδοση του ως προς την επίλυση του προβλήματος που καλείται να επιτελέσει.

Κάθε έγγραφο μπορεί να θεωρηθεί από μία ακολουθία λέξεων και συμβόλων. Στην ακολουθία αυτή μπορεί να υπάρχουν σύμβολα ή χαρακτήρες που να δημιουργούν θόρυβο στα δεδομένα. Η μέθοδος του **tokenization** χωρίζει το κείμενο σε μικρότερα μέρη (tokens), που αυτά μπορεί να είναι λέξεις, σημεία στίξης ή άλλα σύμβολα. Για παράδειγμά, η φράση « Hey , Ho, let's go !" μετατρέπεται στα tokens "Hey" ,",", "Ho" ,"," ,"let's", "go","!" .

Στη συνέχεια οι λέξεις υφίστανται **case folding**, δηλαδή όλοι οι χαρακτήρες που τις απαρτίζουν να μετασχηματίζονται σε κεφαλαίους ή πεζούς (συνήθως το δεύτερο).

Μία επίσης συνήθης μέθοδος που χρησιμοποιείται είναι η **απομάκρυνση των λειτουργικών λέξεων** (stop words). Τέτοιες συχνές λέξεις, ακριβώς λόγω της συχνότητας και του συνδετικού μόνο ρόλου τους, δεν σημασιοδοτούν το κείμενο και άρα μπορούν να αφαιρεθούν για τον σκοπό της κατηγοριοποίησης. Λειτουργικές λέξεις αποτελούν συνήθως μέρη του λόγου όπως: προθέσεις, σύνδεσμοι, βοηθητικά ρήματα, άρθρα και άλλα μέρη του λόγου τα οποία εμφανίζονται αρκετά συχνά, αλλά δεν μας χρειάζονται για την εκπαίδευση του ταξινομητή, αφού δεν προσφέρουν καμία επιπλέον πληροφορία στο μοντέλο. Μαζί με τα stop words μπορεί να γίνεται και απομάκρυνση συμβόλων ή χαρακτήρων που αποτελούν θόρυβο για τα δεδομένα μας. Και στις δύο όμως περιπτώσεις πρέπει να συνυπολογίζεται η επιδραστικότητά τους στο σύνολο δεδομένων που καλούμαστε να ταξινομήσουμε. Έτσι για παράδειγμα οι λέξεις «to» , «be» , «or», «not» , που θα μπορούσαν να θεωρηθούν λέξεις άνευ σημασίας για την πλειοψηφία των συνόλων δεδομένων, θα έπαιζαν αρκετά σημαντικό ρόλο για την ταξινόμηση έργων του Shakespeare.

Μια επιπλέον ευρέως χρησιμοποιούμενη μέθοδος είναι αυτή της **αποκατάληξης** (Stemming**).** Η μέθοδος αυτή, που υλοποιείται με τη βοήθεια ειδικών αλγορίθμων (Lovins, Porterstemmer [19] κ.α.), περικόπτει τις καταλήξεις των διαφορετικών μορφών των λέξεων που μπορεί να συναντώνται στο σύνολο δεδομένων, επιστρέφοντας το στέλεχός των λέξεων. Σημειώνεται πως το στέλεχος αυτό μπορεί να μην είναι πάντα μια κανονική λέξη. Για παράδειγμα, οι λέξεις «training», «train», «trainer» επιστρέφουν τη ρίζα «train». Αυτή η μέθοδος είναι πολύ χρήσιμη στον τομέα της ανάλυσης συναισθήματος ώστε να μειώνονται οι διαστάσεις του training set, δηλαδή να μειώνονται οι λέξεις. Ο λόγος είναι ότι οι διαφορετικές γραμματικές μορφές μιας λέξης πολλές φορές δεν αλλάζουν το νόημά της. Παρόμοια μέθοδος είναι αυτή της **Λημματοποίησης** (Lemmatization), όπου επιστρέφει το λήμμα από το οποίο προέρχονται οι λέξεις. Έτσι, για τα ουσιαστικά μπορεί να

επιστρέφεται ο ενικός αριθμός («cars» → «car») ή για τα ρήματα το απαρέμφατό τους («going» , «go», «went», « gone» –> « go»).

2.2 Αναπαράσταση Κειμένων και Εξαγωγή Χαρακτηριστικών

Διάφορα μοντέλα χρησιμοποιούνται για την αναπαράσταση των υπό κατηγοριοποίηση κειμένων, μοντέλα τα οποία χρησιμοποιούνται για την εξαγωγή των χαρακτηριστικών προκειμένου να εκπαιδευτούν οι αλγόριθμοι ταξινόμησης.

2.2.1 Bag of Words

Το πιο απλοϊκό, πλην όμως διαδεδομένο μοντέλο είναι ο **Σάκος Λέξεων** (Bag of Words - BoW), όπου το κείμενο αναπαρίσταται ως το σύνολο των λέξεων που το αποτελούν. Σύμφωνα με το συγκεκριμένο μοντέλο τα κείμενα μετατρέπονται σε διανύσματα, έτσι ώστε κάθε διάνυσμα να αναπαριστά την συχνότητα όλων των διακριτών λέξεων που υπάρχουν στον διανυσματικό χώρο εγγράφων για το συγκεκριμένο έγγραφο. Το κάθε κείμενο, τελικά, αντιμετωπίζεται ως ένα σύνολο από τις λέξεις που το αποτελούν, παραβλέποντας τη σημασιολογία, τη σύνταξη, τα γραμματικά μέρη και τη σειρά των λέξεων, αλλά κρατώντας μόνο την παρουσία (κωδικοποιείται ως 1) ή την απουσία τους (κωδικοποιείται ως 0), ή και την συχνότητα εμφάνισης τους.

Για παράδειγμα, έστω τα εξής κείμενα:

“John and Jane live in a big house in Athens”

“John married Jane back in 2015”

Λεξιλόγιο: {John, and, Jane, live, in, a, big, house, Athens, married, back, 2015}

BoW αναπαράσταση σύμφωνα με τη συχνότητα εμφάνισης των λέξεων στις προτάσεις

	John	**and**	**Jane**	**live**	**in**	**Athens**	**a**	**big**	**house**	**married**	**back**	**2015**
(1)	1	1	1	1	2	1	1	1	1	0	0	0
(2)	1	0	1	0	1	0	0	0	0	1	1	1

BoW αναπαράσταση σύμφωνα με την εμφάνιση των λέξεων στις προτάσεις.

	John	**and**	**Jane**	**live**	**in**	**Athens**	**a**	**big**	**house**	**married**	**back**	**2015**
(1)	1	1	1	1	1	1	1	1	1	0	0	0
(2)	1	0	1	0	1	0	0	0	0	1	1	1

Το γεγονός αυτό καταδεικνύει αυτόματα μία αβελτηρία του συγκεκριμένου μοντέλου. Δύο προτάσεις που αποτελούνται από τις ίδιες λέξεις, αλλά με εντελώς διαφορετικό νόημα λόγω της σειράς των λέξεων, αντιμετωπίζονται από το μοντέλο με τον ίδιο τρόπο. Επίσης, ένα ακόμη μειονέκτημα του συγκεκριμένου μοντέλου είναι ότι συνυπολογίζει λέξεις που παρουσιάζονται αρκετά συχνά στο κείμενο, χωρίς να προσδίδουν κάτι επιπλέον στο νόημα του (άρθρα, προθέσεις κτλ) και δεν έχουν ήδη παραληφθεί κατά τη διαδικασία της απομάκρυνσης των stopwords. Μολαταύτα, το μοντέλο αυτό χρησιμοποιείται αρκετά συχνά, αφού τα διανύσματα που παράγει

είναι γραμμικώς διαχωρίσιμα και επομένως SVM ταξινομητές με γραμμικό πυρήνα επιτυγχάνουν πολύ καλή απόδοση.

2.2.2 N-grams

Αναφερθήκαμε προηγουμένως στο μειονέκτημα του μοντέλου BoW σχετικά με την παράβλεψη της σειράς των λέξεων που απαρτίζουν το κείμενο. Το πρόβλημα αυτό έρχεται να το αντιμετωπίσει το μοντέλο n-gram, το οποίο είναι μια συνεχόμενη ακολουθία n αντικειμένων που απαρτίζουν το κείμενο και μπορούμε να πούμε ότι αποτελεί μία γενίκευση του μοντέλου BoW (συγκεκριμένα το BOW είναι ένα n-gram μοντέλο για n=1, εφόσον αναφερόμαστε σε λέξεις) Τα αντικείμενα αυτά μπορεί να είναι λέξεις, συλλαβές ή ακόμη και χαρακτήρες. Χρησιμοποιώντας μάλιστα ένα μοντέλο n-gram χαρακτήρων αντιπαρερχόμαστε προβλήματα που μπορεί να οφείλονται σε ορθογραφικά λάθη ή σε λάθος γραφή των λέξεων.

Αν στις προτάσεις του προηγούμενου παραδείγματος θέλαμε να σχηματίσουμε τα 3-grams (trigrams) που προκύπτουν, θα προέκυπτε το παρακάτω λεξιλόγιο:

{ John and Jane, and Jane live, Jane live in, live in a, in a big, a big house, big house in, house in Athens, John married Jane, married Jane back, Jane back in, back in 2015 }

2.2.3 Term Frequency Inverse Document Frequency (tf-idf)

Ένα επίσης μειονέκτημα του BoW μοντέλου, όπως προείπαμε, είναι ότι τα διανύσματα που προκύπτουν βασίζονται αποκλειστικά στη συχνότητα εμφάνισης των λέξεων στο λεξιλόγιο του κειμένου. Συγκεκριμένα, κάποιες λέξεις που εμφανίζονται αρκετά συχνά σε ένα κείμενο μπορεί να μην έχουν καμία σημασία για το νόημα του κειμένου, σε αντίθεση με λέξεις που εμφανίζονται με μικρότερη συχνότητα, αλλά είναι εξαιρετικά πιθανό να νοηματοδοτούν το κείμενο και συνακόλουθα να είναι «βαρύτερης» σημασίας κατά την κατηγοριοποίηση από έναν ταξινομητή. Για το λόγο αυτό χρησιμοποιείται το μοντέλο tf – idf (Term Frequency Inverse Document Frequency), κατά το οποίο αποτιμάται η «σημαντικότητα» κάθε λέξης σύμφωνα με τον αριθμό εμφάνισής της τόσο στο κείμενο, όσο και στο σύνολο των κειμένων. Η τιμή του tf-idf αυξάνεται με τον αριθμό των φορών που ένας όρος εμφανίζεται σε ένα κείμενο, αλλά αντισταθμίζεται και από τις φορές που ο όρος αυτός εμφανίζεται στο σύνολο των κειμένων του συνόλου δεδομένων και αποτελεί ένα από τα πιο συνήθη γνωρίσματα σε προβλήματα κατηγοριοποίησης κειμένων, έχοντας την ίδια επίδραση με τη διαδικασία απαλοιφής των stopwords, που γίνεται κατά την προεπεξεργασία του συνόλου των δεδομένων.

Μαθηματικά, ο δείκτης tf-idf για κάθε όρο είναι το γινόμενο δύο μετρικών, της συχνότητας του όρου στο κείμενο επί της αντίστροφης συχνότητας κειμένου, δηλαδή ***tf-idf = tf x idf***. 9

Για έναν όρο (λέξη) *w* σε ένα κείμενο *d*, η συχνότητα εμφάνισης του όρου *tf(w,d)* θα είναι:

tf(w,d) = f(w,d), όπου *f* είναι η συνάρτηση συχνότητας του όρου.

Η συχνότητα του όρου μπορεί να αποτυπωθεί σαν δυαδικό γνώρισμα, χρησιμοποιώντας το 1 εφόσον ο όρος εμφανίζεται στο κείμενο και το 0 εφόσον αυτός απουσιάζει:

$$tf(w,d) = \begin{cases} 1 \varepsilon \acute{\alpha} \nu f(w,d) > 0 \\ 0 \varepsilon \acute{\alpha} \nu f(w,d) = 0 \end{cases} \quad (1)$$

Η αντίστροφη συχνότητα κειμένου *idf* χρησιμοποιείται για να δοθεί μεγαλύτερο βάρος στους όρους που εμφανίζονται σχετικά λιγότερο συχνά στη συλλογή των κειμένων. Αυτή μαθηματικά ορίζεται ως το πηλίκο *D/df(t)* , όπου *D* είναι ο συνολικός αριθμός των κειμένων που απαρτίζουν το σύνολο δεδομένων και *df(t)* είναι ο αριθμός των κειμένων που περιέχουν τον όρο *t*.

Εξαιτίας του μεγάλου αριθμού κειμένων σε ένα σύνολο δεδομένων, χρησιμοποιείται η λογαριθμική έκφραση της σχέσης. Επομένως η αντίστροφη συχνότητα κειμένου ορίζεται μαθηματικά ως:

$$idf(t) = log\frac{D}{df(t)} \quad (2)$$

Επομένως ο δείκτης *tf-idf* μπορεί να υπολογιστεί με το γινόμενο των δύο επιμέρους όρων *tf* και *idf*, από τις εξισώσεις (1) και (2).

2.2.4 Συντακτικά Χαρακτηριστικά

Το μοντέλο BoW δεν λαμβάνει υπόψη του συντακτική και σημασιολογική πληροφορία που περιέχουν τα κείμενα. Αυτό μπορεί να βελτιωθεί χρησιμοποιώντας τρόπους αναπαράστασης που αποτυπώνουν συντακτικές σχέσεις εξάρτησης μεταξύ των λέξεων και των κειμένων (syntactic dependencies). Οι σχέσεις αυτές εξάγονται μετά από συντακτική ανάλυση των κειμένων και τη δημιουργία δέντρου εξαρτήσεων.

Σε αυτό το μοντέλο αναπαράστασης εντάσσονται διάφορες τεχνικές όπως:

- **Lexical Syntactic Feature (LSF) [16],** όπου επισημαίνονται οι γραμματικές εξαρτήσεις μέσα σε μία πρόταση. Τα γνωρίσματα που προκύπτουν είναι ζεύγη λέξεων που από μόνα τους σημασιοδοτούν την πρόταση.
- **Part-of-Speech tagging (PoS)** κατά την οποία οι λέξεις επισημειώνονται αναλόγως με το μέρος του λόγου στο οποίο ανήκουν, παραδείγματος χάριν, ουσιαστικά, ρήματα, αντωνυμίες, επίθετα, προσδιορισμοί κτλ. (βλέπε *Διάγραμμα* 1 για την αγγλική γλώσσα). Στη διαδικασία λαμβάνονται υπ' όψιν ο ορισμός της λέξης, η θέση της στην πρόταση καθώς και η σχέση της με τις γειτονικές της λέξεις.

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Διάγραμμα 1 - *Part of Speech Tags για την αγγλική γλώσσα(Penn Treebank Project)*[8]

Τέλος, ανάμεσα σε πλειάδες γνωρισμάτων που επιπλέον χρησιμοποιούνται, εξέχουσα θέση κατέχουν η **Word Sense Disambiguation Technique [16]** που υποδηλώνει την αναγνώριση του νοήματος μιας συγκεκριμένης λέξης στο πλαίσιο μιας φράσης όταν η λέξη αυτή μπορεί να έχει πολλαπλές έννοιες καθώς επίσης και η **Ανάλυση Συναισθήματος**, όπου προσδιορίζεται και κωδικοποιείται η πολικότητα των λέξεων, αναλόγως αν έχουν θετική, αρνητική ή ουδέτερη σημασία.

2.3 Αλγόριθμοι Επιβλεπόμενης Μάθησης

Η παρούσα εργασία θα επικεντρωθεί σε πρόβλημα ταξινόμησης (classification). Όπως αναφέρθηκε και παραπάνω το πρόβλημα της ταξινόμησης καλείται να προβλέψει την κατηγορία - ή αλλιώς κλάση - νέων δεδομένων. Ταξινομεί δεδομένα κατασκευάζοντας ένα μοντέλο βασισμένο σε ένα σύνολο δεδομένων εκπαίδευσης (training set) και τις τιμές κλάσεις (ετικέτες κατηγορίας - class labels) για κάθε είσοδο και το χρησιμοποιεί για ταξινόμηση νέων δεδομένων στις αντίστοιχες κλάσεις. Πιο αναλυτικά, το σύστημα πρέπει να «μάθει», δηλαδή να κατασκευάσει ένα νέο μοντέλο

[8] https://skylerkim.tistory.com/entry/nltktagpostagtoken

υπό μορφή μιας **συνάρτησης πρόγνωσης**, η οποία θα απεικονίζει δεδομένες εισόδους σε ήδη γνωστές, εξόδους, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους στις οποίες ζητούμενο είναι η έξοδος τους να παρουσιάζει το ελάχιστο δυνατό σφάλμα πρόβλεψης.

Για τη συνάρτηση πρόγνωσης ισχύουν τα ακόλουθα:

- Κάθε είσοδος που μπορεί να δεχθεί η συνάρτηση χαρακτηρίζεται ως **παράδειγμα**, δημιουργώντας έτσι ένα **σύνολο παραδειγμάτων (dataset)**.
- Οι είσοδοι περιγράφονται με βάση τα **γνωρίσματα** ή **χαρακτηριστικά** (features) που διαθέτουν και έχουν χαρακτηριστεί ως σημαντικά από την αρχή της μελέτης του προβλήματος που καλείται να επιλύσει το σύστημα.
- Οι δεδομένες είσοδοι συγκεντρώνονται από παρατηρήσεις και αποτελούν το λεγόμενο **σύνολο εκπαίδευσης** (training set) που αποτελεί υποσύνολο του συνόλου παραδειγμάτων .
- Το υπόλοιπο μέρος του συνόλου στιγμιότυπων αποτελεί το **σύνολο ελέγχου** (test set) που θα χρησιμοποιηθεί κατά τη φάση πιστοποίησης (validation).
- Η συνάρτηση που απεικονίζει μια είσοδο από το σύνολο εκπαίδευσης στη γνωστή της έξοδο καλείται **συνάρτηση στόχου** (objective/goal function).
- Η τιμή που επιστρέφει η συνάρτηση στόχου για ένα στιγμιότυπο από το σύνολο στιγμιότυπων, δίνεται σε μια μεταβλητή που καλείται **κλάση ταξινόμησης** (classification class).
- Στην επιβλεπόμενη μάθηση, η συμπεριφορά της συνάρτησης στόχου βελτιώνεται μέσω διαδικασιών εκπαίδευσης με τη βοήθεια της **συνάρτησης λάθους** (error ή loss function) που εντοπίζει τη διαφορά της μεταβλητής στόχου από την επιθυμητή έξοδο.

Παρακάτω παρουσιάζονται μερικοί από τους κύριους αλγορίθμους ταξινόμησης. Δεδομένης της ραγδαίας εξέλιξης του κλάδου, η παρουσίαση δεν μπορεί παρά να είναι συνοπτική, με αναφορά στους πιο σύνηθεις αλγόριθμους που συναντάμε στη σχετική βιβλιογραφία.

