ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανίχνευση διαδικτυακού εκφοβισμού με χρήση αλγορίθμων μηχανικής μάθησης

Σοφία Γ. Καλογιαννίδη

Επιβλέπουσα: **Χριστίνα Αλεξανδρή,** Καθηγήτρια

ΑΘΗΝΑ

ΦΕΒΡΟΥΑΡΙΟΣ 2023

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανίχνευση διαδικτυακού εκφοβισμού με χρήση αλγορίθμων μηχανικής μάθησης

Σοφία Γ. Καλογιαννίδη
Α.Μ.: 7115132100002

ΕΠΙΒΛΕΠΟΥΣΑ: **Χριστίνα Αλεξανδρή,** Καθηγήτρια

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Μανόλης Κουμπαράκης,** Καθηγητής

Φεβρουάριος 2023

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία αφορά την εφαρμογή και σύγκριση αλγορίθμων μηχανικής μάθησης για ανάλυση συναισθήματος με σκοπό την ανίχνευση διαδικτυακού εκφοβισμού. Η εφαρμογή των αλγορίθμων πραγματοποιείται σε δύο διαφορετικά σύνολα δεδομένων τα: SOSNet Twitter Dataset και Suspicious Tweets Dataset.

Σκοπός της εργασίας αποτέλεσε εκτός από απλή ανίχνευση του διαδικτυακού εκφοβισμού, να πραγματοποιείται περαιτέρω εύρεση του είδους του εκφοβισμού σύμφωνα με συγκεκριμένα κριτήρια όπως η ηλικία, το φύλο, η εθνικότητα κ.λ.π. Επιπρόσθετα, παρουσιάζονται τα γλωσσολογικά στοιχεία των κειμένων της εκάστοτε κατηγορίας, καθώς και αποτελέσματα άλλων ερευνών σχετικά με τη συχνότητα εμφάνισης διαδικτυακού εκφοβισμού ανάλογα με τα υπό μελέτη προσωπικά χαρακτηριστικά του ατόμου. Ως επεκτάσεις της παρούσας μελέτης τίθενται η δημιουργία ενός συστήματος το οποίο θα λαμβάνει υπόψη προσωπικούς άξονες/ κριτήρια όπως φύλο, εθνικότητα, σεξουαλικός προσανατολισμός κ.α για την ανίχνευση του διαδικτυακού εκφοβισμού. Επιπλέον, τα γλωσσολογικά χαρακτηριστικά συγκεντρώνονται ώστε να υπάρξει αναπροσαρμογή της έρευνας και σε ελληνικά δεδομένα. Η πρώτη κρούση για αυτήν την επέκταση λαμβάνει χώρα στην παρούσα μελέτη.

Τέλος, περιγράφονται αναλυτικά όλες οι μεθοδολογίες που έχουν υλοποιηθεί σε παρεμφερείς έρευνες καθώς και εκείνη που προτιμάται στην τρέχουσα. Τα αποτελέσματα όλων των αλγορίθμων σε κάθε σύνολο δεδομένων παρατίθενται και σχολιάζονται εκτενώς. Η ανίχνευση του διαδικτυακού εκφοβισμού και η σωστή κατηγοριοποίησή του γίνονται με υψηλή ακρίβεια. Ωστόσο, επισημαίνονται κάποιες μικρές αστοχίες και τίθεται ως μελλοντικός στόχος η δημιουργία νευρωνικού δικτύου για ενδεχόμενη βελτίωση αυτών των αστοχιών.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Επεξεργασία Φυσικής Γλώσσας, Μηχανική Μάθηση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Κατηγοριοποίηση κειμένου, Ανάλυση Συναισθήματος, Αλγόριθμοι Μηχανικής Μάθησης

ABSTRACT

This thesis concerns the application and comparison of machine learning algorithms for sentiment analysis in order to detect cyberbullying. The algorithms are applied to two different datasets: SOSNet Twitter Dataset and Suspicious Tweets Dataset.

The purpose of the work was, in addition to simple detection of online bullying, to further find the type of bullying according to specific criteria such as age, gender, nationality, etc. Furthermore, the linguistic elements of the texts of each category are presented, as well as the results of other researches regarding the incidence of cyberbullying according to the personal characteristics of the person under study. As extensions of the present study, the creation of a system which will take into account personal axes/criteria such as gender, nationality, sexual orientation etc. for the detection of online bullying is proposed. In addition, the linguistic features are collected so that the research can be adapted to Greek data as well. The first effort for this extension takes place in the present study.

Finally, all the methodologies that have been implemented in similar research are described in detail, as well as the one preferred in the current one. The results of all algorithms on each data set are listed and commented extensively. The detection of cyberbullying and its correct categorization are done with high accuracy. However, some minor failures are pointed out and a future goal is to create a neural network to potentially improve these failures.

SUBJECT AREA: Natural Language Processing, Machine Learning

KEYWORDS: Text Categorization, Sentiment Analysis, Machine Learning Algorithms

Στην οικογένεια μου.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια, κυρία Αλεξανδρή, για δύο λόγους. Αρχικά, για την ελευθερία που μου έδωσε να ασχοληθώ με ένα θέμα που πραγματικά με ενδιέφερε, χωρίς να θέσει κάποιον περιορισμό στον τρόπο υλοποίησης της μελέτης. Δεύτερον, για την εξαιρετικά άμεση ανταπόκριση στις απορίες μου, κάτι που δεν έχω συναντήσει συχνά στη μέχρι τώρα ακαδημαϊκή μου πορεία.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Η παρούσα Διπλωματική εκπονήθηκε στα πλαίσια του Μεταπτυχιακού Προγράμματος Σπουδών «Τεχνολογίες Πληροφορικής και Επικοινωνιών» του Τμήματος Πληροφορικής και Τηλεπικοινωνιών, της σχολής Θετικών Επιστημών του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών.

Βασικά ερεθίσματα για την επιλογή του συγκεκριμένου θέματος αποτέλεσαν συγκεκριμένα μαθήματα όπως η Υπολογιστική Γλωσσολογία, η Μηχανική Μάθηση και η Τεχνητή Νοημοσύνη, τα οποία κίνησαν το ενδιαφέρον μου κατά τη διάρκεια των σπουδών μου.

1. ΕΙΣΑΓΩΓΗ

Το αντικείμενο μελέτης της παρούσας διπλωματικής είναι η ανάλυση συναισθήματος με σκοπό την ανίχνευση διαδικτυακού εκφοβισμού. Αρχικά, ο διαδικτυακός εκφοβισμός προσεγγίζεται υπό μία θεωρητική σκοπιά και μελετάται η σύνδεση των προσωπικών χαρακτηριστικών του ατόμου όπως η ηλικία,το φύλο κλπ με την πρόκληση ή την αντιμετώπιση διαδικτυακού εκφοβισμού.

Έπειτα χρησιμοποιούνται δύο διαφορετικά σύνολα δεδομένων προκειμένου να εφαρμοστούν γνωστοί αλγόριθμοι μηχανικής μάθησης, να αξιολογηθεί η απόδοσή τους ως προς την ανίχνευση του εκφοβισμού και να πραγματοποιηθεί σύγκριση των επιδόσεων τους. Τα δύο σύνολα δεδομένων που χρησιμοποιούνται είναι τα : SOSNet Twitter Dataset και Suspicious Tweets Dataset. Πρόκειται για δεδομένα που αντλήθηκαν από την πλατφόρμα κοινωνικής δικτύωσης Twitter εξαιτίας της εύκολης πρόσβασης και αναζήτησης που προσφέρει. Ωστόσο, δεδομένα τέτοιας φύσης έχουν κάποια πολυπλοκότητα στην ανάλυση συναισθήματος καθώς η γλώσσα που χρησιμοποείται είναι αρκετά ανεπίσημη, υπάρχει έντονο το στοχείο της ειρωνίας και απαιτείται προεπεξεργασία λόγω ειδικών χαρακτήρων που εμφανίζονται (π.χ. ”@”, “#” κ.λ.π). Μία ακόμη πρόκληση της φύσης των δεδομένων είναι η ποικολομορφία τους καθώς στο SOSNet Twitter Dataset οι κλάσεις είναι ισορροπημένες σε αντίθεση με το Suspicious Tweets Dataset που υπερτερούν τα μη-εκφοβιστικά tweets. Συνεπώς, το ένα σύνολο δεδομένων θέτει το ιδανικό περιβάλλον για εφαρμογή αλγορίθμων μηχανικής μάθησης ενώ το άλλο προσομοιώνει με μεγαλύτερη ακρίβεια το περιβάλλον του διαδικτύου. Η αντικρουόμενη αυτή φύση εξασφαλίζει την αξιοπιστία της τρέχουσας έρευνας.

Ο κύριος λόγος που επιλέχθηκαν τα συγκεκριμένα σύνολα δεδομένων είναι πως δεν διαχωρίζουν απλώς τα κείμενα σε διαδικτυακό εκφοβισμό ή μη, αλλά ταξινομούν περαιτέρω το είδος του διαδικτυακού εκφοβισμού σε σχέση με συγκεκριμένα κριτήρια-άξονες. Το SOSNet Twitter Dataset διαχωρίζει το διαδικτυακό εκφοβισμό σε Ηλικιακό, Θρησκευτικό, με βάση το φύλο και με βάση την Εθνικότητα. Το Suspicious Tweets Dataset προβαίνει σε διαχωρισμό των tweets που αφορούν εκφοβισμό ως ρατσιστικά και σεξιστικά. Η επιλογή συνόλων δεδομένων που διαχωρίζουν το διαδικτυακό εκφοβισμό όσον αφορά προσωπικά χαρακτηριστικά του ατόμου (π.χ. φύλο, εθνικότητα) πραγματοποιήθηκε ως σημείο εκκίνησης για ανάλυση-μελέτη πολυπλοκότερων ζητημάτων που είναι δυσκολότερο να εντοπιστούν και να επεξεργαστούν σε βάθος με πλήρως αυτοματοποιημένες διαδικασίες (π.χ. προσωπικότητα).

Οι αλγόριθμοι που χρησιμοποιούνται για την ανίχνευση διαδικτυακού εκφοβισμού είναι οι: Naïve Bayes, KNN, SVM, Random Forest και Decision Tree. Η παρουσίαση των αποτελεσμάτων πραγματοποιείται με τη βοήθεια έκθεσης ταξινόμησης(classification report) και πίνακα σύγχυσης (confusion matrix) ενώ στο τέλος γίνεται και σύγκριση των πέντε αυτών αλγορίθμων.

Τέλος, κρίνεται σημαντικό να αναφερθεί και το κίνητρο επιλογής του διαδικτυακού εκφοβισμού ως θέματος ανάλυσης συναισθήματος. Ο διαδικτυακός εκφοβισμός μπορεί να έχει πολύ σημαντικές συνέπειες στην ψυχολογική υγεία του ατόμου. Σύμφωνα με τους Raskauskas και Stoltz [128], τα θύματα του διαδικτυακού εκφοβισμού μπορεί να αναπτύξουν κακές γενικές ψυχοκοινωνικές συνθήκες. Στην έρευνά τους διαπίστωσαν ότι το 31% των θυμάτων μαθητών ανέφεραν ότι ήταν πολύ ή εξαιρετικά αναστατωμένοι, το

19% ήταν πολύ ή εξαιρετικά φοβισμένοι και το 18% ήταν πολύ ή εξαιρετικά ντροπιασμένοι από την διαδικτυακή παρενόχληση. Διαπίστωσαν επίσης ότι οι επαναλαμβανόμενες πράξεις διαδικτυακού εκφοβισμού απειλούσαν την υγιή ανάπτυξη της αυτοεκτίμησης και συνέβαλαν σε σχολική αποτυχία, εγκατάλειψη και σε αυξημένα ψυχολογικά συμπτώματα όπως κατάθλιψη και άγχος. Ακόμη, ο διαδικτυακός εκφοβισμός έχει επιπτώσεις και στην ακαδημαϊκή/σχολική επίδοση του θύματος. Σε έρευνα του 2014 [129] επιβεβαιώνεται ότι οι μαθητές που υφίστανται εκφοβισμό στον κυβερνοχώρο συνήθως λαμβάνουν χαμηλότερους βαθμούς και διατρέχουν κίνδυνο χαμηλής ακαδημαϊκής επίδοσης. Με βάση αυτά θεωρήθηκε ότι είναι αρκετά ωφέλιμη η μελέτη και σύγκριση αλγορίθμων για την αναγνώριση του διαδικτυακού εκφοβισμού και η επιτυχής ανίχνευσή του μέσω μίας αυτοματοποιημένης διαδικασίας.

2. ΘΕΩΡΗΤΙΚΟ ΠΛΑΙΣΙΟ

Ο διαδικτυακός εκφοβισμός μπορεί να οριστεί ως «μία επιθετική και σκόπιμη πράξη που διεξάγεται από ομάδα ή άτομο, με χρήση ηλεκτρονικών μορφών επικοινωνίας, επανειλημμένα και με την πάροδο του χρόνου εναντίον ενός θύματος που δεν μπορεί εύκολα να υπερασπιστεί τον εαυτό του» (Prakash,2021) [96]:

[Cyberbullying can be defined as an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself.] Prakash, Sherly. (2021).

Ο διαδικτυακός εκφοβισμός λαμβάνει χώρα σε αρκετές πλατφόρμες όπως άμεσα μηνύματα, μηνύματα κειμένου, μέσα κοινωνικής δικτύωσης και διαδικτυακά παιχνίδια. Σύμφωνα με το statisticbrain.com, οι πιο κοινές πλατφόρμες εμφάνισης διαδικτυακού εκφοβισμού είναι τα κοινωνικά δίκτυα με πρώτο στην κατάταξη το Facebook [97].

Οι ερευνητές έχουν χρησιμοποιήσει διάφορες πηγές δεδομένων. Σε μία έρευνα που διεξήχθη το 2012 από τους Dinakar et al. (2012) [98], αξιολογήθηκαν δεδομένα από τις πλατφόρμες YouTube και Formspring, χρησιμοποιώντας το 50% από το σύνολο δεδομένων του YouTube ως σύνολο εκπαίδευσης, το 20% ως σύνολο επαλήθευσης και το υπόλοιπο 30% ως σύνολο ελέγχου. Σε άλλες έρευνες από τους Van Hee et al. (2012) [99,100] εξήχθησαν δεδομένα από το Ask.fm (βασισμένα σε ερωτήσεις-απαντήσεις) χρησιμοποιώντας το λογισμικό GNU Wget. Επιπλεόν, σε έρευνα του 2015 έρευνες από τους Hosseinmardi et al. (2015) [101] συλλέχθηκαν 316.500 δεδομένα από το Instagram συμπεριλαμβανομένων εικόνων και σχολίων από 25.000 χρήστες. Επίσης, σε έρευνα από τους Zhao et al. (2016) [102] χρησιμοποιήθηκε σαν πηγή δεδομένων το Twitter.

Όσον αφορά τα χαρακτηριστικά που χρησιμοποιούνται στις έρευνες για ανίχνευση διαδικτυακού εκφοβισμού, σύμφωνα με την πηγή Salawu et al. (2020) [103] διακρίνονται σε 4 κατηγορίες:

1) Χαρακτηριστικά που βασίζονται στο περιεχόμενο δηλαδή λεξικολογικά στοιχεία που εξάγονται από το έγγραφο όπως λέξεις-κλειδιά, αντωνυμίες και σημεία στίξης

2) Χαρακτηριστικά βασισμένα σε συναισθήματα, δηλαδή, εκείνα που υποδεικνύουν συγκινησιακά περιεχόμενα όπως λέξεις-κλειδιά, φράσεις και σύμβολα που καθορίζουν το συναίσθημα στο κείμενο

3) Χαρακτηριστικά βασισμένα στο χρήστη όπως ηλικία, φύλο και σεξουαλικός προσανατολισμός και

4) Χαρακτηριστικά βασισμένα στο δίκτυο, όπως ο αριθμός των φίλων, ο αριθμός των ακολούθων, η συχνότητα ανάρτησης περιεχομένου κλπ.

Συχνά γίνεται συνδυασμός των κατηγοριών χαρακτηριστικών κατά τη διεξαγωγή της έρευνας. Για παράδειγμα, στις έρευνες από τους Van Hee et al. (2012) [99,100] πραγματοποιήθηκε συνδυασμός χαρακτηριστικών που βασίζονται στο περιεχόμενο και χαρακτηριστικών που βασίζονται στο συναίσθημα. Στην προσέγγιση που παρουσιάστηκε από τους Hosseinmardi et al. (2015) [101] συνδυάστηκαν χαρακτηριστικά βασισμένα στο περιεχόμενο και χαρακτηριστικά βασισμένα στο δίκτυο ενώ στην προσέγγιση των Zhao et al. (2016) [102] και στην προσέγγιση των Dinakar et al. (2012) [98] χρησιμοποιήθηκαν χαρακτηριστικά βασισμένα στο περιεχόμενο. Για πάραδειγμα οι Zhao et al. (2016) εξήγαγαν χαρακτηριστικά εκφοβισμού (bullying features) βασισμένα στο περιεχόμενο ακολουθώντας την παρακάτω μεθοδολογία:

«Επειδή ορισμένα μηνύματα διαδικτυακού εκφοβισμού συνήθως περιέχουν κατάρα ή υβριστικές λέξεις, αυτές οι λέξεις είναι καλές ενδείξεις της ύπαρξης εκφοβισμού. Επομένως, επιλέγουμε μια λίστα με προσβολές-λέξεις με βάση τις προηγούμενες

γνώσεις μας και κάποιους εξωτερικούς γλωσσικούς πόρους (http://www.noswearing.com/dictionary). Αυτή η λίστα περιέχει 350 λέξεις που υποδεικνύουν- κατάρα ή αρνητικά συναισθήματα, όπως nigga, bitch, fuck, slut, twat κλπ. Στη συνέχεια, συγκρίνουμε τη λίστα λέξεων με τα χαρακτηριστικά μονογράμματα και τα διγράμματα του σώματος κειμένων που χρησιμοποιούνται για να ληφθεί η τομή, η οποία θεωρείται ως insulting seeds(προσβλητικοί σπόροι) σε αυτό το έγγραφο. Για παράδειγμα, η οπτικοποίηση προσβλητικών σπόρων του συνόλου δεδομένων twitter που χρησιμοποιείται στην πειραματική μελέτη φαίνεται στο σχήμα» (Zhao et al., 2016)

Εικόνα 1 Προσβλητικό λεξιλόγιο στη μέθοδο των (Zhao et al., 2016) για εξαγωγή χαρακτηριστικών εκφοβισμού(bullying features) βασισμένα στο περιεχόμενο.

«Αυτές οι προσβλητικές λέξεις χρησιμοποιούνται για την κατασκευή ενός λεξιλογίου, και κάθε κείμενο του συνόλου δεδομένων αντιστοιχίζεται σε ένα διάνυσμα μετρώντας τις φορές εμφάνισης κάθε λέξης που εμφανίζεται. Επεκτείνουμε τους προκαθορισμένους προσβλητικούς σπόρους με βάση τα word embeddings. Για κάθε προσβλητικό σπόρο επιλέγουμε τις h κορυφαίες περισσότερες παρόμοιες λέξεις στο λεξιλόγιο ως εκτεταμένα χαρακτηριστικά εκφοβισμού.» (Zhao et al., 2016) Με αυτόν τον τρόπο, δηλαδή σε δύο βήματα, εξάγονται τελικώς τα χαρακτηριστικά εκφοβισμού. Για παράδειγμα, για τον προσβλητικό σπόρο slut εξάγονται οι ακόλουθες δέκα εκφράσεις-όροι ως πιο παρόμοιες επεκτείνοντας το αρχικό λεξιλόγιο με βάση το περιεχόμενο των κειμένων.

Similar Words	Cosine Similarity Scores
a slut	0.815
whore	0.738
a whore	0.638
hypocrite	0.536
bitch	0.508
puta	0.468
nerd	0.455
bully her	0.451
fat bully	0.440
bully nigga	0.435

Εικόνα 2 Δέκα πιο κοντινοί όροι-εκφράσεις στον προσβλητικό σπόρο slut (Zhao et al., 2016)

Για μια σειρά ζητημάτων που σχετίζονται με την αναγνώριση διαδικτυακού εκφοβισμού, έχει γίνει έρευνα με βάση την εξόρυξη κειμένου όπως, για παράδειγμα, η διαδικτυακή σεξουαλική παρενόχληση [104] και η ανίχνευση ανεπιθύμητων μηνυμάτων [105].

Ειδικότερα, προτάθηκε από τους Yin et al. (2009) [106] μια προσέγγιση εποπτευόμενης μάθησης για τον προσδιορισμό παρενοχλητικών αναρτήσεων σε δωμάτια συνομιλίας και συζητήσεων φόρουμ.

Συγκεκριμένα, τρεις τύποι χαρακτηριστικών, δηλαδή με βάση (α) το περιεχόμενο, (β) το συναίσθημα και (γ) τα συμφραζόμενα χρησιμοποιήθηκαν για την εκπαίδευση ενός SVM ταξινομητή. Επιπλέον, χρησιμοποιήθηκαν n-grams και TFIDF (Συχνότητα όρων – Αντίστροφη Συχνότητα εγγράφου). Αν και τα αποτελέσματα της έρευνας ήταν καλά, δεν χρησιμοποιήθηκε καμία πληροφορία χρήστη και γινόταν αποκλειστική χρήση εποπτευόμενων μεθόδων.

Αρχικά, οι Yin et al. (2009) συνέκριναν την απόδοση των μεθόδων n-grams και TF-IDF. Τα αποτελέσματα που συγκεντρώθηκαν είναι τα παρακάτω:

	Kongregate	Slashdot	MySpace
Precision	0.139	0.179	0.110
Recall	0.140	0.117	0.354
F-measure	0.140	0.141	0.168

	Kongregate	Slashdot	MySpace
Precision	0.289	0.273	0.351
Recall	0.571	0.231	0.217
F-measure	0.384	0.250	0.268

Εικόνα 3 Αριστερά: Αποτελέσματα n-grams στα τρία σύνολα δεδομένων.Δεξιά: Αποτελέσματα TF-IDF στα τρία σύνολα δεδομένων (Yin et al., 2009)

«Τα αποτελέσματα της προσέγγισης TF-IDF είναι καλύτερα από αυτήν των n-grams. Όλες οι μετρήσεις είναι πάνω από 20% και η καλύτερη, ανάκληση (Recall) για το σύνολο δεδομένων Kongregate, φτάνει το 57%. Το TFIDF είναι πολύ περισσότερο αποτελεσματικό ως προσέγγιση από άλλες βασικές μεθόδους για την ανίχνευση της παρενόχλησης. Ωστόσο, η απόδοση του TFIDF απέχει ακόμα πολύ από τις προσδοκίες μας.» (Yin et al., 2009)

Προκειμένου να βελτιώσουν περαιτέρω την απόδοση, οι συγγραφείς (Yin et al., 2009) συνδύασαν τη μέθοδο TF-IDF (περιεχομένου) με χαρακτηριστικά συναισθήματος και συμφραζομένων.

«Η χαμηλή απόδοση των απλών μεθόδων δείχνει ότι απαιτούνται περισσότερο πολύπλοκες μέθοδοι για τον εντοπισμό της παρενόχλησης. Η πειραματική μας προσέγγιση περιλαμβάνει τη χρήση τοπικών χαρακτηριστικών, χαρακτηριστικών συναισθημάτων και συμφραζόμενα χαρακτηριστικά. Για τα συμφραζόμενα χαρακτηριστικά, ορίσαμε την παράμετρο μεγέθους παραθύρου σε k = 3. Για τα χαρακτηριστικά συναισθήματος, το λεξικό είναι το ίδιο με το λεξικό βωμολοχίας που χρησιμοποιείται στο προγενέστερα γλωσσικά μας πειράματα.» (Yin et al., 2009)

Όπως φαίνεται στην παρακάτω εικόνα, ο συνδυασμός των χαρακτηριστικών TF-IDF με χαρακτηριστικά συναισθήματος και συμφραζομένων, βελτίωσε τα αποτελέσματα και στα τρία σύνολα δεδομένων.

	Kongregate	Slashdot	MySpace
Precision	0.352	0.321	0.417
Recall	0.595	0.277	0.250
F-measure	0.442	0.298	0.313

Εικόνα 4 Αποτελέσματα συνδυασμού χαρακτηριστικών TF-IDF με χαρακτηριστικά συναισθήματος και συμφραζομένων στα τρία σύνολα δεδομένων. (Yin et al., 2009)

Γενικά, σε αρκετές έρευνες, μεταξύ των οποίων και οι προαναφερόμενες έρευνες [99,100,102], έχει επιλεγεί ως αλγόριθμος εποπτευόμενης μάθησης ο SVM (Support Vector Machine) διότι έχει φανεί πως είναι κατάλληλος για ταξινόμηση κειμένου υψηλής λοξότητας, όπως για τον εντοπισμό διαδικτυακού εκφοβισμού, χρησιμοποιώντας χαρακτηριστικά που βασίζονται σε περιεχόμενο, όπως επισημαίνουν οι (Desmet & Hoste, 2014) [108].

«Συνολικά, ο SVM χειρίζεται καλύτερα τη συμπερίληψη κακών χαρακτηριστικών. Τα διγράμματα λέξεων, για παράδειγμα, υποβαθμίζουν την απόδοση σε όλες τις δοκιμές και έχουν τον μικρότερο αντίκτυπο στον SVM, πιθανότατα λόγω της εγγενούς επιλογής χαρακτηριστικών του» (Desmet & Hoste, 2014).

Παρ 'όλα αυτά, οι μέθοδοι χωρίς επίβλεψη μπορούν επίσης να αποδειχθούν πολύτιμες. Ειδικότερα, στην μελέτη των (CHISHOLM, 2006), επιχειρήθηκε με το ίδιο σύνολο να προσδιοριστούν clusters (συστάδες) που περιέχουν διαδικτυακό εκφοβισμό, χρησιμοποιώντας έναν αλγόριθμο που βασίζεται σε κανόνες [107].

Όσον αφορά τον εντοπισμό του διαδικτυακού εκφοβισμού, μεταξύ των σχολίων του YouTube, χαρακτηριστικό παράδειγμα αποτελεί η προσέγγιση των Dinakar et al. (2021) [109], που περιέγραψαν μια μέθοδο εντοπισμού και επεξεργασίας. Στη μέθοδό τους, χρησιμοποίησαν μια ποικιλία δυαδικών και πολλαπλών κατηγοριών ταξινομητή σε ένα σύνολο δεδομένων του οποίου οι επισημάνσεις προέκυψαν με χειροκίνητο τρόπο. Επίσης εφάρμοσαν τη γνώση της κοινής λογικής για τον εντοπισμό διαδικτυακού εκφοβισμού. Η χρήση κοινής λογικής μπορεί να βοηθήσει στην παροχή πληροφοριών για τις γνώσεις και τα συναισθήματα των ανθρώπων. Χρησιμοποίησαν ακόμη δύο (2)τύπους χαρακτηριστικών: 1) γενικά χαρακτηριστικά που περιείχαν σταθμισμένα μονογράμματα με TFIDF και 2) ειδικά επισημασμένα χαρακτηριστικά. Η μελέτη τους έδειξε ότι ο δυαδικός ταξινομητής μπορεί να ξεπεράσει σε ακρίβεια την αναγνώριση κειμενικού διαδικτυακού εκφοβισμού σε σύγκριση με ταξινομητές πολλαπλών τάξεων. Οι περιορισμοί της μελέτης τους είναι ότι δεν έλαβαν υπόψη τους τον πραγματισμό του διαλόγου και της συνομιλίας και το κοινωνικό γράφημα δικτύωσης.

	Naïve Bayes		Rule-based Jrip		Tree-based J48		SMO (SVM)	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Sexuality	66%	0.657	**80.20%**	0.598	63.40%	0.573	66.70%	**0.79**
Race	66%	0.789	**68.30%**	0.789	63.50%	0.657	66.70%	**0.718**
Intelligence	72%	0.467	**70.39%**	0.512	70%	0.568	72%	**0.7723**

Εικόνα 5 Δυαδικοί ταξινομητές για ανεξάρτητες ετικέτες (Dinakar et al., 2021)

Mixture	63%	0.445	63%	0.507	61%	0.456	66.70%	0.653

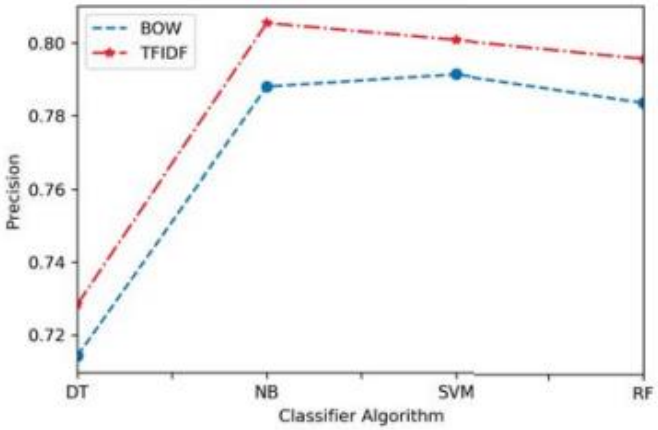
Εικόνα 6 Ταξινομητές πολλαπλών κλάσεων για το συγχωνευμένο σύνολο δεδομένων(Dinakar et al., 2021)

Όπως φαίνεται και από τα αποτελέσματα παραπάνω «οι δυαδικοί ταξινομητές που έχουν εκπαιδευτεί για μεμονωμένες ετικέτες είναι πολύ καλύτεροι από τους ταξινομητές πολλαπλών κλάσεων που έχουν εκπαιδευτεί για όλες τις ετικέτες» (Dinakar et al., 2021).

Επιπλέον, αξίζει να αναφερθεί η μελέτη των Nahar et al. (2013) [110] το 2013, η οποία πρότεινε μια αποτελεσματική προσέγγιση εντοπισμού διαδικτυακού εκφοβισμού από τα μέσα κοινωνικής δικτύωσης. Επιπλέον, στην μελέτη αυτή παρουσίαστηκε ένα μοντέλο γραφήματος για την εξαγωγή του διαδικτυακού εκφοβισμού. Αυτό οδήγησε στον εντοπισμό των πιο ενεργών θυτών και θυμάτων μέσω αλγορίθμου κατάταξης. Το προτεινόμενο μοντέλο γραφήματος θα μπορούσε να χρησιμοποιηθεί για την αναγνώριση του επιπέδου θυματοποίησης διαδικτυακού εκφοβισμού και για λήψη αποφάσεων σε μετέπειτα μελέτες.

Στις περισσότερες μελέτες για ανίχνευση cyberbullying, χρησιμοποιείται επεξεργασία φυσικής γλώσσας - NLP (Natural Language Processing) για την προεπεξεργασία των κειμένων και, στην συνέχεια, αλγόριθμοι μηχανικής μάθησης- ML (Machine Learning) [99,100,102,111,112,109,116,117] για την κατηγοριοποίησή τους. Σαν μέθοδοι για τη δημιουργία feature vectors (διανύσματα χαρακτηριστικών), οι επικρατέστερες είναι η TFIDF (Term Frequency – Inverse Document Frequency) και BoW (Bag of Words). Σύμφωνα με την έρευνα των Islam et al. (2020) [111] η μέθοδος TFIDF είχε καλύτερα αποτελέσματα. Στην ίδια έρευνα συγκρίθηκαν οι αλγόριθμοι ML SVM, Decision Tree και Random Forest, με τα αποτελέσματα να αναδεικνύουν τον SVM ακριβέστερο όπως ειπώθηκε και παραπάνω.

Ενδεικτικά, στην παρακάτω εικόνα βλέπουμε πως η ακρίβεια και των 4 αλγορίθμων μηχανικής μάθησης υπό εξέταση στην έρευνα των (Islam et al., 2020) ήταν ανώτερη στην περίπτωση χρήσης TF-IDF από ότι BoW.



Εικόνα 7 Σύγκριση μεθόδων TF-IDF και BoW ωσ προς την ακρίβεια για τους 4 αλγορίθμους (Islam et al., 2020)

Στη συνέχεια της ίδιας έρευνας, με τη μέθοδο TF-IDF δημιουργήθηκε ένα ROC γράφημα (γράφημα για τη σύγκριση αλγορίθμων ταξινόμησης). Στο γράφημα αυτό φάνηκε καθαρά πως ο SVM απέδωσε καλύτερα συγκριτικά με τους υπόλοιπους αλγορίθμους.

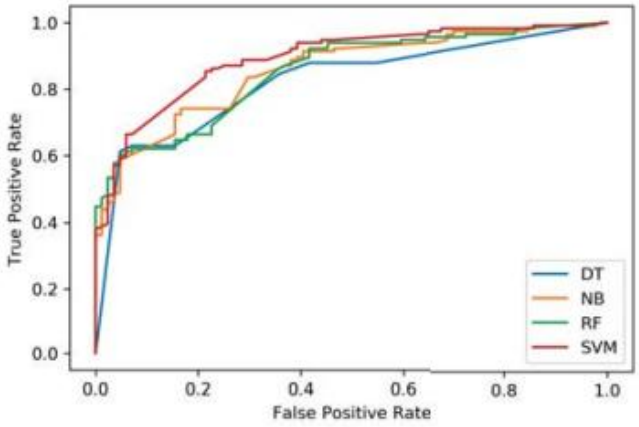


Fig. 5. ROC curve for TF-IDF

Εικόνα 8 Γράφημα ROC των 4 αλγορίθμων μηχανικής μάθησης με τη μέθοδο TF-IDF(Islam et al., 2020)

Εκτός από τους καθιερωμένους αλγορίθμους Machine Learning που αναφέρθηκαν παραπάνω, σε αρκετές έρευνες εξετάζεται η χρήση Deep Learning (Βαθιά Μάθηση) για την αναγνώριση διαδικτυακού εκφοβισμού - cyberbullying. Πρόσφατα μοντέλα βασισμένα σε βαθύ νευρωνικό δίκτυο (DNN) έχουν επίσης εφαρμοστεί για την ανίχνευση του διαδικτυακού εκφοβισμού, όπως παρουσίαζεται από τους Agrawal & Awekar, (2018) και Zhang et al. (2016) [114,115]. Μάλιστα, στην προσέγγιση των Agrawal & Awekar, (2018) [114], χρησιμοποιούνται μοντέλα DNN για την ανίχνευση του διαδικτυακού εκφοβισμού

και έχουν επεκτείνει τα μοντέλα τους σε πολλές πλατφόρμες μέσων κοινωνικής δικτύωσης. Με βάση τα αναφερόμενα αποτελέσματα, τα μοντέλα της προσέγγισης Agrawal & Awekar, (2018) υπερτερούν των παραδοσιακών μοντέλων ML. Εδώ αξίζει σημειωθεί ότι στην εν λόγω μελέτη δηλώθηκε ότι εφαρμόστηκε η μάθηση μεταφοράς (transfer learning), στοιχείο που σημαίνει πως τα μοντέλα που έχουν αναπτύξει για τον εντοπισμό του διαδικτυακού εκφοβισμού μπορούν να προσαρμοστούν και να χρησιμοποιηθούν και σε άλλα σύνολα δεδομένων.

Πιο συγκεκριμένα, όσον αφορά την τεχνική μάθησης μεταφοράς, δοκιμάστηκαν 3 διαφορετικές προσεγγίσεις στην έρευνα των Agrawal & Awekar, 2018.
1. Πλήρης Εκμάθηση Μεταβίβασης (TL1): Σε αυτή τη μορφή, ένα μοντέλο εκπαιδευμένο σε ένα σύνολο δεδομένων χρησιμοποιήθηκε απευθείας για τον εντοπισμό του διαδικτυακού εκφοβισμού σε άλλα σύνολα δεδομένων χωρίς οποιαδήποτε επιπλέον εκπαίδευση. Το TL1 οδήγησε σε σημαντικά χαμηλή ανάκληση υποδεικνύοντας ότι τα τρία σύνολα δεδομένων έχουν διαφορετική φύση διαδικτυακού εκφοβισμού με χαμηλή επικάλυψη.
2. Εκμάθηση Μεταφοράς Επιπέδου Χαρακτηριστικών (TL2): Σε αυτή τη μορφή, εκπαιδεύτηκε ένα μοντέλο σε ένα σύνολο δεδομένων και μόνο οι μαθημένες ενσωματώσεις λέξεων (word embeddings) μεταφέρθηκαν σε άλλο σύνολο δεδομένων για την εκπαίδευση ενός νέου μοντέλου. Σε σύγκριση με το TL1, η βαθμολογία ανάκλησης βελτιώθηκε δραματικά με το TL2.
3. Εκμάθηση μεταφοράς επιπέδου μοντέλου (TL3): Σε αυτή τη μορφή, ένα μοντέλο που εκπαιδεύτηκε σε ένα σύνολο δεδομένων και έμαθε ενσωματώσεις λέξεων, καθώς και βάρη δικτύου, μεταφέρεται σε άλλο σύνολο δεδομένων για εκπαίδευση ενός νέου μοντέλου. Από το TL3 δεν προκύπτει οποιαδήποτε σημαντική βελτίωση σε σχέση με το TL2. Αυτή η έλλειψη βελτίωσης δείχνει ότι η μεταφορά βαρών δικτύου δεν είναι απαραίτητη για τον εντοπισμό διαδικτυακού εκφοβισμού και οι ενσωματώσεις λέξεων που έχουν μάθει είναι η βασική γνώση που αποκτάται από τα μοντέλα DNN.

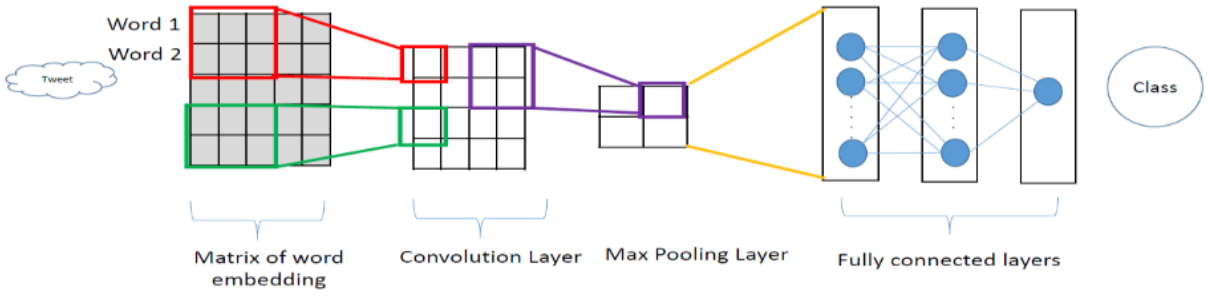
Τα αποτελέσματα των τριών μεθόδων μεταφοράς μάθησης φαίνονται και στην παρακάτω εικόνα :

Metric	Train / Test	F+			T+			W+		
		TL1	TL2	TL3	TL1	TL2	TL3	TL1	TL2	TL3
Precision	F	-	-	-	0.38	0.90	0.88	0.51	0.92	0.85
	T	0.83	0.88	0.90	-	-	-	0.72	0.91	0.90
	W	0.82	0.92	0.91	0.68	0.90	0.91	-	-	-
Recall	F	-	-	-	0.04	0.98	0.98	0.66	0.98	0.99
	T	0.01	0.99	0.99	-	-	-	0.17	0.98	0.99
	W	0.21	0.96	0.96	0.05	0.97	0.96	-	-	-
F1-score	F	-	-	-	0.07	0.95	0.93	0.58	0.95	0.92
	T	0.03	0.93	0.94	-	-	-	0.28	0.94	0.94
	W	0.35	0.94	0.94	0.10	0.94	0.94	-	-	-

Εικόνα 9 Σύγκριση μεθόδων μεταφοράς μάθησης ως προς την ακρίβεια (Agrawal & Awekar, 2018).

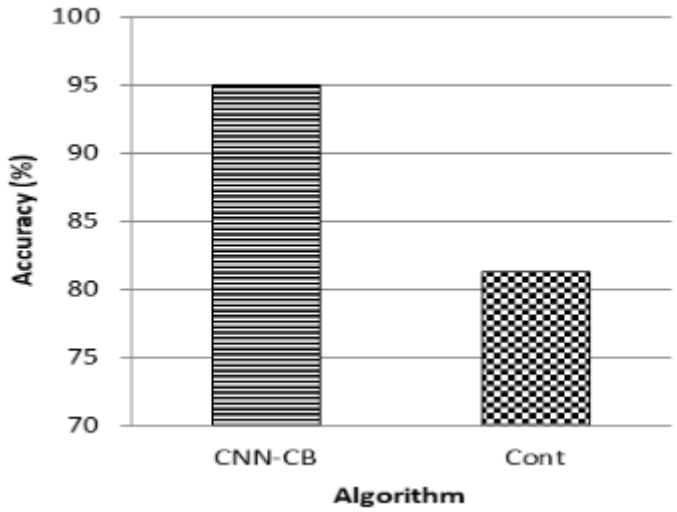
Τέλος, αναφέρεται η προσέγγιση των Al-Ajlan & Ykhlef, 2018 [113], όπου χρησιμοποιούνται word embeddings για την αριθμητική αναπαράσταση των κειμένων,που είναι μία εναλλακτική μέθοδος αντί για TFIDF και BoW, και έπειτα δημιουργείται ένα νευρωνικό δίκτυο CNN (Συνελικτικό Νευρωνικό Δίκτυο) με τη βοήθεια του Keras (διεπαφή προγραμματισμού εφαρμογών βαθιάς μάθησης γραμμένη σε Python), το οποίο συγκρινόμενο με τον SVM αποφέρει καλύτερα αποτελέσματα.

Η αρχιτεκτονική του μοντέλου που χρησιμοποίησαν φαίνεται στην παρακάτω εικόνα:

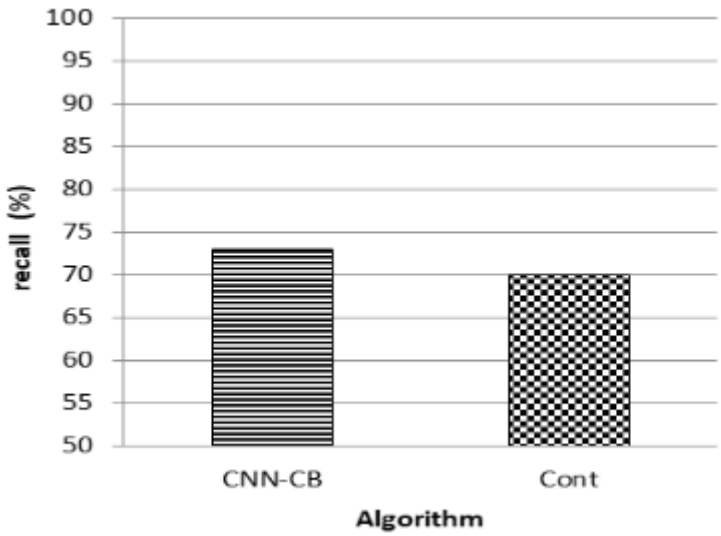


Εικόνα 10 Αρχιτεκτονική CNN μοντέλου των (Al-Ajlan & Ykhlef, 2018)

Το μοντέλο συγκρίθηκε με τον αλγόριθμο SVM ως προς τις μετρικές της ακρίβειας (accuracy) και της ανάκλησης (recall). Τα αποτελέσματα που προέκυψαν φαίνονται στα παρακάτω διαγράμματα:



Εικόνα 11 CNN-CB και Cont SVM ακρίβεια (Al-Ajlan & Ykhlef, 2018)



Εικόνα 12 CNN-CB και Cont SVM ανάκληση (Al-Ajlan & Ykhlef, 2018)

Από τα διαγράμματα είναι εμφανής η υπεροχή του CNN έναντι του αλγορίθμου SVM.