2.3.1 Naïve Bayes

Ο αλγόριθμος Naive Bayes ανήκει σε μια ευρύτερη οικογένεια ταξινομητών (ταξινομητές Bayes) και είναι ένας πιθανοθεωρητικός αλγόριθμος εποπτευόμενης μάθησης που χρησιμοποιεί το θεώρημα του Bayes, κάνοντας μια «αφελή» (naïve) υπόθεση πως κάθε γνώρισμα (feature) είναι ανεξάρτητο από τα άλλα. Στην θεωρία πιθανοτήτων, ανεξάρτητες ονομάζονται δυο τυχαίες μεταβλητές όταν η παρουσία της μιας δεν επηρεάζει την πιθανότητα εμφάνισης της άλλης.

Πιο συγκεκριμένα ζητούμενο είναι να βρεθεί η εκ των υστέρων πιθανότητα (posterior probability) να ανήκει μία παρατήρηση σε μία κλάση, δοθείσης της εκ των προτέρων πιθανότητας (prior probability).

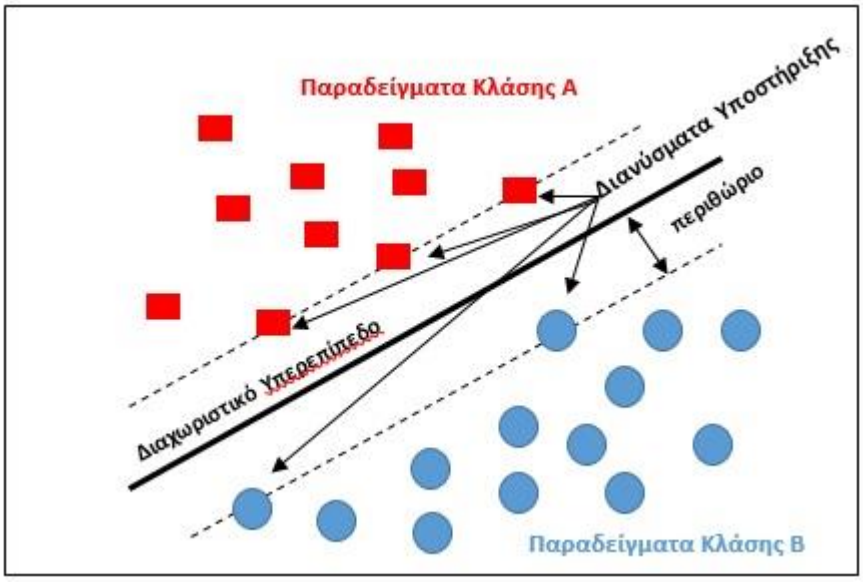
Είναι ένας εξαιρετικά δημοφιλής ταξινομητής, ειδικά σε προβλήματα επεξεργασίας φυσικής γλώσσας, που λόγω της απλότητας του συνήθως συναντάται στη βιβλιογραφία σαν μέθοδος αναφοράς (baseline) για σύγκριση με άλλες ανταγωνιστικές μεθόδους. Σημειώνεται ότι ανάλογα με την κατανομή που χρησιμοποιούν για να μοντελοποιήσουν την εκ των προτέρων πιθανότητας διακρίνονται σε Multinomial, Bernoulli και Gaussian μεταξύ άλλων.

2.3.2 Λογιστική Παλινδρόμηση (Logistic Regression)

Η **παλινδρόμηση** είναι μια ευρέως χρησιμοποιούμενη στατιστική τεχνική μοντελοποίησης που αποσκοπεί στη συσχέτιση μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών, υπολογίζοντας την πιθανότητα συσχέτισης μέσω τροφοδότησης των δεδομένων σε μία λογαριθμική συνάρτηση. Η **Λογιστική Παλινδρόμηση** χρησιμοποιείται όταν η εξαρτημένη μεταβλητή είναι κατηγορική. Επίσης η Λογιστική Παλινδρόμηση χρησιμοποιείται κυρίως για ταξινόμηση σε δύο κλάσεις (binary), χωρίς όμως να μην βρίσκει εφαρμογή σε προβλήματα ταξινόμησης πολλαπλών κλάσεων. Στις περιπτώσεις αυτές χρησιμοποιείται η στρατηγική "one vs. all", όπου κάθε κλάση ταξινομείται απέναντι στις υπόλοιπες ξεχωριστά (δυαδική ταξινόμηση) και εν τέλει επιλέγεται ο ταξινομητής με την καλύτερη απόκριση.

2.3.3 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

Οι Μηχανές Διανυσμάτων Υποστήριξης είναι μέθοδοι που χρησιμοποιούνται για ταξινόμηση και παλινδρόμηση, αν και συνήθως εμπλέκονται στην επίλυση προβλημάτων ταξινόμησης, χρησιμοποιώντας κατά κύριο λόγο μεθόδους βελτιστοποίησης. Η ταξινόμηση των δεδομένων στηρίζεται στην εύρεση ενός βέλτιστου υπερεπιπέδου (hyperplane), μιας γραμμής (επιπέδου σε πάνω από 2 διαστάσεις) που διαχωρίζει τα δεδομένα ανάλογα την κλάση τους, δημιουργώντας το μέγιστο περιθώριο (margin) μεταξύ τους (βλέπε Διάγραμμα 2). Τα πλησιέστερα σημεία στο υπερεπίπεδο, από τα οποία η απόσταση θέλουμε να είναι μέγιστη, ονομάζονται διανύσματα υποστήριξης (support vectors). Στην περίπτωση που ο γραμμικός διαχωρισμός είναι αδύνατος, γίνεται χρήση κατάλληλων απεικονίσεων που μεταφέρουν το σύνολο των δεδομένων σε μεγαλύτερη διάσταση ώστε να επιτευχθεί τελικά ο διαχωρισμός τους. Η ικανότητα γενίκευσης της χρήσης των SVM σε μη γραμμικά δεδομένα στηρίζεται στο τέχνασμα του πυρήνα (kernel trick). Κάθε μηχανή διανυσμάτων υποστήριξης είναι ένας δυαδικός ταξινομητής, έχει δηλαδή τη δυνατότητα διαχωρισμού των δεδομένων σε δύο κλάσεις. Εάν οι κλάσεις είναι περισσότερες, τότε κρίνεται απαραίτητη η χρήση περισσότερων μηχανών διανυσμάτων υποστήριξης και η εφαρμογή διάφορων τεχνικών («one vs rest» , "one vs one").



Διάγραμμα 2 - *Παράδειγμα ενός ταξινομητή SVM*

2.3.4 Δέντρα Αποφάσεων

Τα Δέντρα Απόφασης (Decision Trees) είναι από τα πιο γνωστά μοντέλα κατηγοριοποίησης. Το δέντρο απόφασης έχει τη μορφή ενός γράφου με την κλασική δεντρική δομή, που αποτελείται από: (α) έναν αρχικό κόμβο («ρίζα»), (β) τους εσωτερικούς κόμβους (ή κόμβους εξέτασης) και (γ) τους εξωτερικούς κόμβους («φύλλα», τερματικοί κόμβοι ή κόμβοι απόφασης). Δύο ή περισσότερες ακμές («κλαδιά») μπορούν να αναπτυχθούν από κάθε εσωτερικό κόμβο (όχι από κόμβο «φύλλο»). Κάθε κόμβος αντιστοιχεί σε ένα συγκεκριμένο χαρακτηριστικό και τα «κλαδιά» αντιστοιχούν σε κάποια συνθήκη. Οι τερματικοί κόμβοι αντιστοιχούν σε κάποια κατηγορία (κλάση). Υπάρχουν διάφοροι αλγόριθμοι βασισμένοι στη λογική του Δέντρου Απόφασης, οι πιο ευρέως χρησιμοποιούμενοι σε προβλήματα ταξινόμησης είναι οι ID3 και η εξέλιξή του C4.5, ενώ ο αλγόριθμος CART χρησιμοποιείται τόσο σε προβλήματα ταξινόμησης όσο και παλινδρόμησης.

2.3.5 Τεχνητά Νευρωνικά Δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks, ANN), είναι υπολογιστικά μοντέλα, τα οποία αποτελούνται από υπολογιστικούς κόμβους, που ονομάζονται νευρώνες (neurons), οι οποίοι συνδέονται μεταξύ τους σχηματίζοντας ένα δίκτυο. Τα μοντέλα αυτά είναι εμπνευσμένα από τη λειτουργία του ανθρώπινου εγκεφάλου, την οποία προσπαθούν να προσομοιάσουν. Κάθε διασύνδεση των νευρώνων μεταξύ τους σταθμίζεται με κάποιο βάρος, το οποίο προσαρμόζεται κατά τη διαδικασία εκπαίδευσης του μοντέλου.

Υπάρχουν τρεις κατηγορίες νευρώνων. Οι νευρώνες εισόδου, οι υπολογιστικοί νευρώνες (οι οποίοι στα απλά ANN βρίσκονται σε ένα κρυφό επίπεδο) και οι νευρώνες εξόδου. Οι νευρώνες εισόδου δεν εκτελούν κανέναν υπολογισμό, είναι αυτοί που δέχονται τα δεδομένα στην είσοδο του συστήματος και τροφοδοτούν με αυτά το επόμενο στρώμα κρυφών νευρώνων. Οι νευρώνες εξόδου διοχετεύουν στο περιβάλλον τις τελικές αριθμητικές εξόδους του δικτύου. Οι υπολογιστικοί/κρυφοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο συναπτικό βάρος και υπολογίζουν το ολικό αποτέλεσμα των γινομένων. Το άθροισμα αυτό τροφοδοτείται ως όρισμα στη συνάρτηση ενεργοποίησης, την οποία υλοποιεί εσωτερικά κάθε κόμβος. Η τιμή που λαμβάνει η συνάρτηση για το εν λόγω όρισμα είναι και η έξοδος του νευρώνα για τις τρέχουσες εισόδους και βάρη. Στην περίπτωση της *Βαθιάς Μάθησης* (deep learning), τα ANN διαθέτουν περισσότερα από ένα κρυφά επίπεδα και αυτό το χαρακτηριστικό τα διαχωρίζει από τα απλά ANN.

2.4 Αξιολόγηση Μοντέλων Ταξινόμησης

Η απόδοση των μοντέλων ταξινόμησης αποσκοπεί στην ποσοτικοποίηση της ευστοχίας που έχουν οι προβλέψεις τους στα νέα σημεία δεδομένων που αυτοί καλούνται να ταξινομήσουν. Τις περισσότερες φορές, γίνεται μέτρηση της απόδοσης σε ένα σύνολο δεδομένων που αποτελείται από δεδομένα (δεδομένα ελέγχου ή δεδομένα πιστοποίησης) τα οποία δεν χρησιμοποιήθηκαν για να εκπαιδεύσουν τον ταξινομητή. Αυτό το σύνολο δεδομένων αποτελείται από αρκετές παρατηρήσεις και τις αντίστοιχες κλάσεις τους. Αφού εξαχθούν τα γνωρίσματα με τον ίδιο τρόπο που γινόταν στην εκπαίδευση του μοντέλου, τα γνωρίσματα αυτά τροφοδοτούν το ήδη εκπαιδευμένο μοντέλο και παρατηρούνται οι προβλέψεις του για κάθε σημείο δεδομένων. Αυτές οι προβλέψεις

συγκρίνονται με τις πραγματικές κλάσεις για την εκτίμηση της ικανότητας και της ακρίβειας των προβλέψεων του μοντέλου.

Προκειμένου λοιπόν να αξιολογηθεί το εκάστοτε μοντέλο, χρησιμοποιούνται κάποιες μετρικές, οι οποίες θα αναλυθούν στη συνέχεια. Πριν ορίσουμε τις μετρικές που θα χρησιμοποιήσουμε ας δούμε κάποιες μεταβλητές που τις καθορίζουν:

Έστω για ένα πρόβλημα ταξινόμησης δύο κλάσεων α, β:

- ***TN*** *(True Negative)*: Ο αριθμός των παραδειγμάτων που ανήκουν στην κλάση α και ο αλγόριθμός τα ταξινομεί στην κλάση α.
- ***TP*** *(True Positive)*: Ο αριθμός των παραδειγμάτων που ανήκουν στην κλάση β και ο αλγόριθμός τα ταξινομεί στην κλάση β
- ***FN*** *(False Negative)*: Ο αριθμός των παραδειγμάτων που ανήκουν στην κλάση β και ο αλγόριθμός τα ταξινομεί στην κλάση α
- ***FP*** *(False Positive)*: Ο αριθμός των παραδειγμάτων που ανήκουν στην κλάση α και ο αλγόριθμός τα ταξινομεί στην κλάση β.

Αντίστοιχα ορίζονται οι μεταβλητές αυτές και για περισσότερες κλάσεις. Οι μετρικές που χρησιμοποιούνται συνήθως περιγράφονται στο εξής.

2.4.1 Accuracy (Συνολική Ακρίβεια)

Εκφράζει το ποσοστό των σωστών προβλέψεων του μοντέλου, δηλαδή το κλάσμα των σωστών προβλέψεων προς το σύνολο των προβλέψεων του ταξινομητή.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Η συνολική ακρίβεια εκφράζει τη συνολική απόδοση ενός ταξινομητή, χωρίς να δίνει έμφαση στις κλάσεις ξεχωριστά.

2.4.2 Precision (Ακρίβεια)

Η συγκεκριμένη μετρική αναφέρεται σε κάθε κλάση του προβλήματος ταξινόμησης και εκφράζει το ποσοστό των προβλέψεων που έγιναν και είναι πραγματικά σωστές σε σχέση με όλες τις προβλέψεις για τη συγκεκριμένη κλάση.

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

2.4.3 Recall (Ανάκληση) ή Sensitivity

Η συγκεκριμένη μετρική αναφέρεται και αυτή σε κάθε κλάση του προβλήματος ταξινόμησης και εκφράζει το ποσοστό των σωστών προβλέψεων για τη συγκεκριμένη κλάση προς τον συνολικό αριθμό των προβλέψεων που αναφέρονται στα παραδείγματα που το μοντέλο ταξινομεί σωστά, συν τα παραδείγματα που λανθασμένα ταξινομούνται ως άλλης κλάσης.

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

Η Ανάκληση αξιολογεί με λίγα λόγια την ικανότητα του ταξινομητή να αναγνωρίζει τα παραδείγματα που αναφέρονται σε μία συγκεκριμένη κλάση.

2.4.4 F1 Score

Αποτελεί ένα μέτρο συνδυασμού της Precision και Recall και δεν είναι τίποτε άλλο παρά ο αρμονικός μέσος των δύο παραπάνω μετρικών.

$$F1score = \frac{2xPrecisionxRecall}{Precision+Recall}$$

Συναντώνται τρεις παραλλαγές της συγκεκριμένης μετρικής, συγκεκριμένα macro F1, micro F1 και weighted F1, οι οποίες χρησιμοποιούνται κυρίως σε ταξινομήσεις πολλαπλών κλάσεων:

- Weighted F1: υπολογίζει την τιμή του F1 για κάθε κλάση ξεχωριστά, και τις προσθέτει πολλαπλασιάζοντας κάθε μία με ένα βάρος που προκύπτει από τον αριθμό των αληθώς θετικών (TP) κάθε κλάσης, ως εκ τούτου η κλάση με τα περισσότερα παραδείγματα υπερτερεί σε σχέση με τις άλλες.
- Micro F1: λαμβάνει υπόψη του τον ολικό αριθμό των αληθώς θετικών (TP), ψευδώς αρνητικών (FN) και ψευδώς θετικών (FP) όλων των κλάσεων και υπολογίζει απευθείας το F1, χωρίς να ευνοεί κάποια κλάση.
- MacroF1: υπολογίζεται προσθέτοντας τα F1 κάθε κλάσης.

2.4.5 Καμπύλες ROC

Επίσης σύγκριση μεταξύ μοντέλων μπορεί να γίνει σχηματίζοντας τις καμπύλες ROC του κάθε ταξινομητή. Οι καμπύλες αυτές συσχετίζουν την αληθή θετική τιμή (TP_{rate} = TP/TP+FP) με την ψευδώς θετική τιμή (FP_{rate} = FP/FP+TN). Το διάγραμμα της καμπύλης ROC σχηματίζεται σχεδιάζοντας τις τιμές TP και FP, και κάθε σημείο του διαστήματος ROC αντιστοιχεί στην απόδοση ενός απλού ταξινομητή για μία δεδομένη κατανομή. Η αριθμητική σύγκριση γίνεται με χρήση του εμβαδού της καμπύλης (AUC, Area Under Curve). Η AUC, που παίρνει τιμές από 0-1, εκφράζει την ικανότητα του ταξινομητή να αγνοεί λάθος ταξινόμηση.

Τέλος, ένας τρόπος παρουσίασης των επιδόσεων ανά κλάση ενός ταξινομητή είναι με τη χρήση του **Πίνακα Σύγχυσης** (confusion matrix). Ο Πίνακας Σύγχυσης είναι ένας δισδιάστατος πίνακας, όπου οι στήλες αντιστοιχούν στις προβλέψεις και οι γραμμές στις πραγματικές τιμές κλάσης. Στα κελιά του πίνακα αναγράφονται οι αληθινές θετικές, οι αληθινές αρνητικές, οι ψευδείς θετικές και οι ψευδείς αρνητικές προβλέψεις. Π.χ. για ένα πρόβλημα δυαδικής ταξινόμησης (κλάση Α και Β) ο Πίνακας σύγχυσης θα είχε την εξής μορφή:

	Πρόβλεψη Κλάσης Α	**Πρόβλεψη Κλάσης Β**
Πραγματική Κλάση Α	True A	False B
Πραγματική Κλάση Β	False A	True B

Πίνακας 1 - Υπόδειγμα Πίνακα Σύγχυσης

3 *Επισκόπηση Σχετικής Ερευνητικής Βιβλιογραφίας*

Στο κεφάλαιο αυτό παρατίθεται επισκόπηση και συζήτηση δημοσιευμένων μεθόδων για την αναγνώριση ρητορικής μίσους σε σύνολα δεδομένων φυσικής γλώσσας. Σημειώνεται ότι η επισκόπηση εστιάζει σε έρευνες με σύνολα δεδομένων στην Αγγλική γλώσσα και με σύνολα δεδομένων που προέρχονται από μέσα κοινωνικής δικτύωσης (κυρίως Twitter). Ο λόγος της εστίασης στα τελευταία είναι διττός: α) λόγω της ευρείας εξάπλωσης της χρήσης τους, η αυτοματοποίηση της ανίχνευσης ρητορικής μίσους σε σύνολα δεδομένων κοινωνικών μέσων είναι κοινωνική προτεραιότητα (high impact) και β) λόγω της επιλογής ανωνυμίας σε αυτά ή/και τη δυνατότητα χρήσης τους από αυτοματοποιημένα προφίλ (bots), τα κοινωνικά μέσα αποτελούν κατάλληλες πλατφόρμες για τη διάδοση πρακτικών ή/και ιδεών μίσους.

3.1 Αναλυτική Παρουσίαση Μελετών Βιβλιογραφίας

Οι **Waseem και Hovy [1]** ασχολήθηκαν με την ανίχνευση γνωρισμάτων μίσους σε ένα σύνολο δεδομένων που περιέχει 16.914 tweets, όλα στην Αγγλική γλώσσα. Σημειώνεται ότι η δημοσίευση τους – πέραν των ευρημάτων τους – αποτελεί κοινή αναφορά στη σχετική βιβλιογραφία, λόγω της ανοικτότητας των δεδομένων που χρησιμοποιήθηκαν. Δεδομένου ότι η πρόσβαση σε μεγάλο όγκο, μη τυχαίας σειράς, tweets δεν είναι ελεύθερη, καθώς και του κόστους επισημείωσης (labelling/annotation) τους, η ανοικτότητα των δεδομένων για την επιστημονική κοινότητα είναι σημαντική. Στον παρακάτω πίνακα φαίνονται οι κλάσεις και το πλήθος των κειμένων (tweets) ανά κλάση.

Tweets	**Κλάσεις**
3383 (από 613 χρήστες)	Σεξιστικό
1972 (από 9 χρήστες)	Ρατσιστικό
11559 (από 614 χρήστες)	Τίποτα από τα δύο

Πίνακας 2 - Χαρακτηριστικά Συνόλου Δεδομένων Waseem/Hovy [1]

Στη μελέτη **[1]**, το κάθε tweet υφίσταται αφαίρεση λειτουργικών λέξεων (stop-words) και σημείων στίξης, ως βήματα προεπεξεργασίας κειμένου. Επίσης, αφαιρούνται τα retweets. Οι συγγραφείς λαμβάνουν υπόψη τους γνωρίσματα των χρηστών (φύλο και γεωγραφική περιοχή των χρηστών), ως γνωρίσματα. Εξάγονται οι 10 πιο συχνές λέξεις ανά κλάση (ρατσιστική, σεξιστική). Στην περίπτωση των ρατσιστικών tweets αυτό φαίνεται να λειτουργεί ως προς την κατηγοριοποίηση

των tweets, σε αντίθεση με τη σεξιστική κλάση. Επίσης, ως χαρακτηριστικό λαμβάνεται υπόψη το μέγεθος και του κειμένου και των προσωπικών πληροφοριών του κάθε χρήστη. Τέλος, χρησιμοποιούνται (σε ξεχωριστά πειράματα αντίστοιχα) τα n-grams (n=1..4) σε επίπεδο χαρακτήρων και λέξεων αντίστοιχα.

Η μελέτη **[1]** χρησιμοποιεί τη Λογιστική Παλινδρόμηση ως μέθοδο ταξινόμησης των κειμένων στις τρεις προαναφερθείσες κλάσεις, με χρήση 10-fold cross validation κατά τη διαδικασία εκμάθησης. Τα αποτελέσματα των πειραμάτων συνοψίζονται στον παρακάτω πίνακα:

		Γνωρίσματα			
		char n grams	**+φύλο**	**+φύλο+τοποθεσία**	**word n grams**
Μετρικές	**F1**	73,89	73,93	73,62	64,58
	Precision	72,87	72,93	72,58	64,39
	Recall	77,75	77,74	77,43	71,93

Πίνακας 3 - Συνοψη Αποτελεσμάτων Μεθόδων Waseem/Hovy [1]

Αξίζει να σημειώσουμε τις εξής παρατηρήσεις:

- Γενικά η έλλειψη ενός ακριβή και ενιαίου ορισμού του τι συνιστά ρητορική μίσους (Hate Speech – HS) στην επιστημονική κοινότητα μηχανικής μάθησης είναι προβληματική. Για παράδειγμα, στη συγκεκριμένη μελέτη, ρατσιστικά σχόλια (ιδιαίτερα κατά των μουσουλμάνων) και σχόλια ενάντια στο γυναικείο φύλο, θεωρούνται – δικαίως – ρητορική μίσους, ενώ το ίδιο δεν ισχύει σε όλες τις μελέτες.
- Τα δημογραφικά γνωρίσματα, ιδιαίτερα το φύλο, δείχνει να βελτιώνει την απόδοση της ταξινόμησης, όμως αυτά είναι δύσκολο να εξαχθούν. Επίσης δεν είναι φερέγγυα ως προς τη γνησιότητά τους.
- Η διασπορά των δεδομένων σε χρήστες δεν είναι ισορροπημένη (balanced) ως προς τις τρεις κλάσεις ταξινόμησης. Συγκεκριμένα, η μία κλάση (ρατσιστικά σχόλια) προέρχονται από μόνο 9 χρήστες. Αυτό επηρεάζει γνωρίσματα όπως το φύλο ή η γεωγραφική τοποθεσία.

Οι **Davidson et al. [2]** ασχολήθηκαν με την ταξινόμηση ενός συνόλου δεδομένων που περιέχει 24.802 tweets (ανοιχτό σύνολο δεδομένων) σε τρεις κλάσεις (Ρητορική μίσους, Προσβλητική γλώσσα, Ουδέτερα). Το συγκεκριμένο σύνολο δεδομένων θεωρείται από τα πλέον αξιόπιστα στη βιβλιογραφία και έχει χρησιμοποιηθεί κατά κόρον σε άλλα πειράματα. Οι συγγραφείς αναφέρουν αρκετά αναλυτικά τις μεθόδους προεπεξεργασίας κειμένου, η οποία περιέχει πεζοποίηση των χαρακτήρων, stemming των λέξεων με τη χρήση του αλγόριθμου Porterstemmer [19]. Δημιουργούνται επίσης n-grams (1-3) καθώς και στάθμιση τους βάσει tfidf (term frequency–inverse document frequency). Κάθε n-gram που δημιουργείται συνοδεύεται από μια ετικέτα μέρους του λόγου (PoS - Part of speech). Επίσης χρησιμοποιούνται εργαλεία κατανοησιμότητας, αποδίδεται βαθμός συναισθήματος σε κάθε tweet. Λαμβάνονται τέλος υπόψιν τα hashtags, retweets, mentions, URLs καθώς και ο αριθμός των χαρακτήρων, λέξεων και συλλάβων κάθε tweet.

Η μελέτη χρησιμοποιεί αρχικά τη Λογιστική Παλινδρόμηση (L1 κανονικοποίηση) για τη μείωση της διαστασιμότητας των δεδομένων. Εν συνεχεία οι συγγραφείς πειραματίζονται με

διάφορους «παραδοσιακούς» ταξινομητές, όπως Λογιστική Παλινδρόμηση, Naïve Bayes, Δέντρα Απόφασης, Random Forests και γραμμικό SVM. Για την αποφυγή overfitting κάθε μοντέλο δοκιμάζεται με χρήση 5 fold cross validation, ενώ το 10% του δείγματος χρησιμοποιείται για την αξιολόγηση του μοντέλου. Το αποτέλεσμα είναι ότι οι ταξινομητές Λογιστικής Παλινδρόμησης και γραμμικού SVM να εμφανίζουν καλύτερα αποτελέσματα και ως εκ τούτου οι συγγραφείς χρησιμοποιούν στο τελικό τους μοντέλο τη Λογιστική Παλινδρόμηση (L2 κανονικοποίηση). Η προσέγγιση που ακολουθείται είναι η one-versus-rest, όπου κάθε ταξινομητής εκπαιδεύεται για κάθε κλάση του συνόλου δεδομένων και η κλάση με τη μεγαλύτερη προβλεπόμενη πιθανότητα συνυπολογιζόμενων όλων των ταξινομητών αποδίδεται σε κάθε tweet. Το καλύτερο βάσει αποτελεσμάτων μοντέλο παρουσιάζει τα εξής συνολικά αποτελέσματα: **Precision 0.91, Recall 0.90, F1 0.90.** Βάσει όμως του Πίνακα Σύγχυσης, διαπιστώνεται όμως πως σχεδόν το 40 % των tweets που αναφέρονται σε ρητορική μίσους ταξινομούνται λανθασμένα. Για τη συγκεκριμένη κλάση (Hate), Precision = 0,44 και Recall = 0,61.

Οι συγγραφείς αναδεικνύουν τη δυσκολία διαχωρισμού της ρητορικής μίσους από τη προσβλητική γλώσσα, για διαφόρους λόγους όπως η υποκειμενική αρχική επισημείωση των tweets από εξωτερικούς αξιολογητές, το γεγονός ότι δεν λαμβάνεται υπόψη το σημασιολογικό περιεχόμενο του κάθε tweet, καθώς βάρος δίνεται μόνο στην παρουσία συγκεκριμένων λέξεων, κάτι από επισημαίνεται και από τους **Warner and Hirschberg [3].**

Στη μελέτη των **Malmasi και Zempieri [4],** οι συγγραφείς προσπαθούν να διαχωρίσουν τη ρητορική μίσους (HS) από την προσβλητική-υβριστική γλώσσα (Offensive Language), ένα ζήτημα που τέθηκε και κατά την συζήτηση του προηγούμενου άρθρου. Οι συγγραφείς χρησιμοποιούν ένα ανοιχτό σύνολο δεδομένων που εμπεριέχει 14.509 tweets ήδη επισημειωμένα από εξωτερικούς αξιολογητές.[9] Το σύνολο δεδομένων είναι αρκετά ανομοιογενές και συγκεκριμένα:

Κλάση	**Αριθμός**	**% ποσοστό**
Hate Speech	2399	16,5
Offensive Language	4836	33,3
Ok	7274	30,2
Συνολικά tweets	14509	

Πίνακας 4 - Χαρακτηριστικά Συνόλου Δεδομένων Malmasi/Zempieri [4]

Οι συγγραφείς μετατρέπουν τους χαρακτήρες σε πεζούς, απομακρύνουν URLs και emojis, ενώ σαν γνωρίσματα χρησιμοποιούν char n-grams(1-8), word n-grams (1-3) καθώς και word skip grams (1,2,3 word bigrams).

[9] Το συγκεκριμένο dataset είναι «αμφιλεγόμενο». Βασίζεται στο dataset των Davidson et al. Σε προσωπική επικοινωνία με τους συγγραφείς επισημάνθηκε πως το dataset περιλαμβάνει επιλεγμένα tweets από το αρχικό και διανεμήθηκε χωρίς την άδεια τους και αποσύρθηκε αμέσως. Οι Malmasi και Zempieri ισχυρίζονται πως αυτό είναι το αρχικό dataset και στη συνέχεια εμπλουτίστηκε. Σε κάθε περίπτωση δεν είναι πια διαθέσιμο προς χρήση.

Το πείραμα χρησιμοποιεί ένα γραμμικό SVM ταξινομητή, ο οποίος εκπαιδεύεται κάθε φορά με ένα συγκεκριμένο σύνολο γνωρισμάτων, όπως επίσης και με όλα τα γνωρίσματα μαζί. Η αποτελεσματικότητα του αξιολογείται χρησιμοποιώντας την μετρική της συνολικής ακρίβειας (Accuracy). **Τη μεγαλύτερη Συνολική Ακρίβεια (78,0%) τη δίνει το μοντέλο που χρησιμοποιεί 4-grams χαρακτήρων.**

Και το συγκεκριμένο πείραμα αναδεικνύει τη δυσκολία ταξινόμησης μεταξύ Ρητορικής Μίσους και Υβριστικής/Προσβλητικής Γλώσσας. Συγκεκριμένα, 1113 (από τα 2399) tweets που ανήκουν στην κατηγορία Ρητορικής μίσους ταξινομούνται ως Υβριστική Γλώσσα.

Μία μελέτη που δίνει εντυπωσιακά αποτελέσματα είναι αυτή των **Gaydhani et al. [5].** Εδώ χρησιμοποιείται ένα εξαιρετικά μεγάλο σύνολο δεδομένων που συμπεριλαμβάνει αυτό τον Waseem and Hovy, Davidson και Malmasi, με το μειονέκτημα που έχει ήδη αναφερθεί σχετικά με το τελευταίο dataset. Επίσης σχετικά με το σύνολο δεδομένων των Waseem and Hovy, οι συγγραφείς ενοποιούν τα ρατσιστικά μαζί με τα σεξιστικά tweets, επισημειώνοντας τα σαν Ρητορική Μίσους (Hate) .

Κατά την προεπεξεργασία του συνόλου δεδομένων επιλέγονται οι κάτωθι πρακτικές: μετατροπή χαρακτήρων σε πεζούς, απομάκρυνση space pattern, URLs, mentions, retweets, λειτουργικών λέξεων (stop-words). Επίσης το σύνολο δεδομένων υφίσταται stemming με χρησιμοποίηση αλγόριθμου Porterstemmer.

Οι αλγόριθμοι που χρησιμοποιούνται είναι η Λογιστική Παλινδρόμηση (LR), ο Naïve Bayes (NB) και ο SVM. Αρχικά συγκρίνονται τα τρία μοντέλα σε διαφορετικούς συνδυασμούς γνωρίσματών (n-grams (1-3) + tfidf (L1 και L2 κανονικοποίηση)). Ο συνδυασμός n-grams (1-3) με L2 κανονικοποίηση της tfidf δίνει τα καλύτερα αποτελέσματα (Accuracy) και στα τρία μοντέλα με NB (0,926) > LR (0,918) > SVM (0,901) .

Στη συνέχεια παραμετροποιούνται τα μοντέλα NB και LR. Ρυθμίζοντας την παράμετρο *α* (alpha) στο NB ώστε να έχουνε τη μέγιστη ακρίβεια (Accuracy), καταλήγουνε σε ακρίβεια (Accuracy) 93,4%.Αντίστοιχα η μέγιστη ακρίβεια της LR μετά από τη βελτιστοποίηση των παραμέτρων είναι 95,1%.

Βάσει των παραπάνω, επιλέγεται η LR σαν τελικό μοντέλο η οποία παρουσιάζει τα κάτωθι αποτελέσματα (ανά κλάση και μέσο όρο) και συνολική ακρίβεια/Accuracy = 95,6%.

	Precision	**Recall**	**Fscore**
HS	0,94	0,96	0,95
OL	0,96	0,93	0,94
clean	0,96	0,98	0,97
Average	0,96	0,96	0,96

Πίνακας 5 - Συνοψη Αποτελεσμάτων Μεθόδων Gaydhani et al. [5]

	Ταξινόμηση		
	HS	**OL**	**Clean**
HS	0,965	0,021	0,014
OL	0,048	0,926	0,026
clean	0,01	0,013	0,977

Πίνακας 6 - Πίνακας Σύγχυσης Μεθόδων Gaydhani et al. [5]

Οι **Kwok and Wang [6]** προχωρούν σε δυαδική (binary) ταξινόμηση ενός αγνώστου συνόλου δεδομένων που απαριθμεί 24.582 tweets, χαρακτηρισμένα σαν racist/non racist. Κατά την προεπεξεργασία κειμένου, γίνεται μετατροπή χαρακτήρων σε πεζούς, απομάκρυνση URLs, λειτουργικών λέξεων (stop-words), mentions, σημείων στίξης και διόρθωση εναλλακτικών μορφών προσβλητικών λέξεων στην κανονική τους μορφή. Ως γνωρίσματα, χρησιμοποιούνται word unigrams (9437 μοναδικές λέξεις στα ρατσιστικά tweets, 8401 λέξεις στα όχι ρατσιστικά αναφορικά με τα σύνολα εκπαίδευσης του αλγορίθμου).

Ο Αλγόριθμος που χρησιμοποιείται είναι ο Naïve Bayes και το τελικό του αποτέλεσμα είναι Accuracy = 76% με μέσο λάθος 24%. Τονίζεται το γεγονός ότι από μεμονωμένες λέξεις (unigrams) δεν μπορεί να ανιχνευθεί ρητορική μίσους σε κείμενα και συνεπώς χρειάζεται να εμπλακούν σαν γνωρίσματα και bigrams, ανάλυση συναισθήματος των κειμένων καθώς και αποσαφήνιση των εννοιών των λέξεων.

Στη μελέτη των **Lee et al. [7],** οι συγγραφείς συγκρίνουν τόσο παραδοσιακά μοντέλων μηχανικής μάθησης, όσο και μοντέλα Νευρωνικών Δικτύων, ιδιαίτερα για την ανίχνευση Abusive Language.

Το σύνολο δεδομένων που χρησιμοποιούν περιέχει 70.904 tweets (από σύνολο 99.996) και είναι ένα μερικώς ανοικτό σύνολο δεδομένων (δεν κατέστη δυνατή η ανάκτηση του κειμένου φυσικής γλώσσας των tweets) που στη βιβλιογραφία φέρει τον τίτλο “Hate and Abusive Speech on Twitter”. Τα tweets που εμπεριέχει ταξινομούνται σε 4 κλάσεις, επισημειωμένα ως εξής

Normal	**Spam**	**Hate**	**Abusive**
60.5% (42932)	13.8% (9757)	4.4% (3100)	21.3%(15115)

Πίνακας 7 - Χαρακτηριστικά Συνόλου Δεδομένων Lee et al. [7]

Ως γνωρίσματα, χρησιμοποιούνται word n-grams (1-3), char n-grams (3-8) και bag-of-words σταθμισμένο με tfidf. Οι στατιστικοί ταξινομητές που συγκρίνονται είναι οι Naïve Bayes (NB), Logistic Regression (LR), SVM, Random Forests (RF), Gradient Boosted Trees (GBT), ενώ στο άρθρο δίνονται και οι παραμετροποιήσεις τους για τα πειράματα.

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα των μοντέλων (χρήση char n-grams και χρήση word n- grams) που αξιολογήθηκαν σε αυτή τη μελέτη.

Λαμβάνοντας υπόψιν τα συνολικά αποτελέσματα των μετρικών, διαπιστώνουμε ότι ο αλγόριθμος της Λογιστικής Παλινδρόμησης με χρήση char n-grams αποδίδει καλύτερα σε σχέση με όλα τα άλλα μοντέλα. Επίσης ο αλγόριθμος Λογιστικής Παλινδρόμησης με χρήση char n-grams παρουσιάζει τα καλύτερα αποτελέσματα όσον αφορά την ανίχνευση Abusive γλώσσας (F1=0,860).

Βλέποντας την κλάση Hateful διαπιστώνουμε ότι ο SVM παρουσιάζει καλύτερα αποτελέσματα όσον αφορά την Ακρίβεια (Precision), όμως στην Ανάκληση (Recall), όπως και στο F-Score, η λογιστική παλινδρόμηση (LR) υπερτερεί.