Στην παρούσα ανάλυση και υλοποίηση θα χρησιμοποιηθεί η προσέγγιση των αλγορίθμων μηχανικής μάθησης και όχι βαθιάς μηχανικής μάθησης. Συγκεκριμένα,θα εφαρμοστούν οι αλγόριθμοι Naïve Bayes, Random Forest, Decision Tree, SVM και KNN και θα συγκριθούν ως προς την απόδοση στην ανίχνευση πολυταξικού διαδικτυακού εκφοβισμού.

Όσον αφορά τα σύνολα δεδομένων, στην παρούσα ανάλυση και υλοποίηση θα χρησιμοποιηθούν δύο σύνολα δεδομένων, το «SOSNet Twitter Dataset» και το «Suspicious Tweets Dataset», τα οποία παρουσιάζονται στη συνέχεια.

3. ΔΕΔΟΜΕΝΑ

Τα σύνολα δεδομένων που χρησιμοποιήθηκαν στην παρούσα ανάλυση και υλοποίηση είναι, όπως προαναφέρθηκε, το «SOSNet Twitter Dataset» και το «Suspicious Tweets Dataset». Η επιλογή αυτών των δύο συνόλων έγινε βάσει ευκολίας πρόσβασης σε αυτά και διότι προσέφεραν πολυταξική ταξινόμηση του διαδικτυακού εκφοβισμού γεγονός που αποτελούσε ζητούμενο. Επιπλέον, το SOSNet Twitter Dataset έχει ισορροπημένες κλάσεις, σε αντίθεση με το Suspicious Tweets Dataset στο οποίο τα μη-εκφοβιστικά tweets υπερτερούν. Συνεπώς, η φύση τους είναι διαφορετική και ο συνδυασμός τους κρίνεται ωφέλιμος για την ασφαλέστερη διεξαγωγή συμπερασμάτων ως προς την απόδοση των αλγορίθμων. Στην ανάλυση και επεξεργασία του περιεχομένου των συνόλων «SOSNet Twitter Dataset» και «Suspicious Tweets Dataset», απαραίτητη είναι η παρουσίαση και περιγραφή γλωσσολογικών στοιχείων των κειμένων, καθότι περιέχουν πληροφορίες καθοριστικής σημασίας για την εφαρμογή των αλγορίθμων και την ανάλυση συναισθήματος.

3.1 SOSNet Twitter Dataset

Καθώς η χρήση των μέσων κοινωνικής δικτύωσης γίνεται ολοένα και πιο διαδεδομένη σε κάθε ηλικιακή ομάδα, η συντριπτική πλειοψηφία των πολιτών βασίζεται σε αυτό το ουσιαστικό μέσο για την καθημερινή επικοινωνία. Η πανταχού παρουσία των μέσων κοινωνικής δικτύωσης σημαίνει ότι ο διαδικτυακός εκφοβισμός μπορεί να επηρεάσει αποτελεσματικά οποιονδήποτε ανά πάσα στιγμή και οπουδήποτε, και η σχετική ανωνυμία του Διαδικτύου καθιστά πιο δύσκολο να σταματήσουν τέτοιες προσωπικές επιθέσεις σε σύγκριση με τον παραδοσιακό εκφοβισμό.

Στις 15 Απριλίου 2020, η UNICEF εξέδωσε μια προειδοποίηση ως απάντηση στον αυξημένο κίνδυνο διαδικτυακού εκφοβισμού κατά τη διάρκεια της πανδημίας COVID-19 λόγω του εκτεταμένου κλεισίματος των σχολείων, του αυξημένου χρόνου οθόνης και της μειωμένης κοινωνικής αλληλεπίδρασης πρόσωπο με πρόσωπο. Τα στατιστικά στοιχεία του διαδικτυακού εκφοβισμού είναι εντελώς ανησυχητικά: το 36,5% των μαθητών γυμνασίου και λυκείου έχουν αισθανθεί διαδικτυακό εκφοβισμό και το 87% έχουν παρατηρήσει διαδικτυακό εκφοβισμό, με επιπτώσεις που κυμαίνονται από μειωμένη ακαδημαϊκή επίδοση έως κατάθλιψη με αυτοκτονικές σκέψεις.

Υπό το πρίσμα όλων αυτών, το πρώτο σύνολο δεδομένων το οποίο χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία, αντλήθηκε από το κοινωνικό δίκτυο του Twitter. Πιο συγκεκριμένα, πρόκειται για 47.692 κείμενα χρηστών του Twitter τα οποία έχουν ταξινομηθεί σε κάποια κατηγορία διαδικτυακού εκφοβισμού ή δεν έχουν συνδεθεί καθόλου με διαδικτυακό εκφοβισμό. Το SOSNet Twitter Dataset ανακτήθηκε από τον παρακάτω σύνδεσμο:

https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification

3.1.1 Τρόπος δημιουργίας SOSNet Twitter Dataset

Το σύνολο δεδομένων-dataset παρήχθησε το 2020 από τους J. Wang, K. Fu και C.T. Lu [118]. Για τη δημιουργία του τελικού dataset του οποίου η δομή περιγράφεται αναλυτικά σε επόμενες ενότητες χρειάστηκαν άλλα 6 datasets από άλλες έρευνες για το cyberbullying [119,114,120,121,122,123,124]. Στην παρακάτω εικόνα παρουσιάζεται μία σύνοψη δεδομένων (dataset summary) στην οποία φαίνεται η συνεισφορά των 6 συνόλων δεδομένων στο τελικό που χρησιμοποιείται στην παρούσα διπλωματική.

TABLE I: Dataset Summary

Name	Total	CB	Not CB	Age	Ethnicity	Gender	Religion	Other
Agrawal [30]	16050	5963	10087	0	13	2841	1922	187
Brettschneider [33]	4475	183	4292	49	29	34	0	71
Chatzakou [29]	1500	1278	222	17	100	115	10	1036
Davidson [31]	205	181	24	5	27	121	1	27
Waseem [34], [35]	12899	8900	3999	0	86	3339	0	5475
WISC [32]	4095	1078	3024	94	17	39	0	921
Collective Before DQE	39224	17583	21648	165	272	6489	1933	7717
Collective After DQE	69767	50468	19299	10010	12730	10277	9367	8084

Εικόνα 13 Κατανομή των 6 επιμέρους συνόλων δεδομένων-datasets για τη δημιουργία του SOSNet Twitter Dataset (Wang et. al, 2020)

Τα σύνολα δεδομένων Chatzakou, Waseem,Bretschneider και WISC παρείχαν μόνο τα αναγνωριστικά των tweets (Tweet IDs), επομένως οι συγγραφείς [118] χρησιμοποίησαν το API του Twitter για να ανακτήσουν το περιεχόμενο κειμένου των tweets. Δεδομένου ότι πολλά tweets έχουν αφαιρεθεί ή καταργηθεί από τη δημοσίευση αυτών των συνόλων δεδομένων, επιτεύχθηκε η ανάκτηση του 45,6%, 41,8%, 54,9% και 51,4% των tweets από αυτά τα αντίστοιχα σύνολα δεδομένων. Επειδή η έρευνά επικεντρώθηκε στη λεπτομερή ταξινόμηση των tweets που προκαλούν εκφοβισμό στον κυβερνοχώρο, συνεχίστηκε η ταξινόμηση των περιπτώσεων διαδικτυακού εκφοβισμού αυτών των έξι συνόλων δεδομένων χειροκίνητα και ομαδοποιημένα tweets της ίδιας κατηγορίας (λόγω περιορισμένου χρόνου και ανθρώπινου δυναμικού, μόνο τα πρώτα 1500 tweets από το σύνολο δεδομένων Chatzakou και τα πρώτα 4475 tweets από τα δεδομένα του Davidson) επισημάνθηκαν περαιτέρω και χρησιμοποιήθηκαν. Η κύρια συνεισφορά της έρευνας [118] ήταν η χρήση μιας τροποποιημένης επέκτασης δυναμικού ερωτήματος, για να αυξηθεί ο αριθμός των δειγμάτων κάθε τάξης με ημιεποπτευόμενο τρόπο.

3.1.2 Χρησιμοποιούμενη γλώσσα SOSNet Twitter Dataset

Προκειμένου να υπάρχει μία σαφής εικόνα της προέλευσης των tweets, είναι χρήσιμο να εξεταστούν τα 6 επιμέρους datasets που απαρτίζουν το τελικό.

Το dataset Waseem [122] αρχικά δημιουργήθηκε με τη βοήθεια της αναζήτησης-search του Twitter API σε συγκεκριμένα hashtags όπως #MKR που περιείχαν σεξιστικά tweets. Εν συνεχεία, από το σώμα κειμένων αφαιρέθηκαν όσα δεν ήταν στην Αγγλική γλώσσα. Γι' αυτό, όπως φαίνεται και από τη σύνοψη δεδομένων-dataset summary που δόθηκε παραπάνω, η συντριπτική πλειοψηφία των tweets από αυτό το dataset κατηγοριοποίηθηκαν από τους συγγραφείς [118] ως gender cyberbullying.

Το Davidson dataset [120] αρχικά ξεκίνησε με ένα λεξικό ρητορικής μίσους, το οποίο περιείχε λέξεις και φράσεις προσδιοριζόμενες ως ρητορική μίσους από το Hatebase.org. Χρησιμοποιώντας το Twitter API οι συγγραφείς αναζήτησαν για tweets που περιέχουν όρους από το λεξικό, με αποτέλεσμα δείγμα tweets από 33.458 χρήστες Twitter. Εξήγαγαν το χρονοδιάγραμμα για κάθε χρήστη, με αποτέλεσμα ένα σύνολο 85,4 εκατομμυρίων tweets. Από αυτό το σώμα, στη συνέχεια, πήραν ένα τυχαίο δείγμα 25 χιλιάδων tweets που περιείχαν όρους από το λεξικό και κωδικοποίηθηκαν χειροκίνητα από τους εργαζόμενους του CrowdFlower (CF).

Το Agrawal dataset [114] είναι ένα αρκετά ιδιαίτερο dataset γιατί προέρχεται από πολλαπλές πλατφόρμες. Συγκεκριμένα περιέχει 12K ζεύγη ερωταπαντήσεων από την πλατφόρμα FormSpring, 16K tweets από το Twitter (δανεισμένα από το Waseem [122]) και 100K σχόλια από τη Wikipedia.

Το Bretschneider dataset [121] δημιουργήθηκε από tweets που συλλέχθηκαν μεταξύ της 20-10-2012 και 30-12-2012. Το WISC dataset [124] περιέχει tweets από το δημόσια διαθέσιμο 2011 TREC Microblog (16 εκατομμύρια tweets δειγματοληπτημένα μεταξύ της

23ης Ιανουαρίου και της 8ης Φεβρουαρίου 2011). Τα tweets αυτά δεν είναι μόνο στα Αγγλικά.
Τέλος, όσον αφορά το Chatzakou dataset [119] συλλέχθηκαν tweets από διαφορετικά topics με την κατανομή να φαίνεται αναλυτικά στην παρακάτω εικόνα:

Table 1. An Overview of the Datasets Used with Respect to Number of Tweets, Users Involved, Period of Collection, and Where in the Article Each Dataset Was Used

	period	tweets	#users	size (cleaned)	users (cleaned)	reference sections
Baseline	June–August 2016	1M	610k	70%	73%	3, 4, 5, 6, 7
Gamergate	June–August 2016	600k	312k	69%	58%	3, 4, 5, 6, 7
NBA	July 2017	400k	202k	57%	66%	3, 7.1, 7.2
BBC gender pay	July 2017	100k	64k	69%	75%	3, 7.1, 7.2

Εικόνα 14 Κατανομή tweets στο σύνολο δεδομένων Chatzakou (Chatzakoy et al,2019)

Επομένως, με βάση τις παραπάνω λεπτομέρειες για τα 6 datasets που συνθέτουν αυτό που θα χρησιμοποιηθεί για τη σύγκριση των αλγορίθμων, συμπεραίνεται πως:

1. Η γλώσσα που χρησιμοποιείται είναι η **Αγγλική**. Επειδή τα δεδομένα έχουν αντληθεί από πλατφόρμα κοινωνικής δικτύωσης, η γλώσσα που χρησιμοποιείται είναι αργκό.
2. Τα δεδομένα προέρχονται από εκατομμύρια χρήστες του Twitter, από πολλά διεθνή hashtags που αφορούν ριάλιτι-reality shows, NBA κ.α. Συνεπώς οι χρήστες δεν είναι αποκλειστικά κάποιας συγκεκριμένης εθνικότητας αλλά **Αγγλόφωνοι** από διάφορα μέρη.

3.1.3 Κατηγορίες Διαδικτυακού Εκφοβισμού SOSNet Twitter Dataset

Όπως προαναφέρθηκε, το πρώτο σύνολο δεδομένων στο οποίο θα εφαρμοστούν οι αλγόριθμοι στο πλαίσιο της εργασίας αυτής δεν προβαίνει σε έναν απλό δυαδικό διαχωρισμό των tweets στις κατηγορίες διαδικτυακός εκφοβισμός (cyberbullying)/όχι διαδικτυακός εκφοβισμός (not cyberbullying). Η ταξινόμηση που πραγματοποιείται είναι πολυταξική. Οι κατηγορίες Διαδικτυακού Εκφοβισμού στις οποίες κατανέμονται τα κείμενα είναι οι ακόλουθες:

1. Ηλικιακός διαδικτυακός εκφοβισμός -Age Cyberbullying
2. Διαδικτυακός εκφοβισμός εθνικότητας-Ethnicity Cyberbullying
3. Διαδικτυακός εκφοβισμός φύλου-Gender Cyberbullying
4. Θρησκευτικός διαδικτυακός εκφοβισμός-Religion Cyberbullying
5. Άλλος τύπος διαδικτυακού εκφοβισμού-Other type of cyberbullying
6. Όχι διαδικτυακός εκφοβισμός -Not cyberbullying

Στα πλαίσια της εν λόγω έρευνας, όπως θα περιγραφεί και σε επόμενη ενότητα, θα αγνοηθεί η κατηγορία 5 και θα αφαιρεθούν όλα τα κείμενα της κατηγορίας αυτής. Συνεπώς, τα μοντέλα και οι αλγόριθμοι ταξινόμησης που θα εφαρμοστούν, θα κατατάσσουν τα κείμενα σε μία από τις εναπομείνασες 5 κατηγορίες.

Σημαντικό στοιχείο αποτελεί και η κατανομή των κειμένων ανά κατηγορία. Τα δεδομένα έχουν εξισορροπηθεί ώστε να περιέχουν ~8000 κείμενα από κάθε κατηγορία. Συγκεκριμένα, η κατανομή των κειμένων του συνόλου δεδομένων ανά κατηγορία, παρουσιάζεται αναλυτικά στον παρακάτω πίνακα:

Πίνακας 1 Κατηγορίες Διαδικτυακού Εκφοβισμού στο SOSNet Twitter Dataset

Κατηγορία Εκφοβισμού	Πλήθος κειμένων
Ηλικία (Age)	7.992
Εθνικότητα (Ethnicity)	7.961
Φύλο (Gender)	7.973
Θρησκεία (Religion)	7.998
Όχι διαδικτυακός εκφοβισμός (Not cyberbullying)	7.945

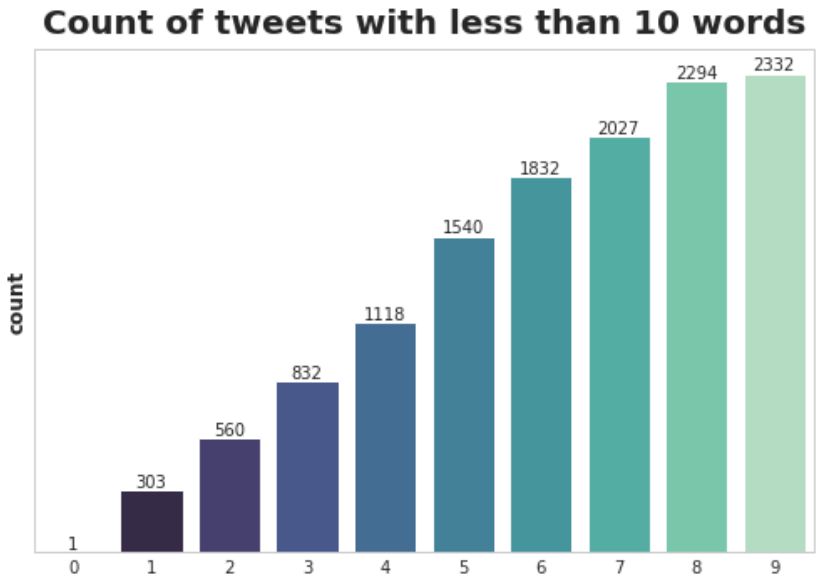
3.1.4 Γλωσσολογικά Στοιχεία Κειμένων SOSNet Twitter Dataset

Σε αυτήν την ενότητα θα περιγραφούν αναλυτικά τα γλωσσολογικά στοιχεία των κειμένων τόσο συνολικά όσο και ειδικότερα ανά κατηγορία διαδικτυακού εκφοβισμού. Συγκεκριμένα, θα γίνει λεξιλογική ανάλυση σχετικά με τις συχνότερα χρησιμοποιούμενες λέξεις στο σώμα των κειμένων, μελέτη του πλήθους των λέξεων που συναντώνται στα tweets καθώς και παρουσίαση συννεφόλεξου για κάθε κατηγορία.

3.1.4.1 Μήκος κειμένων SOSNet Twitter Dataset

Ένα βασικό στοιχείο των κειμένων προς χρήση είναι το μήκος τους σε λέξεις. Είναι ιδιαίτερα σημαντικό να εξεταστεί εάν υπάρχουν tweets με πολύ χαμηλό ή πολύ υψηλό αριθμό λέξεων καθώς και πόσα είναι αυτά ώστε να αξιολογήσουμε μετέπειτα αν είναι εφικτό να θέσουμε άνω και κάτω όρια στο πλήθος των λέξεων ενός κειμένου προκειμένου να το συμπεριλάβουμε τελικά στη φάση της εκπαίδευσης.

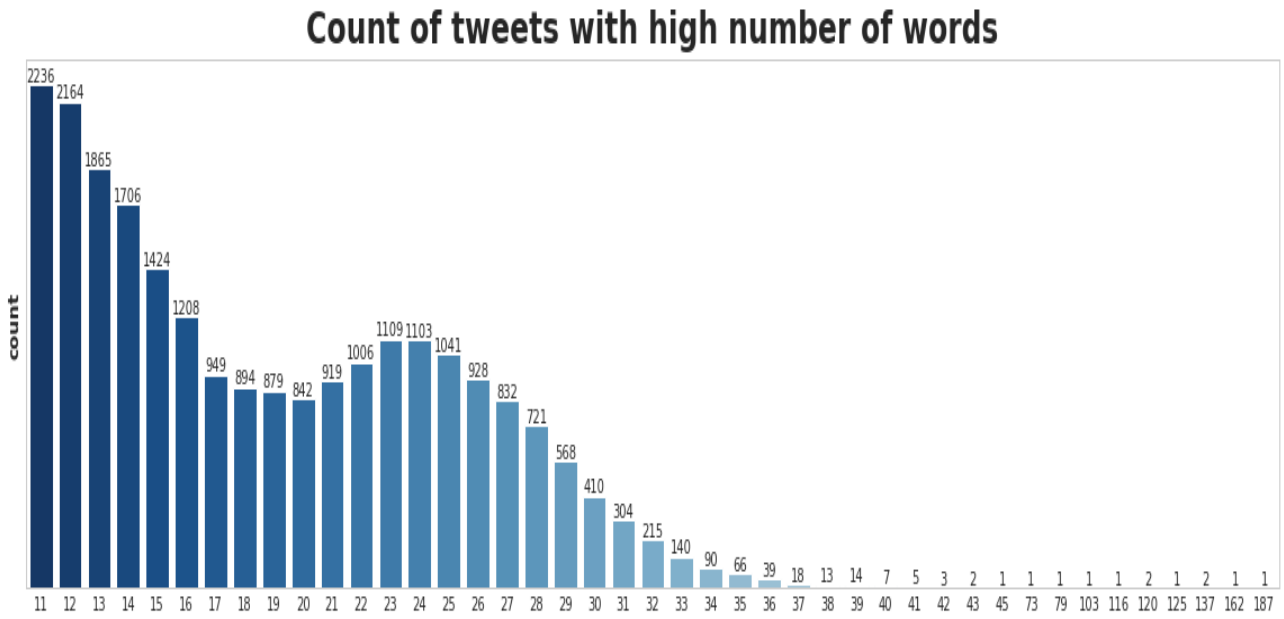
Αρχικά, ένας λογικός αριθμός για κάτω φράγμα είναι οι 10 λέξεις. Με τον κατάλληλο κώδικα, ανακτήθηκε ραβδόγραμμα του πλήθους των κειμένων (κατακόρυφος άξονας) με ακριβώς i λέξεις όπου i=0,1….,10 (οριζόντιος άξονας).



Εικόνα 15 Πλήθος κειμένων στο SOSNet Twitter Dataset με 0 έως 9 λέξεις

Αρχικά, από το ραβδόγραμμα φαίνεται πως υπάρχει μόνο 1 κείμενο με 0 λέξεις. Η καταμέτρηση των λέξεων έγινε μετά από σχετική προεπεξεργασία που αναλύεται σε επόμενη ενότητα. Συνεπώς, εξηγείται λογικά η ύπαρξη προεπεξεργασμένου κειμένου χωρίς λέξεις εφόσον το αρχικό περιείχε μόνο σημεία στίξης, emojis ή απλά κάποια αναφορά (χαρακτήρας @). Η επόμενη αξιοσημείωτη παρατήρηση είναι πως υπάρχουν αρκετά κείμενα με μικρό αριθμό λέξεων. Εάν επιλεχθεί πράγματι ο αριθμός 10 ως κάτω όριο λέξεων, τότε θα αγνοηθούν συνολικά 12.839 κείμενα. Αυτός ο αριθμός είναι πολύ μεγάλος αν αναλογιστεί κάποιος ότι το αρχικό πλήθος των κειμένων είναι 47.692. Συγκεκριμένα, με αυτήν την απόφαση ως κάτω όριο λέξεων θα αφαιρούνταν το 26.92% των κειμένων του αρχικού σώματος κατά τη διάρκεια της εκπαίδευσης. Κρίνεται λοιπόν αναγκαίο, να προσαρμοστεί το κάτω όριο λέξεων ενός κειμένου ώστε να χαθούν όσο το δυνατόν λιγότερα κείμενα γίνεται.

Με βάση το ραβδόγραμμα, φαίνεται πως μία καλή επιλογή κάτω φράγματος θα ήταν οι 3 λέξεις καθώς από τις 4 και πάνω παρατηρείται μεγαλύτερη συγκέντρωση κειμένων με τα πλήθη να γίνονται τετραψήφια. Με την επιλογή του αριθμού 3, θα αφαιρεθούν μόλις 1.696 κείμενα. Επαναλαμβάνεται η διαδικασία και για τα κείμενα με υψηλό αριθμό λέξεων.



Εικόνα 16 Πλήθος κειμένων στο SOSNet Twitter Dataset με υψηλό αριθμό λέξεων

Όπως εύκολα παρατηρείται από το παραπάνω διάγραμμα και σε συνδυασμό με το προηγούμενο, η πλειοψηφία των κειμένων έχει πλήθος λέξεων 7-15 λέξεις.

Το μεγαλύτερο κείμενο που συναντάται έχει 187 λέξεις ενώ μόλις 29 κείμενα έχουν από 40 λέξεις και πάνω. Ωστόσο, με τη λογική πως ένα κείμενο 40 λέξεων δεν είναι μεγάλο οπότε ενδεχομένως να εμπεριέχει σημαντική πληροφορία, η επιλογή ενός άλλου αριθμού ως άνω όριο θα ήταν πιο ιδανική. Με βάση το ραβδόγραμμα, μία καλή επιλογή φαίνεται να είναι οι 100 λέξεις αφού μόλις 9 κείμενα έχουν πάνω από 100 λέξεις και είναι περισσότερο πιθανό να υπάρχει περιττή πληροφορία που δεν θα φανεί χρήσιμη στην έρευνα σε κείμενα τέτοιας έκτασης.

3.1.4.2 Μέσο μήκος ανά κατηγορία

Στην παραπάνω ενότητα παρουσιάστηκε το πλήθος των κειμένων με λίγες και πολλές λέξεις. Κρίνεται σημαντικό να εξεταστεί το μέσο πλήθος λέξεων των κειμένων ανά κατηγορία διαδικτυακού εκφοβισμού. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα:

Πίνακας 2 Μέσο μήκος κειμένων στο SOSNet Twitter Dataset ανά κατηγορία Διαδικτυακού Εκφοβισμού

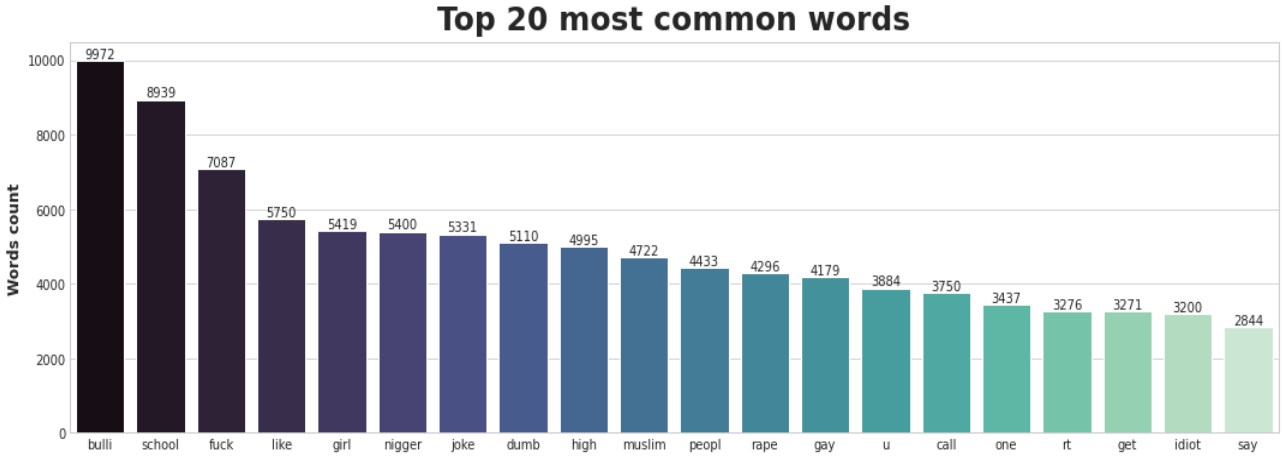
Κατηγορία	Μέσο πλήθος λέξεων
Διαδικτυακός εκφοβισμός με βάση τη θρησκεία-Religion Cyberbullying	17.78
Ηλικιακός διαδικτυακός εκφοβισμός-Age Cyberbullying	16.88
Διαδικτυακός εκφοβισμός με βάση το φύλο-Gender Cyberbullying	13.19

Διαδικτυακός εκφοβισμός εθνικότητας-Ethnicity Cyberbullying	14.71
Όχι διαδικτυακός εκφοβισμός-Not Cyberbullying	7.58

Τα πλήθη που προέκυψαν είναι δεκαδικοί αριθμοί λόγω της φύσης του μέσου όρου και της συνάρτησης mean() που χρησιμοποιήθηκε. Ένα ενδιαφέρον στοιχείο που γίνεται αντιληπτό είναι ότι στην κατηγορία των tweets που δεν παρατηρείται διαδικτυακός εκφοβισμός το μέσο πλήθος των λέξεων που χρησιμοποιούνται είναι περίπου το μισό σε σύγκριση με όλες τις υπόλοιπες κατηγορίες. Συνεπώς φαίνεται ότι τα μικρά κείμενα δεν συνδέονται με τον διαδικτυακό εκφοβισμό. Από την άλλη πλευρά, τα πιο μακροσκελή κείμενα εμφανίζονται στην κατηγορία του θρησκευτικού εκφοβισμού με περίπου 18 λέξεις κατά μέσο όρο.

3.1.4.3 Συνηθέστερες λέξεις στο SOSNet Twitter Dataset

Προτού παρουσιαστούν οι συνηθέστερες λέξεις ανά κατηγορία διαδικτυακού εκφοβισμού, παρουσιάζονται σε αυτήν την ενότητα οι 20 πιο χρησιμοποιημένες λέξεις στο σύνολο των κειμένων. Τα ανακτηθέντα αποτελέσματα αποτυπώνονται στο διάγραμμα που ακολουθεί.



Εικόνα 17 Συχνότητα εμφανίσεων των 20 συνηθέστερων λέξεων στο SOSNet Twitter Dataset

Κατά φθίνουσα σειρά εμφανίσεων οι λέξεις είναι: “bulli”, “school”, “fuck”, “like”, “girl”, “nigger”, “joke”, “dumb”, “high”, “muslim”, “peopl”, “rape”, “gay”, “u”, “call”, “one”, “it”, “get”, “idiot”, “say”. Οι λέξεις αυτές έχουν ανακτηθεί και πάλι έπειτα από σχετική προεπεξεργασία των κειμένων οπότε έχουν αφαιρεθεί οι καταλήξεις (stemming). Το γεγονός αυτό εξηγεί την ύπαρξη λέξεων όπως “peopl” ή “bulli”. Η συνηθέστερη λέξη με 9.972 εμφανίσεις στο σύνολο των κειμένων είναι η λέξη “bulli”, αποτέλεσμα που μπορεί να χαρακτηριστεί ως αναμενόμενο σε σχέση με το θέμα που εξετάζεται στην παρούσα εργασία. Παρατηρώντας πιο προσεκτικά τα αποτελέσματα, εμφανίζονται λέξεις που συνδέονται η καθεμία με διαφορετικό είδος εκφοβισμού. Για παράδειγμα συναντάται η λέξη “muslim” που πιθανώς θα σχετίζεται με την κατηγορία Θρησκευτικός διαδικτυακός εκφοβισμός-Religion Cyberbullying και η λέξη “nigger” που χρησιμοποιείται σαν όρος προσβολής των έγχρωμων ανθρώπων,κυρίως Αφροαμερικανών. Η συγκεκριμένη λέξη λοιπόν, το πιθανότερο είναι να εμφανίζεται σε tweets που συνδέονται με το Διαδικτυακό εκφοβισμό εθνικότητας- Ethnicity Cyberbullying. Αντίστοιχα και για τις υπόλοιπες λέξεις με εξαίρεση κάποιες γενικές όπως “u”, “one”, “get”, “say” κλπ που δεν συνδέονται με οποιουδήποτε είδους επίθεση αλλά είναι ιδιαίτερα συχνοί όροι στον προφορικό και γραπτό λόγο.

3.1.4.4 Αριθμοί και σημεία στίξης στο SOSNet Twitter Dataset

Σε αυτό το σημείο είναι χρήσιμο να παρουσιαστεί το πλήθος των σημείων στίξης, των αριθμών και των stopwords που υπάρχουν στα αρχικά κείμενα πριν την προεπεξεργασία τους. Για αυτόν τον λόγο τα κείμενα χωρίζονται σε 2 μεγάλες κατηγορίες Διαδικτυακός Εκφοβισμός-Cyberbullying (συμπεριλαμβάνονται και τα 4 είδη) και 'Όχι διαδικτυακός εκφοβισμός-Not cyberbullying και καταμετρώνται τα παραπάνω στοιχεία για την κάθε κατηγορία. Τα αποτελέσματα συγκεντρώνονται στον παρακάτω πίνακα:

Πίνακας 3 Εμφανίσεις αριθμών, σημείων στίξης και stopwords σε κείμενα του SOSNet Twitter Dataset που σχετίζονται και δεν σχετίζονται αντίστοιχα με διαδικτυακό εκφοβισμό.

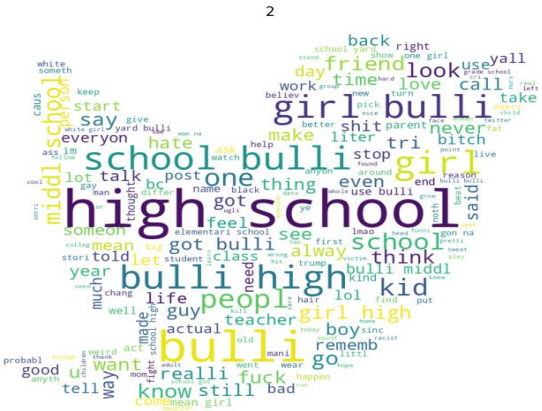
	Διαδικτυακός εκφοβισμός-Cyberbullying	Όχι διαδικτυακός εκφοβισμός- Not cyberbullying
Αριθμοί	4.010	692
Stop words	402.641	40.167
Σημεία Στίξης	37.212	8.303

Είναι λογικό οι αριθμοί της κατηγορίας Διαδικτυακός εκφοβισμός-Cyberbullying να είναι αρκετά υψηλότεροι γιατί συμπεριλαμβάνει τα $^4/_5$ των κειμένων. Όπως γίνεται αντιληπτό, υπάρχουν πάρα πολλοί αριθμοί, stopwords και σημεία στίξης στο σύνολο των αρχικών κειμένων που δεν προσδίδουν κάποια πληροφορία σχετικά με το εάν πρόκειται ή όχι για εκφοβισμό και αυτός είναι ο λόγος που πραγματοποιήθηκε προεπεξεργασία των κειμένων αυτών.

3.1.5 Συννεφόλεξα SOSNet Twitter Dataset

Σε αυτήν την ενότητα, θα παρουσιαστούν τα συννεφόλεξα για καθεμία από τις 4 κατηγορίες διαδικτυακού εκφοβισμού και θα αναδειχθούν οι συνηθέστερες λέξεις ανά κατηγορία. Συγκεκριμένα, θα παρουσιαστούν μετά τα συννεφόλεξα οι 10 πιο συνηθισμένες λέξεις (unigrams) και οι 10 πιο συνηθισμένες φράσεις των 2 λέξεων (bigrams). Όλα τα συννεφόλεξα έχουν δημιουργηθεί με χρήση πλαισίου τη χαρακτηριστική εικόνα του Twitter.

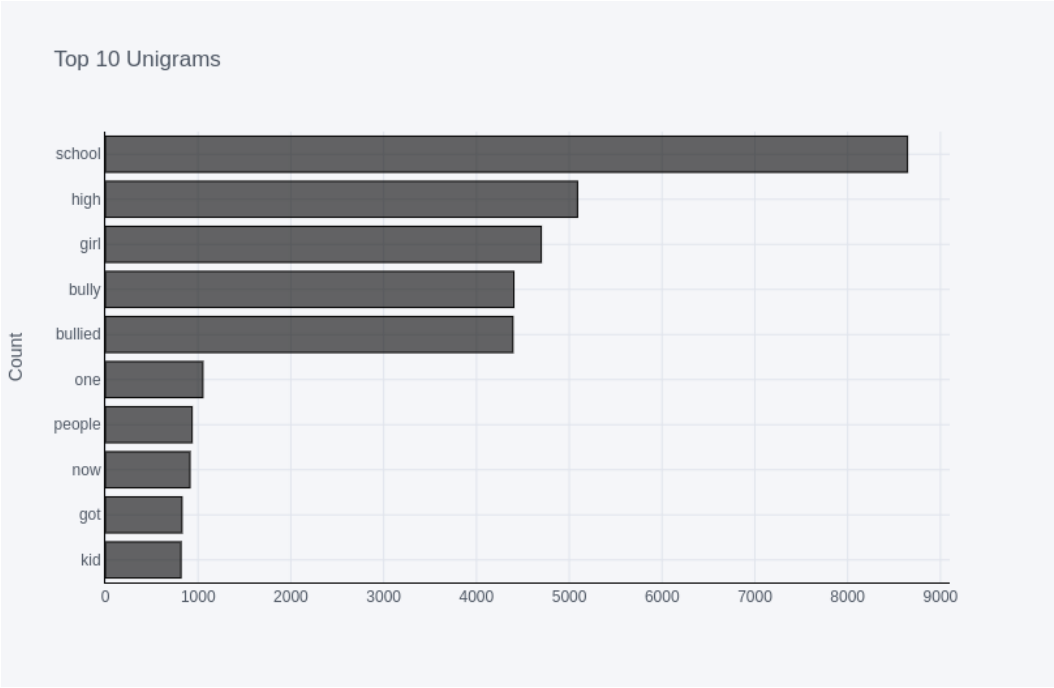
3.1.5.1 Κατηγορία Ηλικία-Age SOSNet Twitter Dataset



Εικόνα 18 Συννεφόλεξο όπως ανακτήθηκε από την κατηγορία Ηλικιακός διαδικτυακός εκφοβισμός-Age Cyberbullying του SOSNet Twitter Dataset

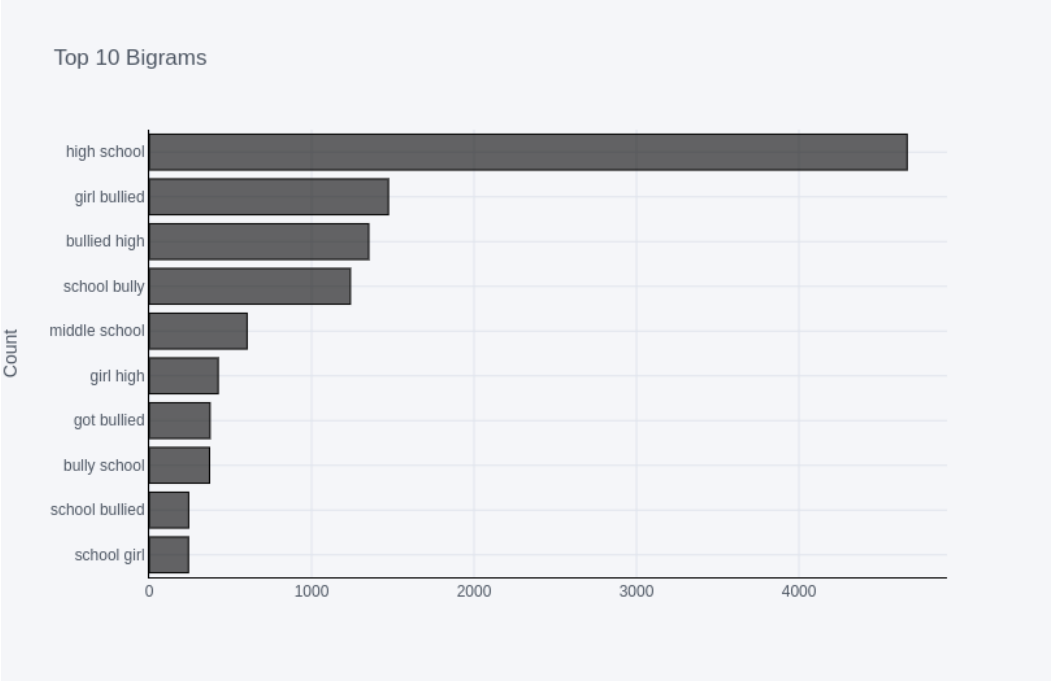
Στην κατηγορία αυτή συναντώνται οι 2 συνηθέστερες λέξεις στο σύνολο των κειμένων οι λέξεις “bulli” και “school”. Συγκεκριμένα η λέξη “school” φαίνεται να παρουσιάζεται μέσα σε φράσεις όπως για παράδειγμα “middl school” και “high school” που χαρακτηρίζουν σχολικές βαθμίδες και κατά συνέπεια πραγματοποιούν ηλικιακή διάκριση των ατόμων δικαιολογώντας την εμφάνιση τους στη συγκεκριμένη κατηγορία. Επιπλέον, εμφανίζονται και άλλες λέξεις που σχετίζονται με ηλικιακό διαχωρισμό όπως “kid”, “boy”, “girl” κλπ με αρκετά μεγάλη συχνότητα.

Το συννεφόλεξο είναι ενδεικτικό για τις λέξεις που χρησιμοποιούνται, παρόλα αυτά στην παρακάτω εικόνα αποτυπώνεται η συχνότητα εμφάνισης των 10 συνηθέστερων λέξεων της κατηγορίας.



Εικόνα 19 Πλήθος εμφανίσεων των 10 συνηθέστερων λέξεων της κατηγορίας Ηλικιακός διαδικτυακός εκφοβισμός-Age cyberbullying του SOSNet Twitter Dataset

Η λέξη “school” είναι μακράν η πιο συχνή λέξη της κατηγορίας Age cyberbullying αφού χρησιμοποιείται σχεδόν 9000 φορές με τα κείμενα της εν λόγω κατηγορίας να είναι 7992. Οι 10 πιο συνηθισμένες φράσεις 2 λέξεων (bigrams) της κατηγορίας αυτής εμφανίζονται επίσης στο παρακάτω ραβδόγραμμα.



Εικόνα 20 Πλήθος εμφανίσεων 10 συνηθέστερων bi-grams της κατηγορίας Ηλικιακός διαδικτυακός εκφοβισμός-Age cyberbullying του SOSNet Twitter Dataset

Με την παραπάνω εικόνα επιβεβαιώνεται η αρχική διαίσθηση που ανακτήθηκε μέσω του συννεφόλεξου και εξηγείται η πολύ συχνή εμφάνιση της λέξης “ school” αφού εμφανίζεται σε πολλές εκφράσεις των 2 λέξεων όπως “ high school”, “school bully”, “middle school”, “bully school” ,”school bullied” και “school girl”.