	Normal			Spam			Hateful			Abusive			Total		
Model	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
NB (word)	.776	.916	.840	.573	.378	.456	.502	.034	.063	.828	.744	.784	.747	.767	.741
NB (char)	.827	.805	.815	.467	**.609**	.528	.452	.061	.107	.788	.832	.803	.752	.751	.744
LR (word)	.807	.933	.865	.616	.365	.458	.620	.161	.254	.868	.844	.856	.786	.802	.780
LR (char)	.808	.934	.866	.618	.363	.457	.636	.183	.283	**.873**	.848	.860	.788	.804	.783
SVM (word)	.757	.967	.850	**.678**	.190	.296	**.836**	.034	.065	.865	.757	.807	.773	.775	.730
SVM (char)	.763	**.968**	.853	**.680**	.198	.306	**.805**	.070	.129	**.876**	.775	.822	.778	.781	.740
RF (word)	.776	.945	.853	.581	.213	.311	.556	.109	.182	.852	.819	.835	.757	.781	.745
RF (char)	.793	.934	.857	.568	.252	.349	.563	.150	.236	.853	.856	.854	.765	.789	.760
GBT (word)	.806	.921	.860	.581	.320	.413	.506	.194	.279	.854	.863	.858	.772	.794	.773
GBT (char)	.807	.913	.857	.560	.346	.428	.472	.187	.267	.859	.859	.859	.770	.791	.772

Πίνακας 8 - Συνοπτικά Αποτελέσματα Μεθόδων Lee et al.[7]

Οι **Park et al. [8],** ξεκινώντας από τη θεώρηση πως η ανίχνευση της Υβριστικής Γλώσσας είναι αρκετά δύσκολο εγχείρημα, αφού αφενός επαφίεται κατά πολύ στην υποκειμενικότητα και στην έλλειψη επιστημονικής επάρκειας των αξιολογητών που αναλαμβάνουν την επισημείωση του αρχικού συνόλου δεδομένων, αφετέρου ότι δεν λαμβάνεται υπόψη το περιεχόμενο του μηνύματος, αναπτύσσουν μια μέθοδο δύο φάσεων. Η πρώτη φάση έχει σαν στόχο την αρχική ταξινόμηση σε υβριστικά tweets και εν συνεχεία την ταξινόμηση των συγκεκριμένων αυτών tweets σε δύο κλάσεις. Παράλληλα εκτελούν και ένα πείραμα ενός βήματος όπου για το ίδιο σύνολο δεδομένων επιχειρούν ταξινόμηση 3 κλάσεων και το χρησιμοποιούν ως σύγκριση για το πείραμα των δύο φάσεων.

Στη μελέτη τους αναπτύσσουν ένα νέο μοντέλο (Hybrid CNN), χρησιμοποιώντας ως baselines παραδοσιακά μοντέλα Λογιστικής Παλινδρόμησης και SVM.

Το σύνολο δεδομένων που χρησιμοποιούν αποτελείται από τον συνδυασμό του Waseem and Hovy και ενός νέου από τον Waseem, το οποίο εμπλουτίζει το αρχικό σύνολο δεδομένων. Όπως προαναφέραμε εκτελούνται 2 πειράματα. Στο πρώτο βήμα, γίνεται ταξινόμηση σε 3 κλάσεις (sexist, racist, none). Στο δεύτερο βήμα, γίνεται αρχικά ταξινόμηση σε 2 κλάσεις, ώστε να διαχωριστούν τα "προσβλητικά» (abusive) tweets από τα υπόλοιπα και στη συνέχεια ταξινόμησή των συγκεκριμένων tweets σε 2 κλάσεις, "sexist" και "racist". Τα αποτελέσματα φαίνονται στους κάτωθι πίνακες.

Μοντέλο	Precision	Recall	F1
Λογιστική Παλινδρόμηση	0,825	0,824	0,814
SVM	0,814	0,808	0,973
Hybrid CNN	**0.827**	**0.827**	**0.827**

Πίνακας 9 - Συνοπτικά Αποτελέσματα Όλων των Κλάσεων Πρώτου Βήματος Park et al. [8]

Εξετάζοντας τη μετρική Recall, διαπιστώνουμε ότι είναι εξαιρετικά χαμηλή όσον αφορά τα σεξιστικά και ρατσιτικά tweets.

Μοντέλο	Sexist Recall	Racist Recall
Λογιστική Παλινδρόμηση	0.556	0.598
SVM	0.483	0.531
Hybrid CNN	0.679	0.766

Πίνακας 10 - Συνοπτικά Αποτελέσματα Ακρίβειας Κλάσεων Μίσους Πρώτου Βήματος Park et al. [8]

Μοντέλο	Precision	Recall	F1
Λογιστική Παλινδρόμηση	**0.828**	**0.831**	**0.824**
SVM	0.816	0.815	0.803
Hybrid CNN	0.807	0.809	0.807
Hybrid CNN+ Λογιστική Παλινδρόμηση	0.821	0.817	0.818

Πίνακας 11 - Συνοπτικά Αποτελέσματα Όλων των Κλάσεων Δευτέρου Βήματος Park et al. [8]

Μοντέλο	Prec	Rec	F1
Λογιστική Παλινδρόμηση	**0.954**	**0.953**	**0.952**
SVM	**0.954**	**0.953**	**0.952**
Hybrid CNN	0.951	0.950	0.950

Πίνακας 12 - Συνοπτικά Αποτελέσματα Ακρίβειας Κλάσεων Μίσους Δευτέρου Βήματος Park et al. [8]

Συμπερασματικά, η χρησιμοποίηση δύο δυαδικών ταξινομητών (στο πείραμα των δύο φάσεων), παρουσιάζει συγκρίσιμα αποτελέσματα με την ταξινόμηση σε μία φάση. Ο συνδυασμός δύο ταξινομητών Λογιστικής Παλινδρόμησης είναι εξίσου αποτελεσματικός όσον αφορά τα αποτελέσματα με την εφαρμογή ενός μοντέλου deep learning (HybridCNN) για ταξινόμηση ενός

βήματος. Επίσης ο συνδυασμός HybridCNN με Λογιστική Παλινδρόμηση παρουσιάζει επίσης πολύ καλά αποτελέσματα

Μία περίπου αντίστοιχη μελέτη που βασίζεται σε δύο φάσεις είναι αυτή των **Juan et al. [9].** Σε μη δημοσιευμένο σύνολο δεδομένων, κατά την πρώτη φάση χρησιμοποιείται ένα RNN, συγκεκριμένα LSTM, που αποσκοπεί στην προεπεξεργασία και στην μοντελοποίηση των γνωρισμάτων και στη συνέχεια ένας ταξινομητής Λογιστικής Παλινδρόμησης που εκτελεί την ταξινόμηση. Οι συγγραφείς χρησιμοποιούν το ίδιο σετ δεδομένων και σε απλά μοντέλα ταξινομητών (SVM και Naïve Bayes) και με διαφορετικά γνωρίσματα (1 και 2 word n-grams). Τα αποτελέσματα των πειραμάτων με μετρική σύγκρισης τη Συνολική Ακρίβεια (Accuracy) φαίνονται στον κάτωθι πίνακα:

	Αριθμός δεδομένων εκπαίδευσης		
	240	360	480
LSTM+LR	**0,887**	**0,901**	**0,910**
SVM(1gram)	0,521	0,725	0,713
SVM(2gram)	0,736	0,756	0,765
Naïve Bayes (1gram)	0,827	0,839	0,860
Naïve Bayes (2gram)	0,852	0,875	0,870

Πίνακας 13 - Αποτελέσματα Συνολικής Ακρίβειας Συνόλου Κλάσεων των Μεθόδων Juan et al. [9]

Η μελέτη των **Robinson et al. [10]** είναι εξαιρετικά ενδιαφέρουσα αναδεικνύοντας την αποτελεσματικότητα της επιλογής γνωρίσματών (feature selection), σε σχέση με την απλή χρησιμοποίηση γνωρίσματών (feature engineering).

Ο αλγόριθμός που χρησιμοποιούν είναι ο SVM. Χρησιμοποιώντας 7 dataset (κάποια αυτούσια όπως βρίσκονται στη βιβλιογραφία π.χ. Waseem and Hovy, Davidson, συνδυασμό κάποιων εξ αυτών κτλ), εκτελούν πειράματα σε SVM αλγορίθμους που χρησιμοποιούν διαφορετικό σύνολο γνωρίσματών. Παράλληλα, εισάγουν ένα νέο μοντέλο όπου τα σύνολα γνωρίσματών υφίστανται προεπεξεργασία βάσει **Λογιστικής Παλινδρόμησης** (με L1 κανονικοποίηση), έτσι ώστε να επιλεχθούν τα πιο «κατάλληλα» (βάσει ενός σκορ που τους αποδίδεται). Στη συνέχεια εφαρμόζονται αυτά τα νέα σετ γνωρίσματών σε SVM αλγορίθμους (SVM_{fs} και SVM^{+}_{fs}). Τα αποτελέσματά τους βάσει του micro F1 (και σε σχέση με CNN+GRU μοντέλο deep learning που αναπτύσουν σε άλλη εργασία τους) είναι:

Dataset	**SVM**	**SVM_{fs}**	**SVM+**	**SVM^{+}_{fs}**	**CNN+GRU**
1	0,71	0,81	0,74	**0,81**	0,82
2	0,86	0,87	0,91	**0,90**	0,92
3	0,89	0,90	0,90	**0,91**	0,92

4	0,86	0,91	0,87	**0,90**	0,93
5	0,72	0,81	0,73	**0,81**	0,82
6	0,87	0,89	0,86	**0,90**	0,94
7	0,86	0,89	0,88	**0,89**	0,92

Πίνακας 14 - Σύνοψη Αποτελεσμάτων Μεθόδων Robinson et al. [10]

Παρατηρούμε ότι η επιλογή γνωρίσματών με χρήση αλγορίθμου Λογιστικής Παλινδρόμησης συμβάλει στην βελτιστοποίηση του μοντέλου και μάλιστα με αποτελέσματα συγκρίσιμα με μοντέλο deep learning.

Στο άρθρο των **Badjatiya et al [11] ,** οι συγγραφείς χρησιμοποιούν τόσο παραδοσιακά μοντέλα ταξινόμησης, όσο και μοντέλα deep learning. Η ταξινόμηση αφορά 3 κλάσεις. Το dataset που χρησιμοποιούν είναι αυτό των Waseem and Hovy. Στα παραδοσιακά μοντέλα (Λογιστική Παλινδρόμηση, SVM, GBDT) πειραματίζονται με διάφορα γνωρίσματα (char n-grams, tfidf, Bag of Words Vectors). Επίσης πειραματίζονται με μοντέλα deep learning, όπως επίσης και με συνδυασμό των δύο κατηγοριών. Το μοντέλο που υπερέχει είναι αυτό που χρησιμοποιεί έναν συνδυασμό RNN για να δημιουργήσει συσσωματώσεις λέξεων (word embeddings) (Φάση 1) και Δέντρα απόφασης (GBDT) που χρησιμοποιούν την έξοδο της πρώτης φάσης προκειμένου να ταξινομήσουν τα tweets. Το μοντέλο αυτό επιτυγχάνει 93% F1 σκορ.

Σε άρθρο τους οι Arango et al. [12], υλοποιούν τα συγκεκριμένο πείραμα, βρίσκοντας ακριβώς το ίδιο αποτέλεσμα. Παρολαυτά, αμφισβητούν την αποτελεσματικότητα του μοντέλου, επισημαίνοντας πως το υποσύνολο του dataset που χρησιμοποιείται για την εκπαίδευση του μοντέλου, χρησιμοποιείται και για την αξιολόγησή του. Ξανατρέχοντας το ίδιο πείραμα, μόνο που κατά την αξιολόγηση του αλγορίθμου χρησιμοποιείται μόνο το υποσύνολο του dataset που αναφέρεται στο test και όχι στο train (το υποσύνολο train χρησιμοποιείται μόνο για την εκπαίδευση του αρχικού αλγορίθμου), οι συγγραφείς βρίσκουν F1=73,1% (20 μονάδες χαμηλότερο). Καθώς ο στόχος της παρούσας εργασίας είναι να περιοριστεί στη μελέτη και συζήτηση στατιστικών ταξινομητών, δεν επεκτείνεται περισσότερο η μελέτη των εργασιών που χρησιμοποιούν βαθέα τεχνητά νευρωνικά δίκτυα. Ωστόσο, σημειώνεται ότι βάσει της απόδοσης τους σε πολλές εργασίες (tasks) τα καθιστούν εργαλεία αιχμής.

3.2 Σύνοψη και Συγκριτική Συζήτηση Δημοσιευμένων Μεθόδων

Ο Πίνακας 8 συνοψίζει τις μελέτες στις οποίες εστιάζει η παρούσα επισκόπηση συν κάποιες επιπρόσθετες, συμπεριλαμβανομένων των κύριων αποτελεσμάτων τους, καθώς και των κύριων γνωρίσματών τους. Μαζί με την παραπομπή στη μελέτη, τις μεθόδους ταξινόμησης που αξιολογήθηκαν και συνοπτικά αποτελέσματα precision/recall/F1/Accuracy, για κάθε μελέτη συνοψίζονται:

- Ο τύπος συνόλου δεδομένων που χρησιμοποιήθηκαν από την εκάστοτε μελέτη για αξιολόγηση των τεχνικών ταξινόμησης, καθώς αυτός μπορεί να αποτελέσει παράμετρο που επηρεάζει την αποτελεσματικότητα μεθόδων που βασίζονται σε φυσική γλώσσα. Για παράδειγμα, ο περιορισμός 200 χαρακτήρων σε μηνύματα Twitter μπορεί να οδηγήσει στην

αλλοίωση γλωσσικών γνωρίσματών (π.χ. πλήρους σύνταξης) τα οποία χρησιμοποιούν συγκεκριμένα μοντέλα (π.χ. τεχνικές που χρησιμοποιούν word2vec),

- Η ανοικτότητα ή μη των δεδομένων που χρησιμοποιήθηκαν από την εκάστοτε μελέτη για αξιολόγηση των τεχνικών ταξινόμησης. Η πληροφορία αυτή είναι χρήσιμη για τον σκοπό της δυνατότητας επαναληψιμότητας των πειραμάτων, για παράδειγμα με σκοπό τη σύγκριση μια τεχνικής με ανταγωνιστικές τεχνικές,
- Επιπρόσθετα γνωρίσματα που δεν είναι συνήθη στην επίλυση του προβλήματος, δηλαδή πέραν των γνωρισμάτων που προκύπτουν από συνήθη γλωσσικά γνωρίσματα. Ένα τέτοιο παράδειγμα είναι η χρήση δημογραφικών γνωρίσματών (φύλο, ηλικία, γεωγραφική προέλευση), και
- Το σύνολο των τεχνικών προ-επεξεργασίας γνωρισμάτων (feature engineering) που χρησιμοποιήθηκαν από τη μελέτη για αξιολόγηση των τεχνικών ταξινόμησης. Οι τεχνικές επισκοπούνται εδώ συνοπτικά και παρουσιάζονται αναλυτικότερα στο υπόλοιπο του κεφαλαίου.

Σημειώνεται ότι τα ποσοτικά αποτελέσματα στον Πίνακα Πίνακας 8 παρουσιάζονται ενδεικτικά ως μια σύνοψη της συγκριτικής αποδοτικότητας των διαφόρων τεχνικών στη βιβλιογραφία. Ωστόσο, καθώς οι συνθήκες των πειραμάτων μεταξύ διαφορετικών μελετών διαφέρουν (π.χ. ως προς το σύνολο δεδομένων, την παραμετροποίηση των μεθόδων ταξινόμησης και την τεχνική εκμάθησης), η αυστηρή σύγκριση των μεθόδων βάσει ποσοτικών αποτελεσμάτων συστήνεται να μην αξιολογηθεί.

Συγγραφείς	**Έτος**	**Σύνολο δεδομένων**	**Επιπρόσθετα Γνωρίσματα**	**Αριθμός Κλάσεων**	**Γνωρίσματα/Τεχνικές Επεξεργασίας Γνωρισμάτων**	**Μέθοδοι ταξινόμησης**	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Accuracy (%)**
Davidson et al [2]	2017	tweets - Davidson dataset	-	3	n-grams (1-3), Tfidf , PoS, mentions, RT, URLs, αριθμός χαρακτήρων, λέξεων και συλλαβών	Λογιστική Παλινδρόμηση	91	90	90	-
Waseem, Hovy [1]	2016	tweets - Waseem+Hovy dataset	Δημογραφικά Γνωρίσματα (φύλο, γεωγραφική περιοχή), Μέγεθος Κειμένου, Πλήθος στοιχείων χρήστη	3	Αφαίρεση σημείων στίξης και λειτουργικών λέξεων (stop words). Γνωρίσματα βασισμένα σε char n-grams (n=1..4)	Λογιστική Παλινδρόμηση	72.87	77.75	73.89	-

Waseem, Hovy [1]	2016	tweets - Waseem+Hovy dataset	Δημογραφικά Γνωρίσματα (φύλο, γεωγραφική περιοχή), Μέγεθος Κειμένου, Πλήθος στοιχείων χρήστη	3	Αφαίρεση σημείων στίξης και λειτουργικών λέξεων (stop words). Γνωρίσματα βασισμένα σε word n-grams (n=1..4)	Λογιστική Παλινδρόμηση	64.39	71.93	64.58	-
Kwok et al. [6]	2013	Δεν Αναφέρεται	-	2	Word Unigrams	Naïve Bayes	-	-	-	0.76
Malmasi et al. [4]	2017	tweets - Davidson dataset + άγνωστο	-	3	char n-grams (n=4)	SVM	-	-	-	0.78
Gaydhani et al. [5]	2018	Συνένωση τριών: Davidson+ Waseem, Hovy + Malmasi	-	3	char n-grams (n=1,2,3), tfidf	Λογιστική Παλινδρόμηση	96	96	96	-
Lee et al. [7]	2018	tweets - Founta	-	4	word n-grams (n=1-3)	Naive Bayes	74.7	76.7	74.1	-
			-			Λογιστική Παλινδρόμηση	78.6	80.2	78	-
			-			SVM	77.3	77.5	73	-
			-			Random Forest	75.7	78.1	74.5	-
			-			GBT	77.2	79.4	77.3	-
			-		char n-grams (n=3-8)	Naive Bayes	75.2	75.1	74.4	-
			-			Λογιστική Παλινδρόμηση	78.8	80.4	78.3	-
			-			SVM	77.8	78.1	74	-
			-			Random Forest	76.5	78.9	76	-
			-			GBT	77	79.1	77.2	-
Park et al [9]	2017	tweets - Waseem+Hovy dataset + renewed Waseem	-	3	Char n-grams	Λογιστική Παλινδρόμηση (1 step)	82.5	82.4	81.4	-

		tweets - Waseem+Hovy dataset + renewed Waseem	-	2	Char n-grams	Λογιστική Παλινδρόμηση (2 step)	82.8	83.1	82.4	-
Badjatiya et al. [11]	2017	tweets- Waseem+Hovy dataset	-	3	Char n-grams	Λογιστική Παλινδρόμηση	72.9	77.8	75.3	-
			-		tfidf	SVM	81.6	81.6	81.6	-
			-		tfidf	GBDT	81.9	80.7	81.3	-
			-		BoW Vector (Glove)	SVM	79.1	78.8	78.9	-
			-		BoW Vector (Glove)	GBDT	80	80.2	80.1	-
			-		Random Embeding	LSTM+GBDT	93	93	93	-
Arango et al. [12]	2019	tweets - Waseem+Hovy dataset	-		Random Embeding, αλλαγη στο test set	LSTM+GBDT	81.6	68.9	73.1	-
Indurthi et al. [17]	2019	tweets - Hateval dataset	-	2	Sentence embedings- Universal Encoder	SVM (RBF kernel)	-	-	65.1	-
Yuan et al. [9]	2016	tweets - Άγνωστο	-	2	n-grams (n=2)	Naïve Bayes	-	-	-	87
			-		n-grams (n=1)	Naïve Bayes	-	-	-	86
			-		n-grams (n=2)	SVM	-	-	-	76.5
			-		n-grams (n=1)	SVM	-	-	-	71.3
			-		word embedings, 2 phases	LSTM + LR	-	-	-	91
Watanabe et al. [15]	2018	Συνένωση τριών: Davidson Waseem +Hovy, Malmasi	-	2	Sentiment based, semantic, unigram, pattern	WEKA J48graft	88	87.4	87.5	87.4
			-	3			79.3	78.4	78.4	78.4