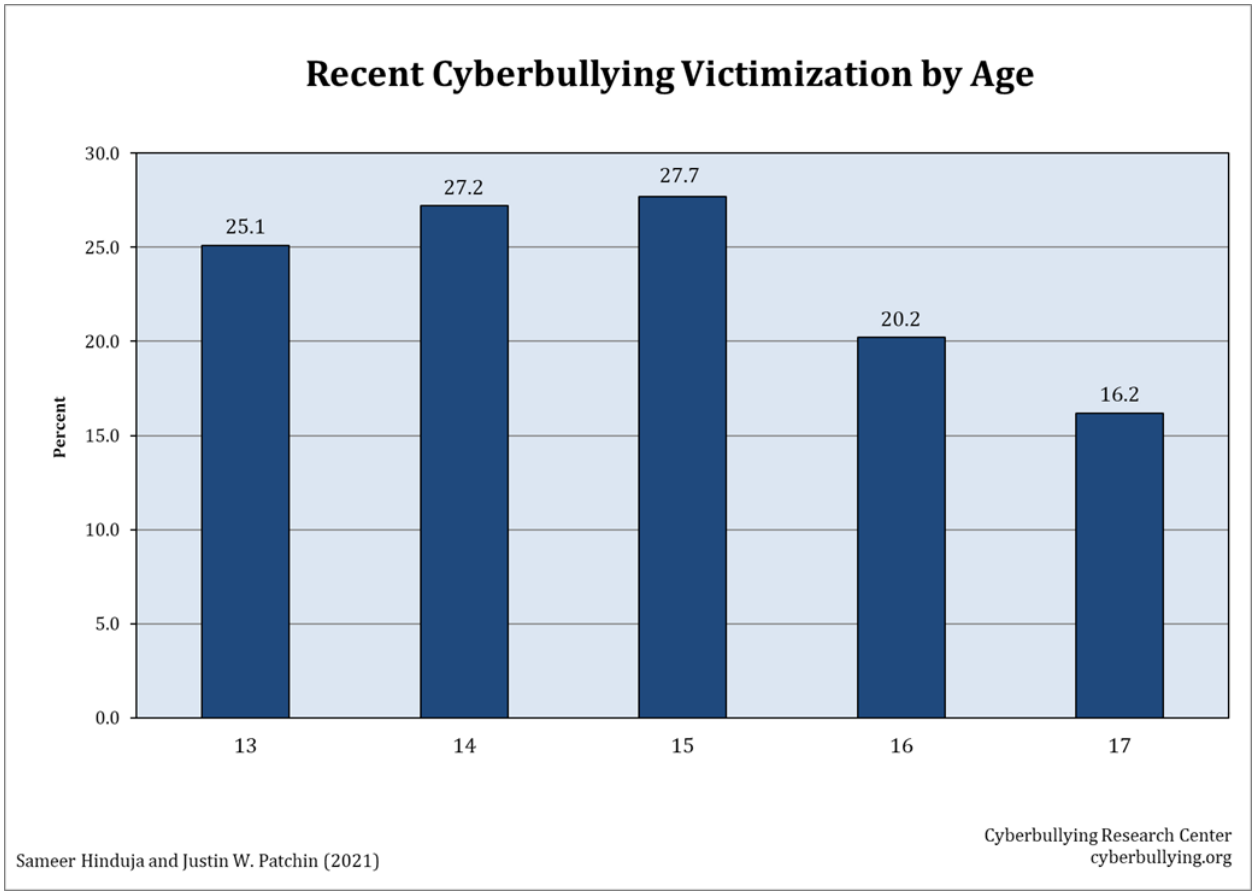
Η έρευνα για τον παραδοσιακό εκφοβισμό δείχνει ότι τα ποσοστά επικράτησης του εκφοβισμού κορυφώνονται κατά τη διάρκεια του γυμνασίου, όπως οι νέοι εργάζονται για να εδραιώσουν τη θέση τους στην κοινωνική ιεραρχία [36]

Ομοίως, ο διαδικτυακός εκφοβισμός είναι ιδιαίτερα διαδεδομένος μεταξύ των παιδιών του γυμνασίου [37]. Ωστόσο, ακόμα και μεταξύ των παιδιών του γυμνασίου υπάρχουν αναπτυξιακές παραλλαγές. Για παράδειγμα, υπάρχουν έρευνες που δείχνουν πως το cyberbullying αυξάνεται μετά την πέμπτη τάξη και κορυφώνεται κατά τη διάρκεια της όγδοης τάξης [38]. Αυτές οι βαθμίδες αφορούν το Ηνωμένο Βασίλειο και σύμφωνα με τα ελληνικά δεδομένα, η πέμπτη τάξη αναφέρεται στις ηλικίες 10-11 και η όγδοη στα παιδιά ηλικίας 13-14 ετών. Ωστόσο, άλλοι ερευνητές προτείνουν ότι οι διαφορές ηλικίας εξαρτώνται από τη μέθοδο με την οποία εμφανίζεται ο διαδικτυακός εκφοβισμός. Στην περίπτωση του εκφοβισμού μέσω άμεσων γραπτών μηνυμάτων και φωτογραφιών παρατηρήθηκε μεγαλύτερη συχνότητα εμφάνισης στη νεολαία λίγο μεγαλύτερης ηλικίας από ότι σε μικρότερα ηλικιακά άτομα [39].

Συνολικά, οι έφηβοι υποστηρίζεται ότι εμπλέκονται πολύ περισσότερο από τους ενήλικες σε συμβάντα εκφοβισμού, δεδομένου του μειωμένου επιπέδου ωριμότητάς τους σε σχέση με ικανότητες όπως η αναζήτηση συγκίνησης, ο έλεγχος των παρορμήσεων, η πίεση των συνομηλίκων, η ευαισθησία ανταμοιβής, η γνωστική επεξεργασία και η ορθολογική λήψη αποφάσεων καθώς και ο μακροχρόνιος χρονοπρογραμματισμός [40]. Επιπλέον, είναι πιο δύσκολο για τους εφήβους να κατανοήσουν τη σχέση μεταξύ της συμπεριφοράς τους και των συνεπειών της. Όπως ειπώθηκε και προηγουμένως, η κορύφωση του εκφοβισμού είναι στο Γυμνάσιο ενώ προς το τέλος του Λυκείου τείνει να

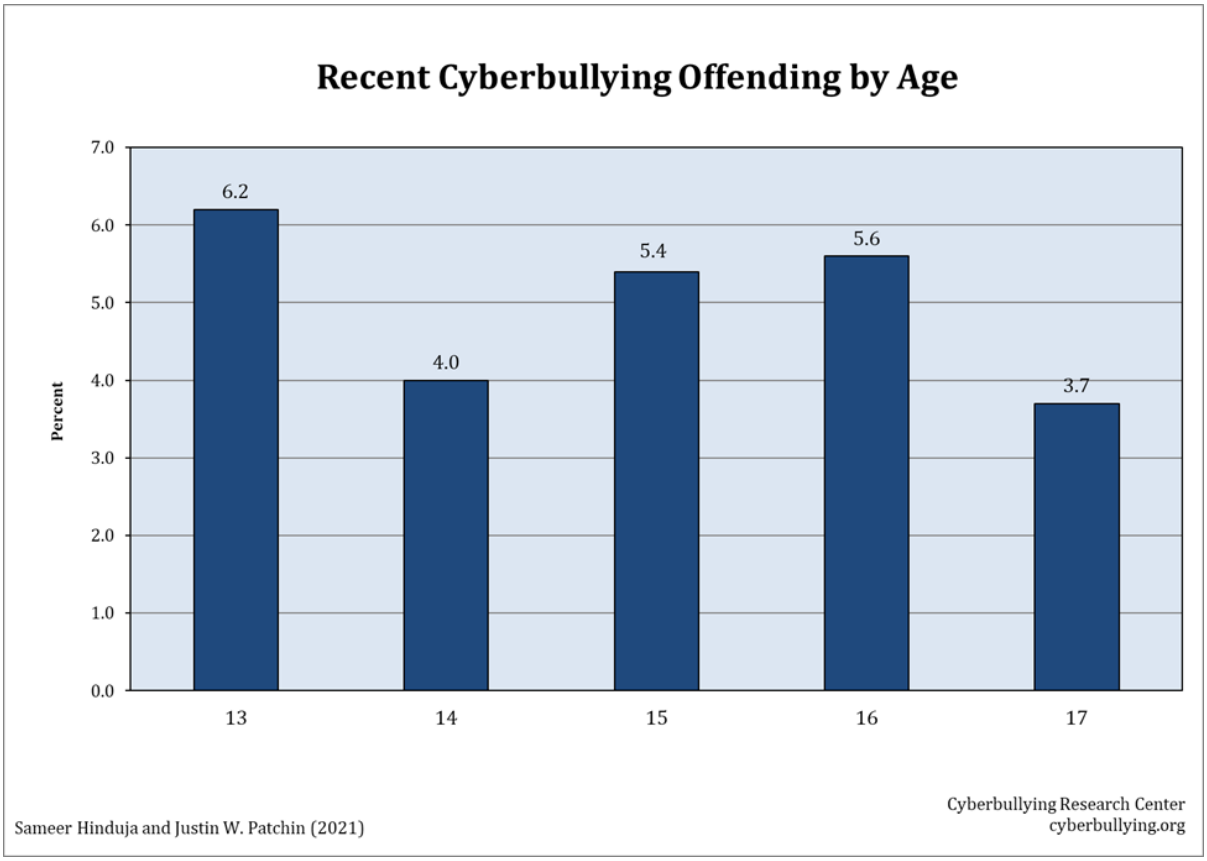
μειώνεται [41]. Όσον αφορά τις μορφές εκφοβισμού, η αύξηση της ηλικίας φαίνεται να σχετίζεται με μια μετατόπιση από τον σωματικό εκφοβισμό στον έμμεσο και ψυχολογικό εκφοβισμό [42].

Σύμφωνα με μία πρόσφατη έρευνα το 2021 [60] διαπιστώθηκε ότι ο διαδικτυακός εκφοβισμός τείνει να κορυφώνεται περίπου στην ηλικία των 14 και 15 ετών πριν μειωθεί κατά τα τελευταία χρόνια της εφηβείας. Με τα μέσα κοινωνικής δικτύωσης και τις πλατφόρμες τυχερών παιχνιδιών που απαιτούν τεχνικά από τους χρήστες να είναι τουλάχιστον 13 ετών, είναι αξιοσημείωτο ότι ένας στους τέσσερις (25,1%) από αυτούς τους πολύ νεαρούς εφήβους έχει υποστεί διαδικτυακό εκφοβισμό πρόσφατα (δηλαδή τις τελευταίες 30 ημέρες). Τα αποτελέσματα αυτά αποτυπώνονται στην παρακάτω εικόνα:



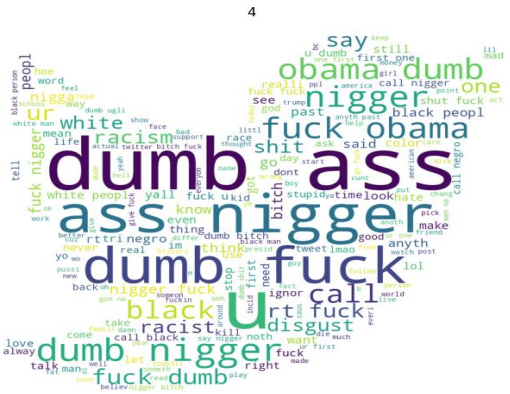
Εικόνα 21 Ποσοστά θυματοποίησης διαδικτυακού εκφοβισμού ανά έτος εφηβείας (Hinduja&Patchin,2021)

Όσον αφορά τη διάπραξη διαδικτυακού εκφοβισμού [60], βλέπουμε και πάλι ανησυχητικούς αριθμούς σχετικά με τη μικρότερη ηλικιακή ομάδα που μελετήθηκε. Εδώ, το υψηλότερο ποσοστό (6,2%) των νέων που ασκούσαν διαδικτυακό εκφοβισμό σε άλλους ήταν ηλικίας 13 ετών. Οι αριθμοί ήταν σχετικά παρόμοιοι με τις υπόλοιπες ηλικιακές ομάδες που μελετήθηκαν, αλλά βλέπουμε ότι μια αποκλιμάκωση της συμμετοχής εμφανίζεται στην προχωρημένη εφηβεία.



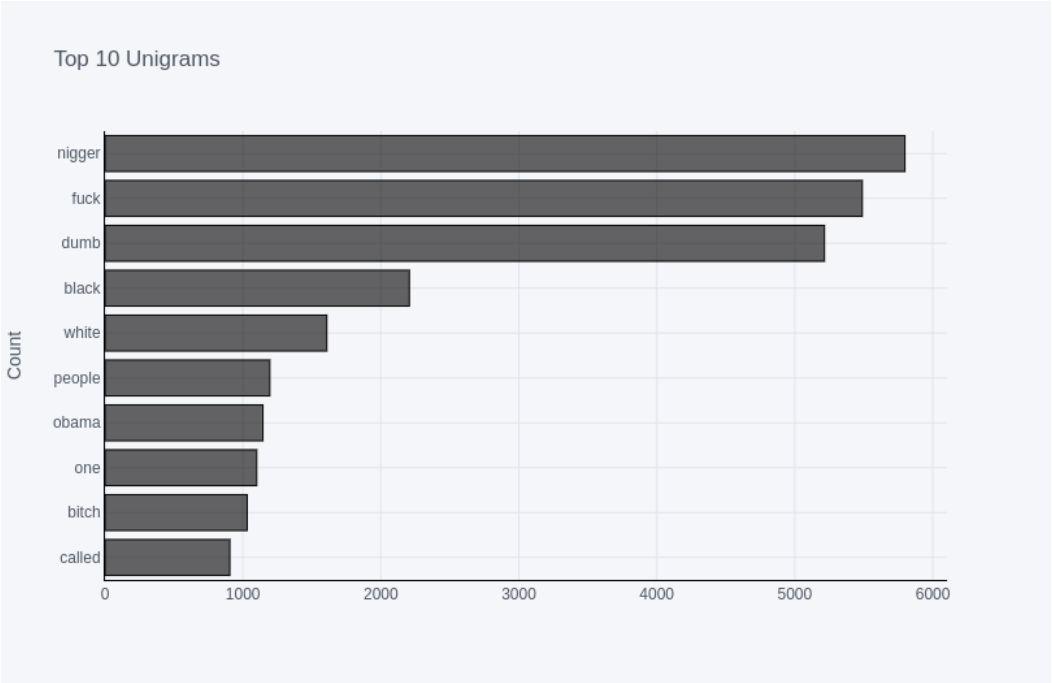
Εικόνα 22 Ποσοστά διάπραξης διαδικτυακού εκφοβισμού ανά έτος της εφηβείας (Hinduja&Patchin,2021)

3.1.5.2 Κατηγορία Εθνικότητα-Ethnicity SOSNet Twitter Dataset



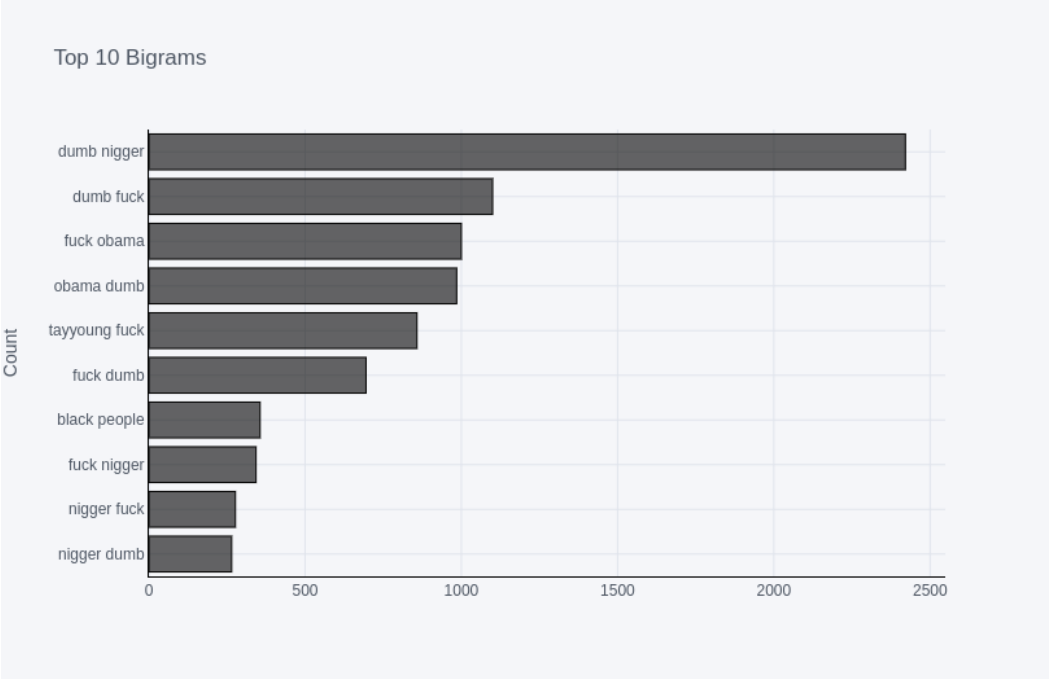
Εικόνα 23 Συννεφόλεξο όπως ανακτήθηκε από την κατηγορία Διαδικτυακός εκφοβισμός εθνικότητας-Ethnicity cyberbullying του SOSNet Twitter Dataset

Όπως φαίνεται από το παραπάνω συννεφόλεξο, ο προηγούμενος ισχυρισμός που συσχέτιζε τη λέξη “nigger” με την κατηγορία Ethnicity Cyberbullying επιβεβαιώνεται άμεσα. Επιπρόσθετα, εμφανίζονται και άλλες λέξεις που μπορούν να ταυτιστούν με ρατσιστική επίθεση όπως “black”, “racist”, “racism”. Πιο συγκεκριμένα, είναι δυνατό να εξεταστεί ο ακριβής αριθμών εμφανίσεων των 10 πιο συνηθισμένων λέξεων της κατηγορίας στο παρακάτω ραβδόγραμμα.



Εικόνα 24 Πλήθος εμφανίσεων 10 συνηθέστερων λέξεων της κατηγορίας Διαδικτυακός εκφοβισμός εθνικότητας- Ethnicity cyberbullying του SOSNet Twitter Dataset

Πράγματι η λέξη “nigger” είναι η πιο συνηθισμένη λέξης αυτής της κατηγορίας διαδικτυακού εκφοβισμού-cyberbullying με το πλήθος των εμφανίσεων της να αγγίζει κατά προσέγγιση τις 6000 φορές. Εδώ, το χάσμα της πρώτης με τη δεύτερη λέξη δεν είναι τόσο μεγάλο όσο στην προηγούμενη κατηγορία, αφού ακολουθεί η λέξη “fuck” με περίπου 5.500 εμφανίσεις και η λέξη “dumb” είναι επίσης κοντά ως τρίτη με 5200. Δημιουργώντας το ραβδόγραμμα για τις φράσεις 2 λέξεων- bigrams, η ανακτηθείσα εικόνα είναι η ακόλουθη:

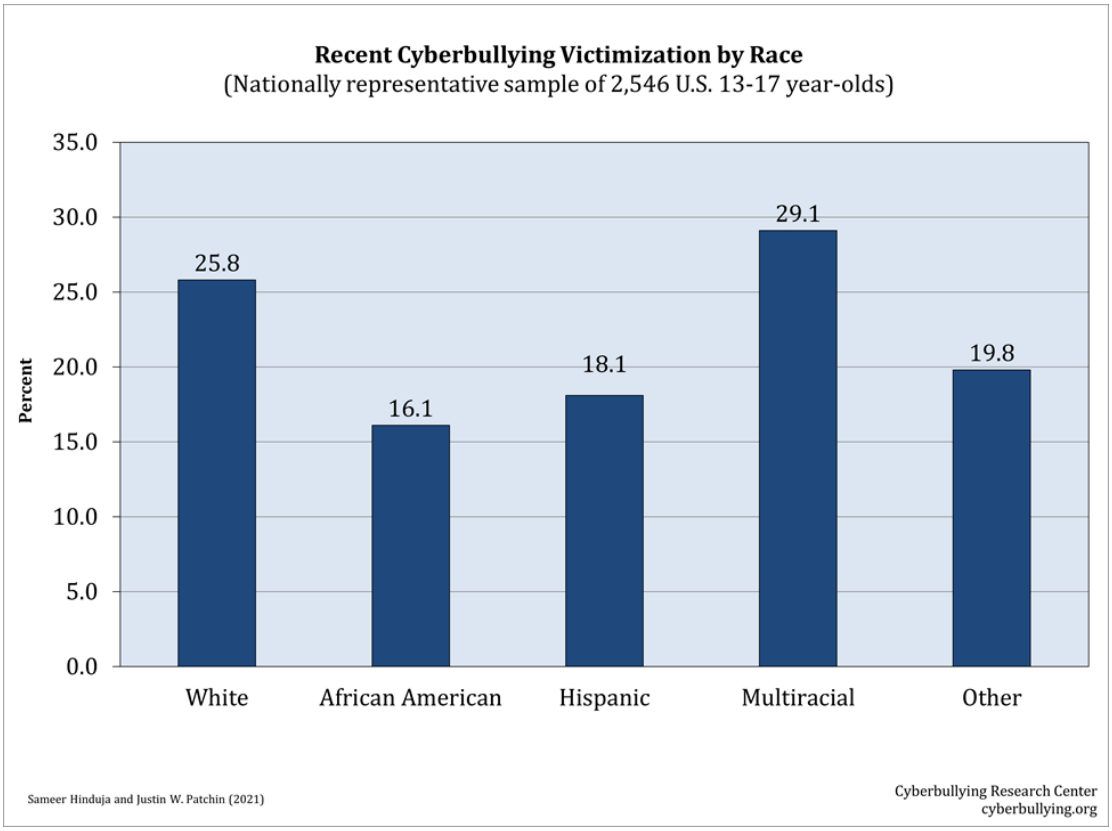


Εικόνα 25 Πλήθος εμφανίσεων 10 συνηθέστερων φράσεων με 2 λέξεις- bigrams της κατηγορίας Διαδικτυακός εκφοβισμός εθνικότητας- Ethnicity cyberbullying του SOSNet Twitter Dataset

Όπως είναι εμφανές, οι 3 συνηθέστερες λέξεις συμμετέχουν σχεδόν σε όλα τα 10 συνηθέστερα bigrams δικαιολογώντας το ισόποσο πλήθος εμφανίσεων τους.

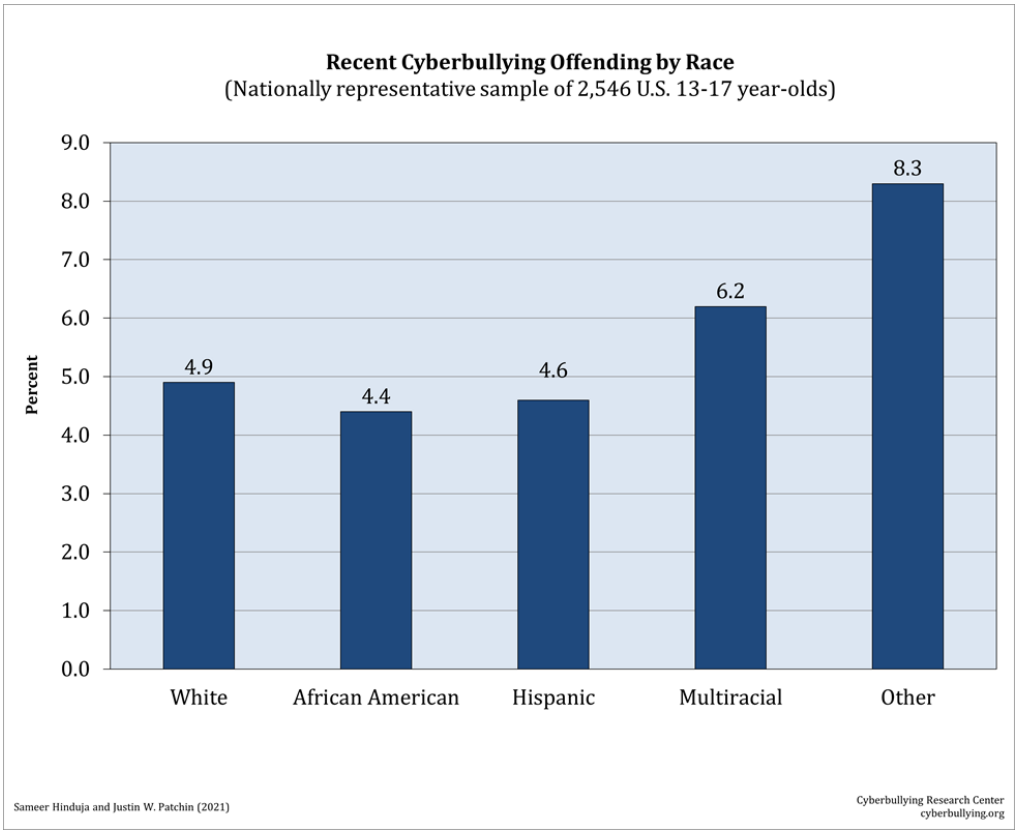
Είναι σημαντικό να προσδιοριστεί επίσης εάν όσοι ανήκουν σε ορισμένες φυλετικές κατηγορίες έχουν περισσότερες πιθανότητες να βιώσουν διαδικτυακό εκφοβισμό - ή είναι πιο πιθανό να εκφοβίσουν άλλους με βάση τη βιβλιογραφία. Ως σημείο αναφοράς, μια ανασκόπηση 15 μελετών [59] που αφορούσαν τον διαδικτυακό εκφοβισμό και τη φυλή/εθνικότητα βρήκε ένα ευρύ φάσμα ποσοστών θυματοποίησης (Λευκοί-White: 18-30% Μαύροι-Black: 4-17% Ισπανόφωνοι-Hispanic: 6-13%) και πρόκλησης εκφοβισμού (Λευκοί-White: 4-42% Μαύροι-Black: 7-11% Ισπανόφωνοι-Hispanic 16-18%).

Μεταξύ των νέων 13-17 ετών στις Ηνωμένες Πολιτείες, βρέθηκε ότι εκείνοι που ήταν Λευκοί (25,8%) και εκείνοι που ήταν πολυφυλετικοί (29,1%) ήταν πιο επιρρεπείς στη θυματοποίηση του διαδικτυακού εκφοβισμού [60]. Τα τρέχοντα ευρήματα έρχονται σε συμφωνία με αυτά που έχουν βρεθεί σε μερικές μελέτες όπου Ισπανόφωνοι και Έγχρωμοι νέοι έχουν αναφέρει λιγότερη θυματοποίηση διαδικτυακού εκφοβισμού από τους συνομηλίκους τους [59,61].



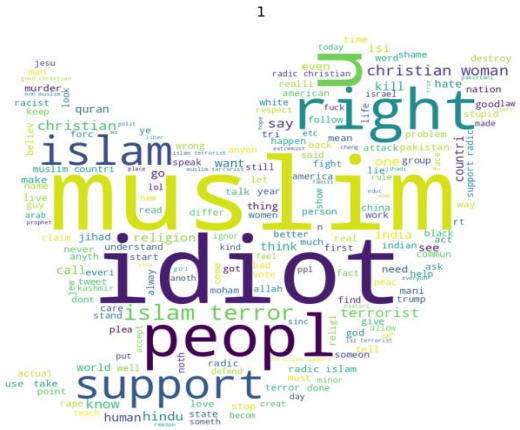
Εικόνα 26 Ποσοστά θυματοποίησης διαδικτυακού εκφοβισμού σύμφωνα με την εθνικότητα/φυλή για το 2021 στις Ηνωμένες Πολιτείες (Hinduja & Patchin,2021)

Όσον αφορά την προσβολή, εκείνοι που ανέφεραν τα υψηλότερα ποσοστά διάπραξης [60]διαδικτυακού εκφοβισμού τις τελευταίες τριάντα ημέρες ταξινομήθηκαν ως «Άλλοι» όταν ρωτήθηκαν για τη φυλή (8,3%). Μεταξύ των υπόλοιπων κατηγοριών, όσοι ήταν πολυφυλετικοί ανέφεραν το μεγαλύτερο ποσοστό διαδικτυακού εκφοβισμού άλλων (6,2%).



Εικόνα 27 Ποσοστά διάπραξης διαδικτυακού εκφοβισμού με βάση την εθνικότητα (Hinduja&Patchin,2021)

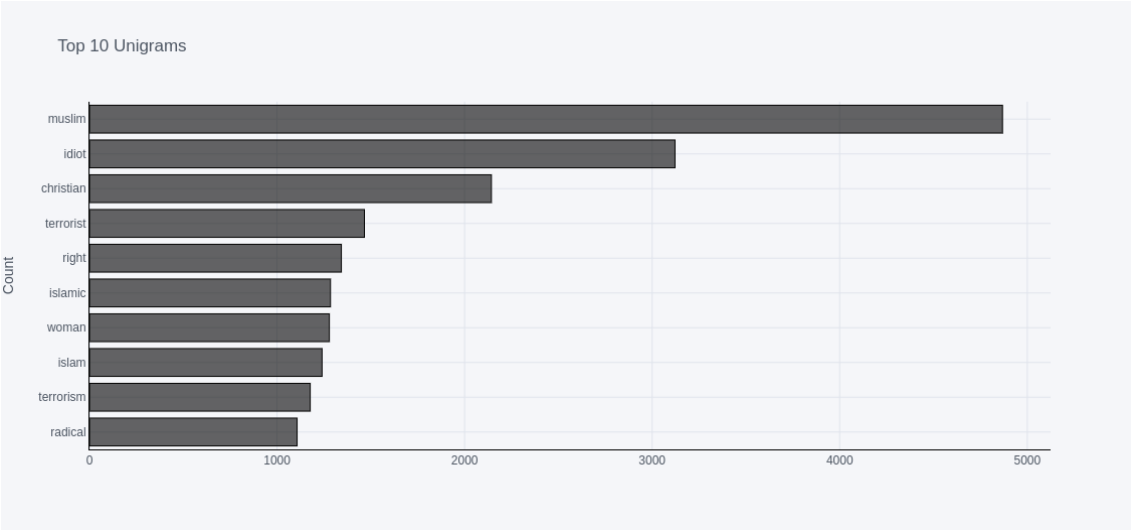
3.1.5.3 Κατηγορία Θρησκεία-Religion SOSNet Twitter Dataset



Εικόνα 28 Συννεφόλεξο όπως ανακτήθηκε από την κατηγορία Θρησκευτικός διαδικτυακός εκφοβισμός-Religion cyberbullying του SOSNet Twitter Dataset

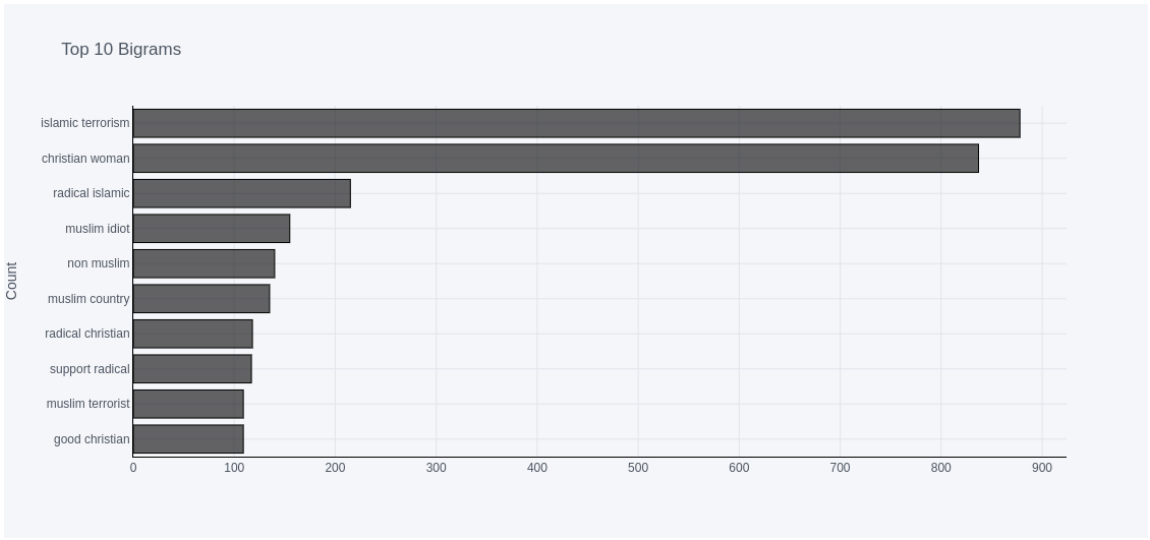
Και σε αυτήν την περίπτωση επιβεβαιώνεται η αρχική υπόθεση πως η λέξη “muslim” που ήταν στις 20 συνηθέστερες θα συνδεόταν με τη συγκεκριμένη κατηγορία. Ακόμα, εμφανίζονται συχνά και οι όροι “islam”,”islam terror”, “christian” που όλοι συνδέονται με

τη θρησκεία. Όπως στην προηγούμενη κατηγορία εκφοβισμού έγινε εμφανές ότι η ομάδα στόχου των tweets ήταν οι Αφροαμερικανοί, έτσι και από το συννεφόλεξο που προέκυψε σε αυτήν την κατηγορία φαίνεται ότι τα κείμενα που συνδέονται με θρησκευτικό εκφοβισμό επικεντρώνονται κατά κύριο λόγο στη Μουσουλμανική θρησκεία. Το πλήθος εμφανίσεων των 10 συνηθέστερων λέξεων της κατηγορίας παρουσιάζεται παρακάτω:



Εικόνα 29 Πλήθος εμφανίσεων 10 συνηθέστερων λέξεων της κατηγορίας Θρησκευτικός διαδικτυακός εκφοσιμός- Religion cyberbullying του SOSNet Twitter Dataset

Πράγματι η λέξη “muslim” κυριαρχεί με λίγο λιγότερες από 5000 εμφανίσεις ενώ αμέσως επόμενη “idiot” είναι αρκετά πιο πίσω με περίπου 3000 εμφανίσεις. Οι 10 πιο συνηθισμένες φράσεις των 2 λέξεων-bigrams της κατηγορίας αποτυπώνονται επίσης στο ακόλουθο ραβδόγραμμα:

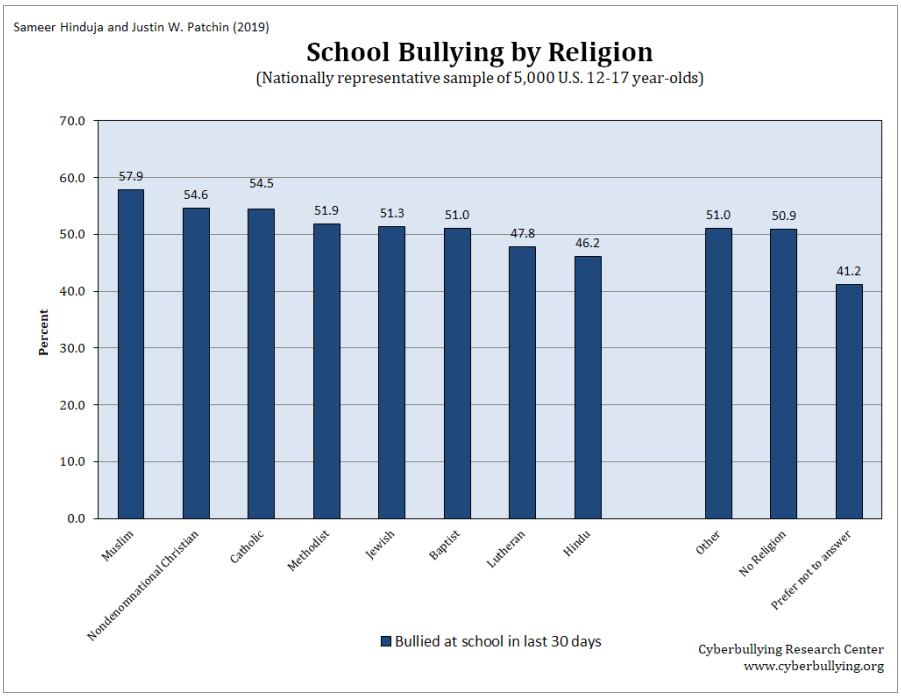


Εικόνα 30 Πλήθος εμφανίσεων 10 συνηθέστερων φράσεων 2 λέξεων-bigrams της κατηγορίας Θρησκευτικός διαδικτυακός εκφοβισμός-Religion cyberbullying του SOSNet Twitter Dataset

Η συνηθέστερη λέξη αυτήν τη φορά δεν αποτελεί μέρος της πιο συνηθισμένης φράσης 2 λέξεων- bigram αλλά εμφανίζεται σε 4 τέτοιες εκφράσεις (“muslim idiot”, “non muslim” , “muslim country”, “muslim terrorist”). Παρατηρείται επίσης ότι η λέξη “Christian” εμφανίζεται σε μη-υβριστικές και ήπιες εκφράσεις όπως “christian woman” ή ακόμα και σε θετικές όπως “good christian”, σε αντίθεση με τις “islamic” και “muslim” που συνδέονται άμεσα με τη λέξη “terrorism” και γενικά με λέξεις με αρνητική σημασία.

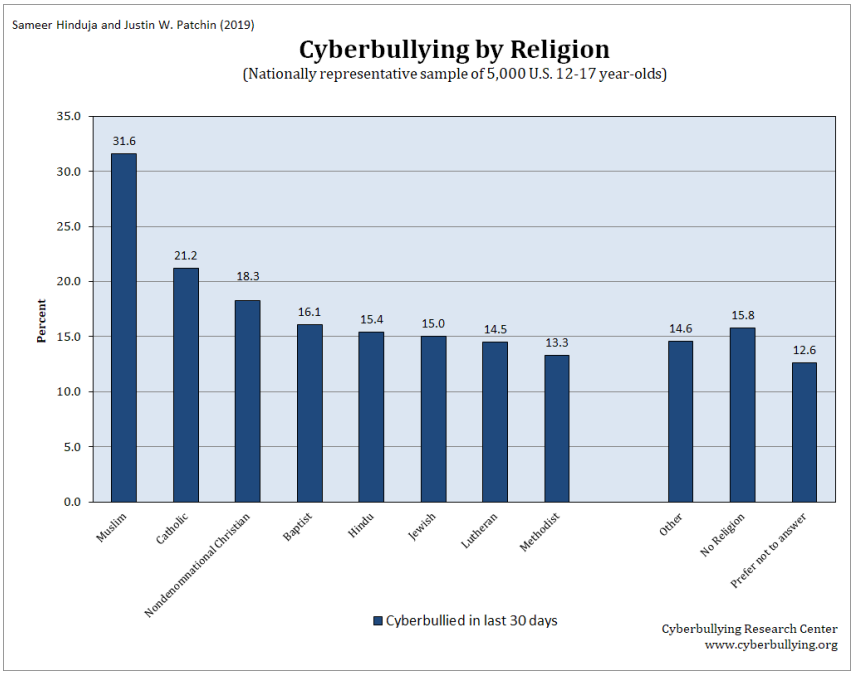
Σε έρευνα που πραγματοποιήθηκε το 2019 [62], συλλέχθηκαν δεδομένα από ένα εθνικά αντιπροσωπευτικό δείγμα 5.000 μαθητών γυμνασίου και γυμνασίου σε όλη την Αμερική.

Όπως είναι εμφανές από το παρακάτω διάγραμμα, μαθητές διαφόρων θρησκειών φαίνεται να εκφοβίζονται στο σχολείο σχετικά εξίσου.



Εικόνα 31 Θρησκείες ατόμων που υπόκεινται σε εκφοβισμό (Hinduja& Patchin,2019)

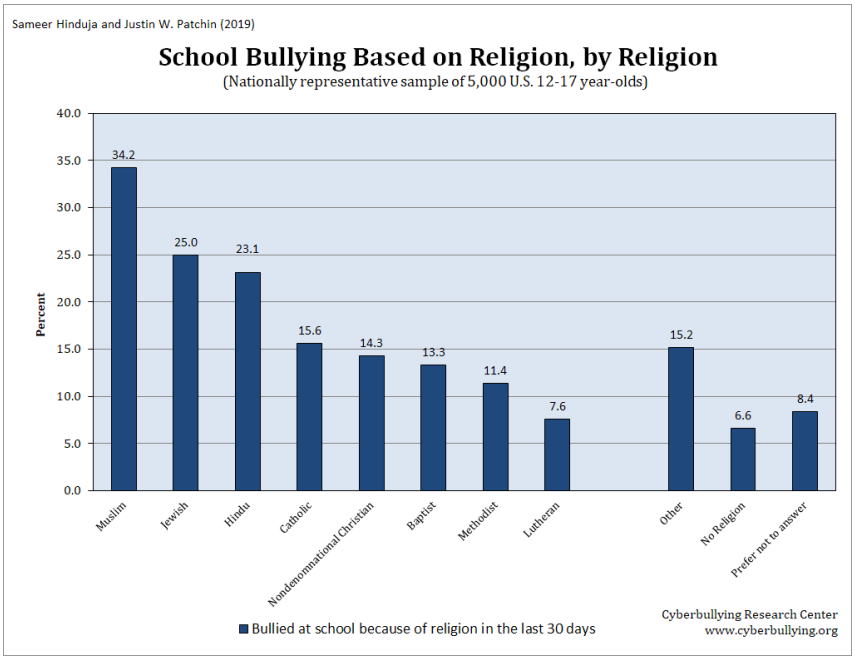
Ωστόσο, όσον αφορά το διαδικτυακό εκφοβισμό, περισσότεροι μουσουλμάνοι νέοι είπαν ότι στοχοποιήθηκαν από εκείνους άλλων θρησκειών.



Εικόνα 32 Θρησκείες ατόμων που υπόκεινται σε διαδικτυακό εκφοβισμό (Hinduja& Patchin,2019)

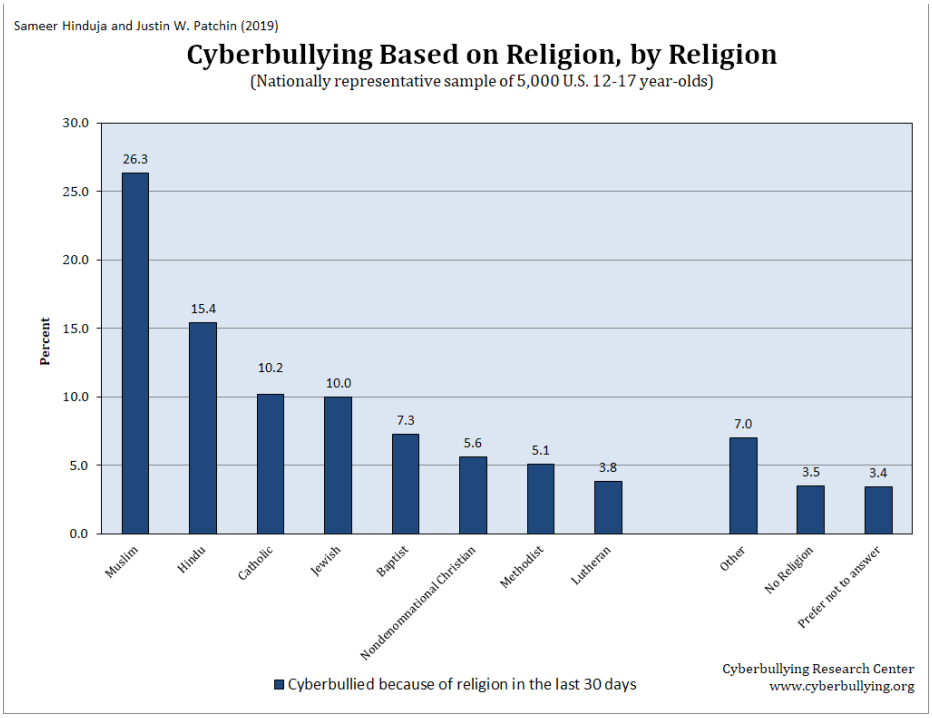
Αυτά τα δεδομένα ωστόσο, αφορούν τις θρησκείες που ασπάζονται τα άτομα που πέφτουν θύματα παραδοσιακού ή διαδικτυακού εκφοβισμού γενικά για οποιοδήποτε αίτιο.
Τώρα, ας στρέψουμε την προσοχή μας στον εκφοβισμό **με βάση** τη θρησκεία [62]. Όπως φαίνεται στο παρακάτω διάγραμμα, το 34,3% των Μουσουλμάνων νέων, το 25% των

Εβραίων και το 23,1% των Ινδουιστών νέων λένε ότι έχουν στοχοποιηθεί στο σχολείο τις τελευταίες 30 ημέρες λόγω της πίστης τους.



Εικόνα 33 Θρησκείες ατόμων που εκφοβίζονται στο σχολείο εξαιτίας της θρησκείας τους (Hinduha&Patchin,2019)

Όσον αφορά τον διαδικτυακό εκφοβισμό, το 26,3% των μουσουλμάνων μαθητών ανέφερε ότι στοχοποιήθηκε τις τελευταίες 30 ημέρες, όπως και το 15,4% των Ινδουιστών μαθητών.



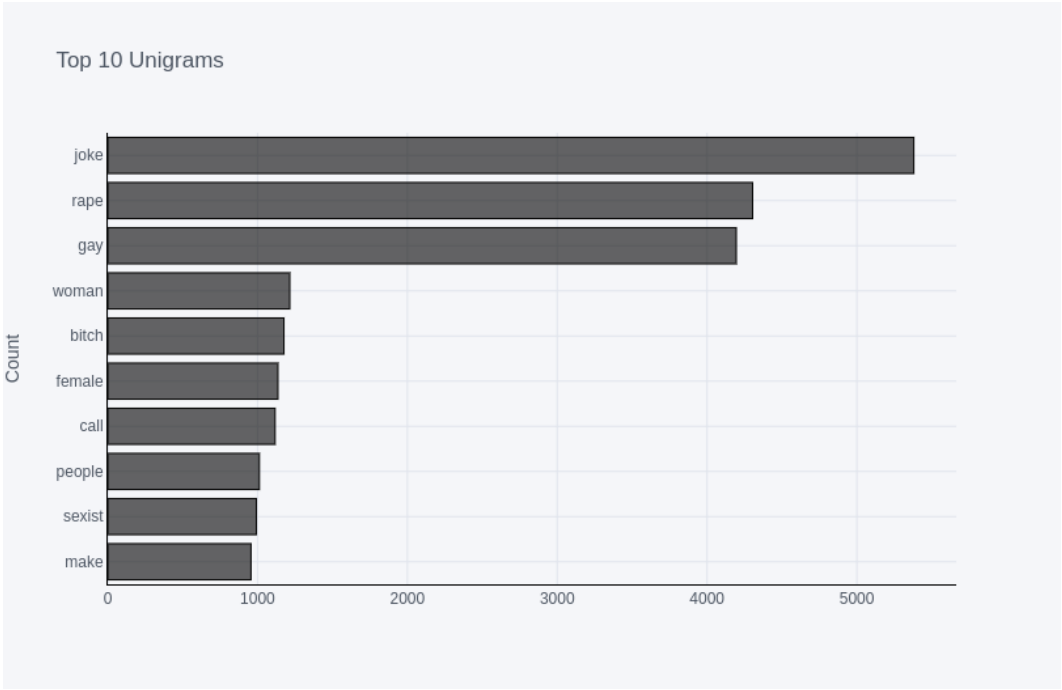
Εικόνα 34 Θρησκείες ατόμων που εκφοβίζονται διαδικτυακά εξαιτίας της θρησκείας τους (Hinduja&Patchin,2019)

Τα στοιχεία αυτά έρχονται σε συμφωνία με την αρχική διαίσθηση που αποκομίθηκε από τη μελέτη συννεφόλεξου για τη συγκεκριμένη κατηγορία εκφοβισμού.

3.1.5.4 Κατηγορία Φύλο-Gender SOSNet Twitter Dataset

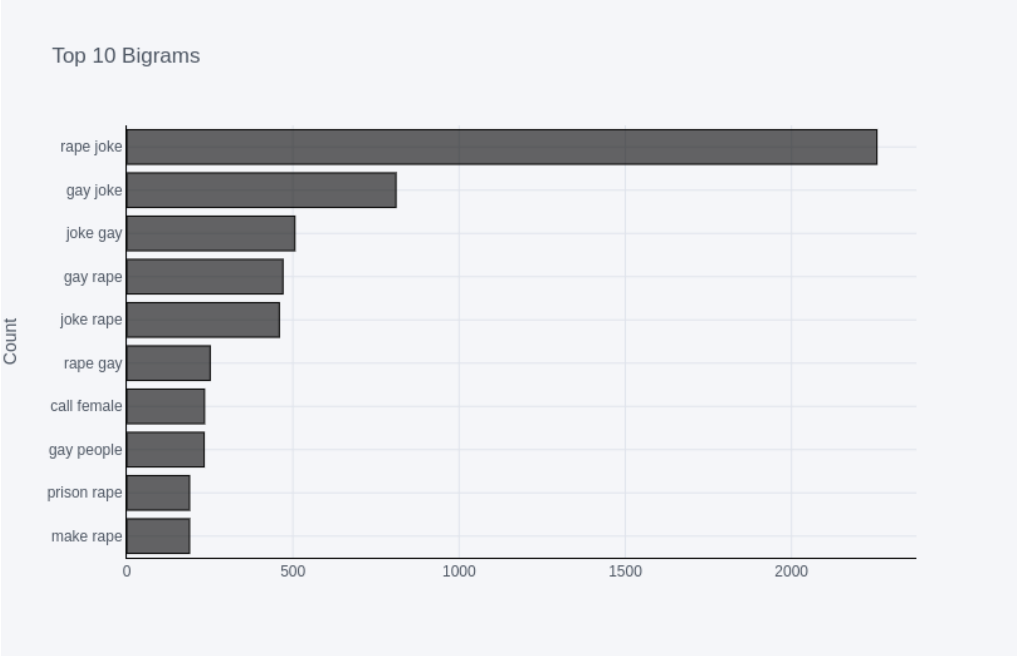
Εικόνα 35 Συννεφόλεξο όπως ανακτήθηκε από την κατηγορία Διαδικτυακός εκφοβισμός με βάση το φύλο-Gender cyberbullying του SOSNet Twitter Dataset

Σε αυτήν την κατηγορία δεν συναντάται κάποια λέξη που να ξεχωρίζει τόσο όσο στις προηγούμενες 3. Ωστόσο, οι λέξεις “gay”, “rape”, “joke”, “bitch” και “women” φαίνεται να κυριαρχούν. Το ραβδόγραμμα εμφανίσεων των 10 συνηθέστερων λέξεων που φαίνεται παρακάτω επιβεβαιώνει άμεσα τον παραπάνω ισχυρισμό.



Εικόνα 36 Πλήθος εμφανίσεων 10 συνηθέστερων λέξεων της κατηγορίας Διαδικτυακός εκφοβισμός με βάση το φύλο-Gender Cyberbullying του SOSNet Twitter Dataset

Είναι ενδιαφέρον να φανεί πώς οι λέξεις “rape” και “joke” εμφανίζονται σε φράσεις 2 λέξεων-bigrams. Από την παρακάτω εικόνα, γίνεται φανερό ότι η πιο συνηθισμένη έκφραση 2 λέξεων περιλαμβάνει και τις 2 αυτές λέξεις (“rape joke”) και παρουσιάζει περίπου 2500 εμφανίσεις. Αμέσως μετά ακολουθεί η φράση “gay joke” με αρκετά λιγότερες εμφανίσεις. Με βάση αυτά τα στοιχεία, φαίνεται πως στην κατηγορία Διαδικτυακός εκφοβισμός με βάση το φύλο-Gender cyberbullying τα θύματα εκφοβισμού στο σύνολο δεδομένων-dataset που εργαζόμαστε είναι κυρίως τα ομοφυλόφιλα άτομα, οι γυναίκες καθώς και τα θύματα βιασμού.



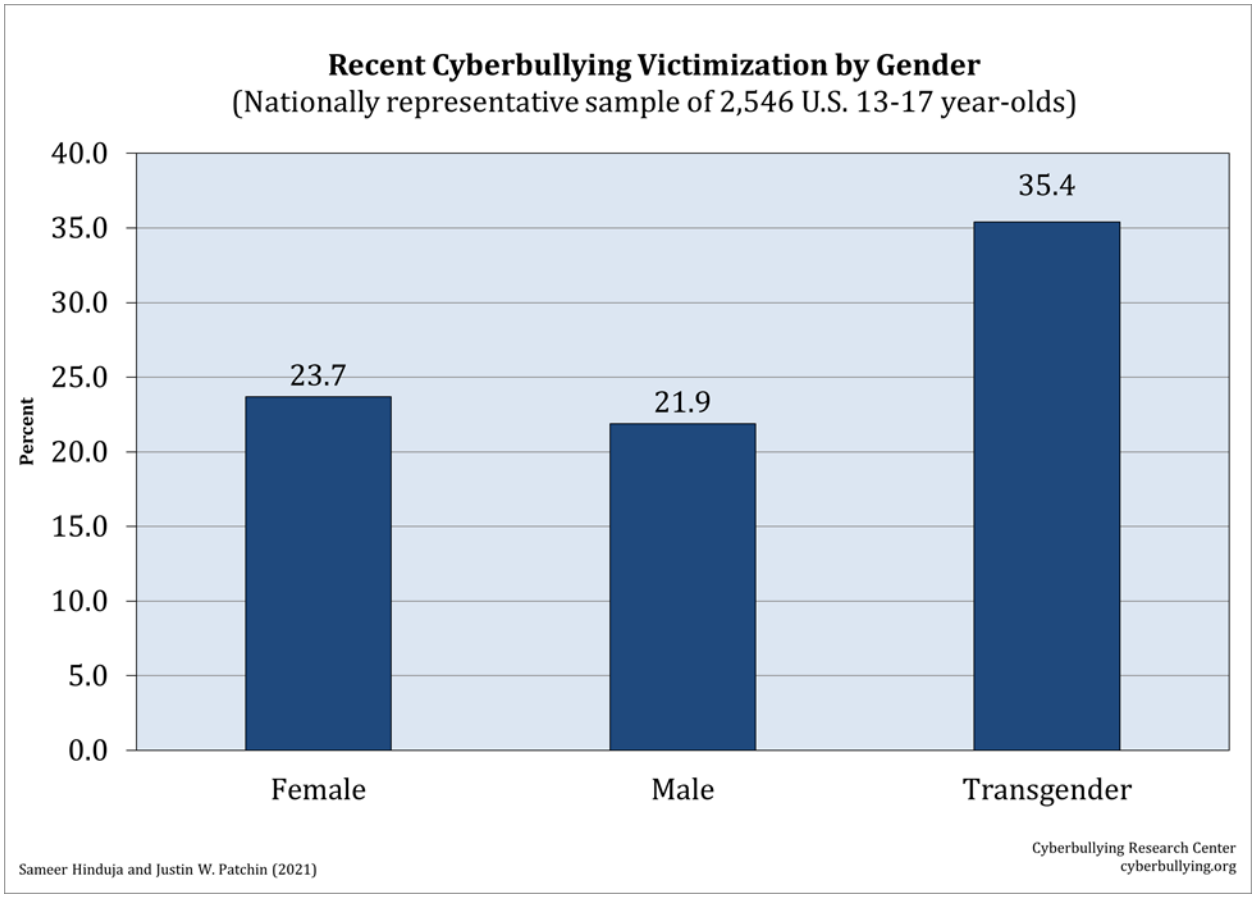
Εικόνα 37 Πλήθος εμφανίσεων 10 συνηθέστερων bigrams της κατηγορίας Διαδικτυακός εκφοβισμός με βάση το φύλο-Gender cyberbullying του SOSNet Twitter Dataset

Η έρευνα για τον παραδοσιακό εκφοβισμό έχει δείξει σταθερά ότι τα αγόρια επιδίδονται σε εκφοβισμό σε μεγαλύτερο βαθμό από ότι τα κορίτσια [43] και η επιθετικότητα είναι

πιο συχνά άμεσης φύσης (ενώ τα κορίτσια πιο συχνά μπλέκονται σε έμμεσους τύπους επιθετικότητας [44]). Ο διαδικτυακός εκφοβισμός είναι μια μορφή έμμεσης επιθετικότητας, η οποία μπορεί να οδηγήσει κάποιον να καταλήξει στο συμπέρασμα ότι τα κορίτσια θα ήταν πιο πιθανό από τα αγόρια να βιώσουν διαδικτυακό εκφοβισμό τόσο ως θύματα όσο και ως θύτες. Αν και κάποιες έρευνες υποστηρίζουν αυτήν την υπόθεση [45] άλλες έρευνες δεν έχουν βρει στατιστικά σημαντική διαφορά μεταξύ κοριτσιών και αγοριών σε ποσοστά διαπράξεων διαδικτυακού εκφοβισμού ή θυματοποίηση [46,47]. Ακόμη άλλη έρευνα διαπιστώνει ότι τα αγόρια είναι πιο πιθανό από τα κορίτσια να διαπράξουν τον διαδικτυακό εκφοβισμό, αλλά ότι δεν υπάρχουν διαφορές μεταξύ των φύλων στα ποσοστά θυματοποίησης μεταξύ ανδρών και γυναικών [48]. Άλλες μελέτες έχουν δείξει ότι τα αγόρια είναι πιο πιθανό να διαπράξουν από τα κορίτσια διαδικτυακό εκφοβισμό, αλλά και όσον αφορά τη θυματοποίηση, τα κορίτσια είναι πιο πιθανό να πέσουν θύματα διαδικτυακού εκφοβισμού-cyberbullying [49].

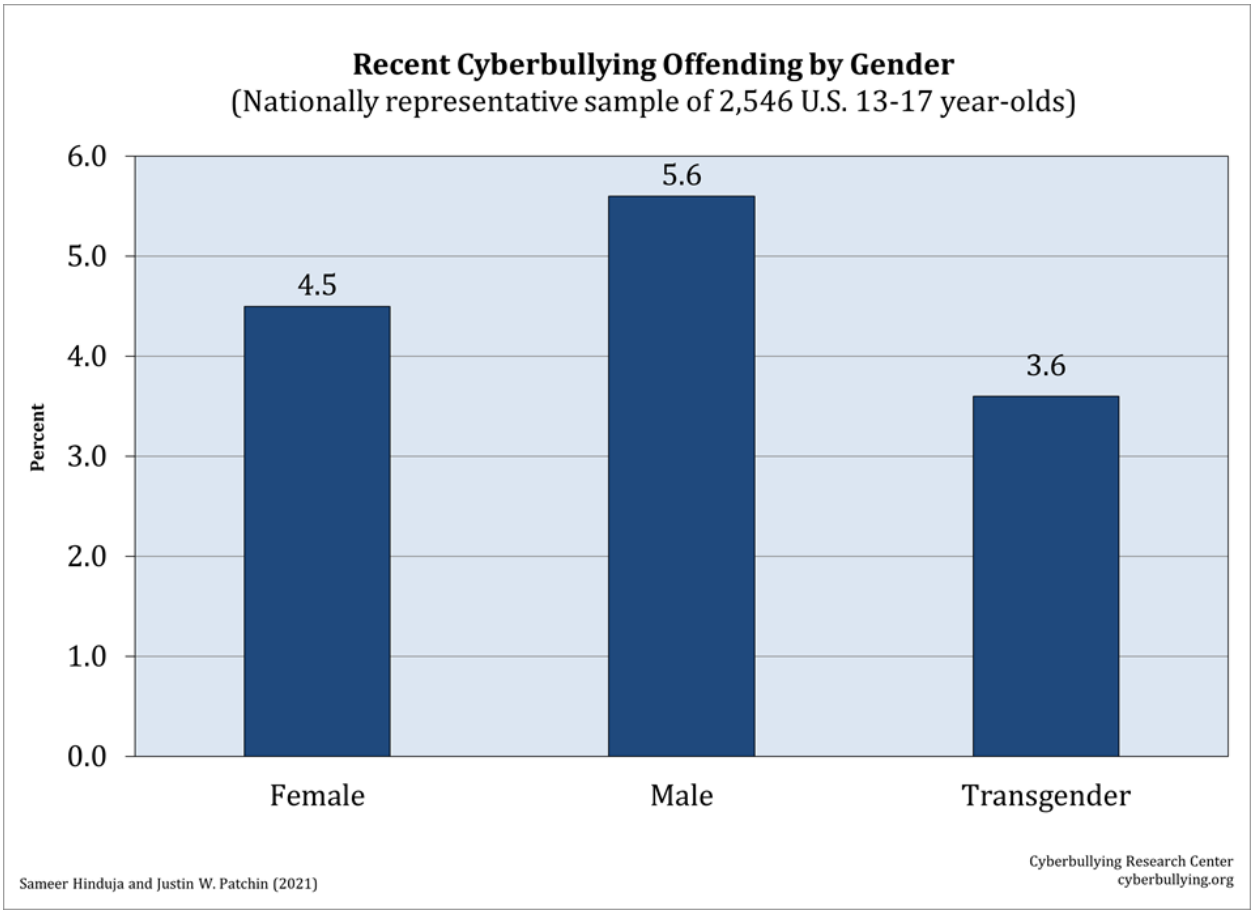
Μια τελευταία ομάδα ερευνητών προτείνει ότι οι διαφορές των φύλων εξαρτώνται από τον τόπο διεξαγωγής και εμφάνισης διαδικτυακού εκφοβισμού. Για παράδειγμα, τα κορίτσια φαίνεται να στοχοποιούνται μέσω e-mail συχνότερα από τα αγόρια [46], ενώ τα αγόρια υφίστανται εκφοβισμό μέσω γραπτών μηνυμάτων συχνότερα από τα κορίτσια [47].

Σύμφωνα με πρόσφατη έρευνα που πραγματοποιήθηκε το 2021 [60] στις Ηνωμένες Πολιτείες με δείγμα 2.546 εφήβους ηλικίας 13-17 ετών όπου αποδεδειγμένα όπως επισημάνθηκε ο εκφοβισμός φθάνει στην κορύφωσή του, εξετάζοντας τα δύο μεγάλα φύλα και τις εμπειρίες τους από εκφοβισμό, διαπιστώθηκε ότι το 23,7% των κοριτσιών και το 21,9% των αγοριών μεταξύ 13 και 17 ετών αναφέρουν ότι υπέστησαν διαδικτυακό εκφοβισμό, ενώ το 35,4% των transexual εφήβων ανέφεραν ότι υπέστησαν διαδικτυακό εκφοβισμό.



Εικόνα 38 Ποσοστά θυματοποίησης διαδικτυακού εκφοβισμού σύμφωνα με το φύλο (Hinduja&Patchin,2021)

Κατά την εξέταση του τρόπου με τον οποίο το φύλο σχετίζεται με τη διάπραξη διαδικτυακού εκφοβισμού, το 5,6% των αγοριών, το 4,5% των κοριτσιών και το 3,6% των διεμφυλικών νέων είχαν εκφοβίσει κάποιον άλλο στον κυβερνοχώρο τις τελευταίες 30 ημέρες. Αυτό το αποτέλεσμα είναι σύμφωνο με όσα έχουν βρεθεί στην πλειοψηφία των ερευνών, πως οι άνδρες είναι πιο πιθανό να εμπλέκονται σε πράξεις διαδικτυακού εκφοβισμού από τις γυναίκες, παρόλο που όπως αναφέρθηκε υπάρχουν και ερευνητές που το αμφισβητούν.



Εικόνα 39 Ποσοστό διάπραξης διαδικτυακού εκφοβισμού σύμφωνα με το φύλο (Hinduja&Patchin,2021)

Όσον αφορά το διαδικτυακό εκφοβισμό ατόμων σύμφωνα με τη σεξουαλική τους προτίμηση, γίνεται εκτενής αναφορά στην ενότητα Κατηγορία Sexism που μελετάται στο Suspicious Tweets Dataset.

3.2 Suspicious Tweets Dataset

Το δεύτερο σύνολο δεδομένων-dataset που χρησιμοποίηθηκε περιέχει και πάλι δεδομένα από την πλατφόρμα κοινωνικής δικτύωσης Twitter. Ο σύνδεσμος από τον οποίο ανακτήθηκε το συγκεκριμένο dataset είναι ο παρακάτω:

https://www.kaggle.com/datasets/munkialbright/classified-tweets?fbclid=IwAR2iqEdLn-iB2l7dY6d7S0mim0ik6F_t7JcoXQ6zoAA0jZ5O3JXvpa4jrsc

Το σύνολο δεδομένων περιέχει 19.934 κείμενα το καθένα από τα οποία διακρίνεται στις εξής κατηγορίες:

- Ύποπτο-Suspicious με τιμές 0 και 1
- Διαδικτυακός εκφοβισμός -Cyberbullying με τιμές 0,1 και 2
- Μίσος -Hate με τιμές 0 και 1
- Αυτοκτονικό -Suicidal με τιμές 0 και 1

Ο σκοπός της παρούσας διπλωματικής είναι η ανάλυση συναισθήματος με σκοπό τον εντοπισμό του διαδικτυακού εκφοβισμού και η σύγκριση διαφορετικών αλγορίθμων ως προς την τελική απόδοση. Προκειμένου να εξυπηρετηθεί αυτό το ζητούμενο, στα πλαίσια της έρευνας που πραγματοποιήθηκε, αγνοήθηκαν οι υπόλοιπες στήλες (με αφαίρεση από το σύνολο δεδομένων-dataset) και κρατήθηκε μόνο το αρχικό κείμενο (tweet) και η στήλη

Διαδικτυακός εκφοβισμός-Cyberbullying που το καταττάσσει σε μία κατηγορία διαδικτυακού εκφοβισμού όπως ακριβώς δηλαδή πραγματοποιήθηκε και στο προηγούμενο σύνολο δεδομένων.

3.2.1 Τρόπος δημιουργίας Suspicious Tweets Dataset

Το σύνολο δεδομένων αποτελεί μία τροποποίηση του συνόλου δεδομένων του Munki Albright [125] το οποίο απλά χώριζε τα tweets σε ύποπτα-suspicious(1) και μη-ύποπτα-non-suspicious(0). Και αυτό το dataset ωστόσο αποτελεί δείγμα του συνόλου δεδομένων του Syed Abbas Raza Zaidi [126] το οποίο χώριζε τα tweets σε ύποπτα-suspicious και μη -ύποπτα-non-suspicious και τα ύποπτα-suspicious διακρίνονταν περαιτέρω σε απειλητικά-threatening, τρομοκρατία-terrorism και διαδικτυακός εκφοβισμός-Cyberbullying. Στο σύνολο δεδομένων που χρησιμοποιείται ως δεύτερο στο πλαίσιο της παρούσας διπλωματικής, λοιπόν, αντί για τα αρχικά 60.000 κείμενα, χρησιμοποιείται ένα δείγμα 20.000 κειμένων τα οποία χωρίζονται στις κατηγορίες που αναφέρονται στην παραπάνω υποενότητα.

3.2.2 Χρησιμοποιούμενη γλώσσα Suspicious Tweets Dataset

Όπως και στο SOSNet Twitter Dataset, όλα τα datasets είναι στην Αγγλική γλώσσα. Τα κείμενα προέρχονται από διαφορετικούς χρήστες με αναζήτηση μέσω διεθνών hashtags όπως πριν. Συνεπώς προέρχονται από ένα αγγλόφωνο διεθνές κοινό και σε αυτήν την περίπτωση.

3.2.3 Κατηγορίες Διαδικτυακού Εκφοβισμού Suspicious Tweets Dataset

Στο συγκεκριμένο dataset οι κατηγορίες στις οποίες διακρίνεται ο διαδικτυακός εκφοβισμός καθώς και τα κείμενα που υπάρχουν σε κάθε κατηγορία παρουσιάζονται στον παρακάτω πίνακα:

Πίνακας 4 Κατηγορίες Cyberbullying και κατανομή κειμένων στο Suspicious Tweets Dataset

Κατηγορία Διαδικτυακού εκφοβισμού-Cyberbullying	Κείμενα Αρχικά
Κανένα από τα δύο-Neither	17.256
Ρατσισμός-Racism	945
Σεξισμός-Sexism	1.733

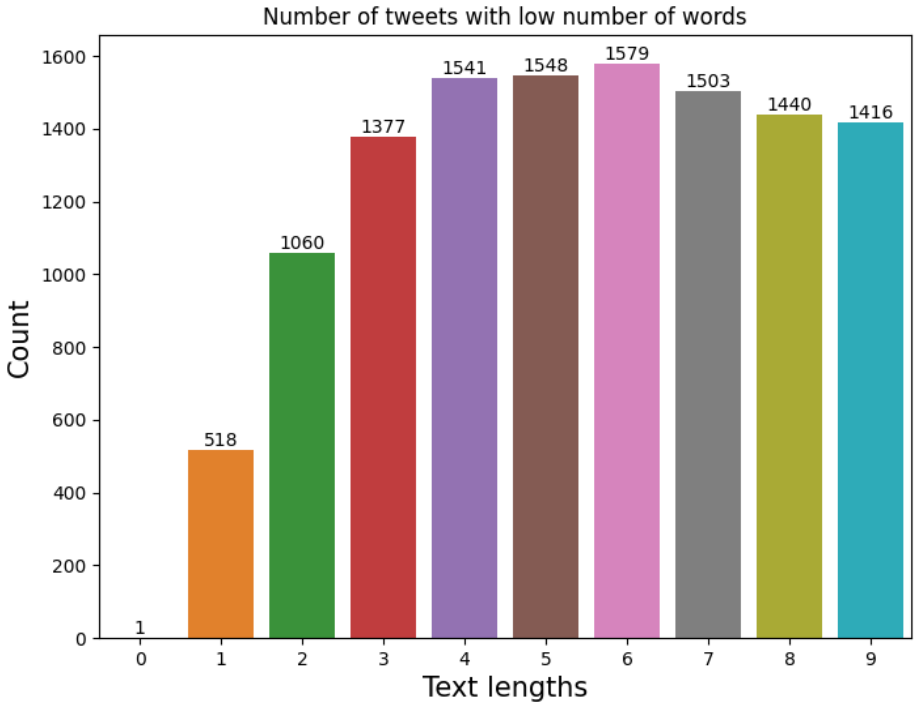
Όπως εύκολα παρατηρείται από τον παραπάνω πίνακα, οι κλάσεις που προκύπτουν δεν είναι ισορροπημένες-balanced όπως στο προηγούμενο dataset. Η κατηγορία Κανέναν από τα δύο-neither διαθέτει τη συντριπτική πλειοψηφία των κειμένων ενώ οι κατηγορίες Ρατσιστικός διαδικτυακός εκφοβισμός-racism cyberbullying και Σεξιστικός διαδικτυακός εκφοβισμός-sexism cyberbullying ακολουθούν με 945 και 1.733 κείμενα αντιστοίχως. Εξετάζονται λοιπόν 2 datasets με ποικιλομορφία, γεγονός που θα προσδώσει παραπάνω προκλήσεις στην έρευνα αφού θα εφαρμοστεί ακριβώς η ίδια μεθοδολογία.

3.2.4 Γλωσσολογικά Στοιχεία Κειμένων Suspicious Tweets Dataset

Σε αυτήν την ενότητα θα παρουσιαστούν τα γλωσσολογικά στοιχεία των κειμένων που απαρτίζουν το Suspicious Tweets Dataset όπως ακριβώς συνέβη και με το SOSNet Twitter Dataset.

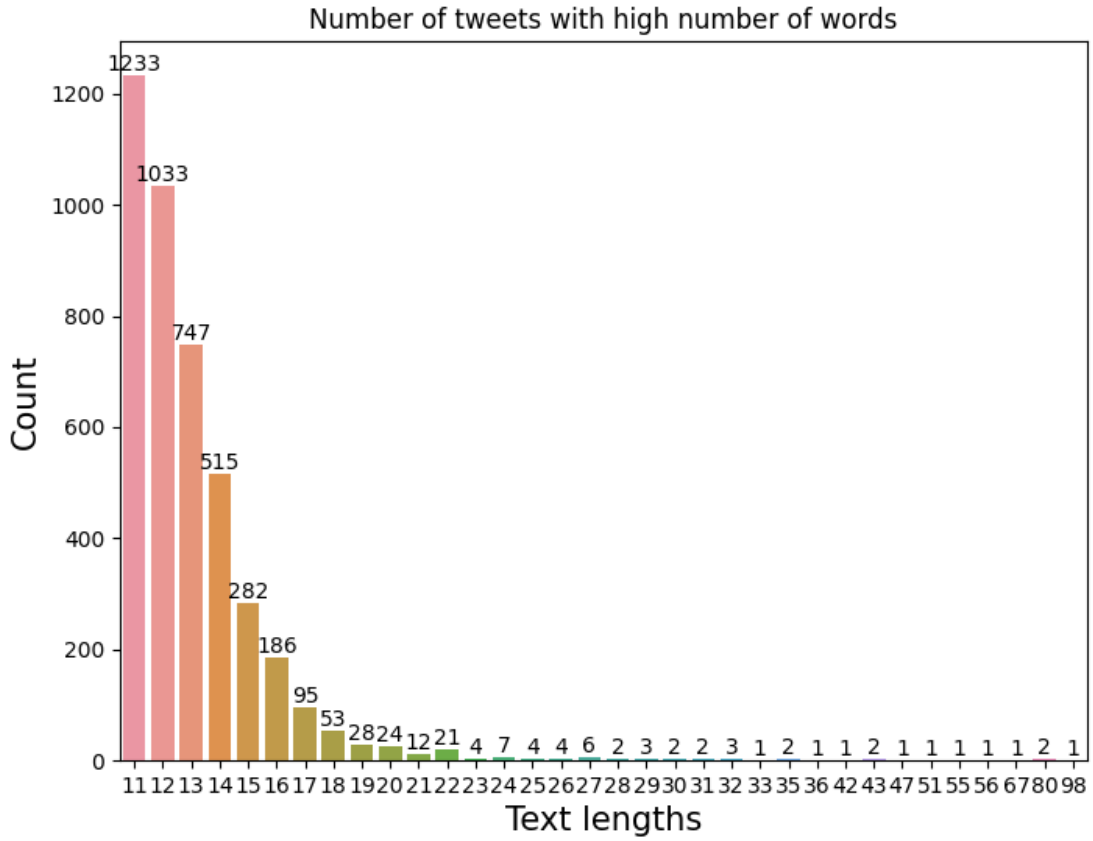
3.2.4.1 Μήκος κειμένων Suspicious Tweets Dataset

Μέσω κώδικα στη γλώσσα προγραμματισμού Python, δημιουργήθηκαν όπως και στο SOSNet Twitter Dataset, ραβδογράμματα που απεικονίζουν στον οριζόντιο άξονα τα μήκη των κειμένων και στον κατακόρυφο το πλήθος των κειμένων με το συγκεκριμένο μήκος. Για λόγους ευκολότερης ερμήνευσης και κατανόησης των διαγραμμάτων, ακριβώς επάνω από κάθε ράβδο αναγράφεται και το ακριβές πλήθος λέξεων.



Εικόνα 40 Ραβδόγραμμα με την κατανομή των κειμένων με αριθμό λέξεων από 1 έως 9 στο Suspicious Tweets Dataset

Στο παραπάνω ραβδόγραμμα αρχικά αποτυπώνονται όλα τα tweets με πλήθος λέξεων από 1 έως και 9. Παρατηρείται πως, όπως και στο προηγούμενο dataset, υπάρχει μεγάλη συγκέντρωση κειμένων με μήκος σε αυτό το διάστημα τιμών. Πιο συγκεκριμένα, και σε αυτό το dataset παρατηρούνται πολλά κείμενα με περισσότερες από 3 λέξεις με την πιο υψηλή ράβδο να αντιστοιχεί στο μήκος 6 σε αντίθεση με το προηγούμενο dataset που ήταν στις 9 λέξεις για το ίδιο διάστημα τιμών. Στο παρακάτω διάγραμμα, παρουσιάζονται και τα κείμενα με υψηλό αριθμό λέξεων.



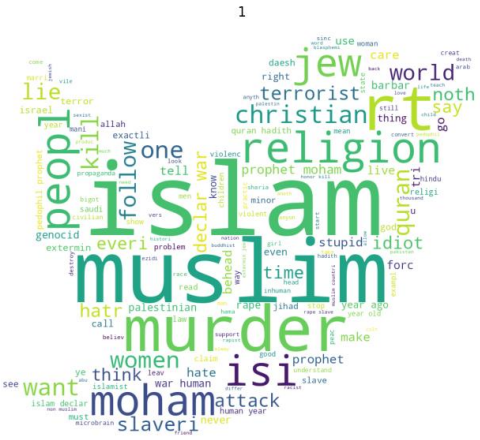
Εικόνα 41 Ραβδόγραμμα απεικόνισης του πλήθος κειμένων με υψηλό αριθμό λέξεων στο Suspicious Tweets Dataset

Σε αυτό το dataset δεν υπάρχει κανένα κείμενο με περισσότερες από 100 λέξεις καθώς, όπως φαίνεται από το διάγραμμα, το πιο μακροσκελές κείμενο που υπάρχει έχει 98 λέξεις. Συνεπώς, η υιοθέτηση ως κάτω ορίου στο μήκος λέξεων τον αριθμό 3 και ως άνω ορίου τον αριθμό 100, όπως επιλέχθηκαν και στο SOSNet Twitter Dataset, θα οδηγήσει στην αφαίρεση πολύ μικρού αριθμού κειμένων αφού δεν υπάρχει κανένα με περισσότερες από 100 λέξεις και μόλις 2.956 κείμενα με λιγότερες από 4 λέξεις. Άρα δεν είναι απαραίτητη η αναπροσαρμογή των ορίων λέξεων σε αυτό το dataset. Σημειώνεται επίσης πως αν ακολουθηθεί ξεχωριστή λογική επιλογής ορίων σε αυτό το dataset θα πρέπει, σύμφωνα με το διάγραμμα, να επιλεγεί ένας αριθμός πλησίον του 30. Ωστόσο, κείμενα με περισσότερες από 30 λέξεις δεν θεωρούνται εξαιρετικά μακροσκελή και είναι πιθανό να περιέχουν χρήσιμες πληροφορίες και, επομένως, η αφαίρεσή τους εγκυμονεί κινδύνους για τα αποτελέσματα της έρευνας.

3.2.5 Συννεφόλεξα Suspicious Tweets Dataset

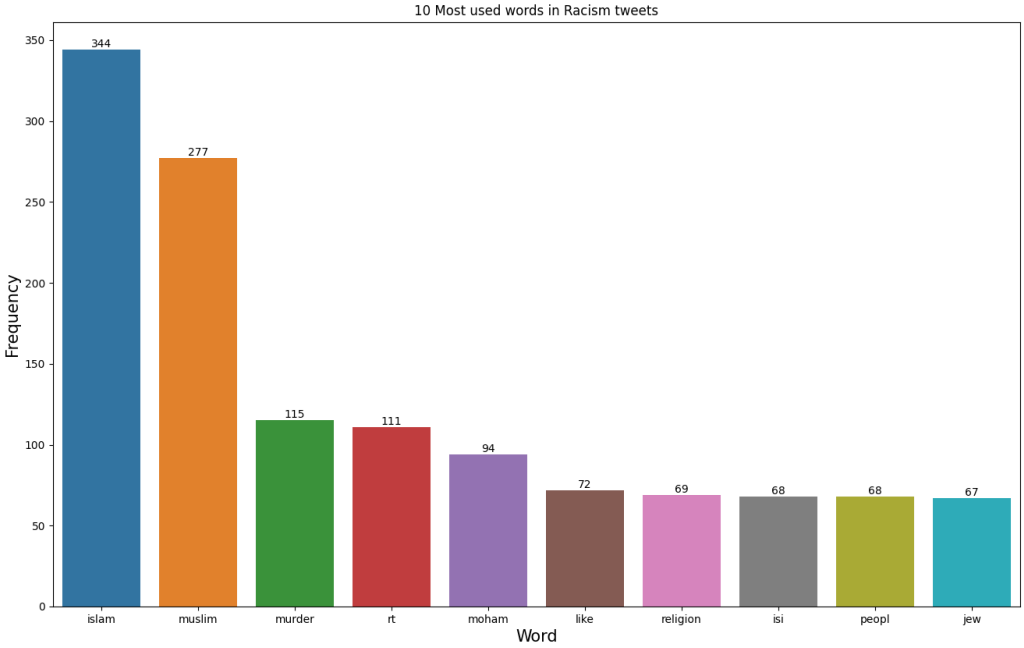
Σε αυτήν την ενότητα θα παρουσιασθούν τα συννεφόλεξα και θα επισημανθούν οι συνηθέστερες λέξεις ανά κατηγορία διαδικτυακού εκφοβισμού.

3.2.5.1 Κατηγορία Ρατσισμός-Racism Suspicious Tweets Dataset



Εικόνα 42 Συννεφόλεξο όπως ανακτήθηκε από τα κείμενα που χαρακτηρίζονται ως Ρατσιστικός διαδικτυακός εκφοβισμός-Racism cyberbullying στο Suspicious Tweets Dataset

Όπως φαίνεται από το συννεφόλεξο, οι λέξεις ‘islam’ και ‘muslim’ αποτελούν τις συνηθέστερες λέξεις σε αυτήν την κατηγορία. Όπως ακριβώς συνέβη και στο προηγούμενο σύνολο δεδομένων-dataset, παρουσιάζονται οι 10 συνηθέστερες λέξεις της κατηγορίας αυτής.

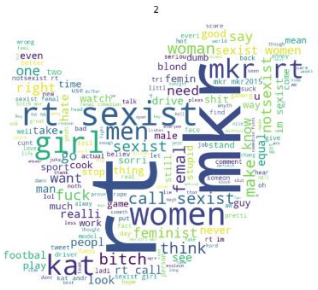


Εικόνα 43 Πλήθος εμφανίσεων 10 συνηθέστερων λέξεων στην κατηγορία Ρατσιστικός διαδικτυακός εκφοβισμός-Racism cyberbullying του Suspicious Tweets Dataset

Στον οριζόντιο άξονα του παραπάνω ραβδογράμματος αναγράφονται οι 10 λέξεις ενώ στον κατακόρυφο οι συχνότητες εμφάνισής τους. Πράγματι, οι λέξεις ‘islam’ και ‘muslim’ ,όπως φάνηκε και από το ανακτηθέν συννεφόλεξο, είναι οι επικρατέστερες με πλήθος εμφανίσεων 344 και 277 αντιστοίχως. Αυτές οι 2 λέξεις βρίσκονταν ανάμεσα στις 10 συνηθέστερες και στο SOSNet Twitter Dataset στην κατηγορία Θρησκευτικός διαδικτυακός εκφοβισμός-Religion cyberbullying. Ωστόσο, δεν υπάρχει κάποια άλλη κοινή λέξη για την πρώτη δεκάδα λέξεων. Σε αυτό το σύνολο δεδομένων-dataset υπάρχει η λέξη ‘moham’ που επίσης συνδέεται με τον Μουσουλμανισμό αλλά και η ‘jew’ που συνδέεται με την Εβραϊκή θρησκεία. Όπως υπογραμμίστηκε σε προηγούμενη ενότητα (Κατηγορία Religion), τα άτομα που ασπάζονται τη Μουσουλμανική θρησκεία υπόκεινται συχνότερα σε διαδικτυακό εκφοβισμό σε σύγκριση με τις υπόλοιπες θρησκείες.

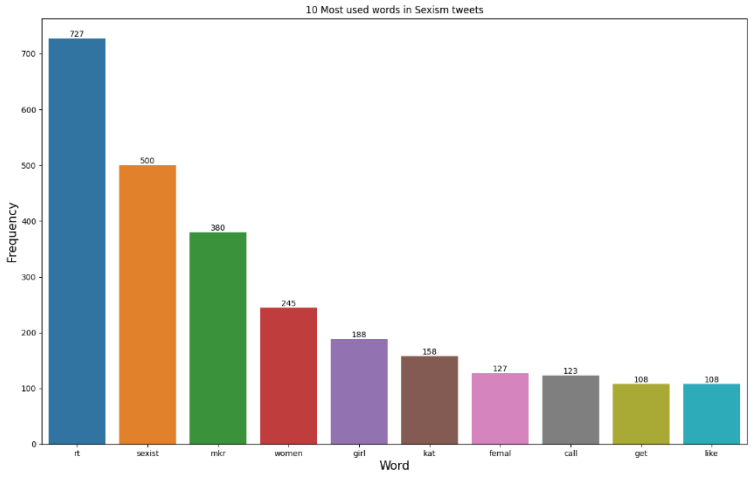
3.2.5.2 Κατηγορία Σεξισμός-Sexism Suspicious Tweets Dataset

Το συννεφόλεξο που ανακτήθηκε από τα tweets της εν λόγω κατηγορίας είναι το ακόλουθο:



Εικόνα 44 Συννεφόλεξο όπως ανακτήθηκε από την κατηγορία Σεξιστικός διαδικτυακός εκφοβισμός-Sexism cyberbullying του Suspicious Tweets Dataset

Προκειμένου να υπάρξει μία σαφέστερη εικόνα, παρουσιάζεται επίσης ,όπως και σε όλες τις προηγούμενες κατηγορίες, ραβδόγραμμα με τη συχνότητα εμφάνισης των 10 συνηθέστερων λέξεων της κατηγορίας. Στον οριζόντιο άξονα επισημαίνεται η λέξη ενώ στον κατακόρυφο η συχνότητα με την οποία παρουσιάσθηκε στα κείμενα. Επιπλέον, για λόγους ευκολότερης ανάγνωσης, πάνω από κάθε ράβδο αναγράφεται και ο ακριβής αριθμός εμφανίσεων.