Πίνακας 15 - Συκριτικός Πίνακας Μεθόδων και Αποτελεσμάτων Μελετών Επισκόπησης Βιβλιογραφίας

Βάσει του παραπάνω πίνακα, αλλά και την αναλυτική παρουσίαση των μελετών στην ενότητα 3.1, παρατηρούμε ότι η Λογιστική Παλινδρόμηση είναι ένα πολύ ισχυρό και ευρέως διαδεδομένο εργαλείο, όσον αφορά τους «παραδοσιακούς» ταξινομητές στη διεθνή βιβλιογραφία για την Ανίχνευση Ρητορικής Μίσους. Επίσης, δείχνει να λειτουργεί εξίσου καλά και σε ensemble μεθόδους. Πρέπει να επισημανθεί βεβαίως ότι τα διάφορα μοντέλα που αναπτύσσονται στη βιβλιογραφία δεν είναι συγκρίσιμα μεταξύ τους για διάφορους λόγους, οπότε δεν μπορεί να υπάρξει μία «αυστηρή» baseline μέθοδος σύγκρισης. Η διαφορετικότητα των σωμάτων δεδομένων, οι ιδιομορφίες που αυτά εμφανίζουν, η ποικιλομορφία τους (σε διαφορετικά άρθρα, το ίδιο σύνολο

δεδομένων εμφανίζεται να έχει διαφορετικά χαρακτηριστικά) καθιστούν την ποσοτική σύγκριση εξαιρετικά επικίνδυνη. Επίσης, η επισημείωση των tweets και ο χαρακτηρισμός τους που λαμβάνεται υπόψη στην εκπαίδευση των μοντέλων παίζουν καθοριστικό ρόλο. Σχετικό άρθρο του Waseem [14], όπου ανάλογα με το προφίλ των επισημειωτών τα αποτελέσματα των μοντέλων για το ίδιο σύνολο δεδομένων διαφέρουν. Η μελέτη των Davidson et al. [13], αναδεικνύει πως σε 7 διαφορετικά σύνολα δεδομένων, η εφαρμογή της Λογιστικής Παλινδρόμησης εμφανίζει διαφορετικά αποτελέσματα. Επίσης, διαπιστώνουμε πως και η επιλογή των γνωρισμάτων καθώς και η προεπεξεργασία αυτών ποικίλουν από μελέτη σε μελέτη, με τη μερίδα του λέοντος να μοιράζονται οι εξής τεχνικές: Bag of Words και συσσωματώσεις λέξεων (word embeddings). Επίσης όπως προκύπτει από τη βιβλιογραφία, οι μεθοδολογίες που βασίζονται σε character n-grams είναι πιο αποτελεσματικές από αυτές που βασίζονται σε word n-grams.

4 *Πειραματική Μέθοδος*

4.1 Στόχοι

Στο πειραματικό της μέρος, η παρούσα διπλωματική εργασία έχει ως στόχο να μελετήσει την απόδοση ταξινομητών ως προς την ανίχνευση μηνυμάτων Twitter που περιέχουν ρητορική μίσους. Στα πειράματα δοκιμάστηκαν οι ταξινομητές Λογιστικής Παλινδρόμησης (LR), γραμμικών Μηχανών Διανυσμάτων Υποστήριξης (SVM με linear kernel), καθώς και Δέντρων Αποφάσεων, συγκεκριμένα Random Forest, τόσο σε δυαδική ταξινόμηση (binary), όπως επίσης και σε ταξινόμηση περισσότερων κλάσεων (multiclass - στη συγκεκριμένη εργασία σε τρεις κλάσεις). Βάσει του πειραματικού περιβάλλοντος και της υλοποίησης που περιγράφονται στις επόμενες ενότητες, γίνεται ποσοτική σύγκριση των μελετηθέντων ταξινομητών.

Επιπροσθέτως, ένας από τους πειραματικούς στόχους της παρούσας εργασίας είναι να μελετήσει την επιρροή συγκεκριμένων γνωρισμάτων στην αποδοτικότητα ταξινομητών κατηγοριοποίησης σωμάτων κειμένου (tweets) ως μίσους. Μεταβάλλοντας μία ανεξάρτητη παράμετρο (γνώρισμα) τη φορά, επαναλαμβάνεται η πειραματική διαδικασία και εξάγονται αποτελέσματα επιρροής του κάθε γνωρίσματος στην αποδοτικότητα ταξινόμησης.

4.2 Σύνολα δεδομένων

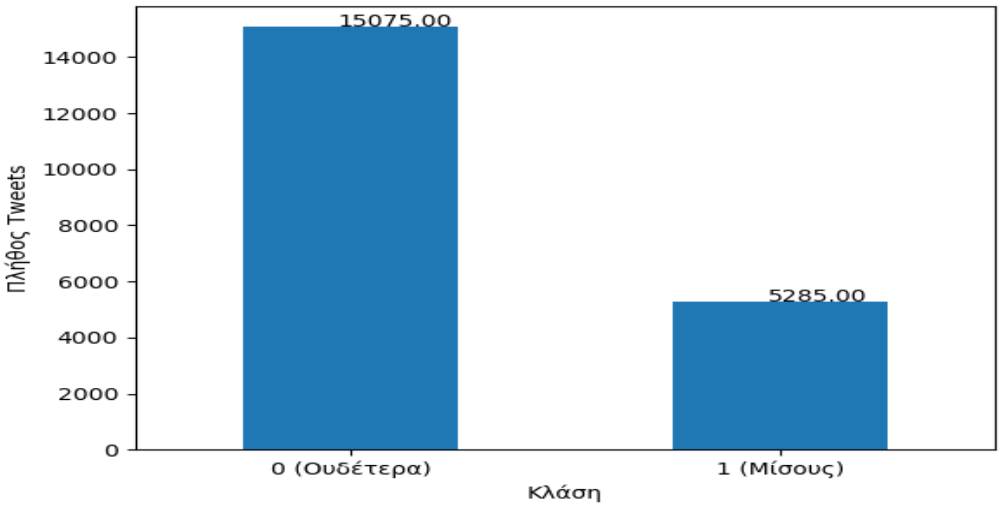
Η θεμέλιος λίθος για τη διεξαγωγή πειραμάτων κατηγοριοποίησης κειμένου είναι η επιλογή του συνόλου δεδομένων (dataset), πάνω στα οποίο ο ταξινομητής καλείται να εκπαιδευτεί και να εκτελέσει την κατηγοριοποίηση. Δυστυχώς, ο αριθμός ανοιχτών σωμάτων δεδομένων που μπορούν να αντληθούν από τη βιβλιογραφία είναι εξαιρετικά περιορισμένος, ενώ η δημιουργία νέου συνόλου δεδομένων είναι μία ακριβή διαδικασία.

Συγκεκριμένα, τα διαθέσιμα ανοιχτά σύνολα δεδομένων Twitter στην αγγλική γλώσσα είναι εξαιρετικά περιορισμένα, σε κάποια εκ των οποίων δεν υπάρχει απευθείας πρόσβαση στο κείμενο του tweet – όπως απαιτείται για τη δημιουργία γνωρισμάτων κατηγοριοποίησης κειμένου - αλλά απαιτείται η εξαγωγή του κειμένου από Twitter API, γεγονός αφενός χρονοβόρο που απαιτεί και εξουσιοδότηση από το Twitter, αφετέρου αρκετά tweets έχουν πια διαγραφεί, οπότε η εγκυρότητά των δεδομένων καθίσταται αμφισβητήσιμη. Επίσης, κατά τη μελέτη της βιβλιογραφίας σχετικά με αντίστοιχες πειραματικές διαδικασίες, διαπιστώθηκε πως κάποια σύνολα δεδομένων, έχουν αλλάξει (έχουν παραληφθεί tweets, δεν υπάρχουν πια διαθέσιμα ή έχουν συγχωνευτεί με άλλα). Δεδομένων των παραπάνω, επιλέξαμε να πειραματιστούμε δύο διαθέσιμα σύνολα δεδομένων: α) αυτό των

Davidson et al.[10], το οποίο ταξινομεί τα tweets σε 3 κλάσεις (Hate, Offensive και Neutral) και παρουσιάζει εξαιρετικά μεγάλη αξιοπιστία, αφού χρησιμοποιείται ατόφιο σε πλειάδα μελετών και σε αυτό των Goldbeck et al. [18], το οποίο είναι σχετικά καινούριο, δεν υπάρχουν πολλές μελέτες που το χρησιμοποιούν και μας απεστάλη μετά από επικοινωνία με τους συγγραφείς και ταξινομεί τα tweets σε 2 κλάσεις (Hate και Neutral).

Σωμα Δεδομένων	**Αρ. Tweets**	**Αρ. Κλάσεων**
Davidson	24783	3
Goldbeck [18]	20360	2

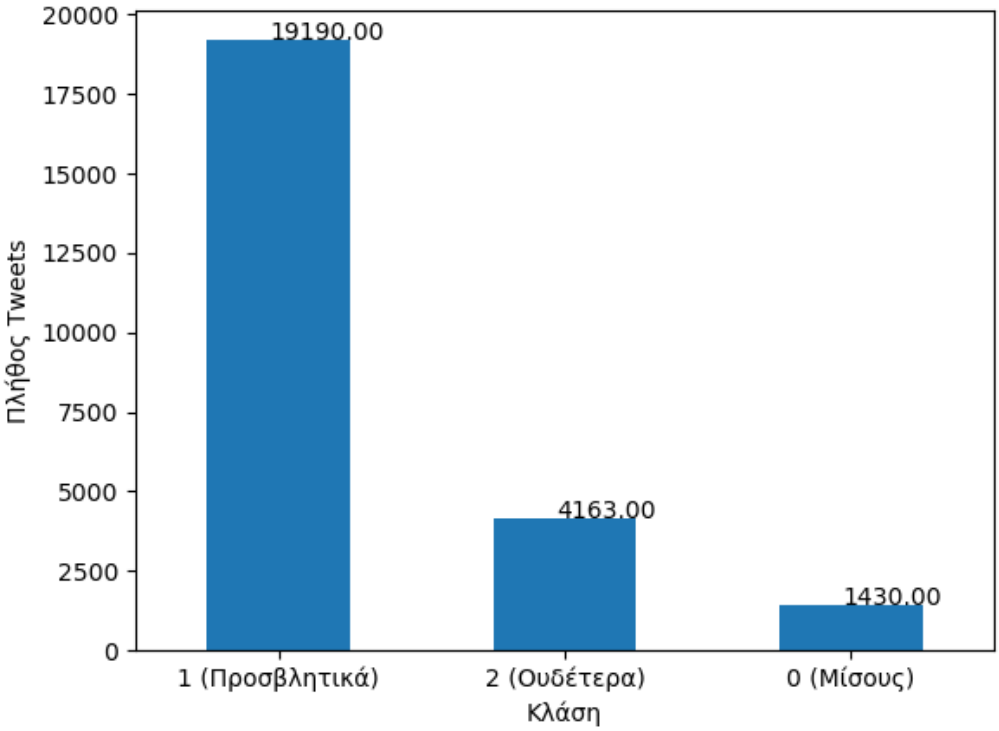
Πίνακας 16 - Χαρακτηριστικά Συνόλων Δεδομένων Πειραμάτων



Διάγραμμα 3 - Ιστόγραμμα κατανομής κλάσεων στο σύνολο δεδομένων Goldbeck

Όσον αφορά το dataset των Goldbeck et al., 15.075 (74%) tweets είναι χαρακτηρισμένα ως Ουδέτερα (Neutral) και 5.285 (26%) tweets ως Μίσους (Hate), όπως φαίνεται στο Διάγραμμα 3. Όσον αφορά το dataset των Davidson et al., 19.190 (77%) tweets είναι χαρακτηρισμένα ως Προσβλητικά (Offensive), 4.163 (17%) ως Ουδέτερα (Neutral) και 1.430 (6%) tweets ως Μίσους (Hate), όπως φαίνεται στο Διάγραμμα .

[10]https://github.com/t-davidson/hate-speech-and-offensive-language



Διάγραμμα 4 - Ιστόγραμμα κατανομής κλάσεων στο σύνολο δεδομένων Davidson

Είναι εμφανής η ανομοιογένεια των κλάσεων και στα δύο dataset, γεγονός που αναμένεται να επηρεάσει και την απόδοση των ταξινομητών.

Αρχική σκέψη ήταν να συσταθεί και ένα μεγαλύτερο σύνολο δεδομένων που να περιλαμβάνει τα tweets και από τα δύο διαφορετικά σύνολα δεδομένων, όμως αποφασίστηκε να μη γίνει πρόσμιξη όλων των δεδομένων σε ένα μεγάλο σύνολο, κυρίως λόγω διαφορετικών επισημειωτών. Οι κλάσεις στο σύνολο δεδομένων των Davidson et al. έχουν προέλθει από εξωτερικούς επισημειωτές του CrowdFlower, όπου κλήθηκαν κατόπιν εκπαίδευσης να χαρακτηρίσουν ένα σύνολο δεδομένων περίπου 25.000 tweets. Κάθε tweet χαρακτηρίστηκε από 3 ή περισσότερους επισημειωτές ως μίσους (hate) ή προσβλητικό (offensive) ή ουδέτερο (neutral) και ο χαρακτηρισμός που πλειοψήφησε του αποδόθηκε ως κλάση. Στο σύνολο δεδομένων των Goldbeck et al., τα tweets χαρακτηρίστηκαν από 2 επισημειωτές ως μίσους (hate) ή ουδέτερο (neutral). Σε περίπτωση διχογνωμίας μεταξύ των επισημειωτών επιστρατεύτηκε και ένας τρίτος ώστε να υπάρχει πλειοψηφία.

Σημειώνεται ότι καθ' όλη τη διάρκεια των πειραμάτων, λαμβάνεται σε κάθε πείραμα ένα τυχαίο δείγμα του 80% των tweets για να απαρτιστεί το σύνολο δεδομένων εκπαίδευσης (training set), ενώ το υπόλοιπο τυχαίο 20% χρησιμοποιείται σύνολο ελέγχου (test set). Αυτή η επιλογή ήταν η κυρίαρχη στη βιβλιογραφία που αναλύθηκε στην Ενότητα 3, πλειοψηφικά. Σημειώνεται ότι καθώς δεν έχει υιοθετηθεί ενιαίο πειραματικό πλαίσιο/περιβάλλον στην ερευνητική κοινότητα, η πιστή αναπαραγωγή πειραμάτων με κάθε δυνατή παραμετροποίηση που αναφέρεται στη βιβλιογραφία κατέστη αδύνατη στο χρονικό προϋπολογισμό μίας διπλωματικής εργασίας. Επομένως, επιλέχθηκε να χρησιμοποιηθεί η πειραματική μέθοδος που απαντήθηκε περισσότερο συχνά στη βιβλιογραφία που μελετήθηκε.

4.3 Αναπαράσταση Κειμένου

Αρχικά αντικαθιστούμε τις διευθύνσεις ιστοσελίδων (urls) και τις αναφορές σε χρήστες (mentions) με το σταθερό αλφαριθμητικό "URLHERE" και το σταθερό αλφαριθμητικό "MENTIONHERE" αντίστοιχα[11]. Επίσης, χρησιμοποιώντας τη βιβλιοθήκη nltk, αφαιρούμε τα γλωσσικά stopwords. Επίσης, αφαιρούνται σημεία στίξης και πολλαπλά κενά ή και σημεία στίξης μετατρέπονται σε ένα. Επιπρόσθετα αφαιρούνται και συγκεκριμένες εκφράσεις που συναντώνται σε tweets (π.χ. "ff", "#ff", "rt"). Κατά την προεπεξεργασία του συνόλου δεδομένων, όλοι οι χαρακτήρες μετατρέπονται σε πεζούς και ακολουθεί Tokenization κάθε μηνύματος (tweet). Η διαδικασία αυτή γίνεται και ως προς την εξαγωγή tokens – λέξεις από τα μηνύματα, αλλά και ως προς τη εξαγωγή tokens – λημμάτων (stems). Η Αποκατάληξη (stemming) των μηνυμάτων γίνεται με τη χρήση του εργαλείου Porterstemmer από τη βιβλιοθήκη nltk. Επίσης, κατά τη συγκεκριμένη διεργασία γίνεται και εξαγωγή των ετικετών μερών του λόγου (PoS tagging) σε κάθε μήνυμα, πάλι με χρήση της βιβλιοθήκης ntlk.

4.4 Γνωρίσματα

Βάσει της βιβλιογραφίας που παρουσιάστηκε στην Ενότητα 3, προκρίθηκε να χρησιμοποιηθούν τα κύρια γνωρίσματα που ενδείκνυνται για το συγκεκριμένο πρόβλημα μηχανικής μάθησης. Εξ' αυτών, αποκλείστηκαν γνωρίσματα λόγω έλλειψης της απαιτούμενης πληροφορίας στα διαθέσιμα σύνολα δεδομένων, όπως για παράδειγμα η γεωγραφική θέση και τα δημογραφικά στοιχεία του συγγραφέα.