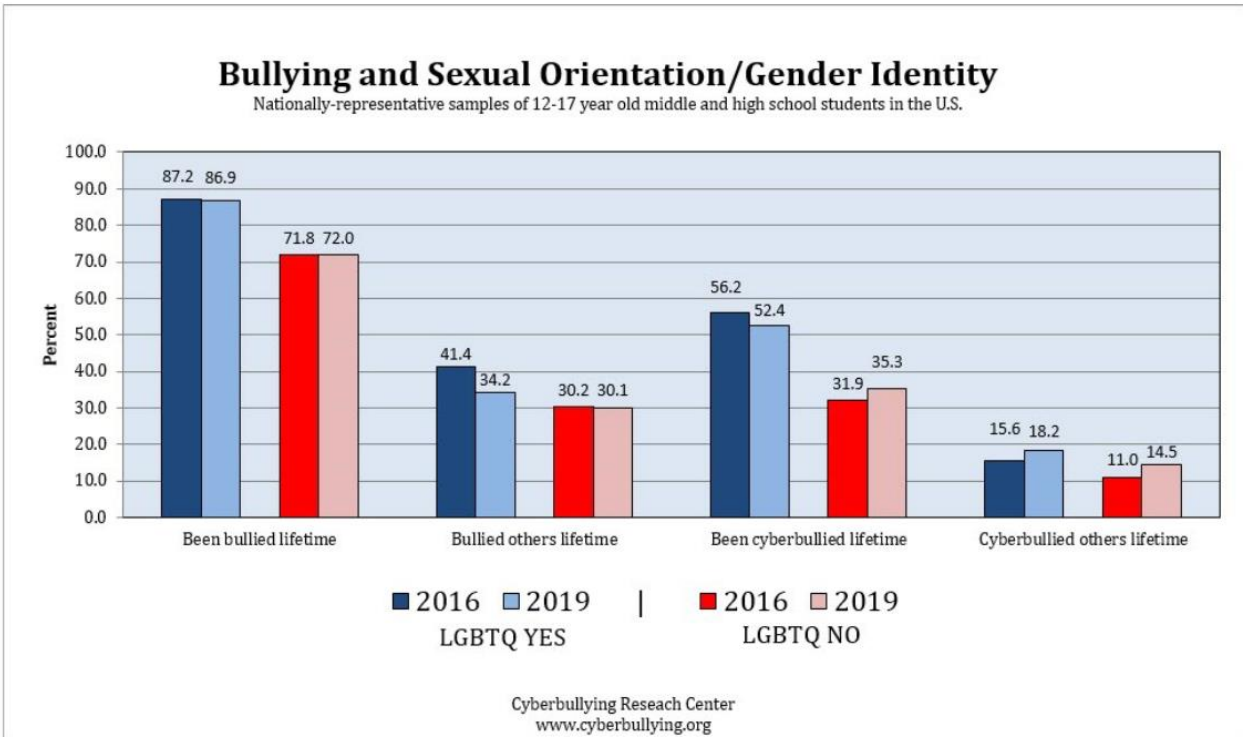


Εικόνα 45 Ραβδόγραμμα συχνοτήτων εμφάνισης 10 συνηθέστερων λέξεων στην κατηγορία Σεξιστικός διαδικτυακός εκφοβισμός-Sexism cyberbullying του Suspicious Tweets Dataset

Η συχνότερη λέξη είναι η ‘rt’ με 727 εμφανίσεις και η αμέσως επόμενη η λέξη ‘sexist’ με 500. Ανάμεσα στις 10 συχνότερες βρίσκονται και οι λέξεις ‘women’, ‘girl’, ‘femal’ ενισχύοντας για ακόμη μία φορά όσα επισημάνθηκαν στην ενότητα Κατηγορία Gender σχετικά με τον εκφοβισμό που δέχονται σύμφωνα με την πλειοψηφία των ερευνητών οι γυναίκες. Στην παρούσα ενότητα, θα παρουσιασθούν στοιχεία σχετικά με τον εκφοβισμό που δέχονται τα άτομα με βάση τον σεξουαλικό τους προσανατολισμό, σύμφωνα με την υπάρχουσα βιβλιογραφία.

Σε τακτική βάση σε πολυάριθμες μελέτες και χρόνια, η έρευνα είναι σαφής ότι εκείνοι οι νέοι που ανήκουν στη σεξουαλική μειονότητα είναι πιο πιθανό να βιώσουν εκφοβισμό και διαδικτυακό εκφοβισμό από τους ετεροφυλόφιλους συνομηλίκους τους [51-54].

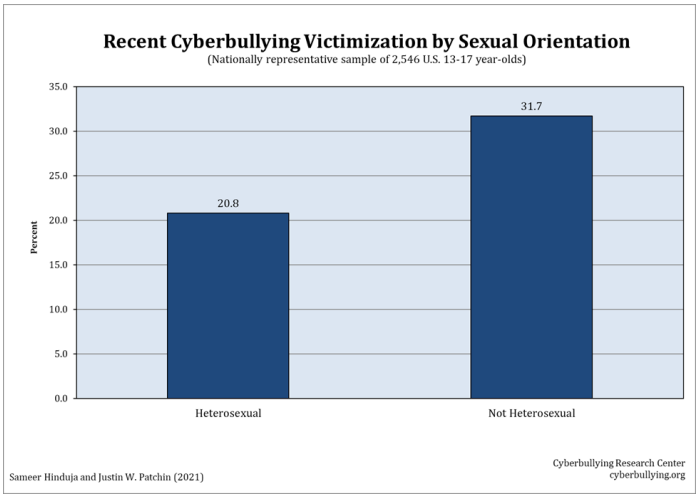
Το 2019, ερευνήθηκε ένα δείγμα 4.500 μαθητών από όλες τις ΗΠΑ. Τα αποτελέσματα αυτής της μελέτης [53] συγκριτικά με τα αντίστοιχα του 2016 φαίνονται στην παρακάτω εικόνα. Μεταξύ των μαθητών LGBTQ, το 87% είχε δεχτεί εκφοβισμό στο σχολείο και το 52% είχε δεχτεί εκφοβισμό στο διαδίκτυο σε ορισμένο σημείο στη ζωή τους (σε σύγκριση με 72% και 35%, εκ νέου αντίστοιχα, για μαθητές που δεν είναι LGBTQ). Είναι σαφές από την έρευνά κατά τη διάρκεια των ετών ότι οι LGBTQ μαθητές βιώνουν περισσότερο εκφοβισμό και διαδικτυακό εκφοβισμό από ότι οι μη-LGBTQ μαθητές.



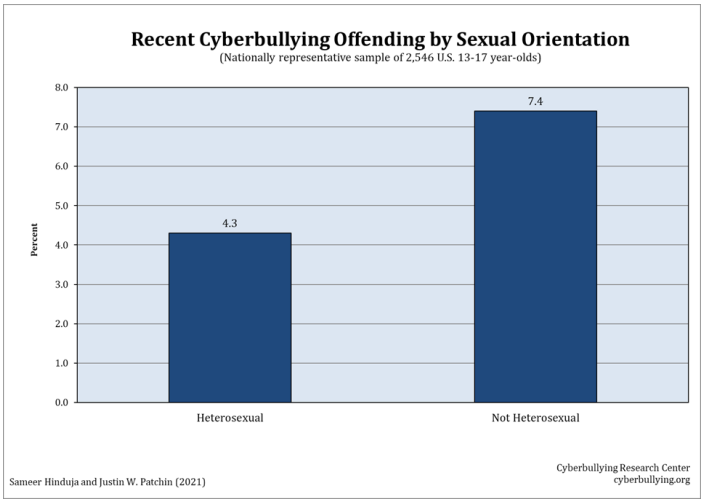
Εικόνα 46 Ποσοστά θυμάτων εκφοβισμού και διαδικτυακού εκφοβισμού για μαθητές που είναι ή δεν είναι μέλη της LGBTQ κοινότητας (Hinduja&Patchin,2020)

Ένα ακόμη εντυπωσιακό στοιχείο που προκύπτει από την εν λόγω έρευνα είναι πως τα μέλη της LGBTQ κοινότητας εκτός από συχνότεροι αποδέκτες του διαδικτυακού και μη-εκφοβισμού, είναι και συχνότερα οι θύτες (18.2% έναντι 14.5% για το 2019).

Το ίδιο ακριβώς αποτέλεσμα προκύπτει και από έρευνα που πραγματοποιήθηκε το 2021 [60]. Όπως φαίνεται στις παρακάτω εικόνες, τα ομοφυλόφιλα άτομα πέφτουν θύματα διαδικτυακού εκφοβισμού συχνότερα από τα ετεροφυλόφιλα (31.7% έναντι 20.8%) αλλά είναι και συχνότερα οι θύτες (7.4% έναντι 4.3%).



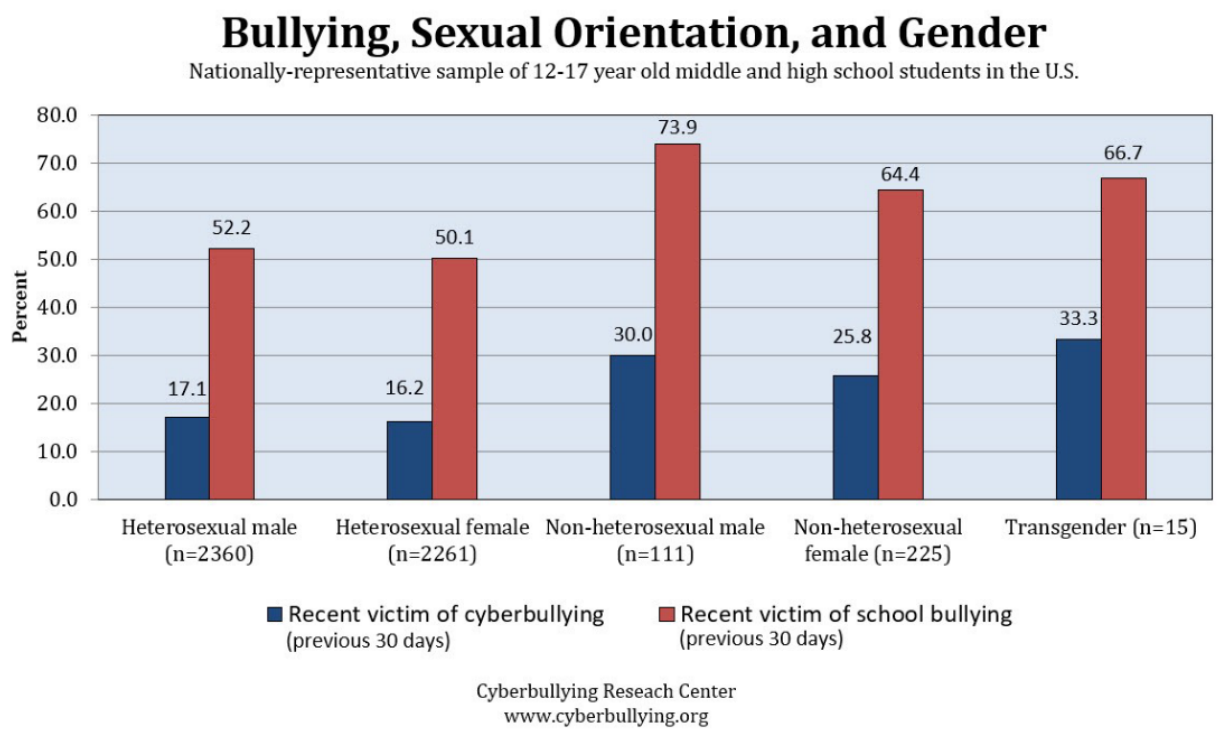
Εικόνα 47 Ποσοστά θυματοποίησης διαδικτυακού εκφοβισμού σύμφωνα με τη σεξουαλική προτίμηση (Hinduja&Patchin,2021)



Εικόνα 48 Ποσοστά διάπραξης διαδικτυακού εκφοβισμού ανάλογα με τη σεξουαλική προτίμηση (Hinduja&Patchin,2021)

Αυτό δεν είναι ασυνήθιστο εύρημα στην έρευνα [55,56] και ορισμένες εξηγήσεις επικεντρώνονται στην προηγούμενη διαπίστωση ότι οι σεξουαλικές μειονότητες είναι πιο πιθανό να στοχοποιηθούν. Δηλαδή, εάν κάποιος στοχοποιηθεί με κάποια μορφή διαδικτυακού εκφοβισμού, μπορεί να το δει ως κανονιστική ή απαραίτητη συμπεριφορά. Θα μπορούσε επίσης να είναι μια μορφή αντιποίνων, η οποία είναι μια συχνή εξήγηση που παρέχεται από εκείνους που παραδέχονται ότι εκφοβίζουν άλλους [57,58]

Σε αυτό το σημείο, γίνεται σύνδεση του sexism cyberbullying με το gender cyberbullying που μελετήθηκε στο προηγούμενο dataset. Όπως προκύπτει αναλύοντας ακόμη περισσότερο τα δεδομένα για το 2019 [53], διαπιστώνεται ότι οι **μη-ετεροφυλόφιλοι άνδρες** ήταν περισσότερο πιθανόν να έχουν υποστεί εκφοβισμό στο σχολείο (73,9%) και online (30%) τις πιο πρόσφατες 30 ημέρες πριν την έρευνα. Η μόνη ομάδα με υψηλότερο ποσοστό διαδικτυακού εκφοβισμού είναι οι trans μαθητές, οι οποίοι ήταν ελαφρώς πιο πιθανό να έχουν υποστεί εκφοβισμό διαδικτυακά (33,3%). Ωστόσο, το μέγεθος του δείγματός ήταν πολύ χαμηλό για αυτήν την ομάδα (n=15) για να εξαχθεί οποιοδήποτε ασφαλές συμπέρασμα ή να πραγματοποιηθεί κάποια χρήσιμη σύγκριση. Οι ετεροφυλόφιλες μαθήτριες ήταν η λιγότερο πιθανή ομάδα να έχει δεχθεί εκφοβισμό στο σχολείο ή στο διαδίκτυο. Τέλος, ένα ακόμη αξιοσημείωτο αποτέλεσμα είναι πως οι μη-ετεροφυλόφιλες γυναίκες έχουν υποστεί σε χαμηλότερα ποσοστά εκφοβισμό και διαδικτυακό εκφοβισμό από τους μη-ετεροφυλόφιλους άντρες (25.8% και 64.4% έναντι 30% και 73.9% αντιστοίχως).



Εικόνα 49 Ποσοστά εκφοβισμού και διαδικτυακού εκφοβισμού εν συναρτήσει του φύλου-gender και σεξουαλικού προσανατολισμού-sexual orientation (Hinduja&Patchin,2020).

3.3 Διαφορές των συνόλων δεδομένων-datasets

Είναι σημαντικό να αποσαφηνιστούν οι διαφορές που υπάρχουν μεταξύ των 2 datasets που επιλέχθηκαν, δηλαδή του SOSNet Twitter Dataset και του Suspicious Tweets Dataset. Οι διαφορές τους είναι οι εξής:

1. Το SOSNet Twitter Dataset έχει προκύψει από 6 διαφορετικά datasets και αποτελείται από 47.692 tweets, σε αντίθεση με το Suspicious Tweets Dataset το οποίο αποτελείται από 19.934 κείμενα και είναι ένα δείγμα ενός και μόνου dataset με 60.000 κείμενα
2. Το SOSNet Twitter Dataset χωρίζει τα tweets σε 5 κατηγορίες. Αυτές είναι: Age cyberbullying (Ηλικιακός διαδικτυακός εκφοβισμός), Religion cyberbullying (Θρησκευτικός διαδικτυακός εκφοβισμός), Gender cyberbullying (Διαδικτυακός εκφοβισμός φύλου), Ethnicity cyberbullying (Διαδικτυακός εκφοβισμός εθνικότητας) και Not cyberbullying (Όχι διαδικτυακός εκφοβισμός). Από την άλλη μεριά το Suspicious Tweets Dataset κατηγοριοποιεί τα tweets σε 3 κατηγορίες: Racism (Ρατσισμός), Sexism (Σεξισμός) και Neither (Κανένα από τα δύο).
3. Στο SOSNet Twitter Dataset οι κατηγορίες είναι ισορροπημένες δηλαδή σε κάθε κατηγορία υπάρχουν περίπου 8.000 κείμενα. Αντίθετα, στο Suspicious Twitter Dataset δεν ισχύει κάτι τέτοιο αφού το 86,56% των κειμένων δεν σχετίζονται με διαδικτυακό εκφοβισμό, το 4,74% σχετίζεται με ρατσισμό και το υπόλοιπο 8,69% με σεξισμό.
4. Το Suspicious Tweets Dataset είναι πιο πρόσφατο αφού η τελευταία του ενημέρωση χρονολογείται στο 2022 ενώ το SOSNet Twitter Dataset δημιουργήθηκε το 2020.

3.4 Λόγοι επιλογής συνόλων δεδομένων-datasets

Σε αυτό το σημείο θα διευκρινιστούν οι λόγοι για τους οποίους επιλέχθηκαν τα συγκεκριμένα σύνολα δεδομένων για την εφαρμογή και σύγκριση των αλγορίθμων. Οι λόγοι επιλογής είναι οι ακόλουθοι:

1. Χρησιμοποιούν και τα 2 δεδομένα από το Twitter. Σύμφωνα με την πηγή [127], το Twitter είναι μία αρκετά καλή πηγή δεδομένων διότι:Α) Είναι το πιο ανοιχτό και προσβάσιμο σε σύγκριση με τις υπόλοιπες πλατφόρμες κοινωνικής δικτύωσης. Αυτό κάνει το Twitter πιο ευνοϊκό για προγραμματιστές που δημιουργούν εργαλεία πρόσβασης σε δεδομένα και κατά συνέπεια αυξάνει τη διαθεσιμότητα λογισμικού και διαδικτυακών εργαλείων στους ερευνητές. Αντίθετα, τα δεδομένα από το Facebook είναι πολύ δύσκολο να ληφθούν και είναι διαθέσιμα μόνο σε συγκεντρωτικό επίπεδο για σκοπούς μάρκετινγκ. Β) Το Twitter διευκολύνει την εύρεση και παρακολούθηση συνομιλιών, καθώς διαθέτει μια δυνατότητα αναζήτησης που επιτρέπει στους χρήστες να αναζητούν tweets και τα tweets εμφανίζονται επίσης στα αποτελέσματα αναζήτησης Google, διευκολύνοντας τον εντοπισμό τους. Το Facebook μπορεί να θεωρηθεί περισσότερο ως μια ιδιωτική πλατφόρμα και δεν εμφανίζονται όλες οι δημόσιες αναρτήσεις στα αποτελέσματα της Αναζήτησης Google. Το Facebook παρέχει επίσης στους χρήστες περισσότερους ελέγχους απορρήτου. Γ) Το Twitter έχει μια ισχυρή κουλτούρα hashtag που διευκολύνει τη συλλογή, την ταξινόμηση και την επέκταση αναζήτησης κατά τη συλλογή δεδομένων. Δ) Το Twitter μπορεί να είναι μια δημοφιλής πλατφόρμα λόγω της προσοχής που μπορεί να λάβει από το mainstream μέσα ενημέρωσης και μπορεί να προσελκύσει περισσότερη έρευνα λόγω της πολιτιστικής του θέσης. Συνεπώς, τα συγκεκριμένα σύνολα δεδομένων έχοντας αξιοποιήσει τις παραπάνω δυνατότητες που παρέχονται από το Twitter καθίστανται ως μία καλή επιλογή για τη συγκέντρωση κειμένων από Αγγλόφωνους χρήστες σε όλα τα μήκη της Γης που σχολιάζουν τη διεθνή επικαιρότητα.
2. Είναι 2 συμπληρωματικές περιπτώσεις. Το SOSNet Twitter Dataset συνθέτει μία αρκετά καλή επιλογή για μελέτη καθώς όλες οι κατηγορίες είναι ισορροπημένες, δηλαδή, περίπου το ίδιο πλήθος Tweets που χαρακτηρίζονται ως Ηλικιακός διαδικτυακός εκφοβισμός, χαρακτηρίζονται και ως μη-διαδικτυακός εκφοβισμός ή ως κάποια από τις υπόλοιπες κατηγορίες. Αυτή η συνθήκη είναι ιδανική για την εφαρμογή αλγορίθμων, αλλά δεν συνθέτει μία ρεαλιστική εικόνα για τα κοινωνικά δίκτυα. Το Suspicious Tweets Dataset συμπληρώνει αυτήν την έλλειψη, καθώς σε αυτό οι κατηγορίες είναι μη-ισορροπημένες με τα Tweets που δεν εμφανίζεται διαδκτυακός εκφοβισμός να υπερισχύουν. Αυτή η κατανομή, σύμφωνα με την πηγή [122], είναι αρκετά πιο αντιπροσωπευτική της πραγματικότητας.
3. Και τα 2 datasets είναι αρκετά πρόσφατα - και ειδικά το Suspicious Tweets Dataset, που ενημερώθηκε τελευταία φορά το 2022.
4. Για τους σκοπούς της παρούσας μελέτης (Διπλωματικής) ήταν επιθυμητό, εκτός από την εφαρμογή και σύγκριση διαφόρων αλγορίθμων, να εξεταστεί ο διαδικτυακός εκφοβισμός ως προς τα χαρακτηριστικά των θυμάτων και θυτών. Επομένως, ήταν επιθυμητό να βρεθούν δεδομένα που δεν θα κάνουν έναν απλό δυαδικό χωρισμό στις κατηγορίες «Διαδικτακός εκφοβισμός» ή «Όχι διαδικτυακός εκφοβισμός» αλλά θα αναλύουν περαιτέρω και το είδος του Διαδικτυακού εκφοβισμόυ σε περίπτωση που αυτός υφίσταται.
5. Τέλος, οι κατηγορίες στις οποίες γίνεται ο επιπλέον διαχωρισμός μπορεί να μην είναι οι ίδιες, όμως είναι αρκετά κοντινές. Για παράδειγμα, το SOSNet Twitter Dataset διαθέτει την κατηγορία «Gender Cyberbullying (Διαδικτυακός εκφοβισμός

με βάση το φύλο)» και το Suspicious Tweets Dataset διαθέτει την κατηγορία «Sexism (Σεξισμός)». Συνεπώς, θεωρήθηκε πιο ωφέλιμος ο συνδυασμός τους σε σύγκριση με τη χρήση ενός πολύ διαφορετικού dataset που θα χώριζε π.χ τα tweets σε Hate (Μίσος), Offensive (Προσβλητικά) κλπ.

4. ΜΕΘΟΔΟΛΟΓΙΑ

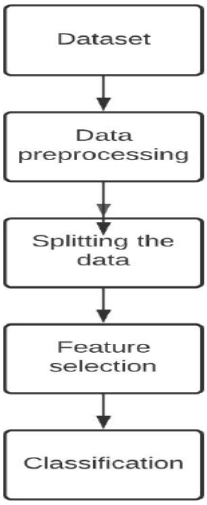
4.1 Γενική Περιγραφή

Στην παρούσα ενότητα, δίνονται περιεκτικά τα βήματα και οι μέθοδοι που ακολουθούνται προκειμένου να πραγματοποιηθεί κατηγοριοποίηση κειμένου και να κατασκευαστεί, εν τέλει, ένα σύστημα ικανό να προβλέπει με μεγάλη ακρίβεια τη σωστή κατηγορία που εντάσσεται το κείμενο.

Σε κάθε πρόβλημα κατηγοριοποίησης ακολουθούνται τα εξής 5 γενικά βήματα [1]:

- Συλλογή δεδομένων (Dataset)
- Προεπεξεργασία Δεδομένων (Data Preprocessing)
- Διαχωρισμός δεδομένων (Splitting the data)
- Επιλογή χαρακτηριστικών (Feature Selection)
- Κατηγοριοποίηση (Classification)

Τα προαναφερθέντα βήματα συνοψίζονται στην παρακάτω εικόνα:



Εικόνα 50 Γενική μεθοδολογία σε κάθε πρόβλημα που περιλαμβάνει κατηγοριοποίηση κειμένου (Mehendale, Rajpara, Shah & Phadtare, 2022)

Καθεμία από αυτές τις τεχνικές σχετίζεται με ποικίλες προσεγγίσεις. Αρχικά, θα παρουσιασθούν οι διάφορες προσεγγίσεις/μεθοδολογίες που προτείνονται από τη σύγχρονη βιβλιογραφία και, εν τέλει, σε ξεχωριστή υποενότητα, η μεθοδολογία που υιοθετήθηκε στην παρούσα διπλωματική.

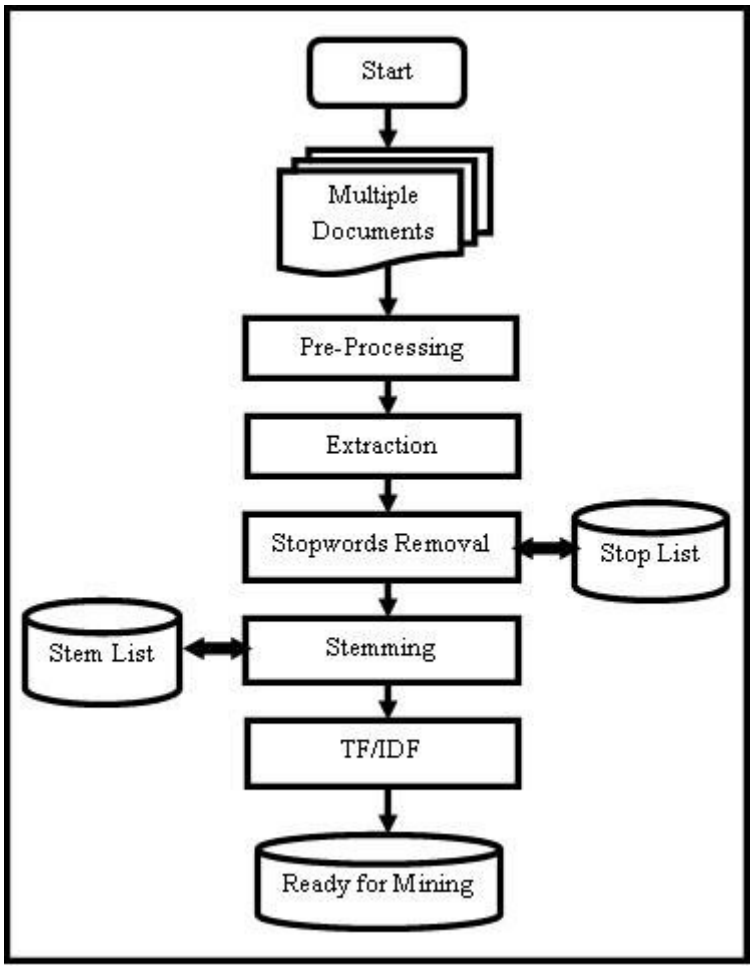
4.1.1 Προεπεξεργασία δεδομένων

Σε ένα NLP σύστημα είναι αναγκαίο να πραγματοποιείται προεπεξεργασία κειμένου [2] προκειμένου να.:

- Μειωθεί το μέγεθος του αρχείου ευρετηρίου (ή δεδομένων) των εγγράφων κειμένου
    1. Τα stopwords αντιστοιχούν στο 20-30% του συνολικού πλήθους λέξεων ενός συγκεκριμένου κειμένου.
    2. Η τεχνική αναγωγής μίας λέξης στο στέλεχός της, αφαιρώντας επιθήματα και προθέματα (stemming) είναι ικανή να μειώσει το μέγεθος του ευρετηρίου κατά 40-50%.
- Βελτιωθεί η αποδοτικότητα και η αποτελεσματικότητα του συστήματος IR (ανάκτησης πληροφορίας).

1. Τα stopwords δεν είναι χρήσιμα για αναζήτηση ή εξόρυξη κειμένου και μπορεί να μπερδέψουν το σύστημα ανάκτησης.
2. Η τεχνική stemming χρησιμοποιείται για ταίριασμα των παρόμοιων λέξεων σε διαφορετικά κείμενα.

Οι βασικές τεχνικές προεπεξεργασίας κειμένων συνοψίζονται στην παρακάτω εικόνα:



Εικόνα 51 Βασικές Τεχνικές Προεπεξεργασίας Κειμένου (Mohan, Vijayarani,2015)

4.1.1.1 Αφαίρεση Stopwords

Τα stopwords είναι μία διαίρεση της φυσικής γλώσσας. Το κίνητρο αφαίρεσης τους είναι πως κάνουν το κείμενο να φαίνεται “βαρύτερο” και λιγότερο σημαντικό για τους αναλυτές. Αφαιρώντας τα stopwords μειώνεται και η διάσταση του χώρου των όρων. Οι πιο συνηθισμένες λέξεις σε κείμενα είναι άρθρα, προθέσεις και αντωνυμίες, κλπ. που δεν δίνουν το νόημα των κειμένων. Αυτές οι λέξεις αντιμετωπίζονται σαν stopwords. Παραδείγματα stopwords της Αγγλικής γλώσσας είναι : the, in, a, an, with κλπ. Αυτές οι λέξεις αφαιρούνται από τα κείμενα διότι δεν θεωρούνται λέξεις-κλειδιά σε εφαρμογές εξόρυξης κειμένου [1].

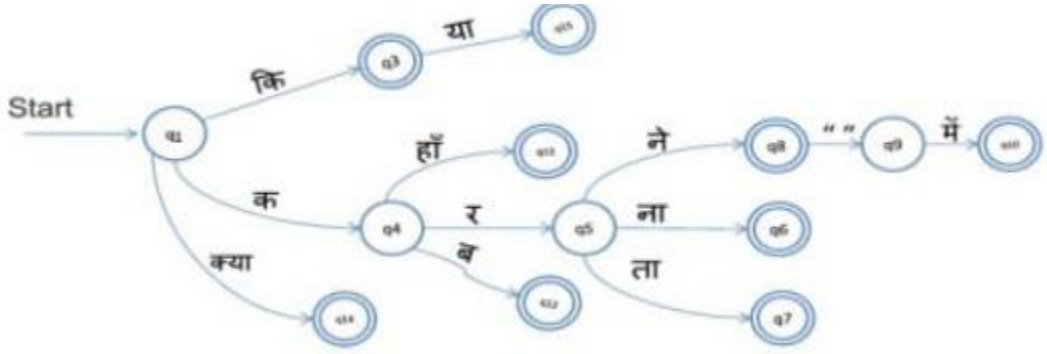
Στην παρούσα υποενότητα, παρουσιάζονται όλες οι τεχνικές εντοπισμού stopwords.

Α. Στατικές μέθοδοι εντοπισμού.

Σε μια στατική προσέγγιση, δημιουργείται μια προκαθορισμένη λίστα ενδιάμεσων λέξεων για τη συγκεκριμένη γλώσσα. Η ίδια η λίστα δεν θα χρειαζόταν να ανανεώνεται ή να αλλάζει αυτόματα [15].

Κλασσική Μέθοδος. Αυτή είναι μια βασική τεχνική [16] στην οποία τα στοχευμένα έγγραφα διακρίνονται σε λέξεις (tokenization). Χρησιμοποιεί μια στατική λίστα από stopwords για να προσδιορίσει τους συγκεκριμένους όρους ανάμεσα στους όρους του εγγράφου. Η λίστα γίνεται χειροκίνητα.

Ντετερμινιστικά πεπερασμένα αυτόματα. Δημιουργεί πεπερασμένα αυτόματα [17] για να προσδιορίσει τα stopwords όπως φαίνεται στην παρακάτω εικόνα. Το DFA αποτελείται από 5 παραμέτρους οι οποίες είναι: Κατάσταση, χαρακτήρας, μετάβαση, αρχική κατάσταση και κατάσταση αποδοχής. Οι συνθήκες εκκίνησης ελέγχονται για κάθε χαρακτήρα και οι μεταβάσεις γίνονται στην επόμενη κατάσταση. Εάν δεν υπάρχουν μεταβάσεις, επιστρέφεται το false. Εάν μεταβάσεις ακολουθίας στα σύμβολα εισαγωγής της λέξης τελειώνουν σε κατάσταση αποδοχής, η λέξη θεωρείται stopword και επιστρέφει true.



Εικόνα 52 Ντετερμινιστικό Πεπερασμένο Αυτόματο για την Αναγνώριση stopwords (Jha, Manjunath, Shenoy, & Venugopal, 2016)

Μέθοδος λεξιλογικής τάξης. Σε αυτή τη μέθοδο [18], με τη βοήθεια ειδικών φωνητικής και βαθιά εκμάθηση της δομής της γλώσσας Γκουτζαράτι, δημιουργείται λίστα με μοναδικές λέξεις και τους ανατίθενται λεξικές τάξεις δηλ. Ουσιαστικό, Ρήμα, Επίρρημα κ.λπ. Στο επόμενο βήμα, μετριέται η συχνότητα των υψηλότερων σε συχνότητα λέξεων σύμφωνα με τις λεξικές τάξεις.

Β.Δυναμικές Μέθοδοι Εντοπισμού

Στη δυναμική προσέγγιση, το stopword εντοπίζεται εν κινήσει. Μια λίστα stopwords θα δημιουργηθεί σύμφωνα με κανόνες ή στατιστικά στοιχεία & δεν προκαθορίζεται από πριν. Ο κατάλογος των λέξεων αποφασίζεται με βάση μία δεδομένη είσοδο κειμένου.

Μέθοδος Προσέγγισης βασισμένη σε κανόνες.

Σε αυτή την τεχνική [19], οι στατικοί κανόνες προκύπτουν μετά από ανάλυση του προτύπου του stopword. Εάν οποιαδήποτε λέξη στο κείμενο ικανοποιεί τον κανόνα, τότε προσδιορίζεται ως stopword.

Μέθοδοι βασισμένες στο νόμο του Zipf.

Ο νόμος του Zipf [20] είναι νόμος σχετικά με τη συχνότητα κατανομής λέξεων σε μια γλώσσα. Για να απεικονιστεί ο νόμος του Zipf, ας υποθέσουμε ότι έχουμε μια συλλογή κειμένου και ας υπάρχουν V (λεξιλόγιο) μοναδικές λέξεις στη συλλογή. Για κάθε όρο στη συλλογή χρειάζεται να υπολογιστεί η συχνότητα δηλαδή Freq(word)= Συνολικές εμφανίσεις του όρου στο κείμενο. Στη συνέχεια, αποδίδουμε κατάταξη στις λέξεις σε φθίνουσα σειρά με βάση τη συχνότητά τους (η λέξη με τη μεγαλύτερη συχνότητα έχει

κατάταξη 1, επόμενη λέξη συχνότητας έχει την κατάταξη 2 και ούτω καθεξής). ο νόμος του Zipf εκφράζει ότι «Δεδομένου ενός μεγάλου δείγματος λέξεων που χρησιμοποιούνται, η συχνότητα οποιασδήποτε λέξης είναι αντιστρόφως ανάλογη με την κατάταξή της στη συχνότητα.

Τυχαία δειγματοληψία με βάση όρους (TBRS)

Αυτή η προσέγγιση εντοπίζει χειροκίνητα τη λέξη τερματισμού από το έγγραφο ιστού. Αξιοποιώντας το μέτρο διαφοράς Kullback-Leibler καθώς επαναλαμβάνεται πάνω σε τυχαία επιλεγμένα ξεχωριστά τμήματα των δεδομένων, κατατάσσει τους όρους πληροφοριών σε κάθε τμήμα ανάλογα με την αξία τους σε πληροφορία.

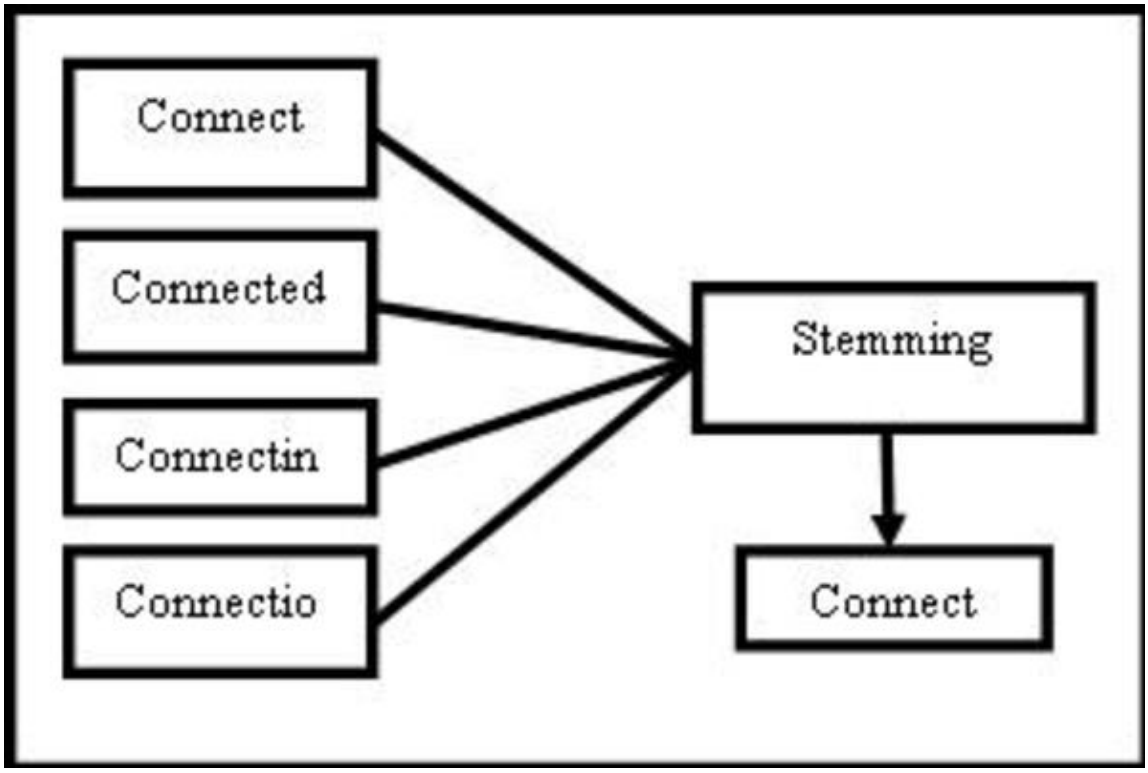
Ακολουθεί πίνακας σύγκρισης των παραπάνω προσεγγίσεων.

References	Technique	Implemented on Language	Advantages	Disadvantages
Saini et al. [21]	Classic Approach	Sanskrit	• Basic Technique • Easy to implement	• Lacks potentially new words • Defined for general purpose i.e. different collection require stopword list • Outdated • Time complexity is high
Rakholia et al. [6]	Lexical Classes Approach	Gujarati	Best for Machine Translation	As of date no tool available for Gujarati language to assign automatic POS to word
Jha et al. [7]	Deterministic Finite Automata	Hindi	Take less time as compare to dictionary-based approach	• Limited word length • Require more space to store data
Rakholia et al. [8]	Rule-based Approach	Gujarati	Dynamic	This algorithm work based on created rule hence, it cannot handle neologism.

Εικόνα 53 Συγκεντρωτικός Πίνακας Πλεονεκτημάτων-Μειονεκτημάτων των μεθόδων αφαίρεσης stopwords (Ladani & Desai,2020).

4.1.1.2 Stemming

Η μέθοδος stemming χρησιμοποιείται για να αναγνωρίσει τη ρίζα/στέλεχος μίας λέξης. Για παράδειγμα, όπως φαίνεται και στην παρακάτω εικόνα, οι λέξεις “connect”, “connected”, “connections” και “connecting” έχουν όλες ως κοινή ρίζα το “connect”.

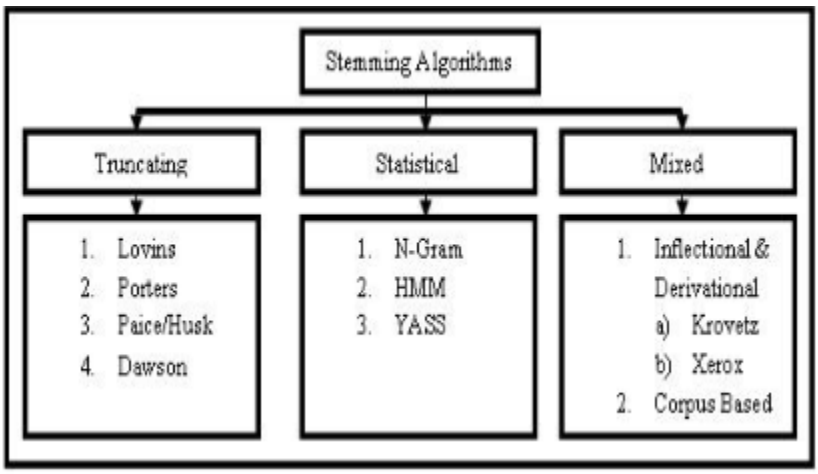


Εικόνα 54 Παράδειγμα της τεχνικής stemming (Mohan, Vijayarani,2015).

Ο σκοπός αυτής της μεθόδου είναι να αφαιρέσει ποικίλες καταλήξεις, προκειμένου να μειώσει τον αριθμό των λέξεων, να έχει ρίζες λέξεων που ταιριάζουν με ακρίβεια και να εξοικονομήσει χρόνο και χώρο μνήμης. Υπάρχουν 2 σημεία που λαμβάνονται υπόψη καθώς χρησιμοποιείται ένας stemmer:

- Λέξεις που δεν έχουν την ίδια σημασία πρέπει να κρατηθούν ξεχωριστά
- Οι μορφολογικές μορφές μιας λέξης υποτίθεται πως έχουν την ίδια βασική σημασία και συνεπώς πρέπει να αντιστοιχηθούν στο ίδιο stem.

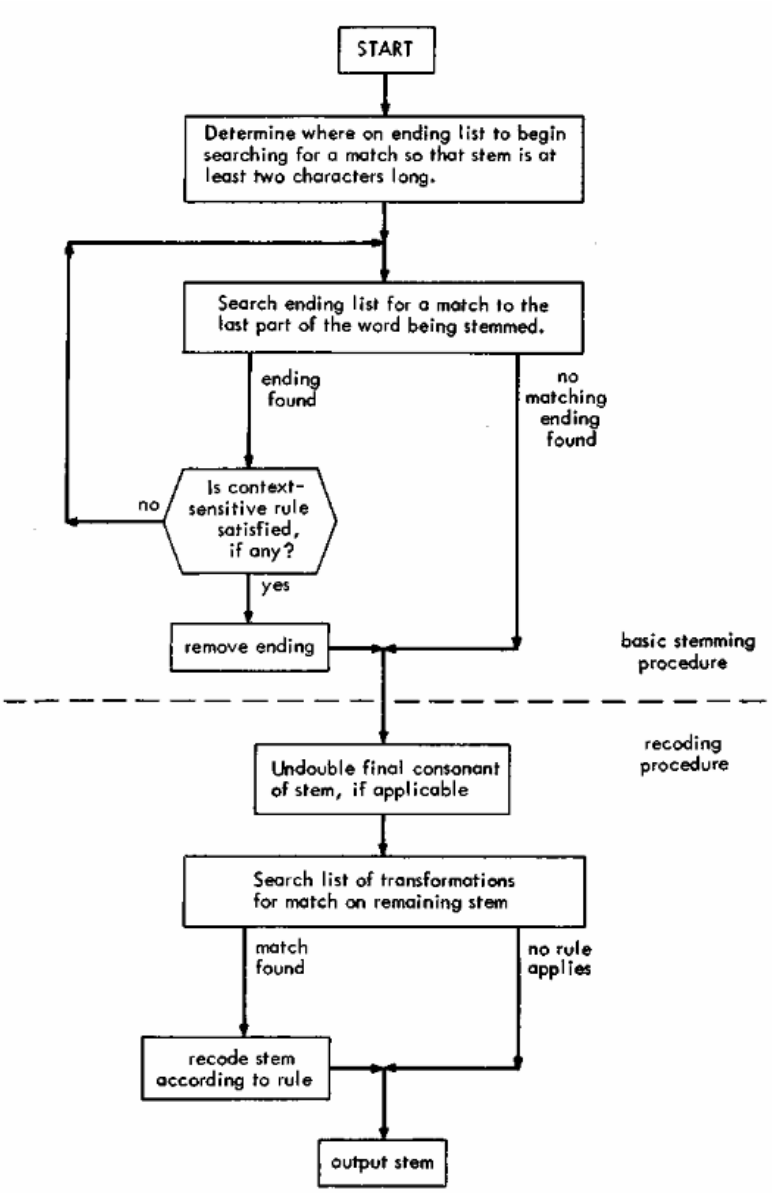
Οι αλγόριθμοι για stemming μπορούν να κατηγοριοποιηθούν σε 3 βασικές ομάδες [21]: μέθοδοι περικοπής (truncating methods), στατιστικές μέθοδοι (statistical methods), μικτές μέθοδοι (mixed methods) όπως παρουσιάζονται και στην ακόλουθη εικόνα.



Εικόνα 55 Κατηγορίες Stemming Αλγορίθμων (Mohan, Vijayarani,2015).

Truncating Methods

1. Lovins Stemmer.Αυτός ήταν ο πρώτος μοντέρνος και αποτελεσματικός stemmer που προτάθηκε από τον Lovins το 1968. Ο Lovins stemmer αφαιρεί το μεγαλύτερο επίθημα από μια λέξη. Μόλις αφαιρεθεί η κατάληξη, η λέξη επανακωδικοποιείται χρησιμοποιώντας διαφορετικό πίνακα που κάνει διάφορες προσαρμογές για να μετατρέψει αυτά τα στελέχη σε έγκυρες λέξεις. Αφαιρεί πάντα το πολύ ένα επίθημα από μια λέξη, λόγω της φύσης του ως αλγόριθμου μεμονωμένης διέλευσης. Τα πλεονεκτήματα αυτού του αλγορίθμου είναι ότι είναι πολύ γρήγορος και μπορεί να χειριστεί την αφαίρεση διπλών γραμμάτων σε λέξεις όπως "getting" (που μετατρέπονται σε "get") και, επίσης, χειρίζεται πολλούς ακανόνιστους πληθυντικούς όπως – “mouse” και “mice”, “index” και “indices» κ.λπ. Ένα μειονέκτημα της προσέγγισης Lovins είναι πως αποτελεί χρονοβόρο αλγόριθμο [3]. Επιπλέον, πολλά επιθέματα δεν είναι διαθέσιμα στον πίνακα των καταλήξεων. Μερικές φορές είναι εξαιρετικά αναξιόπιστο και συχνά αποτυγχάνει να σχηματίσει λέξεις από τους κορμούς ή να ταιριάζει με τους μίσχους λέξεων με το ίδιο νόημα. Τα βήματα του αλγορίθμου αποτυπώνονται και διαγραμματικά στην παρακάτω εικόνα [23]:



Εικόνα 56 Διαγραμματική απεικόνιση των βημάτων αφαίρεσης κατάληξης και επανακωδικοποίησης του αλγορίθμου Lovins Stemmer (Lovins,1968)

2. Porters Stemmer. Αποτελεί έναν από τους πιο δημοφιλείς αλγορίθμους stemming και προτάθηκε το 1980 [22]. Πολλοί μετασχηματισμοί και βελτιώσεις έχουν υλοποιηθεί και προταθεί στο βασικό αλγόριθμο. Βασίζεται στην ιδέα ότι οι καταλήξεις στην Αγγλική γλώσσα (περίπου 1200) αποτελούνται κυρίως από ομαδοποίηση μικρότερων και απλούστερων επιθημάτων. Περιέχει 5 βήματα και ανάμεσα σε κάθε βήμα εφαρμόζονται κανόνες ώσπου ένας από αυτούς να περάσει τις προϋποθέσεις. Αν ένας κανόνας γίνει αποδεκτός, το επίθημα αφαιρείται και το επόμενο βήμα εφαρμόζεται. Το προκύπτον stem στο τέλος του πέμπτου βήματος επιστρέφεται. Ο κανόνας μοιάζει όπως ο παρακάτω [3] :

   <condition> <suffix> → <new suffix>

3. Paice/Husk Stemmer. Ο stemmer Paice/Husk είναι ένας επαναληπτικός αλγόριθμος με έναν πίνακα που περιέχει περίπου 120 κανόνες ευρετηριασμένους από το τελευταίο γράμμα μιας κατάληξης [24] . Προσπαθεί να βρει τον εφαρμοστέο κανόνα από τον τελευταίο χαρακτήρα της λέξης. Κάθε κανόνας καθορίζει είτε

διαγραφή είτε αντικατάσταση μίας κατάληξης. Εάν δεν υπάρχει τέτοιος κανόνας, τερματίζεται. Επίσης, τερματίζεται αν μια λέξη αρχίζει από φωνήεν και απομένουν μόνο δύο ή τρεις χαρακτήρες. Διαφορετικά, εφαρμόζεται ο κανόνας και η διαδικασία επαναλαμβάνεται. Το πλεονέκτημα είναι η απλότητα και πως κάθε επανάληψη φροντίζει τόσο για τη διαγραφή όσο και την αντικατάσταση σύμφωνα με τον κανόνα προς εφαρμογή. Το μειονέκτημα είναι ότι είναι πολύ βαρύς αλγόριθμος και είναι πιθανό να προκύψει υπερβολικά μικρή ρίζα-overstemming.

4. Dawson Stemmer. Αυτός ο Stemmer είναι μία επέκταση της προσέγγισης του Lovins Stemmer εκτός από το ότι καλύπτει πολύ πιο πλήρη λίστα με περίπου 1200 επιθήματα. Όπως και ο Lovins, είναι ένας Stemmer με ένα πέρασμα, συνεπώς, είναι αρκετά γρήγορος. Τα επιθήματα αποθηκεύονται με την αντίστροφη σειρά που ευρετηριάζονται από το μήκος και τελευταίο γράμμα τους. Στην πραγματικότητα οργανώνονται ως ένα σύνολο διαιρεμένων δέντρων χαρακτήρων για γρήγορη πρόσβαση. Το πλεονέκτημα είναι ότι καλύπτει περισσότερα επιθήματα από τον Lovins και είναι γρήγορος στην εκτέλεση. Το μειονέκτημα είναι πως είναι πολύ σύνθετος και δεν διαθέτει μία τυπική επαναχρησιμοποιήσιμη υλοποίηση [21].

Statistical Methods (Στατιστικές Μέθοδοι)

Αυτοί είναι οι stemmers που βασίζονται σε κάποια στατιστική ανάλυση και τεχνικές. Οι περισσότερες μέθοδοι αφαιρούν τα προθήματα, αλλά αφού εκτελέσουν κάποια στατιστική διαδικασία [21].

1) N-Gram Stemmer. Είναι ένας ανεξάρτητος γλώσσας stemmer. Η προσέγγιση ομοιότητας συμβολοσειρών χρησιμοποιείται για τη μετατροπή μίας πληθωρικής λέξης στη ρίζα της. Το N-gram είναι μια συμβολοσειρά n, συνήθως γειτονικών, χαρακτήρων που εξάγονται από ένα τμήμα συνεχούς κειμένου. Το N-gram είναι ένα σύνολο από n συνεχόμενους χαρακτήρες το οποίο εξάγεται από μια λέξη. Η κύρια ιδέα πίσω από αυτήν την προσέγγιση είναι ότι παρόμοιες λέξεις θα έχουν μεγάλη ποσότητα κοινών n-grams. Για n ίσο με 2 ή 3, το οι λέξεις που εξάγονται ονομάζονται διαγράμματα (diagrams) ή τριγράμματα (trigrams), αντίστοιχα (Jivani,2011). Για παράδειγμα, η λέξη INTRODUCTIONS οδηγεί στη δημιουργία των παρακάτω diagrams:

I, IN, NT, TR, RO, OD, DU, UC, CT, TI, IO, ON, NS, S και trigrams **I, *IN, INT, NTR, TRO, ROD, ODU, DUC, UCT, CTI, TIO, ION, ONS, NS*, S**

Όπου το * δηλώνει ένα χώρο παραγεμίσματος για όταν δεν μπορούν να δημιουργηθούν πλήρη N-grams.

Υπάρχουν n+1 και n+2 τέτοια diagrams και trigrams αντίστοιχα για μία λέξη που περιέχει n χαρακτήρες. Συνήθως επιλέγεται n=4 ή n=5. Στη συνέχεια, αναλύεται ένα κείμενο ή ένα έγγραφο σε όλα τα n-grams. Είναι σαφές ότι μια λέξη ρίζα γενικά εμφανίζεται λιγότερο συχνά από τη μορφολογική της μορφή.
Αυτό σημαίνει ότι μια λέξη έχει γενικά ένα επίθεμα που σχετίζεται με αυτήν. Αυτός ο stemmer έχει το πλεονέκτημα πως είναι ανεξάρτητος από τη γλώσσα και, συνεπώς, πολύ χρήσιμος σε αρκετές εφαρμογές. Το μειονέκτημα του είναι πως χρειάζεται μεγάλος όγκος αποθηκευτικού χώρου, για τη δημιουργία και αποθήκευση των n-grams και των ευρετηρίων και συνεπώς δεν αποτελεί μία πρακτική προσέγγιση.

2) HMM Stemmer. Αυτός ο stemmer βασίστηκε στην ιδέα των Hidden Markov Models (HMMs) που είναι αυτόματα πεπερασμένων καταστάσεων όπου οι μεταβάσεις μεταξύ των καταστάσεων κατευθύνονται από συναρτήσεις πιθανότητας. Σε κάθε

μετάβαση, η καινούρια κατάσταση παραλείπει ένα σύμβολο με μία δοθείσα πιθανότητα [26]. Αυτή η μέθοδος βασίζεται σε μη-επιβλεπόμενη μάθηση και δεν απαιτεί καμία γλωσσολογική γνώση του συνόλου δεδομένων. Σε αυτήν τη μέθοδο, η πιθανότητα του κάθε μονοπατιού μπορεί να υπολογιστεί και το πιο πιθανό μονοπάτι βρίσκεται στο γράφο αυτομάτων. Προκειμένου να εφαρμοστεί HMMs στο stemming, μία ακολουθία γραμμάτων που σχηματίζει μία λέξη μπορεί να ληφθεί υπόψη ως το αποτέλεσμα της ένωσης δύο υποακολουθιών : ενός prefix και ενός suffix. Ένας τρόπος να μοντελοποιηθεί αυτή η διαδικασία είναι μέσω ενός HMM όπου οι καταστάσεις χωρίζονται σε δύο ξένα σύνολα: στο αρχικό μπορεί να είναι μόνο οι ρίζες (stems) και στο δεύτερο οι ρίζες ή οι καταλήξεις (suffixes). Οι αρχικές καταστάσεις ανήκουν μόνο στο σύνολο με τα stems – μία λέξη αρχίζει πάντα από ένα stem. Οι μεταβάσεις από καταστάσεις του συνόλου με τα suffixes σε καταστάσεις του συνόλου με τα stems, έχουν πάντοτε κενή πιθανότητα- μία λέξη μπορεί μόνο να είναι η ένωση ενός stem και ενός suffix. Οι τελικές καταστάσεις ανήκουν και στα 2 σύνολα- Ένα stem μπορεί να έχει έναν αριθμό διαφορετικών παραγωγών αλλά μπορεί επίσης να μην έχει καθόλου suffix. Το πλεονέκτημα αυτής της μεθόδου είναι πως είναι μη-επιβλεπόμενη και συνεπώς δεν απαιτείται γνώση της γλώσσας. Το μειονέκτημα είναι ότι είναι λίγο σύνθετη και μπορεί να οδηγήσει σε over-stemming μερικές φορές [26]

3) YASS Stemmer. Σύμφωνα με τους Plisson et al. (2004) [27], η απόδοση ενός stemmer που δημιουργείται από την ομαδοποίηση ενός λεξικού χωρίς καμία γλωσσική εισαγωγή είναι ισοδύναμη με αυτήν που λαμβάνεται από τη χρήση τυπικών, βασισμένους σε κανόνες stemmers όπως του Porter. Αυτός ο stemmer εντάσσεται στην τάξη των στατιστικών και βασισμένων στο corpus stemmers. Δεν στηρίζεται στη γλωσσική εμπειρογνωμοσύνη. Αυτή η προσέγγιση είναι αποτελεσματική για γλώσσες που έχουν κυρίως καταλήξεις στη φύση τους.

Mixed Methods

1) Krovetz Stemmer. Αποτελεσματικά και με ακρίβεια αφαιρεί τα επιθήματα σε 3 βήματα [28]: Α) Μετατροπή του πληθυντικού μίας λέξης στον τύπο της στον ενικό.Β) Μετατροπή του παρελθοντικού χρόνου μίας λέξης σε ενεστώτα. Γ) Αφαίρεση της κατάληξης “ing”. Σε σύγκριση με τον Porter και τον Paice/Husk είναι ένας πολύ ελαφρύς αλγόριθμος [3]. Στοχεύει στην αύξηση της ακρίβειας και της ευρωστίας με την αντιμετώπιση των ορθογραφικών λαθών και των ανούσιων στελεχών. Εάν το μέγεθος του κειμένου εισόδου είναι μεγάλο, αυτός ο stemmer γίνεται αδύναμος και δεν αποδίδει πολύ αποτελεσματικά. Το κύριο ελάττωμα για τέτοιου είδους αλγορίθμους αποτελεί η αδυναμία τους να χειριστούν λέξεις που δεν υπάρχουν στο λεξικό. Δεν παράγει σταθερά μία καλή απόδοση ακρίβειας (precision) και ανάκλησης (recall) [28]
2) Xerox. Η βάση δεδομένων κλίσης μειώνει την επιφάνεια κάθε λέξης στη μορφή που μπορεί να βρεθεί στο λεξικό ως εξής [29]:
    - Ενικός ουσιαστικών (π.χ. children → child)
    - Απαρέμφατο ρημάτων (π.χ. understood → understand)
    - Θετικός βαθμός επιθέτων (π.χ. best → good)
    - Ονομαστική Αντωνυμιών (π.χ. whom → who)

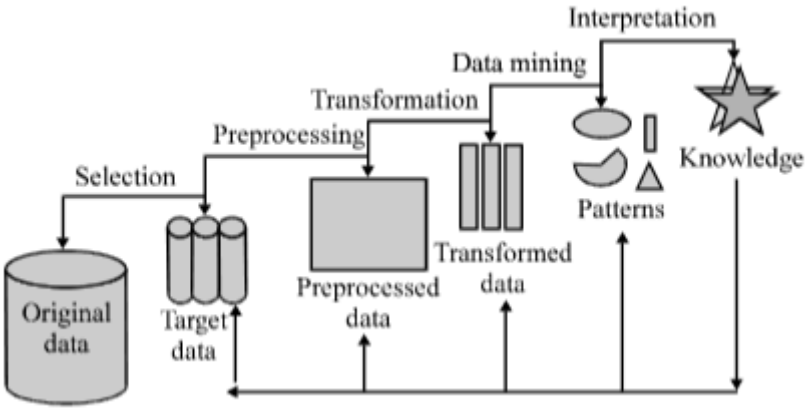
Αυτός ο αλγόριθμος δουλεύει καλά σε μεγάλα κείμενα και αφαιρεί τα προθέματα όπου αυτό είναι εφικτό. Όλα τα stems είναι έγκυρες λέξεις αφού μία λεξιλογική βάση δεδομένων που παρέχει μορφολογική ανάλυση για κάθε λέξη στο λεξικό είναι διαθέσιμη για stemming. Έχει αποδειχθεί πως λειτουργεί καλύτερα από τον Krovetz για μεγάλο σύνολο κειμένων. Το μειονέκτημα είναι πως το αποτέλεσμα εξαρτάται από τη λεξιλογική βάση δεδομένων η οποία μπορεί να μην είναι εξαντλητική. Δεν έχει υλοποιηθεί σε πολλές άλλες γλώσσες. Ακόμη, η εξάρτηση

του stemmer από μία λεξιλογική βάση δεδομένων τον καθιστά εξαρτώμενο από τη γλώσσα.

3) Corpus based Stemmer. Αυτή η προσέγγιση προσπαθεί να ξεπεράσει κάποια από τα μειονεκτήματα του Porter's Stemmer [21]. Για παράδειγμα, οι λέξεις 'policy' και 'police' συγχέονται παρόλο που έχουν διαφορετική σημασία αλλά οι λέξεις 'index' και 'indices' δεν συγχέονται, παρόλο που έχουν την ίδια ρίζα. Επιπλέον, ο Porter stemmer παράγει στελέχη τα οποία δεν είναι πραγματικές λέξεις όπως το 'iteration' που γίνεται 'iter' και το 'general' που γίνεται 'gener'. Αυτό το είδος αλγορίθμου αναφέρεται στην αυτόματη τροποποίηση των τάξεων συγχώνευσης- λέξεις που έχουν οδηγήσει σε ένα κοινό στέλεχος, να ταιριάζουν στα χαρακτηριστικά ενός δεδομένου σώματος κειμένου χρησιμοποιώντας στατιστικές μεθόδους. Το πλεονέκτημα αυτής της μεθόδου είναι πως αποφεύγει συγχωνεύσεις που δεν είναι κατάλληλες για ένα δοθέν σώμα κειμένων και το αποτέλεσμα είναι μία πραγματική λέξη και όχι ένα ελλιπές στέλεχος. Το μειονέκτημα είναι πως χρειάζεται να αναπτυχθεί το στατιστικό μέτρο για κάθε σώμα ξεχωριστά και ο χρόνος επεξεργασίας αυξάνεται, καθώς στο πρώτο βήμα δύο αλγόριθμοι stemming χρησιμοποιούνται πριν χρησιμοποιηθεί αυτή η μέθοδος.

4.1.2 Διανυσματοποίηση

Οι αλγόριθμοι μηχανικής μάθησης εφαρμόζονται σε αριθμητικά δεδομένα, που σημαίνει πως αναμένουν ως είσοδο έναν δισδιάστατο πίνακα, οι γραμμές του οποίου είναι διακεκριμένες περιπτώσεις και οι στήλες είναι τα χαρακτηριστικά (features) [31]. Συνεπώς, προκειμένου να εκτελεστεί μηχανική μάθηση σε κείμενο, είναι αναγκαίο να μετατραπούν τα πηγαία κείμενα σε διανυσματικές αναπαραστάσεις, στις οποίες αριθμητική μηχανική μάθηση θα εφαρμοστεί ακολούθως. Αυτή η διαδικασία ονομάζεται διανυσματοποίηση (vectorization) ή data transformation και αποτελεί ένα πολύ σημαντικό βήμα για την ανάλυση κειμένων.



Εικόνα 57 Βήματα ανακάλυψης της γνώσης (Alasadi & Wesam,2017)

Η διαδικασία αυτή, όπως φαίνεται στην παραπάνω εικόνα, πραγματοποιείται αμέσως μετά τη διαδικασία της προεπεξεργασίας. Στη συνέχεια, τα προεπεξεργασμένα δεδομένα διανυσματοποιούνται και δίνονται ως είσοδος σε κάποιον αλγόριθμο [35] ώστε να πραγματοποιηθεί αναγνώριση προτύπων (pattern recognition).

Η μετατροπή κειμένων στην αριθμητική τους μορφή καθιστά εφικτή την ανάλυσή τους και την εφαρμογή των επιλεγμένων αλγορίθμων μηχανικής μάθησης. Τα έγγραφα (ή οι προτάσεις) μπορούν να έχουν διαφορετικά μεγέθη, τα διανύσματα που ορίζουμε για αυτά, όμως, θα έχουν πάντα το ίδιο μήκος. Κάθε ιδιότητα σε μία διανυσματική αναπαράσταση είναι ένα feature. Στην εν λόγω περίπτωση, θα είναι λέξεις που συμπεριλαμβάνονται στο

κείμενο. Μαζί, όλα αυτά τα features, θα περιγράφουν έναν πολυδιάστατο χώρο χαρακτηριστικών στον οποίον μέθοδοι μηχανικής μάθησης δύναται να εφαρμοστούν.

Επομένως, πρέπει να υπάρξει μετακίνηση από ατομικές προτάσεις και λέξεις σε σημεία σε έναν πολυδιάστατο σημασιολογικό χώρο. Αυτά τα σημεία μπορούν να τοποθετηθούν μακριά ή κοντά το ένα στο άλλο, κατανεμημένα ομοιόμορφα ή, αντιθέτως, τυχαία. Με βάση αυτό, μπορεί να εξαχθεί ως συμπέρασμα πως προτάσεις με παρεμφερές νόημα θα τοποθετηθούν κοντά και οι διαφορετικές, αντιθέτως, μακριά [31]

Φορέας Συχνότητας/Frequency Vectorizer

Ένας τρόπος να διανυσματοποιηθεί το πηγαίο κείμενο είναι να υπολογισθεί η συχνότητα εμφάνισης κάθε λέξης σε κάθε πρόταση και να συσχετισθεί αυτή η τιμή με ολόκληρο το σύνολο λέξεων του αρχικού συνόλου δεδομένων. Η εκκίνηση μπορεί να γίνει με τη δημιουργία ενός dictionary όλων των λέξεων όλων των προτάσεων του συνόλου δεδομένων. Με τον όρο dictionary σε αυτήν την περίπτωση εννοείται μία λίστα των λέξεων που συναντώνται στα κείμενα, όπου κάθε λέξη έχει το δικό της δείκτη. Αυτό επιτρέπει τη δημιουργία ενός διανύσματος για κάθε πρόταση – απλώς παίρνοντας την πρόταση προς διανυσματοποίηση και μετρώντας τις εμφανίσεις κάθε λέξης. Το μήκος του τελικού διανύσματος θα είναι ίσο με το μέγεθος του dictionary και θα περιέχει ως τιμή τον αριθμό των εμφανίσεων της λέξης από το dictionary σε κάθε συγκεκριμένη πρόταση. Ας ληφθεί υπόψη ένα συγκεκριμένο παράδειγμα [31].

```
sentences = ['Сломалась кофемашина на нашем этаже',
             'Лопнула лампочка на восьмом этаже',
             'Кофемашина отремонтирована и работает',
             'Лампочка упала и разбилась']
```

Εικόνα 58 Οι προτάσεις για επίδειξη πάνω στις οποίες θα εφαρμοστεί ο frequency vectorizer (Kozhevnikov, & Pankratova, 2020).

Έπειτα, στο παραπάνω σύνολο προτάσεων, μπορεί να χρησιμοποιηθεί η μέθοδος CountVectorizer() από τη βιβλιοθήκη scikit-learn για να υπάρξει διανυσματοποίηση. Το αποτέλεσμα φαίνεται στην παρακάτω εικόνα:

```
from sklearn.feature_extraction.text import CountVectorizer

vectorizer = CountVectorizer()
vectorizer.fit(sentences)
vectorizer.vocabulary_

Output:
{'восьмом': 0, 'кофемашина': 1, 'лампочка': 2, 'лопнула': 3, 'на': 4,
'нашем': 5, 'отремонтирована': 6, 'работает': 7,  'разбилась': 8,
'сломалась': 9, 'этаже': 11, 'упала': 10}
```

Εικόνα 59 Επίδειξη-Demo για τον τρόπο λειτουργίας της CountVectorizer() στο σύνολο των προτάσεων (Kozhevnikov, & Pankratova, 2020).

Σαν αποτέλεσμα, λαμβάνεται ένα dictionary όλων των μοναδικών λέξεων που είναι διαθέσιμες σε όλες τις προτάσεις που δόθηκαν ως είσοδος στην CountVectorizer().

Στη συνέχεια, είναι εφικτό να χρησιμοποιηθεί η ίδια μέθοδος για να ληφθεί διάνυσμα για κάθε πρόταση που θα φανερώνει τον αριθμό των εμφανίσεων κάθε λέξης από το dictionary κάθε πρόταση.

```
vectorizer.transform(sentences).toarray()

Output:
array([[0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1],
       [1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1],
       [0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0],
       [0, 0, 2, 0, 0, 0, 0, 0, 1, 0, 1, 0]])
```

Εικόνα 60 Αποτέλεσμα διανυσματοποίησης CountVectorizer() για το σύνολο προτάσεων που παρουσιάστηκε παραπάνω (Kozhevnikov, & Pankratova, 2020).

Το αποτέλεσμα δείχνει ότι στην τέταρτη πρόταση (στοιχείο του πίνακα) η λέξη “лампочка” εμφανίστηκε 2 φορές επομένως στην αντίστοιχη θέση του πίνακα υπάρχει ο αριθμός 2. Αυτή η προσέγγιση ονομάζεται ***Bag of words*** και αποτελεί έναν σύνηθη τρόπο για τη μετατροπή κειμένου σε διάνυσμα. Κάθε έγγραφο ή πρόταση παρουσιάζεται σαν ένα ξεχωριστό διάνυσμα.

Τα μειονεκτήματα αυτής της προσέγγισης είναι τα ακόλουθα: Με την αύξηση του μεγέθους του λεξικού, διανύσματα θα μεγαλώσουν και θα γίνουν πιο αραιά, και θα έχουν μέσα τους προβολή μεγάλου αριθμού μηδενικών, λόγω του γεγονότος ότι το καθένα έγγραφο θα περιέχει μόνο έναν μικρό αριθμό λέξεων από το λεξικό. Τέτοια διανύσματα απαιτούν περισσότερη μνήμη και υπολογιστικούς πόρους, οι οποίοι μπορούν να έχουν σημαντικό αντίκτυπο στην απόδοση των μοντέλων. Αλλά αυτό μπορεί να λυθεί με τις ακόλουθες τεχνικές:

- Αφαίρεση stop words (τεχνική που αναλύθηκε σε προηγούμενη ενότητα Αφαίρεση Stopwords) καθώς δεν προσφέρουν σημαντική πληροφορία.
- Τροποποίηση λέξεων στην κανονική τους μορφή με χρήση αλγορίθμων stemming και lemmatization (Stemming)
- Διόρθωση λέξεων που έχουν γραφτεί με λάθος.

- **Άμεση Κωδικοποίηση/Direct coding**

Προς συμπλήρωση του προηγούμενου τρόπου που μετράει τον αριθμό των εμφανίσεων κάθε λέξης του dictionary, υπάρχει και ένας πιο απλός τρόπος διανυσματοποίησης- η προσέγγιση direct coding. Η άμεση (ή δυαδική) κωδικοποίηση είναι μία λογική μέθοδος διανυσματοποίησης η οποία καταγράφει True(ή 1) στην αντίστοιχη θέση του διανύσματος όταν η λέξη υπάρχει στην πρόταση και False(ή 0) στην αντίθετη περίπτωση. Συνεπώς, κάθε στοιχείο στο διάνυσμα υπονοεί την παρουσία ή απουσία της αντίστοιχης λέξης στην περιγραφόμενη πρόταση. Με αυτόν τον τρόπο, το κείμενο απλοποιείται στα συστατικά στοιχεία του. Αυτή η μέθοδος είναι πολύ αποτελεσματική για σύντομα έγγραφα, όπως, για παράδειγμα, tweets που περιέχουν μικρό αριθμό επαναληπτικών στοιχείων. Η προσέγγιση άμεσης κωδικοποίησης χρησιμοποιείται επίσης συχνά σε νευρωνικά δίκτυα, όπου οι συναρτήσεις ενεργοποίησης απαιτούν εισαγωγές τιμών από τις περιοχές [0, 1] ή [-1, 1].

Το αποτέλεσμα αυτής της μεθόδου για το σύνολο προτάσεων που μελετήθηκε είναι το ακόλουθο.

```
vectorizer = CountVectorizer(binary=True)

vectorizer.transform(sentences).toarray()

Output:
array([[0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1],
       [1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1],
       [0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0],
       [0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0]])
```

Εικόνα 61 Αποτέλεσμα μεθόδου Direct coding (Kozhevnikov, & Pankratova, 2020).

- **Tf-idf**

Η προσέγγιση της μέτρησης του αριθμού των εμφανίσεων των λέξεων σε ένα έγγραφο έχει ένα πρόβλημα: λέξεις που είναι πιο πιθανό να εμφανιστούν έχουν υψηλότερη βαθμολογία. Αυτές όμως οι λέξεις μπορεί να έχουν πολύ λίγες χρήσιμες πληροφορίες, όπως μπορεί και λιγότερο συχνές λέξεις. Μια πιο επιτυχημένη προσέγγιση βασίζεται ακριβώς σε σύγκριση της σχετικής συχνότητας ή σπανιότητας λέξεων σε ένα συγκεκριμένο έγγραφο με τη συχνότητά τους σε άλλα έγγραφα. Η κύρια ιδέα αυτής της προσέγγισης είναι ότι το κύριο νόημα είναι κρυμμένο σε αυτές τις λέξεις που είναι λιγότερο κοινές.

Η μέθοδος Tf-idf ομαλοποιεί τη συχνότητα λέξης στο έγγραφο, λαμβάνοντας υπόψη το περιεχόμενο σε ολόκληρη την περίπτωση. Έτσι, αποδεικνύεται ότι αν μια λέξη βρίσκεται συχνά σε ένα συγκεκριμένο έγγραφο, αλλά σπάνια βρίσκεται στα υπόλοιπα, τότε αυτή η λέξη είναι υψηλής σημασίας για αυτό το ίδιο το έγγραφο και τέτοιες λέξεις θα κερδίζουν περισσότερο βάρος σε σύγκριση με άλλες λέξεις του σώματος κειμένων.

Η Tf-IDF χρησιμοποιείται ως ένας συντελεστής στάθμισης στους χώρους της ανάκτησης πληροφορίας και εξόρυξης κειμένου. Η τιμή του αυξάνεται αναλογικά με το πλήθος των φορών που μία λέξη εμφανίζεται σε ένα κείμενο, αλλά εξουδετερώνεται από τη συχνότητα της λέξης στο σώμα των κειμένων [3]. Αυτό μπορεί να βοηθήσει στον έλεγχο της περίπτωσης κατά την οποία κάποιες λέξεις είναι συχνότερα χρησιμοποιούμενες από κάποιες άλλες, όπως, για παράδειγμα, οι stopwords. Η Tf-IDF είναι το γινόμενο των εξής 2 στατιστικών:

Tf=$\frac{nt}{\sum_k nk}$, όπου n_t ο αριθμός των εμφανίσεων του όρου t και ο παρονομαστής είναι ο συνολικών αριθμός λέξεων στο έγγραφο

IDF=log$\frac{N}{\mathrm{DF}t}$,όπου N είναι ο αριθμός των κειμένων στη συλλογή κειμένων και Df_t είναι ο αριθμός των κειμένων από τη συλλογή που υπάρχει ο όρος

.

Τελικά, τα βάρη υπολογίζονται ως:

TF-IDF= TF · IDF

	восьмом	кофемашина	лампочка	лопнула	на	нашем	\
0	0.000000	0.401043	0.000000	0.000000	0.401043	0.508672	
1	0.508672	0.000000	0.401043	0.508672	0.401043	0.000000	
2	0.000000	0.486934	0.000000	0.000000	0.000000	0.000000	
3	0.000000	0.000000	0.744450	0.000000	0.000000	0.000000	

	отремонтирована	работает	разбилась	сломалась	упала	этаже
0	0.000000	0.000000	0.00000	0.508672	0.00000	0.401043
1	0.000000	0.000000	0.00000	0.000000	0.00000	0.401043
2	0.617614	0.617614	0.00000	0.000000	0.00000	0.000000
3	0.000000	0.000000	0.47212	0.000000	0.47212	0.000000

Εικόνα 62 Αποτελέσματα Tf-idf στις αρχικές προτάσεις (Kozhevnikov, & Pankratova, 2020)

- **Word2Vec**

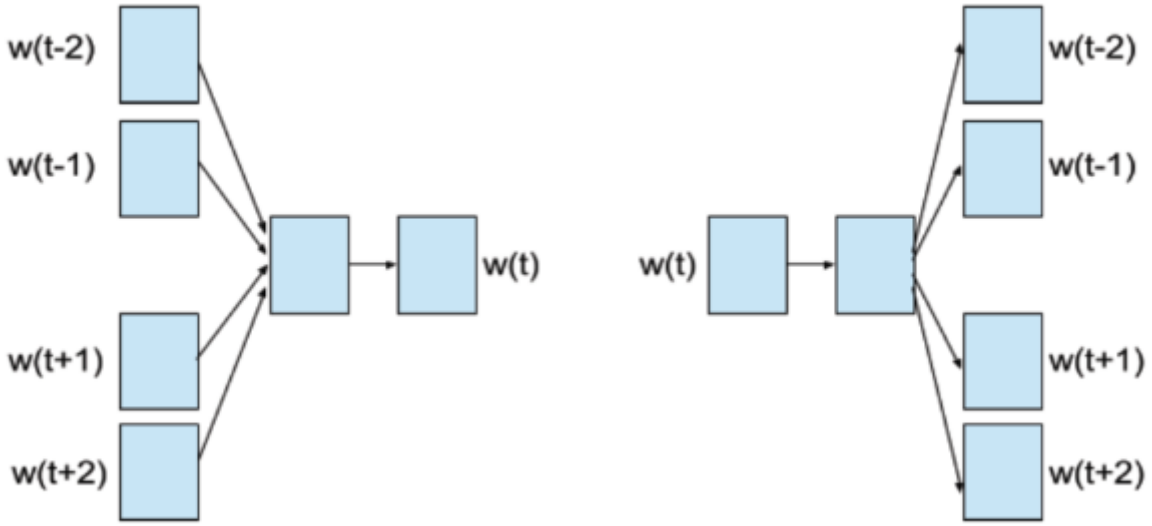
Οι μέθοδοι που περιγράφονται παραπάνω παράγουν διανύσματα μόνο με θετικά στοιχεία, τα οποία δεν επιτρέπουν τη σύγκριση εγγράφων που δεν έχουν κοινές λέξεις, λόγω του γεγονότος ότι δύο διανύσματα που έχουν τιμή συνημιτόνου της μεταξύ τους γωνίας ίση με 1 θα εξακολουθούν να θεωρούνται μακρινά σε νόημα.

Εάν η ομοιότητα μεταξύ εγγράφων παίζει ένα σημαντικό ρόλο στην εφαρμογή των αλγορίθμων μηχανικής μάθησης, τότε τα δεδομένα μπορούν να κωδικοποιηθούν χρησιμοποιώντας τη μέθοδο της κατανεμημένης αναπαράστασης. Με αυτή την προσέγγιση, ένα διάνυσμα δεν είναι απλώς μια χαρτογράφηση των θέσεων των λέξεων στην αριθμητική τους τιμή, αλλά ένα σύνολο χαρακτηριστικών που προσδιορίζει την ομοιότητα των λέξεων. Η πολυπλοκότητα του χώρου χαρακτηριστικών (και το μήκος του διανύσματος) καθορίζεται από τα μαθησιακά χαρακτηριστικά αυτής της αντιπροσώπευσης και δεν σχετίζεται άμεσα με το ίδιο το έγγραφο [31].

Εργαλεία για τη δημιουργία διανυσματικά-σημασιολογικών μοντέλων υπήρχαν και πριν [33] αλλά το word2vec ήταν η πρώτη δημοφιλής υλοποίηση [34] λόγω της ευκολίας χρήσης, της ταχύτητας εργασίας και, το πιο σημαντικό, ανοιχτού κώδικα. Η βασική υπόθεση ήταν πως «λέξεις που συμβαίνουν σε πανομοιότυπα περιβάλλοντα έχουν παρόμοιες σημασίες» [34]. Η εγγύτητα σε αυτό το πλαίσιο μπορεί να γίνει κατανοητή ως το γεγονός ότι μόνο οι λέξεις που ταιριάζουν μπορούν να σταθούν κοντά. Δηλαδή, για παράδειγμα, είναι φυσιολογικό να το ακούμε τη φράση «злой человек», αλλά η φράση «злой холодильник» είναι εντελώς ασυνήθιστη [31].

Το μοντέλο που πρότεινε ο Tomaš [34] είναι αρκετά απλό - η πιθανότητα μιας λέξης θα προβλεφθεί από τα συμφραζόμενα. Δηλαδή, θα εκπαιδεύσουμε το μοντέλο διανυσματοποίησης έτσι ώστε η πιθανότητα που αποδίδεται από το μοντέλο σε μια λέξη να είναι κοντά στην πιθανότητα να συναντηθεί αυτή η λέξη εντός αυτού του περιβάλλοντος σε πραγματικό κείμενο. Αυτή η προσέγγιση ονομάζεται CBOW (Continuous Bag of Words). Ονομάζεται συνεχής, γιατί τροφοδοτούνται τα σύνολα λέξεων από το κείμενο διαδοχικά στην είσοδο του, και BoW, επειδή η σειρά των λέξεων στο πλαίσιο δεν είναι σημαντική. Το στοιχείο εισαγωγής στο νευρωνικό δίκτυο, είναι το σύνολο διανυσμάτων πλαισίου w (t - k), ..., w (t - 1), w (t + 1), ..., w (t + k), και το διάνυσμα

εξόδου είναι το w(t), όπου w(t) είναι το διάνυσμα της προβλεπόμενης λέξεις με βάση το πλαίσιο. Ο Mikolov [34] πρότεινε επίσης μια διαφορετική προσέγγιση, που είναι ακριβώς η αντίθετη από την προσέγγιση CBOW και την ονόμασε skip-gram. Η αρχιτεκτονική του Skip-gram διαφέρει από τη CBOW στο ότι προβλέπει ένα σύνολο λέξεων γύρω, με βάση μια δεδομένη λέξη. Επομένως, αυτήν τη φορά, την είσοδο αποτελεί το διάνυσμα w(t) και την έξοδο το σύνολο διανυσμάτων M={w(t-k), ….,w(t-1),w(t+1),….,w(t+k)}. Κάθε λέξη που αντιστοιχεί σε διανύσματα από το σύνολο M χαρακτηρίζει μια λέξη που αντιστοιχεί σε ένα διάνυσμα-είσοδο. Το σχήμα λειτουργίας που χρησιμοποιούν οι 2 προσεγγίσεις φαίνεται στην παρακάτω εικόνα:



Εικόνα 63 Αριστερά: Τρόπος λειτουργίας αλγορίθμου CBOW. Δεξιά: Τρόπος λειτουργίας αλγορίθμου skip-gram (Kozhevnikov, & Pankratova, 2020).

4.1.3 Επιλογή χαρακτηριστικών-Feature selection

Ένα από τα βήματα της βασικής μεθοδολογίας (Γενική Περιγραφή) αποτέλεσε και η επιλογή χαρακτηριστικών ή αλλιώς feature selection.

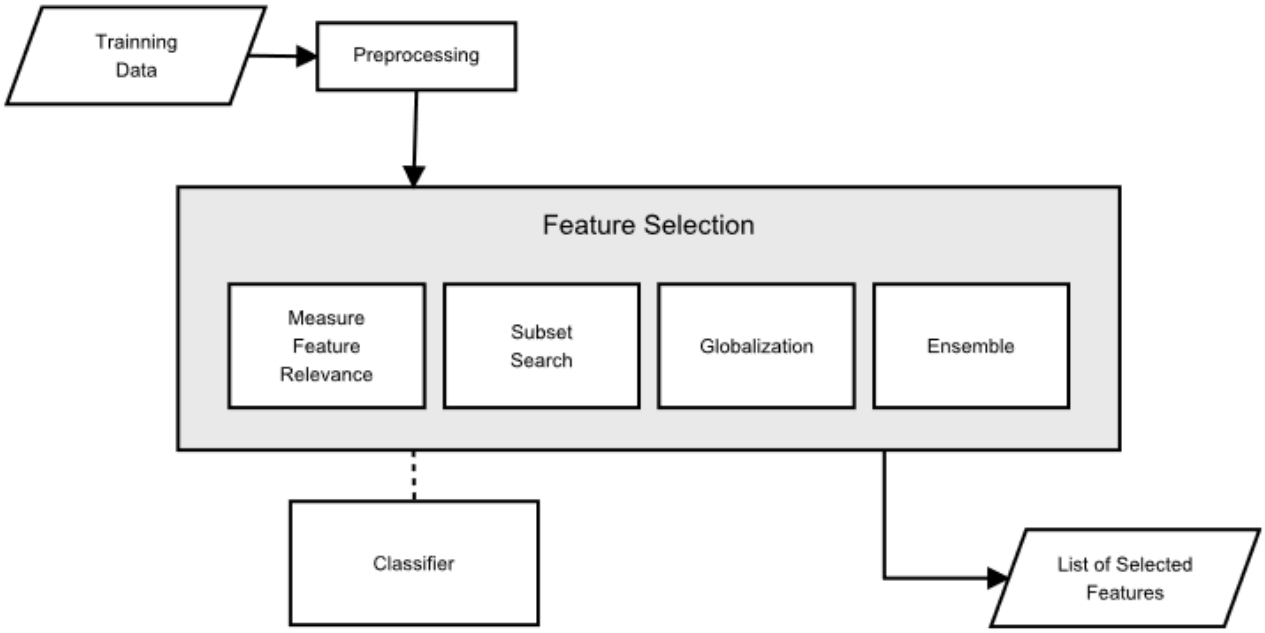
Ακολουθεί εν συντομία μία σύντομη περιγραφή των υπο εργασιών που πρέπει να πραγματοποιηθούν κατά τη διάρκεια του feature selection [6]

- *Μέτρηση συνάφειας χαρακτηριστικών.* Υπάρχουν διάφοροι τρόποι εκτίμησης της συνάφειας των χαρακτηριστικών, όπως π.χ μέτρηση της συσχέτισης με τον στόχο, τη μεταβλητή εντροπίας ή τον υπολογισμό του πλεονασμού των χαρακτηριστικών [30]. Ωστόσο, η βασική ιδέα είναι ότι όσο μεγαλύτερη είναι η συνάφεια ενός χαρακτηριστικού, τόσο μεγαλύτερη πρέπει να είναι η ισχύς για την αύξηση της ακρίβειας του μοντέλου (στη δική μας περίπτωση ταξινομητή κειμένου)
- *Αναζήτηση υποσυνόλου.* Η εργασία αναζήτησης υποσυνόλου στοχεύει στην εύρεση της καλύτερης υποομάδας χαρακτηριστικών ώστε να χρησιμοποιηθεί στην ταξινόμηση.
- *Καθολικοποίηση*.Μπορούν να εφαρμοστούν συνήθως μετρήσεις συνάφειας και μέθοδοι αναζήτησης υποσυνόλων ειδικά για μία κλάση ή ετικέτα του συνόλου

δεδομένων. Επομένως, μια μέθοδος που καθολικοποιεί τα αποτελέσματα κάθε κλάσης/ετικέτας απαιτείται για την κατασκευή ενός τελικού συνόλου χαρακτηριστικών που αντιπροσωπεύει όλες τις κατηγορίες ή τις ετικέτες.

- *Συνδυασμός*. Κάθε μέθοδος FS έχει συγκεκριμένα πλεονεκτήματα και μειονεκτήματα, οπότε ο συνδυασμός δύο ή περισσότερων μεθόδων μπορεί να οδηγήσει σε καλύτερα αποτελέσματα από τη χρήση τους ξεχωριστά.

Όλες οι παραπάνω υπο εργασίες απεικονίζονται σχηματικά στην παρακάτω εικόνα.



Εικόνα 64 Υποεργασίες κατά τη διαδικασία επιλογή χαρακτηριστικών-feature selection (Pintas, Fernandes & Garcia, 2021).

4.1.3.1 Μέθοδοι Επιλογής Χαρακτηριστικών-Feature Selection

Σε αυτήν την ενότητα θα περιγραφούν εν συντομία 10 διαδεδομένες μέθοδοι για Feature selection για κατηγοριοποίηση κειμένου [85].