Ένα από τα **γλωσσικά/σημασιολογικά γνωρίσματα** που χρησιμοποιήθηκε είναι ο δείκτης tfidf, τόσο για λέξεις/λήμματα, όσο και για τα μέρη του λόγου κάθε κειμένου (tweet στην περίπτωση μας). Στα πειράματα που εκτελούνται υπάρχει η παραμετροποιήσιμη δυνατότητα να χρησιμοποιούνται είτε n-grams λέξεων (**word n-grams**), είτε χαρακτήρων (**char n-grams**), βάσει των οποίων υπολογίζεται το tfidf. Υπάρχει η δυνατότητα το κάθε tfidf vector να έχει range (1,n) n-grams. Όσον αφορά τα word n-grams, ο υπολογισμός του tfidf μπορεί να γίνει είτε βάσει της λέξης, είτε βάσει του λήμματος (**stemming**), ανάλογα με τον tokenizer που θα χρησιμοποιηθεί. Υπολογίζονται επίσης τα tfidf επί των μερών του λόγου (POS) n-gram λέξεων (αφού έχουν περάσει την παραπάνω προεπεξεργασία κειμένου), όπου επίσης υπάρχει η δυνατότητα το κάθε tfidf vector να έχει ranges (1,n) n-grams.

Επιπλέον, τα ακόλουθα γλωσσικά και συντακτικά χαρακτηριστικά κάθε tweet (**textstats**) χρησιμοποιήθηκαν ως γνωρίσματα:

- Αριθμός συλλαβών
- Αριθμός λεκτικών χαρακτήρων
- Αριθμός χαρακτήρων (και εκτός λέξεων λεξικού, όπως π.χ. σε urls)
- Αριθμός λέξεων

[11] https://github.com/t-davidson/hate-speech-and-offensive-language

- Αριθμός λέξεων (και εκτός λέξεων λεξικού, όπως π.χ. σε urls)
- Αριθμός μοναδικών λέξεων

Επιπροσθέτως, στα πλαίσια γλωσσικών χαρακτηριστικών, χρησιμοποιήθηκαν δύο δείκτες αναγνωσιμότητας και κατανοησιμότητας κειμένου. Συγκεκριμένα, χρησιμοποιήθηκε ο δείκτης αναγνωσιμότητας Flesch-Kincaid Reading Age (FKRE), ο οποίος μετρά την ελάχιστη εκπαιδευτική βαθμίδα που απαιτείται για την ανάγνωση ενός κειμένου, καθώς και ο δείκτης Flesh-Kincaid Readability Ease (FRE), ο οποίος μετρά την ευκολία ανάγνωσης ενός κειμένου. Οι δείκτες αυτοί κωδικοποιούν μέσω μαθηματικής έκφρασης πόσο εύκολο είναι ένα κείμενο στην αγγλική γλώσσα να διαβαστεί και να κατανοηθεί, βάσει του μέσου όρου των λέξεων σε κάθε πρόταση και του μέσου όρου των συλλαβών κάθε λέξης του κειμένου.

Πέραν των αμιγώς γλωσσικών/συντακτικών χαρακτηριστικών, χρησιμοποιήθηκαν και τα εξής γνωρίσματα:

- Ο αριθμός **tweeter** αναφορών σε κάθε tweet (urls, mentions, hashtags), όπως έχουν επισημανθεί κατά την προεπεξεργασία του κειμένου, που αποτελεί επίσης γνώρισμα κατά την ταξινόμηση των δεδομένων.
- **Μετρικές ανάλυσης συναισθήματος κάθε κειμένου (tweet),** με χρήση του αναλυτή vaderSentiment (sentiment analyzer). Ο αναλυτής δίνει 4 διαφορετικά σκορ: ουδέτερο, αρνητικό, θετικό, μικτό. Το μικτό είναι ένα αποτέλεσμα που μπορεί να θεωρηθεί ότι περιλαμβάνει τα άλλα τρία, αλλά μπορούν και όλοι οι δείκτες να χρησιμοποιηθούν μαζί ως γνωρίσματα για μεγαλύτερη ακρίβεια.

4.5 Αλγόριθμοι

Όπως διαπιστώσαμε και κατά τη μελέτη της βιβλιογραφίας (Κεφάλαιο 3, Πίνακας 15), η Λογιστική Παλινδρόμηση αποτελεί το πιο σύνηθες εργαλείο κατά την ανίχνευση ρητορικής μίσους και παρουσιάζει ως επί το πλείστον τα καλύτερα αποτελέσματα. Επίσης, ένα εξίσου σύνηθες εργαλείο ταξινόμησης μηνυμάτων ρητορικής μίσους αποτελούν οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines). Αποφασίσαμε να χρησιμοποιήσουμε τους συγκεκριμένους αλγορίθμους κατά την εκτέλεση των πειραμάτων μας και επιπρόσθετα έναν ακόμη αλγόριθμο που ανήκει στην ευρεία κατηγορία των Δέντρων Απόφασης, τον Random Forest, ο οποίος είναι συνδυαστικός ταξινομητής και χρησιμοποιείται σε προβλήματα κατηγοριοποίησης και παλινδρόμησης. Επιλέχθηκε ο συγκεκριμένος αλγόριθμος, αφού συνδυάζει πολλά δέντρα απόφασης. Αυτά εκπαιδεύονται σε υποσύνολα του συνόλου δεδομένων και ανεξάρτητα το ένα με το άλλο ταξινομούν τα παραδείγματα στις κλάσεις τους. Στη συνέχεια ο αλγόριθμος ταξινομεί τα παραδείγματά στις κλάσεις τους μετά από ψηφοφορία των αποτελεσμάτων των επί μέρους δέντρων. Χαρακτηριστικό του συγκεκριμένου αλγόριθμου αποτελεί το γεγονός πως είναι εξαιρετικά ανθεκτικός στο φαινόμενο της υπερεκπαίδευσης (overfitting), φαινόμενο που είναι εξαιρετικά επιρρεπή τα δέντρα απόφασης, ιδιαίτερα σε τόσο μεγάλα σύνολα δεδομένων.

Οι παράμετροι που χρησιμοποιούνται είναι οι default παράμετροι που περιέχονται στη βιβλιοθήκη sklearn για τους παραπάνω αλγορίθμους, αφού κρίνεται πως οι συγκεκριμένοι γραμμικοί ταξινομητές εμφανίζουν εξαιρετική ανθεκτικότητα στην επιλογή παραμέτρων.

4.6 Λεπτομέρειες Υλοποίησης

Η υλοποίηση έγινε με χρήση της γλώσσας Python 3, με κριτήριο την πληθώρα υποστήριξης βιβλιοθηκών για τις ανάγκες της εργασίας. Έγινε μερική επανάχρηση και ακολούθως αναπροσαρμογή δημόσιου πηγαίου κώδικα από ανοιχτά αποθετήρια κώδικα[12] – κυρίως για προτυποποιημένη επεξεργασία γνωρισμάτων – και χρήση των βιβλιοθηκών που φαίνονται στον Πίνακας 17.

Βιβλιοθήκη	Σκοπός	URL
pandas	Χρήση για σχηματισμό και υπολογισμό σε dataframes	https://pandas.pydata.org/
numpy	Υπολογισμοί σε dataframes	https://numpy.org/
scikit-learn	Επεξεργασία συνόλου δεδομένων, εκπαίδευση και πρόβλεψη ταξινομητών	https://scikit-learn.org/
nltk	Επεξεργασία φυσικής γλώσσας (stopwords, λημματοποίηση)	https://www.nltk.org/
VADER	Ανάλυση συναισθήματος	https://github.com/cjhutto/vaderSentiment
textstat	Γλωσσικά στατιστικά κειμένου	https://pypi.org/project/textstat/
matplotlib	Οπτικοποίηση δεδομένων/αποτελεσμάτων	https://matplotlib.org/
seaborn	Οπτικοποίηση δεδομένων/αποτελεσμάτων	https://seaborn.pydata.org/

Πίνακας 17 - Βιβλιοθήκες που χρησιμοποιήθηκαν στην υλοποίηση των πειραμάτων

Επενδύθηκε επιπλέον χρόνος κατά το σχεδιασμό και την εκτέλεση της υλοποίησης με τους παρακάτω σκοπούς:

- Αναγνωσιμότητα (επανάχρηση, κατανόηση) του πηγαίου κώδικα μέσω αρθρωτής δομής του κώδικα (συναρτήσεις).
- Επέκταση/Επανάχρηση του κώδικα μέσω αρθρωτής δομής κατά τη δημιουργία και χτίσιμο των γνωρισμάτων, πριν αυτά καταναλωθούν για την εκπαίδευση ταξινομητών.
- Ευχέρεια εκτέλεσης πειραμάτων και δομής/προσπέλασης των αναφορών αποτελεσμάτων μέσω πλήρους παραμετροποίησης των γνωρισμάτων (ως παραμέτρων εισόδου) που θέλει να χρησιμοποιήσει ο χρήστης στο εκάστοτε πείραμα, καθώς και του αρχείο συνόλου δεδομένων εισόδου.

[12] https://github.com/t-davidson/hate-speech-and-offensive-language

5 Πειραματικά Αποτελέσματα

5.1 Σύγκριση Ταξινομητών

Αρχικά επιλέχθηκε να κρατηθούν όλα τα γνωρίσματα (βλέπε Ενότητα 4.4 για την περιγραφή του συνόλου των γνωρισμάτων) και να συγκριθούν οι τρεις ταξινομητές (SVM, Logistic Regression και Random Forest) μεταξύ τους με χρήση word n-grams(1-3) και με τη χρήση char n-grams(1-3). Επιχειρείται η σύγκριση των ταξινομητών ξεχωριστά για κάθε σύνολο δεδομένων. Αρχικά η σύγκριση γίνεται βάσει του μέσου (macro) F1 Score, για το λόγο ότι συνδυάζει και την Ακρίβεια και την Ανάκληση κατά τον υπολογισμό του. Επίσης, χρησιμοποιούμε την macro έκφραση του F1 Score και όχι την weighted, στην οποία αναφερθήκαμε σε προγενέστερο κεφάλαιο, επειδή και τα δύο σύνολα δεδομένων που μελετάμε εμφανίζουν πολύ μεγάλη ανομοιογένεια ως προς τις κλάσεις τους.

5.1.1 Σύνολο δεδομένων Goldbeck

Για το συγκεκριμένο σύνολο δεδομένων που αφορά δυαδική ταξινόμηση (binary classification), κάνοντας χρήση word n-grams (1-3) ως γνώρισμα σημείο αναφοράς, οι τρεις ταξινομητές παρουσίασαν τα αναλυτικά αποτελέσματα που δείχνει ο Πίνακας 18.

	Precision			**Recall**			**F1 score**		
	Neutral	**Hate**	**Macro Avg**	**Neutral**	**Hate**	**Macro Avg**	**Neutral**	**Hate**	**Macro Avg**
LR	0,82	0,37	0,59	0,70	0,52	0,61	0,76	0,43	**0,59**
SVM	0,80	0,35	0,57	0,72	0,45	0,59	0,76	0,39	0,58
RF	0,78	0,60	0,69	0,96	0,18	0,57	0,86	0,28	0,57

Πίνακας 18 - Αποτελέσματα Αξιολόγησης Ταξινομητών στο dataset Goldbeck (word n-grams(1,3))

Επιπλέον, ο Πίνακας 19 απεικονίζει τα αναλυτικά αποτελέσματα με τη χρήση char n-grams (1,3).

	Precision			**Recall**			**F1 score**		
	Neutral	**Hate**	**Macro Avg**	**Neutral**	**Hate**	**Macro Avg**	**Neutral**	**Hate**	**Macro Avg**
LR	0,82	0,38	0,60	0,71	0,54	**0,62**	0,76	0,44	**0,60**
SVM	0,80	0,35	0,57	0,72	0,45	0,59	0,76	0,39	0,58
RF	0,78	0,60	**0,69**	0,96	0,18	0,57	0,86	0,28	0,57

Πίνακας 19 - Αποτελέσματα Αξιολόγησης Ταξινομητών στο σdataset Goldbeck (char n-grams(1,3))

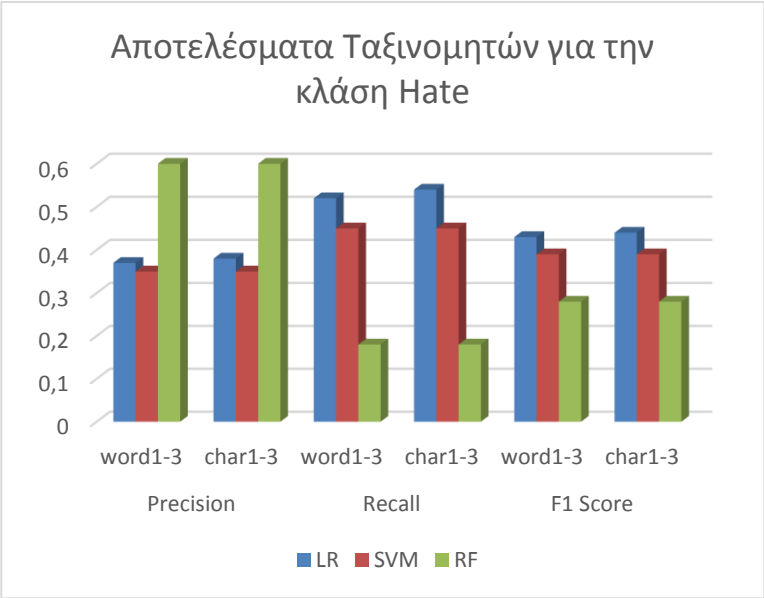
Συγκρίνοντας τους τρεις ταξινομητές βάσει του μέσου F1 Score που παρουσιάζουν, διαπιστώνουμε ότι υπερτερεί ο ταξινομητής της Λογιστικής Παλινδρόμησης.

Εστιάζοντας στην ανίχνευση ρητορικής μίσους στο σύνολο δεδομένων, το κύριο ενδιαφέρον μεταφέρεται πρωτίστως στο κατά πόσο οι ταξινομητές αποκρίνονται στην κατηγοριοποίηση των μηνυμάτων μίσους.

Ο Πίνακας 20 παρουσιάζει τα αποτελέσματα (τα οποία οπτικοποιούνται στο Διάγραμμα 5) των μετρικών που χρησιμοποιήθηκαν (Precision, Recall και F1 Score) αποκλειστικά για την κλάση Hate (η οποία και πρωτίστως μας ενδιαφέρει), τόσο με τη χρήση word- όσο και με τη χρήση char n-grams.

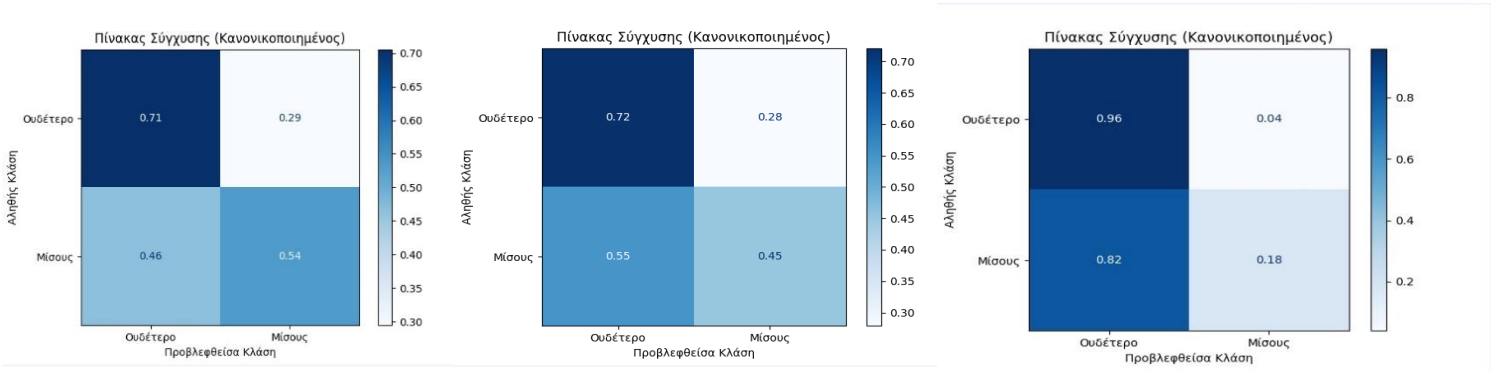
	Precision		**Recall**		**F1 Score**	
	word1-3	**char1-3**	**word1-3**	**char1-3**	**word1-3**	**char1-3**
LR	0,37	0,38	**0,52**	**0,54**	**0,43**	**0,44**
SVM	0,35	0,35	0,45	0,45	0,39	0,39
RF	0,60	0,60	0,18	0,18	0,28	0,28

Πίνακας 20 - Αποτελέσματα Αξιολόγησης Ταξινομητών στο dataset Goldbeck (word, char n-grams(1,3)) για την κλάση Hate



Διάγραμμα 5 – Απεικόνιση Αποτελεσμάτων Αξιολόγησης Ταξινομητών στο dataset Goldbeck

Διαπιστώνουμε ότι ο ταξινομητής Λογιστικής Παλινδρόμησης υπερτερεί στο F1 Score της συγκεκριμένης κλάσης. Επίσης, υπερτερεί και κατά την σωστή κατηγοριοποίηση των μηνυμάτων μίσους στην κλάση Hate, δηλαδή στην Ανάκληση (Recall) που εμφανίζει, όπως φαίνεται και στους Πίνακες Σύγχυσης (Confusion Matrices), που προκύπτουν από την εφαρμογή των μοντέλων με χρήση char 1-3 grams, στο Διάγραμμα (κάτω δεξιό τεταρτημόριο).



Διάγραμμα 6- Απεικόνιση Πινάκων Σύγχυσης για τους τρεις ταξινομητές στο dataset Goldbeck (char (1-3) n-grams). Αριστερά: Λογιστική Παλινδρόμηση, Μέσο: SVM, Δεξιά: RandomForest.

Το μοντέλο που δείχνει να υπερέχει εν τέλει είναι αυτό που χρησιμοποιεί γνωρίσματα char 1-3 n-grams επιτυγχάνοντας τα εξής αποτελέσματα σχετικά με την κατηγοριοποίηση μηνυμάτων μίσους: Precision = 0,38 , Recall = 0.54 και F1 Score = 0.44

Συγκρίναμε τα αποτελέσματα με αυτά των Davidson et al. [13] που αναφέρονται στο ίδιο σύνολο δεδομένων με χρήση Λογιστικής Παλινδρόμησης και διαπιστώσαμε πως το μοντέλο μας εμφανίζει καλύτερα αποτελέσματα από αυτό των Davidson (Precision = 0.41, Recall = 0.19, F1Score = 0.26), χωρίς όμως το μοντέλο των Davidson et al. να αναφέρεται αναλυτικά στη χρήση των n-grams που χρησιμοποιήθηκαν.