1. Συχνότητα εγγράφου-Document Frequency (DF). Το DF είναι μια απλή και αποτελεσματική μέθοδος επιλογής χαρακτηριστικών χωρίς επίβλεψη [86] που βαθμολογεί τα χαρακτηριστικά σύμφωνα με τον αριθμό των εμφανίσεων τους στο έγγραφο [87]. Δηλαδή, τα πιο συχνά χαρακτηριστικά είναι πιο σημαντικά. Το DF υπολογίζεται μετρώντας τον αριθμό των εγγράφων στα οποία εμφανίζεται ένα χαρακτηριστικό. Ένα προφανές μειονέκτημα είναι ότι ορισμένοι όροι υψηλής συχνότητας που δεν βοηθούν στην ταξινόμηση, όπως τα stopwords, θα υπολογίζονται ως χαρακτηριστικά.
2. Κέρδος Πληροφοριών-Information Gain (IG). Η IG είναι μια εποπτευόμενη μέθοδος που έχει σχεδιαστεί για τον προσδιορισμό της συνεισφοράς ενός όρου, σύμφωνα με μια αναλογία που υπολογίζεται μετρώντας την παρουσία ή την απουσία του σε ένα σύνολο εγγράφων [88]. Ο ακριβής υπολογισμός είναι:

$$IG(t) = -\sum_{i=1}^{M} P(Ci)\log(P(Ci)) + P(t)\sum_{i=1}^{M} P(Ci|t)\log\big(P(Ci|t)\big) + P(\bar{t})\sum_{i=1}^{M} P(Ci|\bar{t})\log(P(Ci|\bar{t}))$$

3. Δείκτης Gini-Gini Index (GI). Το GI χρησιμοποιήθηκε αρχικά σε αλγόριθμους δένδρων αποφάσεων, αλλά οι Shang et al. [89] πρότειναν ένα βελτιωμένο GI για την επιλογή χαρακτηριστικών εντός κειμένου. Είναι μια εποπτευόμενη μέθοδος με απλούστερο υπολογισμό από το IG:

$$GI(t) = \sum_{i=1}^{M} P(t|Ci)^2 P(Ci|t)^2$$

4. Διακριτικός επιλογέας χαρακτηριστικών-Distinguishing feature selector (DFS). Προτεινόμενο από τους Uysal και Gunal [90], το DFS είναι μια επιτυχημένη, σχετικά νέα μέθοδος επιλογής χαρακτηριστικών για ταξινόμηση κειμένου. Είναι μια εποπτευόμενη μέθοδος που προορίζεται για την ανάδειξη των πιο αντιπροσωπευτικών χαρακτηριστικών και την αφαίρεση των μη-πληροφοριακών. Υπολογίζεται ως εξής:

$$DFS(t) = \sum_{i=1}^{M} \frac{P(Ci|t)}{P(\bar{t}|Ci) + P(t|\bar{C}\imath) + 1}$$

5. Αναμενόμενη Διασταυρούμενη Εντροπία-Expected Cross Entropy (ECE). Η ECE είναι μια εποπτευόμενη μέθοδος που εξετάζει την παρουσία ενός όρου t και αγνοεί την απουσία του [91]. Η ECE μπορεί να υπολογιστεί ως:

$$ECE(t) = P(t)\sum_{i=1}^{M} P(Ci|t)\log\left(\frac{P(Ci|t)}{P(Ci)}\right)$$

6. Μέτρο ταξικής διάκρισης-Class discriminating measure (CDM). Το CDM είναι μια καθολική μέθοδος επιλογής χαρακτηριστικών που προέρχεται από τον λόγο πιθανοτήτων και προτάθηκε από τους Chen et al. [91]. Υπολογίζεται ως εξής:

$$CDM(t) = \sum_{i=1}^{M} \left| log\frac{P(t|Ci)}{P(t|\bar{\bar{C}}\imath)} \right|$$

7. Χ-Τετράγωνο-Chi-squared (CHI). Το CHI είναι μια εποπτευόμενη, μονόπλευρη μέθοδος επιλογής χαρακτηριστικών που υπολογίζει τη συσχέτιση του όρου t με την κλάση [87]. Το CHI υπολογίζεται ως εξής:

$$\chi^2(t, Ci) = \frac{Nx(aidi - bici)^2}{(ai + bi)x(ai + ci)x(bi + di)x(ci + di)}$$

8. Λόγος πιθανοτήτων-Odds ratio (OR). Το OR είναι μια εποπτευόμενη και μονόπλευρη μέθοδος που λαμβάνεται με τον υπολογισμό της ιδιότητας μέλους και της μη-συμμετοχής σε μια συγκεκριμένη κλάση με τον αριθμητή και τον παρονομαστή της, αντίστοιχα [92]. Υπολογίζεται ως εξής:

$$OR(t, Ci) = log\left(\frac{P(t|Ci)(1 - P(t|\bar{C}\imath))}{(1 - P(t|Ci))P(t|\bar{C}\imath)}\right)$$

9. Αμοιβαία Ενημέρωση-Mutual Information (MI). Το MI είναι μια εποπτευόμενη και μονόπλευρη μέθοδος που αντιπροσωπεύει τη συσχέτιση μεταξύ κλάσεων και χαρακτηριστικών. Υπολογίζεται ως εξής:

$$M(t, Ci) = log \frac{P(t|Ci)}{P(t)}$$

10. Σταθμισμένη αναλογία πιθανότητας καταγραφής-Weighted log likelihood ratio (WLLR). Το WLLR προτείνεται από τους Nigam et al. [93]. Είναι μια εποπτευόμενη και μονόπλευρη μέθοδος και υπολογίζεται από:

$$WLLR(t, Ci) = P(t|Ci) log \frac{P(t|Ci)}{P(t|\overline{Ci})}$$

4.1.4 Αλγόριθμοι Εκπαίδευσης

Μετά από τη συλλογή των δεδομένων, την προεπεξεργασία τους, τον διαχωρισμό τους σε σύνολο εκπαίδευσης και σύνολο ελέγχου καθώς και τη διαδικασία επιλογής χαρακτηριστικών, ακολουθεί η εκπαίδευση του μοντέλου. Σε αυτήν την ενότητα θα περιγραφούν οι πιο διαδεδομένοι αλγόριθμοι που χρησιμοποιούνται για sentiment classification σύμφωνα με τη βιβλιογραφία, οι οποίοι πρόκειται να εφαρμοστούν πάνω στο SOSNet Twitter Dataset και Suspicious Tweets Dataset και μετέπειτα να συγκριθούν ως προς την απόδοσή τους στο επόμενο κεφάλαιο.

4.1.4.1 Αφελής Bayes-Naive Bayes

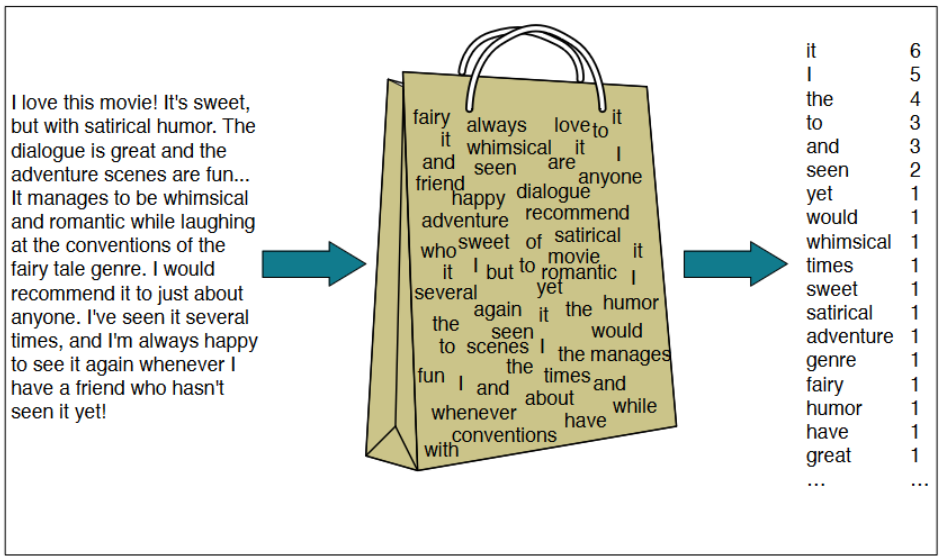
Ο Naive Bayes είναι ένας πιθανοτικός ταξινομητής, που σημαίνει ότι για ένα κείμενο d, από τις κλάσεις c∈C, ο ταξινομητής επιστρέφει την κλάση $\hat{c}$ που έχει τη μεγαλύτερη posterior probability μεταξύ των κλάσεων [7]. Με αυτόν τον συμβολισμό ($\hat{c}$) εννοείται η εκτίμηση που κάνει ο αλγόριθμος για τη σωστή κλάση δηλαδή ισχύει ότι:

$$\hat{c} = {}_{c \in C} argmax P(c|d)$$

(1)

Αυτή η ιδέα του Bayesian συμπεράσματος, έγινε γνωστή το 1763 από τον Bayes [63] και εφαρμόστηκε για πρώτη φορά στην ταξινόμηση κειμένου το 1964 [64]. Ο σκοπός της Bayesian ταξινόμησης είναι να χρησιμοποιήσει τον κανόνα του Bayes για να μετασχηματίσει την Εξ. (1) σε άλλες πιθανότητες που έχουν κάποιες χρήσιμες ιδιότητες.

Ο ταξινομητής έχει αντλήσει το όνομά του από μία αφελή (naive) υπόθεση που πραγματοποιεί, σχετικά με το πώς τα διάφορα χαρακτηριστικά αλληλεπιδρούν μεταξύ τους.



Εικόνα 65 Η διαίσθηση του πολυωνυμικού αφελούς ταξινομητή Bayes εφαρμόστηκε σε μια κριτική ταινίας. Η θέση των λέξεων αγνοείται (Bag of words) και χρησιμοποιείται η συχνότητα κάθε λέξης (Jurafsky & Martin,2013)

Όπως φαίνεται και στο παράδειγμα της παραπάνω εικόνας, ο αλγόριθμος αγνοεί τη θέση των λέξεων και λαμβάνει υπόψη του τη συχνότητα κάθε λέξης. Με απλά λόγια, ένας ταξινομητής Naive Bayes υποθέτει ότι η παρουσία ενός συγκεκριμένου χαρακτηριστικού σε μια κλάση δεν σχετίζεται με την παρουσία οποιουδήποτε άλλου χαρακτηριστικού. Παρά την απλουστευμένη φύση του, ο συγκεκριμένος ταξινομητής συνήθως έχει εντυπωσιακά αποτελέσματα. Ο υπολογισμός της ζητούμενης πιθανότητας γίνεται με τη βοήθεια του κανόνα:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

(2)

Μπορεί τότε να αντικατασταθεί η Εξ. (2) στην Εξ(1) για να ληφθεί η Εξ. (3):

$$\hat{c} = \ c \in C argmax P(c|d) = \ \ c \in C \ argmax \frac{P(d|c)P(c)}{P(d)}$$

(3)

Η παραπάνω ισότητα δύναται να απλοποιηθεί περαιτέρω εφόσον οι πιθανότητες των κλάσεων συγκρίνονται. Συνεπώς, θα έχουν όλες κοινό παρονομαστή που δεν θα χρησιμεύσει στη σύγκριση και μπορεί να απλοποιηθεί δηλαδή ισχύει :

$$\hat{c} = \ _{c \in C} argmax \ \overbrace{P(d|c)}^{likelihood} \overbrace{P(c)}^{prior}$$

(4)

Το Naive Bayes ονομάζεται γενεσιουργό μοντέλο [7] επειδή η (4) μπορεί να διαβαστεί ως αναφορά ενός είδους σιωπηρής υπόθεσης για το πώς δημιουργείται ένα έγγραφο: πρώτα

μια κλάση δειγματοληπτείται από το P(c), και, στη συνέχεια, οι λέξεις δημιουργούνται με δειγματοληψία από το P(d|c). (Στην πραγματικότητα θα μπορούσαμε να φανταστούμε τη δημιουργία τεχνητών εγγράφων, ή τουλάχιστον το πλήθος των λέξεων τους, ακολουθώντας αυτή τη διαδικασία).

Χωρίς βλάβη της γενικότητας, όπως έχει ήδη αναφερθεί, κάθε κείμενο αναπαρίσταται ως ένα σύνολο χαρακτηριστικών $f_1,f_2,.....,f_n$

Άρα ισχύει ότι :

$$\hat{c} = \underset{c \in C}{argmax} \overbrace{P(f1, f2, \ldots, fn|c)}^{likelihood} \overbrace{P(c)}^{prior}$$

(5)

Η πιθανότητα P(c) ονομάζεται prior probability και είναι εύκολο να υπολογιστεί αφού ισχύει πως:

$$p(c) = \frac{N_0}{N}$$

(6)

, όπου N_0 το πλήθος των κειμένων στην κλάση c και N το συνολικό πλήθος των κειμένων στο corpus.

Η likelihood πιθανότητα είναι αρκετά πιο δύσκολο να υπολογιστεί.Για να γίνει ο υπολογισμός πραγματοποιείται η παραδοχή πως όλες οι πιθανότητες $P(f_i|c)$ είναι ανεξάρτητες δοθείσης της κλάσης c και σαν συνέπεια μπορούν να πολλαπλασιαστούν, δηλαδή ισχύει πως:

$$P(f_1, f_2, \ldots, f_n|c) = P(f_1|c) \cdot P(f_2|c) \cdot \ldots \cdot P(f_n|c)$$

(7)

Για να εφαρμόσουμε τον απλό ταξινομητή Bayes στο κείμενο, πρέπει να λάβουμε υπόψη τις θέσεις των λέξεων, απλά περπατώντας κατά ένα δείκτη σε κάθε θέση λέξης στο έγγραφο:

positions ← όλες οι θέσεις των λέξεων σε κείμενο ελέγχου

$$c_{NB} = \underset{c \in C}{argmax} P(c) \prod_{i \in positions} P(w_i|c)$$

(8)

Οι υπολογισμοί Naive Bayes, όπως και οι υπολογισμοί για τη μοντελοποίηση της γλώσσας, γίνονται σε λογαριθμικό χώρο, για να αποφευχθούν τα πάρα πολύ μικρά νούμερα και να αυξηθεί η ταχύτητα. Επομένως η τελική εξίσωση είναι:

$$c_{NB} = argmax \ log P(c) + \sum log P(w_i|c)$$

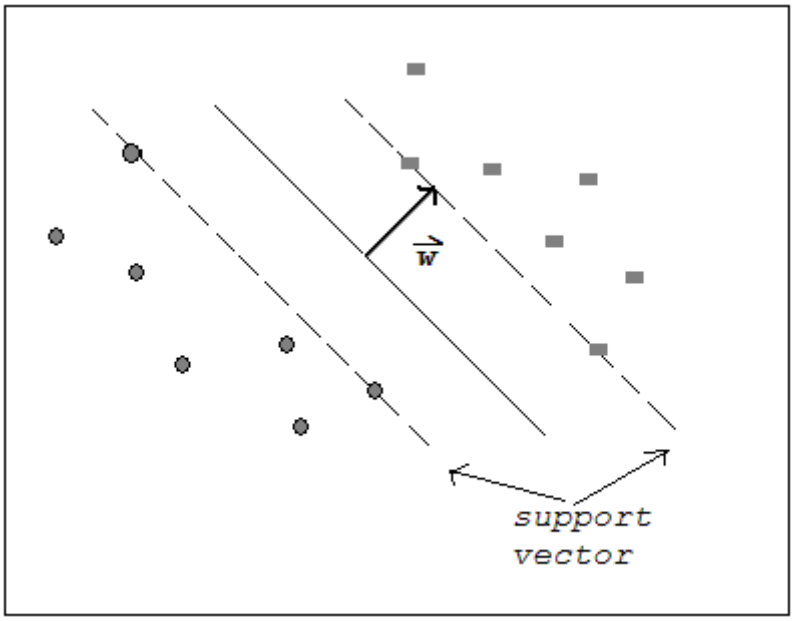
$$c \in C \qquad i \in positions$$

(9)

Οι λόγοι που ο συγκεκριμένος αλγόριθμος είναι τόσο δημοφιλής ,είναι η αρκετά καλή του απόδοση και ο μικρός χρόνος εκπαίδευσης.

4.1.4.2 Διανύσματα Υποστήριξης Μηχανής-SVM

Ο επόμενος διαδεδομένος αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για την κατηγοριοποίηση κειμένων είναι ο γνωστός αλγόριθμος SVM. Προτάθηκε από τον Vapnik [65] για την επίλυση προβλημάτων με 2 κλάσεις. Ο συγκεκριμένος αλγόριθμος βασίζεται στην εύρεση ενός διαχωρισμού μεταξύ υπερεπιπέδων που ορίζονται από κατηγορίες δεδομένων [8], όπως φαίνεται και στο παράδειγμα της παρακάτω εικόνας:



Εικόνα 66 Παράδειγμα του μοτίβου υπερπλάνου SVM (Basu, Walters & Shepherd, 2003)

Αυτό σημαίνει ότι ο SVM αλγόριθμος μπορεί να λειτουργήσει ακόμη και σε αρκετά μεγάλα σύνολα χαρακτηριστικών, αφού ο στόχος είναι να μετρηθεί το περιθώριο διαχωρισμού των δεδομένων αντί για το ταίριασμα των χαρακτηριστικών. Είναι ένας supervised αλγόριθμος, δηλαδή, τα κείμενα πρέπει να είναι προ κατηγοριοποιημένα για τη διαδικασία της εκπαίδευσης.

Ο λόγος δημοφιλίας του συγκεκριμένου αλγορίθμου είναι τα πολύ καλά αποτελέσματα που φαίνεται να έχει στην κατηγοριοποίηση κειμένου. Η έρευνα έχει δείξει [66] ότι ο SVM έχει καλή κλιμάκωση και απόδοση σε μεγάλα datasets. Χρησιμοποιώντας ολόκληρο το λεξιλόγιο σαν σύνολο χαρακτηριστικών, οι Rennie και Rifkin [9] βρήκαν πως ο SVM ξεπέρασε σε αποτελέσματα τον Naive Bayes σε 2 datasets: 19.997 κείμενα ειδήσεων χωρισμένα σε 20 κατηγορίες και 9649 κείμενα στον τομέα της βιομηχανίας χωρισμένα σε 105 κατηγορίες. Όπως ειπώθηκε και παραπάνω, ο Naïve Bayes αλγόριθμος βασίζεται στην υπόθεση ότι οι όροι που χρησιμοποιούνται στα κείμενα είναι ανεξάρτητοι. Τόσο ο Naïve Bayes όσο και ο SVM είναι αλγόριθμοι γραμμικοί, αποτελεσματικοί και κλιμακούμενοι σε μεγάλα σύνολα κειμένων [9]. Ο Joachims [67] χρησιμοποίησε ένα μειωμένο λεξιλόγιο σαν σύνολο χαρακτηριστικών, αφού προηγουμένως είχε εφαρμόσει Stemming και αφαίρεση των πολύ ασυνήθιστων λέξεων από το feature set. Χρησιμοποιώντας 12.902 κείμενα από το Reuters – 21578 κείμενα και 20000 ιατρικά abstracts από το Ohsumed corpus, ο Joachims συνέκρινε την απόδοση αρκετών

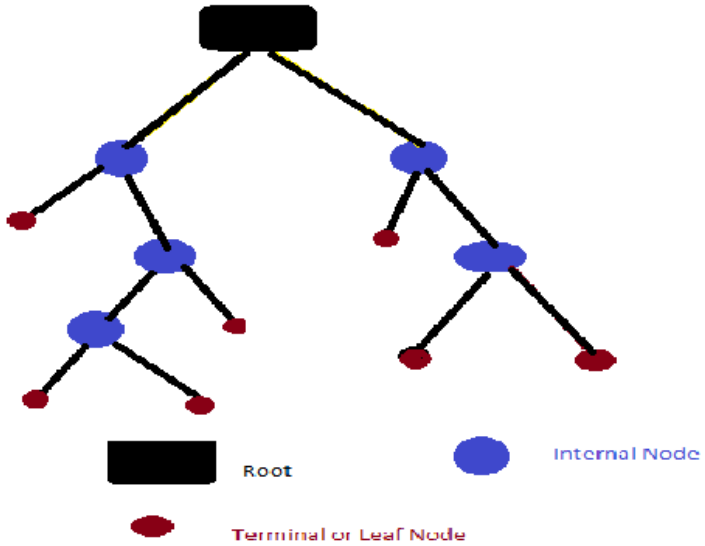
αλγορίθμων συμπεριλαμβανομένων των SVM και Naïve Bayes. Και για τα 2 σύνολα δεδομένων αυτός ο έλεγχος υπέδειξε πως ο SVM απέδωσε καλύτερα.

Σε άλλη έρευνα [68], χρησιμοποιώντας τη συλλογή Reuters- 21578, βρέθηκε ότι ο αλγόριθμος SVM είχε τη μεγαλύτερη ακρίβεια σε ένα σύνολο ελέγχου σε σύγκριση με τους Naïve Bayes και Decision Trees. Συνεπώς, με βάση τη βιβλιογραφία αναμένεται πως η χρήση αυτού του αλγορίθμου και στα 2 datasets που χρησιμοποιούνται στην παρούσα μελέτη (Διπλωματική) είναι πιθανό να επιφέρει αρκετά καλά αποτελέσματα ως προς την ακρίβεια.

4.1.4.3 Δέντρα Απόφασης-Decision Tree (DT)

Τα δέντρα απόφασης ενσωματώνουν μια εποπτευόμενη προσέγγιση ταξινόμησης [69]. Η ιδέα προήλθε από τη δομή του συνηθισμένου δέντρου που αποτελείται από ρίζα και κόμβους (θέσεις όπου τοποθετούνται οι διαιρέσεις των κλάδων), κλάδους και φύλλα. Με παρόμοιο τρόπο, ένα δέντρο απόφασης κατασκευάζεται από κόμβους που αντιπροσωπεύουν κύκλους και οι κλάδοι αντιπροσωπεύονται από τα τμήματα που συνδέουν τους κόμβους.

Τα DT είναι μια από τις τεχνικές ταξινόμησης στην εξόρυξη δεδομένων που χρησιμοποιεί διακλαδώσεις για την απεικόνιση κάθε εφικτού αποτελέσματος κατά τη λήψη μιας απόφασης. Ένα DT περιλαμβάνει τρία είδη κόμβων, τον “root node”, “internal node” και “leaf”. Ένα δέντρο απόφασης ξεκινά από τη ρίζα, κινείται προς τα κάτω και γενικά σχεδιάζεται από τα αριστερά προς τα δεξιά. Ο κόμβος από όπου ξεκινά το δέντρο ονομάζεται κόμβος “ρίζα”. Ο κόμβος όπου τερματίζεται η αλυσίδα είναι γνωστός ως κόμβος "φύλλο". Δύο ή περισσότεροι κλάδοι μπορούν να επεκταθούν από κάθε εσωτερικό κόμβο, π.χ κόμβος που δεν είναι κόμβος φύλλου. Ένας κόμβος αντιπροσωπεύει ένα ορισμένο χαρακτηριστικό, ενώ οι κλάδοι αντιπροσωπεύουν ένα εύρος των τιμών. Αυτά τα εύρη τιμών λειτουργούν ως σημεία κατάτμησης για το σύνολο των τιμών του δεδομένου χαρακτηριστικού [12].
Ο γνωστός ριζικός κόμβος (root node) χρησιμοποιείται ως αρχικό χαρακτηριστικό ή κορυφαίος κόμβος απόφασης σε ένα δέντρο και αντιστοιχεί στον καλύτερο δείκτη για τη λήψη αποφάσεων. Έχει μηδενικές εισερχόμενες άκρες, ενώ οι εσωτερικοί κόμβοι (internal nodes) έχουν τουλάχιστον μία εισερχόμενη και τουλάχιστον μία εξερχόμενη ακμή. Τέλος, οι κόμβοι φύλλα (leaf nodes) δεν διαθέτουν καμία εξερχόμενη ακμή, συνεπώς αναπαριστούν μια κατηγορία. Ουσιαστικά, το δέντρο απόφασης αναλύει ένα σετ (σύνολο) των δεδομένων σε όλο και μικρότερα υποσύνολα ενώ ταυτόχρονα η συσχέτιση της απόφασης χτίζεται σταδιακά. Στην παρακάτω εικόνα παρουσιάζεται η δομή ενός δέντρου απόφασης.



Εικόνα 67 Η δομή ενός δέντρου απόφασης (Ali et al.,2012)

Σε αυτό το σημείο θα παρουσιασθούν τα πλεονεκτήματα και τα μειονεκτήματα των δέντρων απόφασης όπως αποτυπώθηκαν από τη C.Petri [70]

Πλεονεκτήματα

- Γραφικά. Μπορούν να αντιπροσωπεύσουν εναλλακτικές αποφάσεις, πιθανά αποτελέσματα και τυχαία γεγονότα σχηματικά. Η οπτική προσέγγιση είναι ιδιαίτερα χρήσιμη στην κατανόηση διαδοχικών αποφάσεων και εξαρτήσεων από τα αποτελέσματα.
- Αποτελεσματικά. Μπορούν να εκφράσουν γρήγορα περίπλοκες εναλλακτικές με σαφήνεια. Είναι δυνατό να τροποποιηθεί εύκολα ένα δέντρο αποφάσεων καθώς γίνονται διαθέσιμες νέες πληροφορίες. Η δημιουργία ενός δέντρου αποφάσεων διευκολύνει τη σύγκριση των επιδράσεων διαφορετικών τιμών εισόδου στις διάφορες εναλλακτικές αποφάσεις. Ο τυπικός συμβολισμός δέντρου αποφάσεων είναι εύκολο να υιοθετηθεί.
- Αποκαλυπτικά. Μπορούν να συμβάλλουν στη σύγκριση ανταγωνιστικών εναλλακτικών - ακόμη και χωρίς ολοκληρωμένες πληροφορίες-από άποψη κινδύνου και πιθανής αξίας. Ο όρος της αναμενόμενης τιμής (EV) συνδυάζει το σχετικό επενδυτικό κόστος, τις αναμενόμενες αποδόσεις και τις αβεβαιότητες σε μια ενιαία αριθμητική τιμή. Το EV αποκαλύπτει τα συνολικά πλεονεκτήματα των ανταγωνιστικών εναλλακτικών.
- Συμπληρωματικά. Τα δέντρα αποφάσεων μπορούν να χρησιμοποιηθούν σε συνδυασμό με άλλα εργαλεία διαχείρισης έργων. Για παράδειγμα, η μέθοδος του δέντρου αποφάσεων μπορεί να βοηθήσει στην αξιολόγηση χρονοδιαγραμμάτων έργων [71].
- Τα δέντρα απόφασης είναι αυτονόητα και όταν συμπιέζονται είναι επίσης εύκολο να ακολουθηθούν. Με άλλα λόγια, εάν τα δέντρα απόφασης έχουν λογικό αριθμό φύλλων, μπορούν να γίνουν κατανοητά από **μη-επαγγελματίες χρήστες**. Επιπλέον, τα δέντρα απόφασης μπορούν να είναι μετατραπούν σε ένα σύνολο κανόνων. Έτσι, αυτή η αναπαράσταση θεωρείται ως κατανοητή.

- Τα δέντρα αποφάσεων μπορούν να χειριστούν τόσο ονομαστικά όσο και αριθμητικά χαρακτηριστικά.
- Η αναπαράσταση δέντρων απόφασης είναι αρκετά πλούσια ώστε να αντιπροσωπεύει οποιονδήποτε ταξινομητή διακριτής τιμής.
- Τα δέντρα αποφάσεων είναι ικανά να χειρίζονται σύνολα δεδομένων που μπορεί να έχουν σφάλματα.
- Τα δέντρα αποφάσεων είναι ικανά να χειρίζονται σύνολα δεδομένων που μπορεί να έχουν τιμές που λείπουν.
- Τα δέντρα απόφασης θεωρούνται ως μη-παραμετρική μέθοδος. Αυτό σημαίνει ότι τα δέντρα απόφασης δεν έχουν υποθέσεις σχετικά με την κατανομή του χώρου και τη δομή του ταξινομητή.

Μειονεκτήματα

- Οι περισσότεροι από τους αλγόριθμους (όπως ID3 και C4.5) απαιτούν ότι το χαρακτηριστικό-στόχος θα έχει μόνο διακριτές τιμές.

- Καθώς τα δέντρα αποφάσεων χρησιμοποιούν τη μέθοδο «διαίρει και βασίλευε», τείνουν να αποδίδουν καλά, εάν υπάρχουν λίγα υψηλά συσχετιζόμενα χαρακτηριστικά, αλλά λιγότερο καλά εάν υπάρχουν πολλά χαρακτηριστικά με μεγάλη συσχέτιση. Ένας από τους λόγους για αυτό είναι ότι άλλοι ταξινομητές μπορούν να περιγράψουν με συμπαγή τρόπο έναν ταξινομητή που θα ήταν πολύ δύσκολο να αναπαρασταθεί χρησιμοποιώντας ένα δέντρο αποφάσεων.

- Το άπληστο χαρακτηριστικό των δέντρων απόφασης οδηγεί σε ένα άλλο μειονέκτημα που πρέπει να επισημανθεί. Αυτό είναι η υπερβολική ευαισθησία του στο training set, τα άσχετα χαρακτηριστικά και τον θόρυβο [72].

4.1.4.4 Κ-κοντυνότεροι Γείτονες-KNN

Ένας από τους πιο απλούς αλγόριθμους κατηγοριοποίησης στην εξόρυξη δεδομένων και τη μηχανική μάθηση είναι ο K-Nearest Neighbors (KNN). Πρόκειται για την πιο αποδεκτή μέθοδο κατηγοριοποίησης εξαιτίας της ευκολίας και της πρακτικής αποτελεσματικότητάς της: δεν απαιτεί την τοποθέτηση (fitting) σε μοντέλο και έχει αποδειχθεί ότι έχει ανώτερη απόδοση για την ταξινόμηση πολλών τύπων δεδομένων. Όμως, η ανώτερη απόδοση ταξινόμησης του kNN εξαρτάται σε μεγάλο βαθμό από τη μετρική που χρησιμοποιείται για τον υπολογισμό των αποστάσεων μεταξύ των ζευγών σημείων δεδομένων. Μία μετρική απόστασης d έχει τις εξής 3 ιδιότητες [10]:

- $d(x, y) \geq 0$ για κάθε feature vector x, y και $d(x,y)=0$ αν και μόνο αν $x=y$
- $d(x, y)=d(y, x)$
- $d(x, z) \leq d(x, y) + d(y, z)$

Η ιδιότητα 1 μας διαβεβαιώνει ότι η απόσταση είναι πάντα μη-αρνητική, και ο μόνος τρόπος για να μηδενιστεί η απόσταση είναι οι συντεταγμένες (π.χ. στο διάγραμμα διασποράς) να είναι ίδιες.

Η ιδιότητα 2 υποδεικνύει την αντιμεταθετικότητα, έτσι ώστε, για παράδειγμα, η απόσταση από τη Νέα Υόρκη στο Λος Άντζελες είναι ίδια με την απόσταση από το Λος Άντζελες στη Νέα Υόρκη [73].

Η ιδιότητα 3 είναι η τριγωνική ανισότητα, η οποία δηλώνει ότι η εισαγωγή ενός τρίτου σημείου δεν μπορεί ποτέ να μειώσει την απόσταση ανάμεσα σε δύο άλλα σημεία [74].

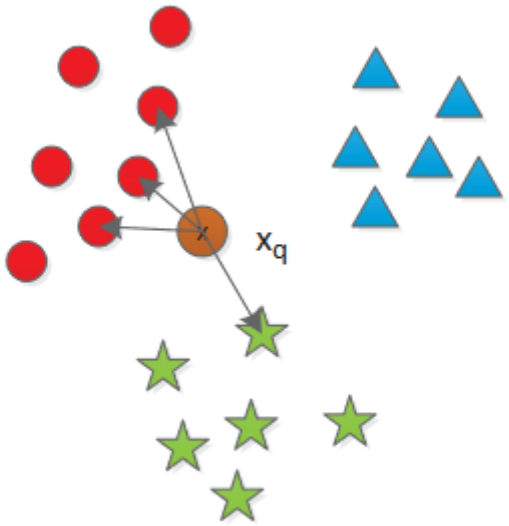
Η πιο κοινή συνάρτηση απόστασης είναι η Ευκλείδεια απόσταση, που αντιπροσωπεύει τον συνήθη τρόπο με τον οποίο σκέφτονται οι άνθρωποι την απόσταση στον πραγματικό κόσμο [10]:

$$deuclidean(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

(10)

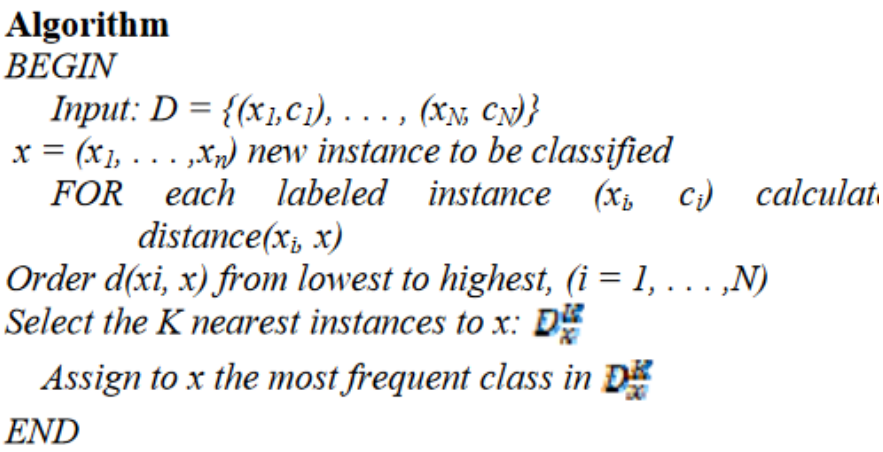
όπου x = x_1, x_2, . . . , x_m, και y = y_1, y_2, . . . , y_m αντιπροσωπεύουν τις τιμές των χαρακτηριστικών m δύο εγγραφών.

Πρακτικά, για να υπολογιστούν οι k πλησιέστεροι γείτονες , οι Ευκλείδειες αποστάσεις χρησιμοποιούνται συχνά ως μετρική ομοιότητας. Κάποιος ερευνητής συχνά χρειάζεται να ανακαλύψει ή να επιλέξει μία καλή μέτρηση απόστασης για την κατηγοριοποίηση δεδομένων υψηλών διαστάσεων σε πραγματικές εφαρμογές. Οι κανόνες ταξινόμησης του kNN δημιουργούνται μόνο από τα δείγματα εκπαίδευσης, χωρίς άλλα πρόσθετα δεδομένα. Σε μία περισσότερο περίπλοκη προσέγγιση, η ταξινόμηση kNN, βρίσκει μια ομάδα των k αντικειμένων στο σετ εκπαίδευσης που είναι πιο κοντά στο αντικείμενο ελέγχου και βασίζει την εκχώρηση μιας ετικέτας στην επικρατέστερη κλάση σε αυτήν τη γειτονιά.



Εικόνα 68 Η βασική αρχή του KNN αλγορίθμου (Zhongguo et al., 2017)

Ο αλγόριθμος k-Nearest Neighbor (kNN) είναι μια μέθοδος για ταξινόμηση αντικειμένων με βάση τα πλησιέστερα παραδείγματα εκπαίδευσης στο χώρο των χαρακτηριστικών. Ο kNN είναι ένας τύπος μάθησης που βασίζεται σε instance-based learning ή lazy learning όπου η συνάρτηση προσεγγίζεται τοπικά και όλοι οι υπολογισμοί αναβάλλονται μέχρι την ταξινόμηση. Ακολουθεί ψευδοκώδικας του αλγορίθμου:



Εικόνα 69 Ψευδοκώδικας του αλγορίθμου KNN (Nikhath et al., 2016)

Η επιλογή του χώρου χαρακτηριστικών, το σύνολο δεδομένων εκπαίδευσης που χρησιμοποιείται και η τιμή του k μπορεί να επηρεάσει εξαιρετικά την ακρίβεια της ταξινόμησης [75]

Αν και οι μελέτες έχουν επικεντρωθεί σε αυτό το θέμα για μεγάλο χρονικό διάστημα, η επιλογή της τιμής k για τον k-NN αλγόριθμο εξακολουθεί να είναι πολύ δύσκολη και να αποτελεί πρόκληση [77,78].

Για παράδειγμα, κάποιοι ερευνητές [79]υποστηρίζουν πως η καλύτερη επιλογή για το k είναι η τιμή $\sqrt{N}$ για τα σύνολα δεδομένων με μέγεθος δείγματος μεγαλύτερο από 100. Ωστόσο, μια τέτοια επιλογή έχει αποδειχθεί ότι δεν είναι κατάλληλη για όλες τις περιπτώσεις συνόλων δεδομένων [80]. Οι Wang et al. [81] εφάρμοσαν μια μέθοδο για την εκτίμηση του βέλτιστου k από τη γωνία της στατιστικής απόφασης. Η μέθοδός τους προσαρμόζει δυναμικά την τιμή k έως ότου ένα έχει επιτευχθεί ένα ικανοποιητικό επίπεδο ακρίβειας.

Επιπλέον, αξίζει να σημειωθούν δύο βασικά σημεία για μερικά πραγματικά σύνολα δεδομένων. Πρώτον, η βέλτιστη τιμή του k στην πλειοψηφία των συνόλων είναι 1 [76] . Δεύτερον, ορισμένα σύνολα δεδομένων δεν είναι ευαίσθητα στην επιλογή της τιμής k.

Στην παρακάτω εικόνα συνοψίζονται όσα ειπώθηκαν για τους αλγορίθμους KNN, SVM και DT και προβαίνει σε σύγκρισή τους [11].

KNN	SVM	DT
Can be used for continuous value inputs.	Can be used for continuous value inputs.	Can be used for continuous and categorical inputs.
Algorithm is simple with straight forward classifier easy understand.	Mathematically complex and hard to build own algorithm.	Data classification with less calculation involved. Easily understand.
It is automatically non-linear, able to detect linear and non-linear distributed data. Perform very well a lot of data points.	Can be used in linear and non-linear ways and good with limited set of points in many dimensions.	It is non-linear classifier. Able to illustrate relationship between independent and dependent variables.
Do classification by determine neighbourhoods.	Do classification by searches for closest points.	Do classification by form a tree.
Computationally expensive.	Computationally expensive to train	Computationally low end.
Not suitable for auto classification.	Suitable for auto classification technique but a bit complex.	Suitable for auto classification technique and less complex.
Time consuming.	Time consuming when processing large amount	Time consuming if involves multiple branches.

Εικόνα 70 Σύγκριση αλγορίθμων KNN,SVM,DT (M.U. Noormanshah, N.E. Nohuddin & Zainol, 2018)

Όπως γίνεται εμφανές, οι αλγόριθμοι δεν είναι γραμμικοί με εξαίρεση τον SVM που μπορεί να χρησιμοποιηθεί και με γραμμικό τρόπο, αν προβληθούν τα διανύσματα σε μεγαλύτερη διάσταση. Ο KNN αλγόριθμος είναι, γενικά, ένας χρονοβόρος αλγόριθμος, αφού απαιτείται, όπως αναλύθηκε, ο υπολογισμός της ευκλείδιας απόστασης μεταξύ του κάθε δυνατού ζευγαριού που απαρτίζεται από το test point και το train point. Οι άλλοι 2 είναι και εκείνοι χρονοβόροι κάτω από συγκεκριμένες συνθήκες. Ωστόσο, Ο KNN είναι πολύ πιο απλός αλγόριθμος και εύκολος στην κατανόηση, σε αντίθεση, κυρίως, με τον SVM, που είναι αρκετά πολύπλοκος αλγόριθμος.

4.1.4.5 Τυχαίο δάσος-Random Forest

Το Random Forest που αναπτύχθηκε από τον Leo Breiman [82] είναι μια ομάδα μη-κλαδεμένων δέντρων ταξινόμησης ή παλινδρόμησης δημιουργημένα από την τυχαία επιλογή δειγμάτων των δεδομένων εκπαίδευσης. Στην επαγωγική διαδικασία επιλέγονται τυχαία χαρακτηριστικά. Η πρόβλεψη στην περίπτωση της ταξινόμησης γίνεται με την τεχνική majority voting, κατά την οποία επιλέγεται η κλάση που συναντήθηκε στην πλειοψηφία των δέντρων που τέθηκε το ερώτημα [12].

Συγκεκριμένα, κάθε δέντρο δημιουργείται [83]:

- Με δειγματοληψία Ν τυχαίων δειγμάτων με αντικατάσταση από τα αρχικά δεδομένα. Αυτό το δείγμα θα χρησιμοποιηθεί ως σετ εκπαίδευσης για την καλλιέργεια του δέντρου

- Για Μ αριθμό μεταβλητών εισόδου, διαλέγονται m<<M τυχαία από τις αρχικές ώστε να δημιουργηθούν οι διακλαδώσεις. Με αυτόν τον τρόπο μειώνεται πολύ η διακύμανση (variance) σε περίπτωση εξαρτημένων μεταβλητών, κάτι που δεν μπορεί να αντιμετωπίσει ένα απλό DT.
- Κάθε δέντρο αναπτύσσεται στο μεγαλύτερο βαθμό. Δεν χρησιμοποιείται η τεχνική pruning (κλαδέματος).

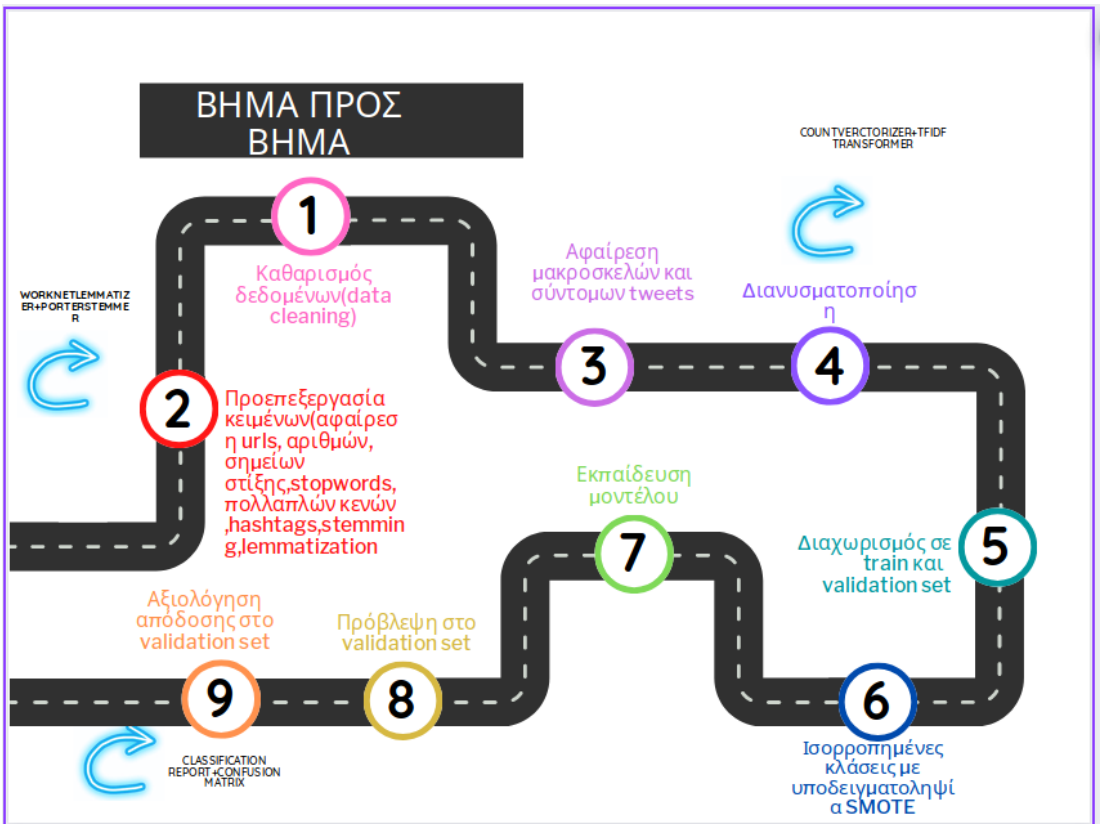
Τα πλεονεκτήματα του Random Forest είναι [84]:

- Ξεπερνάει το πρόβλημα του overfitting.
- Στα δεδομένα εκπαίδευσης, είναι λιγότερο επιρρεπές στις ακραίες τιμές (outliers).
- Οι παράμετροι μπορούν να ρυθμιστούν εύκολα και επομένως εξαλείφεται η ανάγκη για κλάδεμα των δέντρων.
- Η ακρίβεια και σημασία των μεταβλητών παράγεται αυτομάτως.