Το ίδιο σύνολο δεδομένων χρησιμοποιείται στην εργασία των Khirsagar et al. [20], όπου το μοντέλο εμφανίζει στο σύνολό του ένα weighted average F1 Score = 0,68 , ίδια ακριβώς τιμή και με το δικό μας μοντέλο. Ωστόσο, πρέπει να σημειωθεί πως κατά την απόδοση των αποτελεσμάτων μας δεν χρησιμοποιήσαμε τη συγκεκριμένη μετρική, αφού η ανομοιογένεια των δεδομένων την καθιστά μη ενδεικτική ως προς την αξιολόγηση του μοντέλου.

5.1.2 Σύνολο δεδομένων Davidson

Για το συγκεκριμένο σύνολο δεδομένων με τη χρήση word n-grams(1-3) οι τρεις ταξινομητές (LR, SVM, RF) παρουσίασαν τα αναλυτικα αποτελέσματα που παρουσιάζει ο Πίνακας 21.

	Accuracy	**Precision**				**Recall**				**F1 score**			
		Neutral	**Hate**	**Offen.**	**Macro Avg**	**Neutral**	**Hate**	**Offen.**	**Macro Avg**	**Neutral**	**Hate**	**Offen.**	**Macro Avg**
LR	0,84	0,78	0,28	0,97	0,68	0,92	0,70	0,83	0,82	0,84	0,40	0,89	**0,71**
SVM	0,88	0,82	0,32	0,94	0,70	0,87	0,43	0,91	0,74	0,85	0,37	0,93	**0,71**
RF	0,87	0,89	0,50	0,86	0,75	0,58	0,03	0,99	0,53	0,70	0,06	0,92	0,56

Πίνακας 21 - Αποτελέσματα Αξιολόγησης Ταξινομητών στο dataset Davidson (word n-grams(1,3))

Ο Πίνακας 22 απεικονίζει τα αναλυτικά αποτελέσματα με τη χρήση char n-grams (1,3).

	Accuracy	Precision				Recall				F1 score			
		Neutral	Hate	Offen.	Macro Avg	Neutral	Hate	Offen.	Macro Avg	Neutral	Hate	Offen.	Macro Avg
LR	0,84	0,77	0,27	0,96	0,67	0,90	0,61	0,84	0,78	0,83	0,37	0,90	**0,70**
SVM	0,86	0,79	0,29	0,92	0,67	0,79	0,34	0,91	0,68	0,79	0,31	0,92	**0,67**
RF	0,87	0,86	0,41	0,88	0,72	0,66	0,03	0,98	0,55	0,75	0,05	0,92	0,57

Πίνακας 22 - Αποτελέσματα Αξιολόγησης Ταξινομητών στο dataset Davidson(char n-grams(1,3))

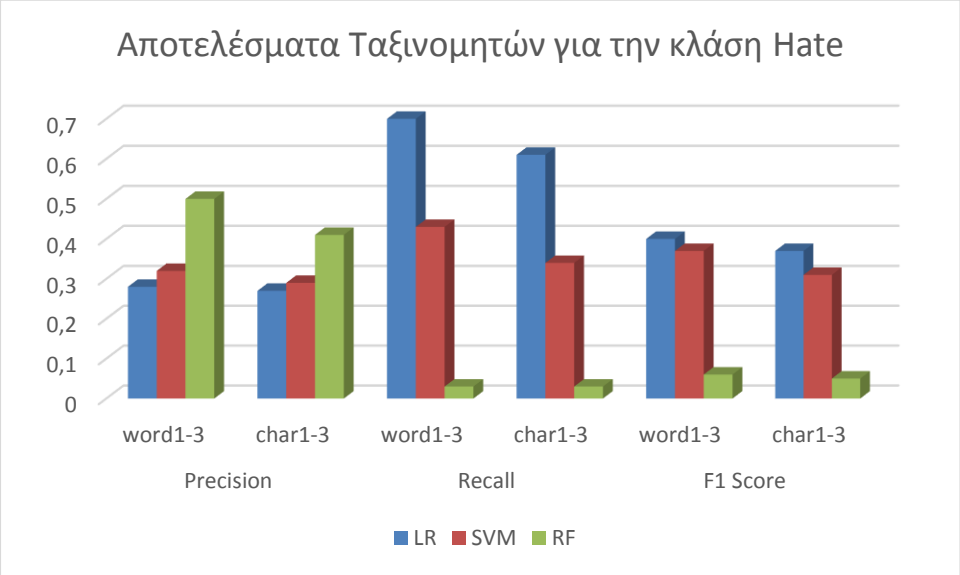
Μελετώντας τα αποτελέσματα, διακρίνουμε ότι κατά τη χρήση των word n-grams ο ταξινομητής της Λογιστικής Παλινδρόμησης εμφανίζει προσεγγιστικά παρόμοιο F1-Score με τον ταξινομητή SVM, υπερέχοντας όμως στην ανάκληση της κλάσης μίσους. Με τη χρήση char n-grams ο ταξινομητής Λογιστικής Παλινδρόμησης εμφανίζεται να υπερτερεί σε σχέση με τον SVM. Το καλύτερο μοντέλο μπορεί να θεωρηθεί ο ταξινομητής Λογιστικής Παλινδρόμησης με χρήση word (1-3) n-grams με **Precision = 0.68, Recall = 0.82 και F1 Score = 0.71.**

Στο άρθρο τους οι Davidson et al. [2] καταλήγουν επίσης ότι ο ταξινομητής Λογιστικής Παλινδρόμησης υπερτερεί σαν μοντέλο συγκριτικά με τον SVM, σημειώνοντας Precision=0,91, Recall = 0,90 και F1Score = 0,90**. Οι τιμές αυτές όμως αναφέρονται σε σταθμισμένες μέσες τιμές (weighted average), οι οποίες λαμβάνουν υπόψη τους τον αριθμό των δεδομένων στις αντίστοιχες κλάσεις. Στον ταξινομητή της παρούσας εργασίας οι αντίστοιχες τιμές (weighted avg) είναι Precision = 0,90, Recall = 0,84, F1Score = 0,86, σε μεγάλο βαθμό συγκρίσιμες με αυτές της δημοσιευμένης έρευνας. Παρόλαυτα θεωρούμε ότι στο συγκεκριμένο σύνολο δεδομένων, από τη στιγμή που οι κλάσεις εμφανίζουν μεγάλη ανομοιογένεια μεταξύ τους, δεν αποτελεί ασφαλές συμπέρασμα η σύγκριση των ταξινομητών χρησιμοποιώντας τις προαναφερθείσες μετρικές.**

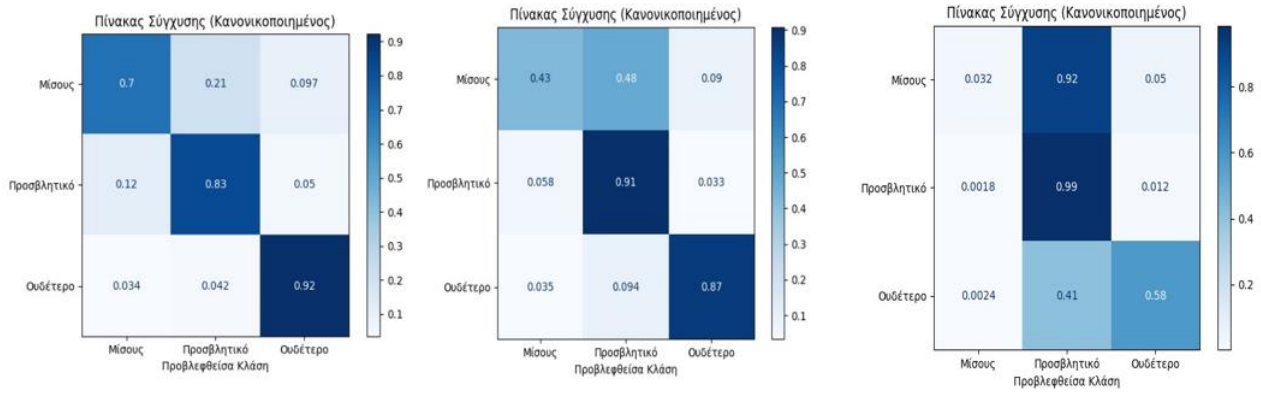
Μελετώντας τα αποτελέσματα των ταξινομητών για την κλάση Hate (Πίνακας 23, οπτικοποίηση στο Διάγραμμα)**, διαπιστώνουμε ότι ο ταξινομητής Λογιστικής Παλινδρόμησης εμφανίζει τα καλύτερα αποτελέσματα ταξινόμησης μηνυμάτων μίσους στη σωστή τους κλάση, με αρκετά μεγάλη διαφορά από τον SVM και ειδικότερα μεγαλύτερη αποτελεσματικότητα εμφανίζει το μοντέλο που χρησιμοποιεί σαν γνωρίσματα n-grams λέξεων,** γεγονός που αποτυπώνεται και στους Πίνακες Σύγχυσης των τριών ταξινομητών (Διάγραμμα).

	Precision		Recall		F1 Score	
	word1-3	**char1-3**	**word1-3**	**char1-3**	**word1-3**	**char1-3**
LR	**0,28**	0,27	**0,70**	0,61	**0,40**	0,37
SVM	0,32	0,29	0,43	0,34	0,37	0,31
RF	0,50	0,41	0,03	0,03	0,06	0,05

Πίνακας 23 - Αποτελέσματα Αξιολόγησης Ταξινομητών στο dataset Davidson (word, char n-grams(1,3)) για την κλάση Hate



Διάγραμμα 7 - Απεικόνιση Αποτελεσμάτων Αξιολόγησης Ταξινομητών στο dataset Davidson



Διάγραμμα 8- Απεικόνιση Πινάκων Σύγχυσης για τους τρεις ταξινομητές στο dataset Davidson (word 1-3 n-grams). Αριστερά: Λογιστική Παλινδρόμηση, Μέσο: SVM, Δεξιά: RandomForest

Το εύρημα αυτό (Recall = 0,70) έρχεται σε πλήρη αντιστοιχία με τα ευρήματα στην μελέτη των Davidson et. al. [2] και μάλιστα τα ξεπερνάει (Recall για Hate class = 0.61). Σημειώνεται βέβαια πως στην συγκεκριμένη μελέτη η επικύρωση των ταξινομητών έγινε με τη μέθοδο 5-fold cross validation, ενώ κατά τη διεξαγωγή των παρόντων πειραμάτων χρησιμοποιήθηκε η Hold out μέθοδος (80-20). Επίσης σε μεταγενέστερο άρθρο των ίδιων συγγραφέων [13], πάλι με εφαρμογή Λογιστικής Παλινδρόμησης στο ίδιο σύνολο δεδομένων , για την κλάση μίσους αναφέρουν τα εξής αποτελέσματα: Precision = 0.32, Recall = 0.53 και F1Score = 0.40, συγκρίσιμα με αυτά των πειραμάτων μας, με την Ανάκληση (Recall) υποδεέστερη, αλλά χωρίς να αναλύουν διεξοδικά το μοντελο n-grams που χρησιμοποίησαν.

5.2 Μελέτη επιρροής μοντέλου n-gram

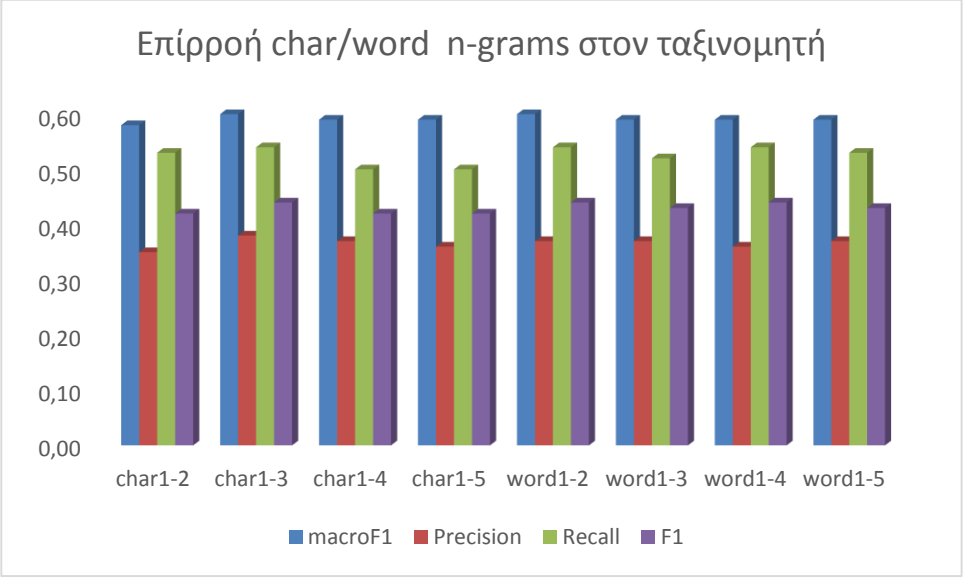
Στη συνέχεια, έχοντας προκρίνει τον ταξινομητή Λογιστικής Παλινδρόμησης ως τον πιο αποτελεσματικό σε σχέση με τους άλλους δύο, επιχειρείται να εκτιμηθεί η επιρροή που έχουν τα n-grams στην ταξινόμηση των tweets. Συγκεκριμένα, εκτελέσαμε πειράματα με χρήση n-grams λέξεων, αλλά και χαρακτήρων, και αποτυπώσαμε τα αποτελέσματα των μετρικών ως προς της κλάση που αναφέρεται σε μηνύματα μίσους.

5.2.1 Σύνολο δεδομένων Goldbeck

Εκτελώντας τα πειράματα με διάφορα n-grams, τόσο λέξεων όσο και χαρακτήρων, όπου οι τιμές του n κυμαίνονται από 1 έως 5, προκύπτουν τα αποτελέσματα που δείχνει ο Πίνακας 24 και αναφέρονται στο μέσο macro F1 του ταξινομητή και στις επιμέρους μετρικές της κλάσης Hate, τα οποία και οπτικοποιούνται στο Διάγραμμα .

		class: Hate		
	macroF1	**Precision**	**Recall**	**F1Score**
char1-2	0,58	0,35	0,53	0,42
char1-3	**0,60**	**0,38**	**0,54**	**0,44**
char1-4	0,59	0,37	0,50	0,42
char1-5	0,59	0,36	0,50	0,42
word1-2	0,60	0,37	0,54	0,44
word1-3	0,59	0,37	0,52	0,43
word1-4	0,59	0,36	0,54	0,44
word1-5	0,59	0,37	0,53	0,43

Πίνακας 24 - Αποτελέσματα αξιολόγησης επιρροής παραμετρικών n-grams (dataset Goldbeck)



Διάγραμμα 9 - Απεικόνιση αξιολόγησης επιρροής παραμετρικών n-grams (dataset Goldbeck)

Παρατηρούμε ότι το μοντέλο που χρησιμοποιεί n-grams χαρακτήρων εύρους 1-3 παρουσιάζει τα καλύτερα αποτελέσματα, άμεσα συγκρίσιμα όμως με το μοντέλο που χρησιμοποιεί n-grams λέξεων εύρους 1-2. Συμπερασματικά, το συγκεκριμένο μοντέλο δεν δείχνει να διαφοροποιείται σημαντικά είτε με τη χρήση char n-grams, είτε με τη χρήση word n-grams.

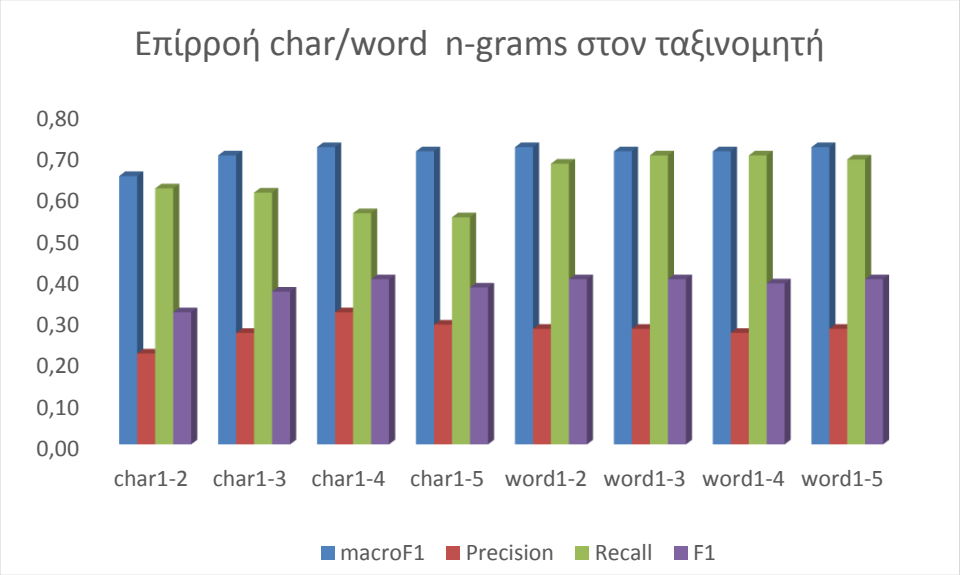
5.2.2 Σύνολο δεδομένων Davidson

Εκτελώντας τα ίδια πειράματα στο συγκεκριμένο σύνολο δεδομένων, καταλήγουμε στα αποτελέσματα που παρουσιάζει ο Πίνακας 25 και οπτικοποιεί το Διάγραμμα . Πειραματικά, διαπιστώνουμε πως η χρήση word n-grams υπερτερεί σημαντικά στην ταξινόμηση ρητορικής μίσους σε σχέση με τη χρήση char n-grams. **Σε συνδυασμό με τα ευρήματα που παρουσιάστηκαν στην ενότητα 5.2.1 για το σύνολο δεδομένων Goldbeck, τα πειραματικά αποτελέσματα**

συνιστούν ότι η χρήση word n-grams (μεγέθους 1-3) για τη δημιουργία γνωρισμάτων (tfidf) αποδίδει τα καλύτερα αποτελέσματα στην κατηγοριοποίηση μηνυμάτων ρητορικής μίσους.

		class: Hate		
	macroF1	**Precision**	**Recall**	**F1Score**
char1-2	0,65	0,22	0,62	0,32
char1-3	0,70	0,27	0,61	0,37
char1-4	0,72	0,32	0,56	0,40
char1-5	0,71	0,29	0,55	0,38
word1-2	0,72	0,28	0,68	0,40
word1-3	**0,71**	**0,28**	**0,70**	**0,40**
word1-4	0,71	0,27	0,70	0,39
word1-5	0,72	0,28	0,69	0,40

Πίνακας 25- Αποτελέσματα αξιολόγησης επιρροής παραμετρικών n-grams (dataset Davidson)



Διάγραμμα 10 - Απεικόνιση αξιολόγησης επιρροής παραμετρικών n-grams (dataset Davidson)

5.3 Μελέτη επιρροής Αποκατάληξης (Stemming)

Στο σώμα της βιβλιογραφίας που μελετήθηκε δεν συναντήσαμε πειράματα που να συγκρίνουν την επιρροή της αποκατάληξης στις αποδόσεις ταξινομητών. Παρότι χρησιμοποιείται κατά κόρον ως εργαλείο προεπεξεργασίας κειμένων, δεν συναντάται σαν επιπλέον γνώρισμα κατά την κατηγοριοποίηση κειμένου. Αποφασίσαμε να εκτελέσουμε πειράματα με και χωρίς stemming σε

διάφορα word n-grams, προκειμένου να διαπιστώσουμε αν επηρεάζει την απόδοση της στο πρόβλημα (machine learning task) της κατηγοριοποίησης μηνυμάτων μίσους.

5.3.1 Σύνολο δεδομένων Goldbeck

Αρχικά εκτελέσαμε πειράματα με χρήση word n-grams για να διαπιστώσουμε αν και σε ποιόν βαθμό η αποκατάληξη (stemming) επηρεάζει τα αποτελέσματα της ταξινόμησης.

	class: Hate		
	Precision	**Recall**	**F1Score**
word1-2 stem	**0,37**	**0,54**	**0,44**
word1-2 no stem	0,36	**0,54**	0,43
word1-3 stem	**0,37**	0,52	0,43
word1-3 no stem	0,35	0,50	0,41
word1-4 stem	0,36	**0,54**	**0,44**
word1-4 no stem	0,36	0,51	0,42

Πίνακας 26 - Απεικόνιση αξιολόγησης επιρροής αποκατάληξης των λέξεων ως γνωρίσματος (dataset Goldbeck)

Διαπιστώνουμε από τα αποτελέσματα (Πίνακας 26) ότι η χρήση του stemming ως χαρακτηριστικού κατά την ταξινόμηση επηρεάζει σε μικρό βαθμό την απόδοση του ταξινομητή, αυξάνοντας το F1 Score. Ιδιαίτερα κατά τη χρήση word n-grams (1-3) και (1-4), αυξάνεται και η ανάκληση της κλάσης Hate, αν και το εύρημα αυτό παρατηρήθηκε ομοιόμορφα στα αποτελέσματα.

5.3.2 Σύνολο δεδομένων Davidson

Κατά τη μελέτη της αποκατάληξης και σε αυτό σύνολο δεδομένων, διαπιστώσαμε πως η επιλογή της αποκατάληξης των λέξεων ως γνωρίσματος, επιφέρει στο μοντέλο μία αμυδρή βελτίωση στις μετρικές αξιολόγησής του.

	class: Hate		
	Precision	**Recall**	**F1Score**
word1-3 stem	0,28	**0,70**	**0,40**
word1-3 no stem	0,26	0,65	0,38
word1-5 stem	0,28	0,69	**0,40**
word 1-5 no stem	0,27	0,67	0,38

Πίνακας 27 - Απεικόνιση αξιολόγησης επιρροής αποκατάληξης των λέξεων ως γνωρίσματος (dataset Davidson)

Συμπερασματικά, παρατηρούμε πειραματικά ότι και για τα δύο σύνολα δεδομένων η χρήση της αποκατάληξης ως γνωρίσματος κατά την ταξινόμηση ρητορικής μίσους επιδρά στην κατηγοριοποίηση των μηνυμάτων μίσους, αυξάνοντας τα αποτελέσματα των μετρικών που τη χαρακτηρίζουν. Ο φαινομενικά μικρός βαθμός βελτίωσης που επιφέρει η χρήση stemming, ενδέχεται να οφείλεται στο γεγονός ότι οι λέξεις «μίσους», πάνω στις οποίες εκπαιδεύεται ο ταξινομητής είναι συγκεκριμένες και εξαιρετικά περιορισμένες σε έναν τόσο

μικρό αριθμό tweets. Έτσι για παράδειγμά, οι λέξεις "jewish", "jews", "jew", έχουν την ίδια ρίζα, "jew". Σίγουρα για να μελετηθεί επαρκώς η επιρροή του stemming, απαιτείται ένα μεγαλύτερο σύνολο δεδομένων, με ένα επίσης μεγαλύτερο φάσμα λέξεων που χαρακτηρίζουν τα μηνύματα ως μίσους.

5.4 Μελέτη επιρροής ανάλυσης συναισθήματος

Σαν τελευταία πειραματική διαδικασία, θελήσαμε να μελετήσουμε την επίδραση της ανάλυσης συναισθήματος στην ανίχνευση της ρητορικής μίσους. Η αρχική ιδέα επιλογής του συγκεκριμένου γνωρίσματος ως πεδίου σύγκρισης προήλθε από τη λογική ότι τα μηνύματα μίσους πιθανόν να εμφανίζουν αρνητικό συναίσθημα κατά τη διατύπωσή τους.

5.4.1 Σύνολο δεδομένων Goldbeck

Εκτελώντας πειράματα σε word n-grams διαπιστώνουμε μία ελαφρά βελτίωση του μοντέλου με τη χρήση γνωρίσματος ανάλυσης συναισθήματος, βάσει των αποτελεσμάτων που δείχνει ο Πίνακας 28.

	class: Hate		
	Precision	**Recall**	**F1Score**
word1-2 sent	0,37	0,54	0,44
word1-2 no sent	0,36	0,52	0,43
word1-3 sent	0,37	0,52	0,43
word1-3 no sent	0,36	0,54	0,43
word1-4 sent	0,36	0,54	0,44
word1-4 no sent	0,37	0,52	0,43
word1-5 sent	0,37	0,53	0,43
word 1-5 no sent	0,36	0,54	0,43

Πίνακας 28 - Απεικόνιση αξιολόγησης επιρροής ανάλυσης συναισθήματος (dataset Goldbeck)

5.4.2 Σύνολο δεδομένων Davidson

Ομοίως με το προηγούμενο σύνολο δεδομένων, διαπιστώνεται πειραματικά ότι και κατά την ταξινόμηση του συγκεκριμένου συνόλου δεδομένων (Πίνακας 29) η ανάλυση συναισθήματος ως γνωρίσματος επιφέρει μικρές διαφορές στην κατηγοριοποίηση μηνυμάτων μίσους.

	class: Hate		
	Precision	**Recall**	**F1Score**
word1-2 sent	0,28	0,68	0,40
word1-2 no sent	0,27	0,70	0,39
word1-3 sent	0,28	0,70	0,40
word1-3 no sent	0,27	0,69	0,39
word1-4 sent	0,27	0,67	0,39
word1-4 no sent	0,27	0,67	0,39
word1-5 sent	0,28	0,69	0,40
word 1-5 no sent	0,27	0,69	0,39

Πίνακας 29 - Απεικόνιση αξιολόγησης επιρροής ανάλυσης συναισθήματος (dataset Davidson)

6 *Συμπεράσματα*

Στην παρούσα διπλωματική επιχειρήθηκε καταρχάς σύγκριση παραδοσιακών ταξινομητών που κυριαρχούν στη διεθνή βιβλιογραφία ως προς την ταξινόμηση μηνυμάτων μίσους. Συγκεκριμένα, κατά τη σύγκριση των ταξινομητών σε δύο διαφορετικά σύνολα δεδομένων που έχουν αντληθεί από το Twitter, διαπιστώθηκε πως ο ταξινομητής της Λογιστικής Παλινδρόμησης εμφανίζεται πιο αποτελεσματικός σε σχέση με τις Μηχανές Διανυσμάτων Υποστήριξης καθώς και των Δένδρων Απόφασης, τόσο σε προβλήματα δυαδικής ταξινόμησης, όσο και σε προβλήματα ταξινόμησης περισσοτέρων κλάσεων, γεγονός που έρχεται σε συμφωνία με τα περισσότερα ευρήματα της βιβλιογραφίας. Επίσης, τα αποτελέσματα που κατέγραψε ο ταξινομητής Λογιστικής Παλινδρόμησης για τα σύνολα δεδομένων που εφαρμόστηκε είναι σε μεγάλο βαθμό συγκρίσιμα με τα αποτελέσματα που παρουσιάζονται σε επιστημονικά άρθρα, σε κάποιες περιπτώσεις μάλιστα υπερτερούν. Πρέπει ωστόσο να σημειωθεί πως στο σύνολο της επιστημονικής έρευνας που αφορά παρόμοια προβλήματα ταξινόμησης απουσιάζει ένα σταθερό μέτρο σύγκρισης, τόσο των μετρικών που χρησιμοποιούνται, όσο και τον μεθόδων εκπαίδευσης και επικύρωσης των ταξινομητών.

Στο παραπάνω πλαίσιο, ένα χρήσιμο εύρημα της παρούσας διπλωματικής σε σχέση με την αξιολόγηση/επιλογή βέλτιστων ταξινομητών για κατηγοριοποίηση ρητορικής μίσους είναι το εξής: ενώ η μερίδα του λέοντος της σχετικής βιβλιογραφίας επικεντρώνεται στην αξιολόγηση για όλες τις κλάσεις, η παρούσα διπλωματική καταδεικνύει ότι επιπλέον βελτιώσεις απαιτούνται συγκεκριμένα για την κλάση μίσους, η οποία είναι και κυρίου ενδιαφέροντος προς τον πλήρη αυτοματισμό σχετικών εργαλείων που μπορούν να αναπτυχθούν. Το σημαντικό αυτό εύρημα της παρούσας διπλωματικής πηγάζει από την αναλυτική μελέτη και συζήτηση της σχετικής επιστημονικής βιβλιογραφίας και επιβεβαιώνεται από τα ευρήματα των εκτενών πειραμάτων που πραγματοποιήθηκαν, σε περισσότερα του ενός σύνολα δεδομένων.

Επίσης επιχειρήθηκε η μελέτη της επιρροής κάποιων συγκεκριμένων γνωρισμάτων (n-grams, stemming και ανάλυση συναισθήματος) στην αποτελεσματικότητα του ταξινομητή που κρίθηκε καταλληλότερος για την ανίχνευση και ταξινόμηση ρητορικής μίσους. Τα ευρήματα έδειξαν πως όσον αφορά τα n-grams, η χρήση word n-grams καθίσταται πιο αποτελεσματική σε σχέση με char n-grams, αφού εμφανίζει καλύτερα αποτελέσματα κατά την ταξινόμηση, αλλά και καταναλώνει λιγότερους πόρους, μειώνοντας τον αριθμό του συνόλου των γνωρισμάτων. Ως προς τη χρήση του stemming ως επιπλέον γνωρίσματος, δεν κατέστη δυνατό να εξαχθεί ένα συμπαγές συμπέρασμα, όπως επίσης και με τη χρήση ή όχι της ανάλυσης συναισθήματος. Όσον αφορά το stemming, αποτελεί άμεση συνάρτηση με το σύνολο των δεδομένων, το φάσμα των λέξεων που χαρακτηρίζουν τα μηνύματα μίσους και το σημασιολογικό περιεχόμενο των μηνυμάτων μίσους. Η ρητορική μίσους μπορεί να χρησιμοποιεί τόσο προσβλητικές λέξεις, όσο και λέξεις που χρησιμοποιούνται στο καθημερινό μας λεξιλόγιο. Το γεγονός αυτό καθιστά την ανίχνευσή της ελλειμματική, αφού είτε

μπορεί να ταξινομηθεί ως απλά υβριστική γλώσσα, είτε ως «κανονική». Το ίδιο ισχύει και κατά τη μελέτη της επίδρασης ανάλυσης συναισθήματος για την ταξινόμησή της. Τα αποτελέσματα των πειραμάτων μας έδειξαν πως παρόλο που η ανάλυση συναισθήματος ενδέχεται να επιφέρει μία βελτίωση στο μοντέλο μας, δεν αποτελεί από μόνη της ένα ισχυρό γνώρισμα ανίχνευσης ρητορικής μίσους, αφού μηνύματα με αρνητικό συναίσθημα ενδέχεται να διατυπώνουν περιεχόμενο που δεν αφορά ρητορική μίσους. Ωστόσο φαίνεται πως αποτελεί ένα χαρακτηριστικό που σε συνδυασμό με επιπλέον γνωρίσματα σημασιολογικού περιεχομένου μπορεί να αποτελέσει εργαλείο για τη σωστή ανίχνευση και ταξινόμηση μηνυμάτων μίσους.

Η παρούσα διπλωματική μελέτησε και συνέκρινε την αποτελεσματικότητα τριών ταξινομητών σε δύο συγκεκριμένα σύνολα δεδομένων, καθώς και την επίδραση που έχουν κάποια συγκεκριμένα γνωρίσματα στην ταξινόμηση μηνυμάτων ρητορικής μίσους. Στα πλαίσια των χρονικών/υπολογιστικών πόρων της παρούσας διπλωματικής, αυτή περιορίστηκε στη μελέτη στατιστικών ταξινομητών. Βάσει της ανάλυσης βιβλιογραφίας, μελλοντικές προσπάθειες αξίζει να επεκταθούν στη μελέτη και χρήση ταξινομητών βαθέων νευρωνικών δικτύων, δεδομένης της ραγδαίας εξέλιξης που έχει επιτευχθεί στον τομέα. Κατά αντιστοιχία, συνίσταται να μελετηθεί η χρήση γνωρισμάτων που προκύπτουν με χρήση νευρωνικών δικτύων, συγκεκριμένα γνωρισμάτων συσσωματώσεων λέξεων ή/και προτάσεων (word/sentence embeddings), καθώς αυτά τα γνωρίσματα έχουν αποδειχθεί αρκετά αποδοτικά σε παρεμφερή προβλήματα μηχανικής μάθησης. Τέλος, ένα σημαντικό «κενό» που κατέδειξε η παρούσα διπλωματική ότι πρέπει να καλυφθεί είναι η ύπαρξη μιας πιο ενιαίας προσέγγισης της σχετικής επιστημονικής κοινότητας, όσον αφορά στον σχηματισμό αξιόπιστων σωμάτων δεδομένων, δεδομένης της σημαντικότητας τους στην αξιολόγηση ταξινομητών και στη δημιουργία αυτοματοποιημένων εργαλείων. Συγκεκριμένες κατευθυντήριες σε αυτή την κατεύθυνση αποτελούν: α) ενοποίηση κλάσεων που επισημειώνονται, ώστε να μπορεί να υπάρξει πειραματική επαναληψιμότητα μελετών που χρησιμοποιούν διαφορετικά σύνολα δεδομένων ή/και ενοποίηση σωμάτων δεδομένων για δημιουργία μεγαλύτερων, β) υιοθέτηση ενός κοινού και ενιαίου ορισμού ρητορικής («μίσους», «ρατσιστική», «σεξιστική», «προσβλητική», «ουδέτερη» κλπ.) που μπορεί να χρησιμοποιείται από επισημειωτές με σκοπό τη δυνατότητα άμεσης σύγκρισης αποτελεσμάτων μελετών που βασίζονται σε διαφορετικά σύνολα δεδομένων, γ) τα σύνολα δεδομένων να παρουσιάζουν μεγαλύτερη ομοιογένεια ως προς την κατανομή των υπό διερεύνηση κλάσεων.

Βιβλιογραφία

[1] Zeerak Waseem and Dirk Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter", in Proceedings of the NAACL Student Research Workshop, p. 88--93, 2016.

[2] Thomas Davidson, Dana Warmsley, Michael Macy and Ingmar Weber, "Automated Hate Speech Detection and the Problem of Offensive Language", in CoRR Journal, 2017.

[3] W.Warner and J.Hirschberg, " Detecting Hate Speech on World wide Web", Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012), pages 19–26,Montreal, Canada, June 7, 2012.

[4] S. Malmasi Shervin and M. Zampieri, "Detecting Hate Speech in Social Media", in Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP), Varna, Bulgaria, 2017.

[5] Gaydhani, V. Doma, S. Kendre and L. Bhagwat, "Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach", in IEEE International Advance Computing Conference, 2018.

[6] Irene Kwok and Yuzhou Wang, "Locate the hate: detecting tweets against blacks", in Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI'13). AAAI Press 1621-1622, 2013.

[7] Y. Lee, S.Yoon and K. Jung, "Comparative Studies of Detecting Abusive Language on Twitter", in Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Brussels, Belgium, 2018.

[8] Ji Ho Park, Pascale Fung, "One-step and Two-step Classification for Abusive Language Detection on Twitter", in Proceedings of the First Abusive Language Workshop at the Annual Meeting for the Association for Computational Linguistics, 2017.

[9] S.Juan, X.Wu, Y.Xiang, "A two phase deep learning model for identifying discrimination from tweets", 19th Inter-national Conference on Extending Database Technology (EDBT), March15-18, 2016.

[10] David Robinson, Ziqi Zhang and Jonathan Tepper, "Hate Speech Detection on Twitter: Feature Engineering v.s. Feature Selection", in Proceedings of the European Semantic Web Conference (ESWC), 2018.

[11] P.Badjatiya, S.Gupta, M.Gupta,V.Varma "Deep Learning for Hate Speech Detection in Tweets" , In Proceedings of the 26th International Conference on World Wide Web Companion, pages 759–760. International World Wide Web Conferences Steering Committee, 2017.

[12] Arango, J. Pérez, B. Poblete: “Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation”, SIGIR'19 Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019.

[13] Thomas Davidson, Debasmita Bhattacharya, Ingmar Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets", in Proceedings of the Third Abusive Language Workshop at the Annual Meeting for the Association for Computational Linguistics, 2019.

[14] Zeerak Waseem “Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter”, in Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science, pages 138–142, Austin, TX, November 5, 2016

[15] H.Watanabe, M.Bouazizi, T.Ohsuki, “Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection “, IEEE Access, Volume 6, 2018.

[16] P.Fortuna, S.Nunes, “A Survey on Automatic Detection of Hate Speech in Text”, ACM Computing Surveys, vol 51, No.4, Article 85, July 2018.

[17] V.Indurthi, B.Syed, M.Shrivastava, N. Chakravartula, M.Gupta, V. Varma : “Fermi at SemEval2019 Task 5: Using Sentence embeddings to identify Hate Speech against Immigramts and Women on Twitter” , Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 70–74, 2019.

[18] Golbeck, Jennifer, et al. "A Large Labeled Corpus for Online Harassment Research.", in Proceedings of the 2017 ACM on Web Science Conference, ACM, 2017.

[19] M. F. Porter, “An Algorithm for Suffix Stripping,” Program, Vol. 14, No. 3, 1980, pp. 130-137.

[20] R. Kshirsagar, T.Cukuvac, K.McKeown, S.McGregor, “Predictive Embeddings for Hate Speech Detection on Twitter, in Proceedings of the Second Workshop on Abusive Language Online (ALW2), pages 26-32, 2018.

[21] A.Schmidt, M. Wiegand. “A survey on hate speech detection using natural language processing.” In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pages 1–10, 2017.

[22] D. Jurafsky, J.Martin “Speech and Language Processing. An introduction to Natural Language Processing,Computational Linguistics and speech Recognition”, Third Edition draft, 2019

[23] Sokolova M., Lapalme G. “ A systematic analysis of performance measures for classification tasks” , 2009. Information Processing and Management 45, (427-437.)