Το Random Forest, όχι μόνο διατηρεί τα οφέλη που επιτυγχάνονται από τα Decision Trees, αλλά με τη χρήση Bagging σε δείγματα, το σύστημα ψηφοφορίας του [82] μέσω του οποίου λαμβάνεται η απόφαση και ένα τυχαίο υποσύνολο μεταβλητών, τις περισσότερες φορές επιτυγχάνει καλύτερα αποτελέσματα από τα Decision Trees.

4.1.5 Επιλεγμένη μεθοδολογία

Μετά από την παρουσίαση των προτεινόμενων από τη βιβλιογραφία τεχνικών και μεθοδολογιών, σε αυτήν την ενότητα παρουσιάζεται η μεθοδολογία που ακολουθήθηκε στην παρούσα μελέτη (Διπλωματική) για καθέναν από τους αλγορίθμους που πρόκειται να εφαρμοστούν και συγκριθούν στα 2 σύνολα δεδομένων.

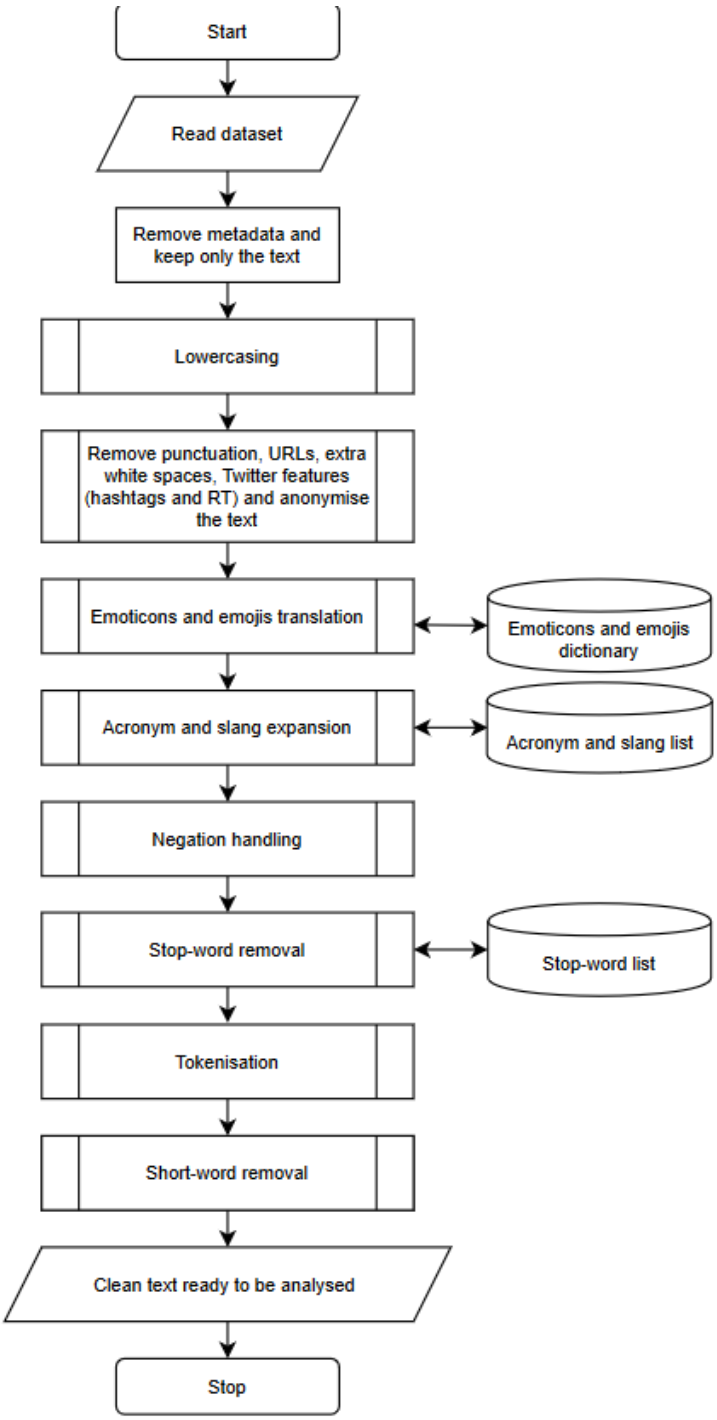


Εικόνα 71 Μεθοδολογία βήμα προς βήμα

Το παραπάνω διάγραμμα ροής αποτυπώνει οπτικά την επιλεγόμενη μεθοδολογία βήμα προς βήμα. Όπως φαίνεται και από το διάγραμμα, τα βήματα που πραγματοποιήθηκαν ήταν τα ακόλουθα:

- Καθαρισμός δεδομένων-Data cleaning. Πιο συγκεκριμένα, αφαιρέθηκαν τα διπλότυπα (duplicates) και οι γραμμές που περιείχαν ελλείψεις σε τιμές (missing values).
- Προεπεξεργασία κειμένων. Συγκεκριμένα, πραγματοποιήθηκε α)lower casing δηλαδή μετατροπή όλων των κεφαλαίων γραμμάτων στα αντίστοιχα πεζά τους β) αφαίρεση σημείων στίξης, hashtags και urls από τα κείμενα καθώς λόγω της φύσης των δεδομένων του Twitter αυτοί οι χαρακτήρες συναντώνται αρκετά συχνά, γ) αφαίρεση όλων των emojis με τη βοήθεια της βιβλιοθήκης emoji δ)Stemming με PorterStemmer και ε) Lemmatization με WordNetLemmatizer.
- Διανυσματοποίηση των κειμένων. Δημιουργήθηκε ένα pipeline με CountVectorizer και TfidfTransformer να εφαρμόζονται στα προεπεξεργασμένα κείμενα.
- Διαχωρισμός των δεδομένων σε σύνολο εκπαίδευσης (80%) και σύνολο επαλήθευσης (20%).
- Υποδειγματοληψία SMOTE. Όπως αναφέρθηκε και αιτιολογήθηκε στην ενότητα Μήκος κειμένων , προκειμένου να διατηρηθεί ένα tweet στο dataset τέθηκε σαν κάτω όριο λέξεων ο αριθμός 3 και σαν άνω όριο ο αριθμός 100. Οι αριθμοί αυτοί, φυσικά, ποικίλουν ανάλογα το dataset και επιλέχθηκαν έπειτα από σχετική διερεύνηση. Η βασική ιδέα είναι πως πολύ μακροσκελή ή πολύ σύντομα tweets δεν θα περιέχουν σημαντικές πληροφορίες σχετικά με το διαδικτυακό εκφοβισμό, δηλαδή το υπό εξέταση θέμα. Ωστόσο, μετά την αφαίρεση αυτών των tweets οι κλάσεις παύουν να είναι ισορροπημένες (στο SOSNet Twitter Dataset), οπότε πραγματοποιήθηκε resampling με τη βοήθεια της συνάρτησης SMOTE.

Οι τεχνικές που περιγράφηκαν στις προηγούμενες ενότητες εφαρμόζονται σε οποιαδήποτε εφαρμογή για text classification. Ωστόσο, στην παρούσα μελέτη, τα δεδομένα που χρησιμοποιούνται έχουν την ιδιαιτερότητα ότι προέρχονται από τα κοινωνικά δίκτυα και συγκεκριμένα το Twitter. Για τα συγκεκριμένα δεδομένα η μεθοδολογία που προτείνεται από τη σύγχρονη βιβλιογραφία [4] παρουσιάζεται στην αμέσως επόμενη εικόνα:



Εικόνα 72 Χειρισμός δεδομένων που προέρχονται από το Twitter (Palomino & Aider, 2022)

Συνεπώς, προκειμένου να γίνει η τελική επιλογή της μεθοδολογίας, μελετήθηκαν τόσο οι κοινές πρακτικές σε εφαρμογές text classification αλλά ελήφθη σοβαρά υπόψη, η φύση των δεδομένων του Twitter και η κατάλληλη αντιμετώπισή τους.

5. ΑΝΑΛΥΣΗ

Σε αυτήν την ενότητα θα δοθούν τα αποτελέσματα κατηγοριοποίησης κειμένου για κάθε έναν από τους αλγόριθμους που περιεγράφηκαν στην ενότητα Αλγόριθμοι Εκπαίδευσης και θα πραγματοποιηθεί σύγκριση των αποτελεσμάτων. Η παρουσίαση των αποτελεσμάτων κάθε αλγορίθμου πραγματοποιείται με την παρουσίαση των εξής:

- Έκθεση Ταξινόμησης-Classification report
- Πίνακας σύγχυσης-Confusion matrix

όπως ακριβώς προέκυψαν από την εκτέλεση.

Πριν όμως την παρουσίαση των αποτελεσμάτων, πραγματοποιείται μία συνοπτική παρουσίαση των μετρικών αξιολόγησης ώστε να είναι πιο εύκολη η μετέπειτα ερμηνεία των τιμών τους.

5.1 Μετρικές Έκθεσης Ταξινόμησης-Classification report

Οι χρησιμοποιούμενες μετρικές για το classification report είναι οι : ακρίβεια accuracy, ακρίβεια precision, ανάκληση recall, σκορ f1 -f1-score.

5.1.1 Ακρίβεια Accuracy

Ισχύει πώς:

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

1. TN / True Negative: όταν μια περίπτωση είναι αρνητική και προβλέπεται ως αρνητική
2. TP / True Positive: όταν ένα κρούσμα είναι θετικό και προβλέπεται ως θετικό
3. FN / False Negative: όταν ένα κρούσμα είναι θετικό αλλά προβλέπεται ως αρνητικό
4. FP / False Positive: όταν ένα κρούσμα είναι αρνητικό αλλά προβλέπεται θετικό

Οι 4 παραπάνω όροι [94] αφορούν δυαδική ταξινόμηση με κλάσεις θετικό/αρνητικό.

Δηλαδή, με τον όρο αυτό εννοείται το κλάσμα με αριθμητή το πλήθος των σωστών προβλέψεων και παρονομαστή το συνολικό πλήθος προβλέψεων. Για το πρώτο υπό εξέταση dataset - που δεν πρόκειται απλώς για ταξινόμηση σε 2 κλάσεις στον αριθμητή - θα υπήρχε το άθροισμα των κειμένων που κατηγοριοποιήθηκαν σωστά στις 5 κατηγορίες (5 όροι αθροίσματος) και αντίστοιχα και ο παρονομαστής θα περιείχε 10 όρους (σωστές και λάθος προβλέψεις ανά κατηγορία/κλάση). Αντιστοίχως, για το Suspicious Tweets Dataset που οι κατηγορίες είναι 3 στον αριθμό, στον αριθμητή της μετρικής accuracy, θα υπάρχουν 3 όροι (πλήθος κειμένων που κατηγοριοποιήθηκαν ορθά ανά κατηγορία) και στον παρονομαστή 6 όροι (σωστές και λάθος κατηγοριοποιήσεις ανά κατηγορία).

Η ακρίβεια είναι ένα μέτρο για το πόσες σωστές προβλέψεις έκανε το μοντέλο για το πλήρες σύνολο δεδομένων δοκιμής. Η ακρίβεια είναι μια καλή βασική μέτρηση για τη

μέτρηση της απόδοσης του μοντέλου. Σε μη-ισορροπημένα σύνολα δεδομένων, η ακρίβεια γίνεται κακή μέτρηση [95]

5.1.2 Ανάκληση Recall

Όσον αφορά τον όρο recall ισχύει ότι [94]:

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

Δηλαδή, recall για την κλάση i είναι, από όλες τις τιμές που ανήκουν πραγματικά στην κλάση i (παρονομαστής), πόσες προβλέπονται σωστά ως κλάση i (αριθμητής).

5.1.3 Ακρίβεια Precision

Ο όρος precision εκφράζει από όλες τις προβλεπόμενες τιμές μίας κλάσης i (παρονομαστής), πόσες ανήκουν πραγματικά στην κλάση i (αριθμητής), δηλαδή, με βάση την ορολογία TP,FP,TN,FN, η μετρική precision μίας κλάσης i ορίζεται ως:

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

5.1.4 F1-score

Για να έχουμε ένα συνδυασμένο αποτέλεσμα precision και recall, χρησιμοποιούμε το F1-score. Το F1-score είναι ο αρμονικός μέσος όρος των μετρικών precision και recall. Συνεπώς, ορίζεται ως :

$$F1 - score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

5.2 Πίνακας σύγχυσης-Confusion Matrix

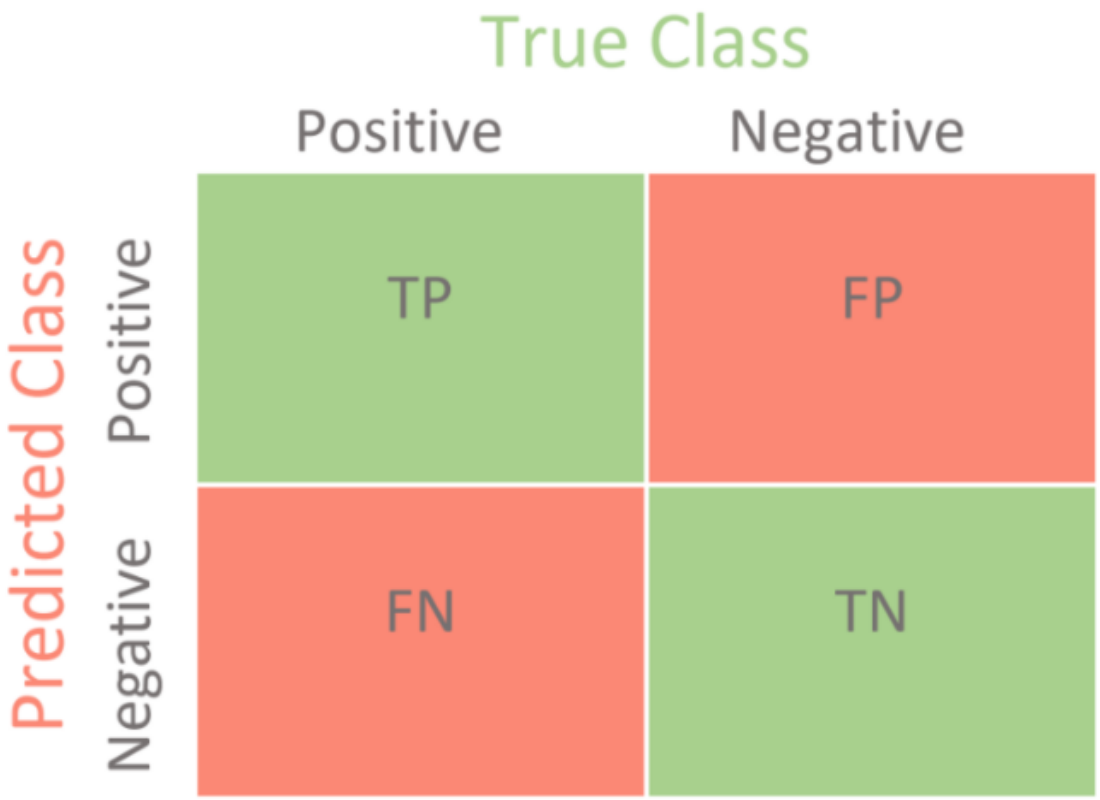
- Ο πίνακας σύγχυσης (confusion matrix) είναι ένα εργαλείο για την προγνωστική ανάλυση στη μηχανική μάθηση [95]. Προκειμένου να ελεγχθεί η απόδοση ενός μοντέλου μηχανικής εκμάθησης που βασίζεται σε ταξινόμηση, αναπτύσσεται ο συγκεκριμένος πίνακας/μήτρα.

- Επίσης, μπορούμε να πούμε ότι ο confusion matrix είναι ένας συνοπτικός πίνακας του αριθμού των σωστών και λανθασμένων προβλέψεων που παράγονται από έναν ταξινομητή (ή μοντέλο ταξινόμησης) για εργασίες δυαδικής κυρίως ταξινόμησης .

- Ένας πίνακας σύγχυσης είναι ένας πίνακας N x N που χρησιμοποιείται για την αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης, όπου N είναι ο αριθμός

των στόχων. Με την οπτικοποίηση του πίνακα σύγχυσης, ένας παρατηρητής θα μπορούσε να προσδιορίσει την ακρίβεια του μοντέλου παρατηρώντας τις διαγώνιες τιμές για τη μέτρηση του αριθμού ακριβούς ταξινόμησης.

Στην παρακάτω εικόνα παρουσιάζεται η δομή του confusion matrix



Εικόνα 73 Η δομή του πίνακα σύγχυσης-confusion matrix (Karimi,2021)

Ο πίνακας σύγχυσης έχει τη μορφή τετραγωνικού πίνακα όπου η στήλη αντιπροσωπεύει τις πραγματικές τιμές και η σειρά απεικονίζει την προβλεπόμενη τιμή του μοντέλου και αντίστροφα.

Μπορεί ο confusion matrix να χρησιμοποιείται κυρίως για περιπτώσεις δυαδικής ταξινόμησης, ωστόσο είναι εφικτό να επεκταθεί και σε περισσότερες κλάσεις με παρόμοια λογική. Για παράδειγμα, στην περίπτωση του δεύτερου dataset που υπάρχουν 3 κατηγορίες, ο confusion matrix θα είναι και πάλι ένας τετραγωνικός πίνακας αυτήν τη φορά 3Χ3 όπου και πάλι η διαγώνιος θα αντιπροσωπεύει το πλήθος των στοιχείων που κατηγοριοποιήθηκαν σωστά και τα υπόλοιπα στοιχεία θα αφορούν τις λανθασμένες προβλέψεις. Ομοίως για το SOSNet Twitter Dataset, ο confusion matrix θα είναι 5Χ5 με αντίστοιχη δομή.

5.3 Αποτελέσματα Naive Bayes

Σε αυτήν την ενότητα θα παρουσιασθούν και θα σχολιασθούν τα αποτελέσματα των αλγορίθμων εκπαίδευσης στο σύνολο επαλήθευσης (validation set) και για τα δύο datasets.

5.3.1 SOSNet Twitter Dataset

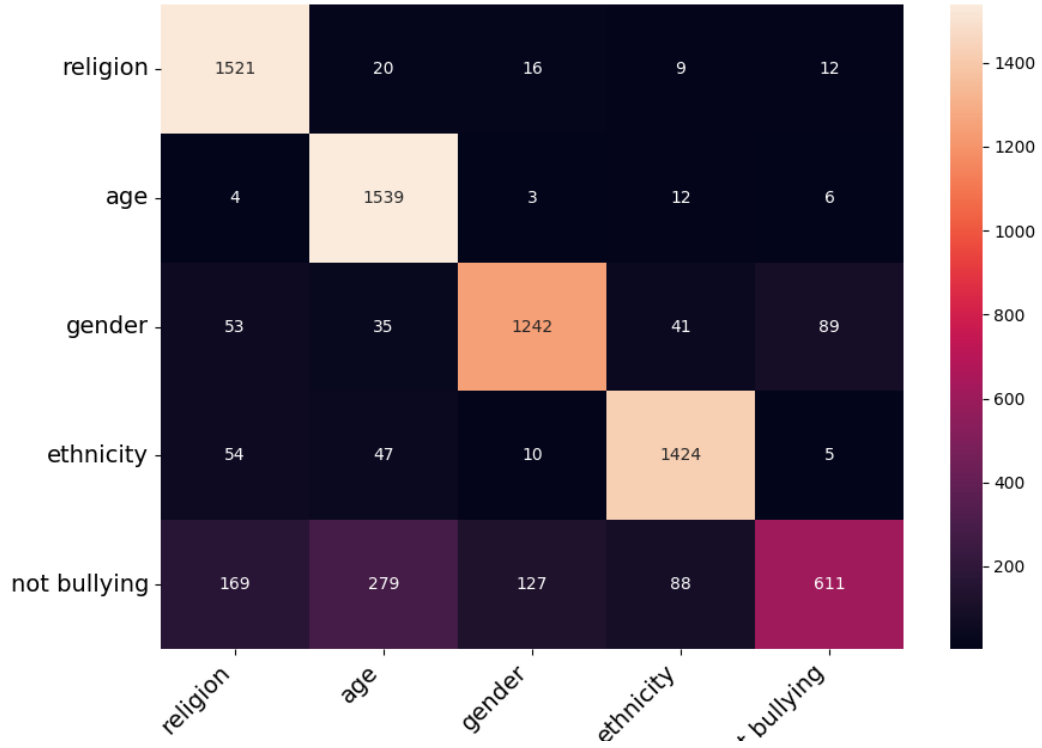
Classification Report for Naive Bayes:

	precision	recall	f1-score	support
religion	0.85	0.96	0.90	1578
age	0.80	0.99	0.88	1564
gender	0.89	0.85	0.87	1460
ethnicity	0.90	0.92	0.91	1540
not bullying	0.85	0.49	0.62	1274
accuracy			0.86	7416
macro avg	0.86	0.84	0.84	7416
weighted avg	0.86	0.86	0.85	7416

Εικόνα 74 ΄Εκθεση ταξινόμησης-Classification report αλγορίθμου Naive Bayes για το SOSNet Twitter Dataset

Όπως φαίνεται από την παραπάνω εικόνα, αλγόριθμος Naive-Bayes παρήγαγε αρκετά καλά αποτελέσματα κατά την κατηγοριοποίηση κάθε κειμένου. Πιο συγκεκριμένα, επετεύχθη συνολική ακρίβεια 86%. Εστιάζοντας σε κάθε κατηγορία ξεχωριστά τόσο στην έκθεση ταξινόμησης-classification report όσο και στον πίνακα σύγχυσης- confusion matrix, γίνεται αντιληπτό πως τα χειρότερα αποτελέσματα παρατηρήθηκαν για την κατηγορία “Not cyberbullying (Όχι διαδικτυακός εκφοβισμός)”. Το recall για τη συγκεκριμένη κατηγορία είναι 49%, που σημαίνει ότι ο αλγόριθμος προβλέπει σωστά μία στις 2 περιπτώσεις για αυτήν την κατηγορία (απόδοση ίδια με έναν random classifier). Αυτό είναι και το μικρότερο ποσοστό που παρατηρείται. Από την άλλη πλευρά για την κατηγορία “Age cyberbullying (Ηλικιακός διαδικτυακός εκφοβισμός)” ο αλγόριθμος εμφανίζει recall 99%, δηλαδή, ανιχνεύει σχεδόν όλα τα κείμενα αυτής της κατηγορίας. Εστιάζοντας στη μετρική f1-score που είναι πιο αντιπροσωπευτική των αποτελεσμάτων, καθώς συνοψίζει το accuracy και το recall, παρατηρείται καλύτερη απόδοση με 91% στην κλάση “Ethnicity cyberbullying (Διαδικτυακός εκφοβισμός εθνικότητας)” και χειρότερη όπως αναμενόταν στην κατηγορία “Not cyberbullying (Όχι διαδικτυακός εκφοβισμός)”.

Στην παρακάτω εικόνα παρουσιάζεται και ο confusion matrix, για τον συγκεκριμένο αλγόριθμο. Όπως αναφέρθηκε και στην ενότητα Confusion Matrix, ο πίνακας είναι 5X5 καθώς οι κατηγορίες διαδικτυακού εκφοβισμού που μελετώνται είναι συνολικά 5. Παρατηρώντας τα διαγώνια στοιχεία του πίνακα γίνεται εύκολα αντιληπτό πως το υψηλότερο TP συναντάται στην κατηγορία “Age cyberbullying (Ηλικιακός διαδικτυακός εκφοβισμός)” με 1539 κείμενα. Αυτό είναι αναμενόμενο, καθώς, όπως φάνηκε και από το classification report η συγκεκριμένη κατηγορία έχει το υψηλότερο recall (99%) και ο αριθμητής της μετρικής αυτής είναι ο παράγοντας TP (ενότητα Recall).



Εικόνα 75 Confusion matrix αλγορίθμου Naive Bayes για το SOSNet Twitter Dataset

Στην κατηγορία «Not cyberbullying (Όχι διαδικτυακός εκφοβισμός)», τα αποτελέσματα είναι τα χειρότερα, όπως φάνηκε και από το classification report. Από το confusion matrix λαμβάνονται επιπρόσθετες πληροφορίες σχετικά με αυτήν την κακή επίδοση. Για παράδειγμα, φαίνεται πως στις περισσότερες περιπτώσεις που έγινε λάθος κατηγοριοποίηση σε αυτού του είδους τα κείμενα, η κατηγορία που εντάχθηκαν ήταν η “Age Cyberbullying (Ηλικιακός διαδικτυακός εκφοβισμός)” (279 περιπτώσεις).

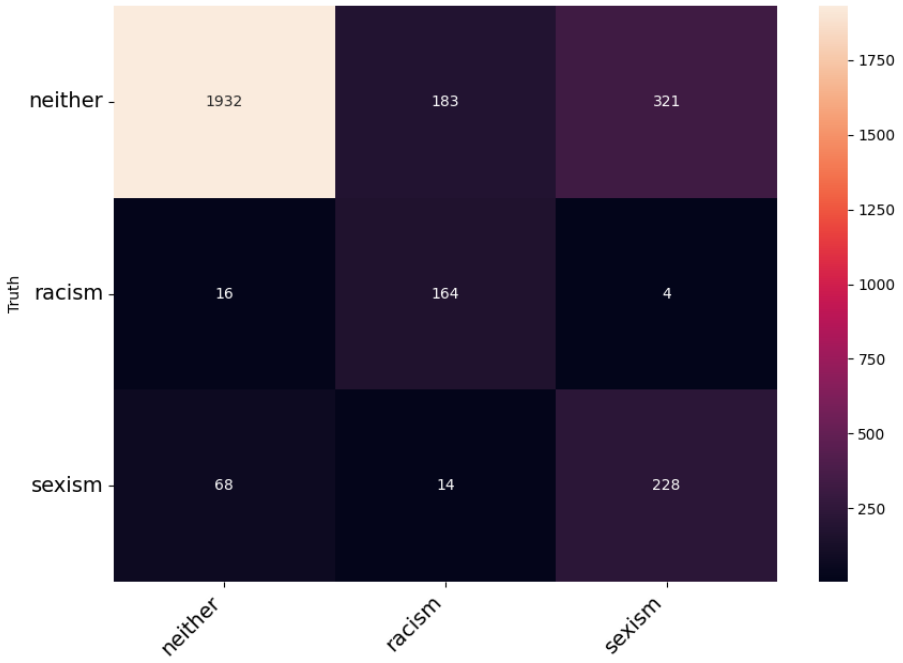
5.3.2 Suspicious Tweets Dataset

Classification Report for Naive Bayes:

	precision	recall	f1-score	support
neither	0.96	0.79	0.87	2436
racism	0.45	0.89	0.60	184
sexism	0.41	0.74	0.53	310
accuracy			0.79	2930
macro avg	0.61	0.81	0.67	2930
weighted avg	0.87	0.79	0.82	2930

Εικόνα 76 Classification report αλγορίθμου Naive Bayes για το Suspicious Tweets Dataset

Για το Suspicious Tweets Dataset, η συνολική ακρίβεια που σημειώθηκε για τον αλγόριθμο Naïve Bayes είναι 79% έναντι 86% του πρώτου. Σε αυτό το σημείο, αξίζει να ειπωθεί πως οι κλάσεις είναι unbalanced στο δεύτερο dataset, επομένως είναι λογικό να υπάρχει μία μικρή μείωση στην απόδοση των αλγορίθμων. Η πρώτη παρατήρηση που μπορεί να γίνει από το classification report είναι πως υπάρχουν μεγάλες διαφορές μεταξύ των μετρικών precision και recall. Στις 2 κλάσεις που υπάρχει διαδικτυακός εκφοβισμός, η μετρική recall είναι αρκετά υψηλότερη, υποδηλώνοντας πως, από όλες τις τιμές που ανήκουν πραγματικά στην κλάση, ένα μεγάλο μέρος προβλέπεται σωστά (χαμηλό FN). Από την άλλη πλευρά, η μετρική precision είναι χαμηλή ,δείχνοντας πως από όλες τις προβλεπόμενες τιμές της κλάσεις, ένα μικρό μέρος ανήκει πράγματι στην κλάση (υψηλό FP). Η δεύτερη παρατήρηση είναι πως η κλάση “neither (Κανένα από τα 2)”, σε αντίθεση με τις υπόλοιπες 2, έχει πολύ μεγαλύτερο precision από ότι recall. Συνολικά, η κλάση αυτή έχει μακράν το υψηλότερο f1-score, γεγονός που αναμένεται αφού είναι η κλάση που υπερτερεί σε κείμενα.



Εικόνα 77 Confusion matrix αλγορίθμου Naive Bayes για το Suspicious Tweets Dataset

Ο confusion matrix επιβεβαιώνει τις παρατηρήσεις. Εδώ, κάθετα φαίνονται οι αληθείς τιμές (Truth) και οριζόντια οι προβλεπόμενες. Πράγματι, η κλάση racism (ρατσισμός) έχει το υψηλότερο recall καθώς το FN είναι μόλις 20 κείμενα (16 προβλέφθηκαν neither (κανένα από τα 2) και 4 ως sexism (σεξισμός)). Από την άλλη μεριά για την κλάση neither (κανένα από τα δύο), το FP που σημειώνεται είναι μικρό αναλογικά με το πλήθος των κειμένων(16 ήταν racism (ρατσισμός) και 68 neither (κανένα από τα δύο)). Συνεπώς, αυτό εξηγεί το πολύ υψηλό precision της εν λόγω κατηγορίας.

5.4 Αποτελέσματα SVM

5.4.1 SOSNet Twitter Dataset

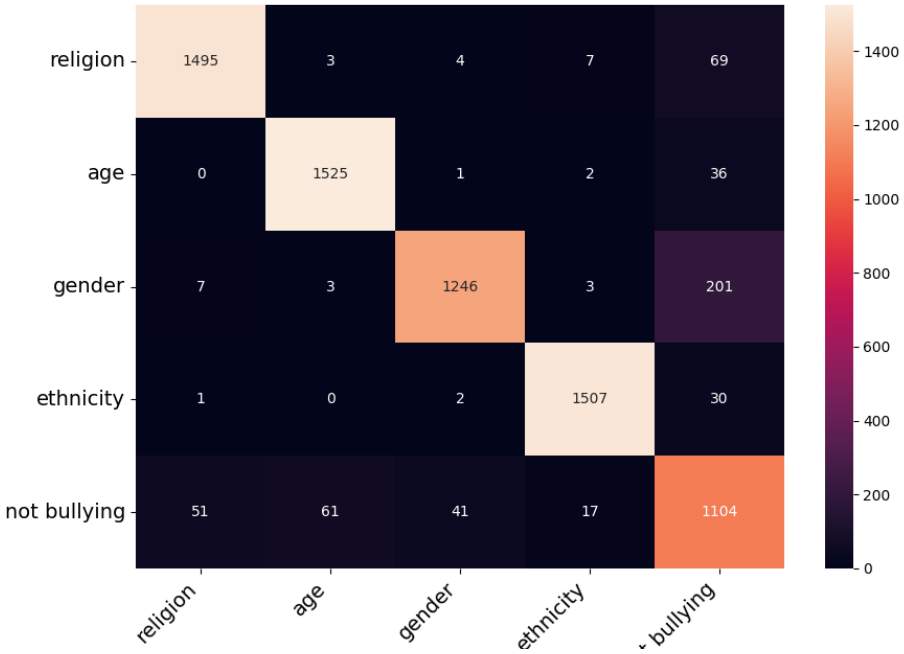
Classification Report for Support Vector Machine:

	precision	recall	f1-score	support
religion	0.96	0.95	0.95	1578
age	0.96	0.98	0.97	1564
gender	0.96	0.85	0.91	1460
ethnicity	0.98	0.98	0.98	1540
not bullying	0.77	0.87	0.81	1274
accuracy			0.93	7416
macro avg	0.93	0.92	0.92	7416
weighted avg	0.93	0.93	0.93	7416

Εικόνα 78 Classification report αλγορίθμου SVM για το SOSNet Twitter Dataset

Όπως φαίνεται από τα αποτελέσματα, ο αλγόριθμος SVM αποδίδει καλύτερα από τον Naive-Bayes, γεγονός που έρχεται σε συμφωνία με την υπάρχουσα βιβλιογραφία [9]. Η συνολική ακρίβεια του αλγορίθμου είναι 93%. Επιπλέον, σε όλες τις κατηγορίες τα f1-

scores που σημειώνονται είναι πολύ υψηλά, ξεπερνώντας το 90% με εξαίρεση την κατηγορία “Not cyberbullying”, που, και πάλι, το ποσοστό είναι υψηλό, φθάνοντας το 81%, σε αντίθεση με το 62% του προηγούμενου αλγορίθμου. Το recall της ίδιας κατηγορίας κυμαίνεται αυτήν τη φορά στο 87% σε αντίθεση με το 49% της προηγούμενης περίπτωσης. Οι κατηγορίες “Ethnicity cyberbullying (Διαδικτυακός εκφοβισμός εθνικότητας) ” και “Age cyberbullying (Ηλικιακός διαδικτυακός εκφοβισμός)” είναι αυτές που και πάλι αποδίδουν καλύτερα.



Εικόνα 79 Confusion matrix αλγορίθμου SVM για το SOSNet Twitter Dataset

Από το confusion matrix επιβεβαιώνεται και πάλι πως τα αποτελέσματα είναι βελτιωμένα σε σχέση με τον Naïve Bayes σε όλες τις κατηγορίες, με τη μεγαλύτερη διαφορά στην κατηγορία “Not cyberbullying (Όχι διαδικτυακός εκφοβισμός)” όπου, αυτήν τη φορά, TP=1104 σε σύγκριση με 611 στον Naïve Bayes. Τα κείμενα που προβλέφθηκαν σωστά ως μη-σχετιζόμενα με διαδικτυακό εκφοβισμό αυξήθηκαν κατά 493, με φυσικό επακόλουθο τη ραγδαία αύξηση των μετρικών precision και recall.

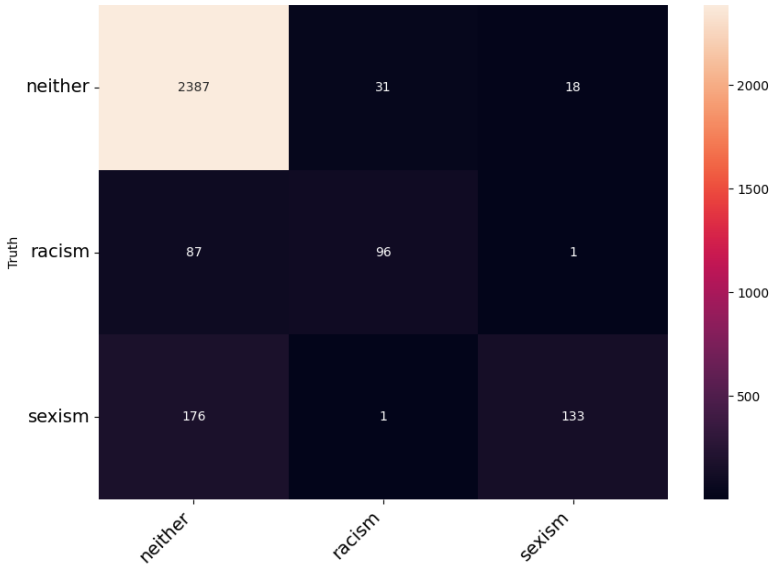
5.4.2 Suspicious Tweets Dataset

Classification Report for Support Vector Machine:

	precision	recall	f1-score	support
neither	0.90	0.98	0.94	2436
racism	0.75	0.52	0.62	184
sexism	0.88	0.43	0.58	310
accuracy			0.89	2930
macro avg	0.84	0.64	0.71	2930
weighted avg	0.89	0.89	0.88	2930

Εικόνα 80 Classification report αλγορίθμου SVM για το Suspicious Tweets Dataset

Η συνολική ακρίβεια του αλγορίθμου SVM για το Suspicious Tweets Dataset είναι 89% δηλαδή 10% πάνω σε σχέση με τον αλγόριθμο Naïve Bayes για το ίδιο dataset. Παρατηρώντας και τα f1-scores για κάθε κατηγορία, υπάρχει βελτίωση για όλες. Στην κλάση neither (κανένα από τα δύο) όλες οι μετρικές κυμαίνονται πάνω από 90% με τη μετρική recall να βρίσκεται στο 98% υποδηλώνοντας πολύ χαμηλό FN. Το χειρότερο score παρατηρείται στη μετρική recall της κλάσης sexism (σεξισμός), δηλαδή από όλα τα κείμενα που πραγματικά σχετίζονται με διαδικτυακό εκφοβισμό αυτής της κατηγορίας, λιγότερο από 1 στα 2 κείμενα προβλέπονται σωστά.



Εικόνα 81 Confusion matrix αλγορίθμου SVM για το Suspicious Tweets Dataset

Το recall της τάξης 98% για την κατηγορία neither (κανένα από τα δύο) επιβεβαιώνεται και από το confusion matrix, αφού μόλις 31 κείμενα προβλέπονται λανθασμένα ως racism (ρατσισμός) και μόλις 18 ως sexism(σεξισμός) από τα συνολικά 2456 κείμενα αυτής της κατηγορίας για το validation set. Επίσης, αξιοσημείωτο είναι πως μόλις ένα κείμενο της κατηγορίας racism (ρατσισμός) προβλέπεται ως sexism (σεξισμός) και μόλις ένα κείμενο της κατηγορίας sexism (σεξισμός) προβλέπεται ως racism (ρατσισμός), δηλαδή ο αλγόριθμος δεν συγχέει τα είδη διαδικτυακού εκφοβισμού.

5.5 Αποτελέσματα KNN

5.5.1 SOSNet Twitter Dataset

Σε αυτόν τον αλγόριθμο υπάρχει η παράμετρος k η οποία είναι δυνατό να επηρεάσει σημαντικά την απόδοση. Αρχικά λοιπόν, διερευνήθηκε ποια είναι η βέλτιστη τιμή του k για το συγκεκριμένο dataset . Πιο συγκεκριμένα, δοκιμάστηκαν όλες οι τιμές του k από το 1 μέχρι και το 40 κατά τη διάρκεια της εκπαίδευσης. Έπειτα, σχεδιάστηκε μία γραφική παράσταση του error rate σε συνάρτηση με το k. Η γραφική παράσταση είναι η ακόλουθη:

s

Παρατηρείται εύκολα ότι η χαμηλότερη τιμή του error rate είναι 0.2833 και σημειώνεται για k=3. Συνεπώς, με βάση αυτό το αποτέλεσμα επιλέχθηκε ως k η τιμή 3 για το SOSNet Twitter Dataset. Τα αποτελέσματα του αλγορίθμου για αυτήν την τιμή είναι τα εξής:

Classification Report for K-nearest neighbours:

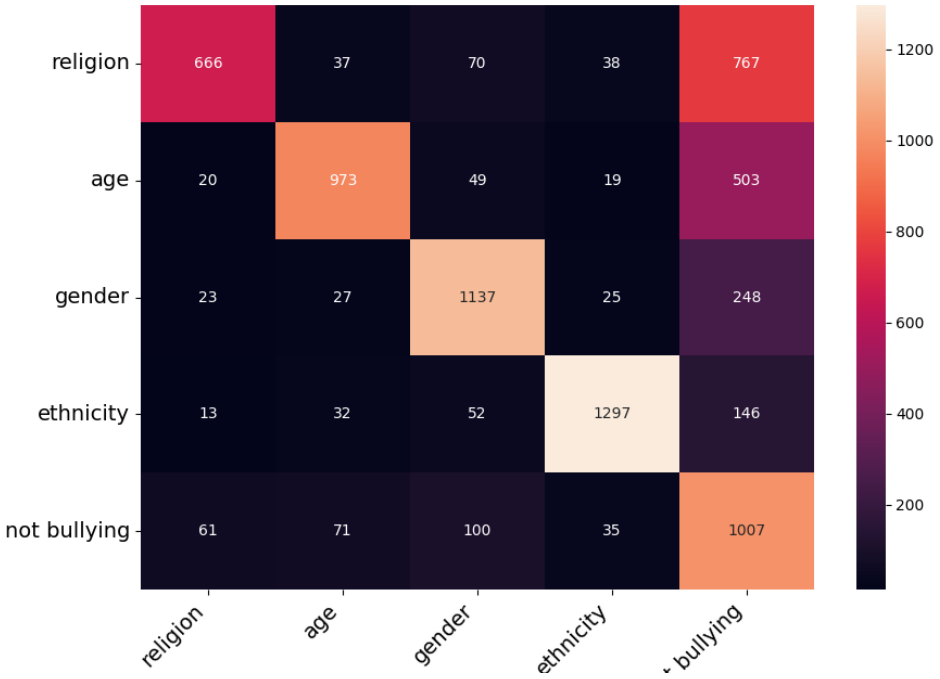
	precision	recall	f1-score	support
religion	0.84	0.42	0.57	1578
age	0.85	0.69	0.76	1564
gender	0.78	0.77	0.77	1460
ethnicity	0.92	0.83	0.87	1540
not bullying	0.39	0.77	0.52	1274
accuracy			0.69	7416
macro avg	0.76	0.70	0.70	7416
weighted avg	0.77	0.69	0.70	7416

Εικόνα 82 Έκθεση ταξινόμησης-Classification report αλγορίθμου KNN για το SOSNet Twitter Dataset

Ο KNN παρά την επιλογή της βέλτιστης τιμής ως k αποδίδει χειρότερα στο dataset σε σχέση με τους προηγούμενους 2 αλγόριθμους. Η συνολική ακρίβεια είναι 69%. Επιπλέον, σε καμία κατηγορία δεν παρατηρείται ιδιαίτερα υψηλό f1-score. Μάλιστα, τόσο στην κατηγορία “Religion cyberbullying (Θρησκευτικός διαδικτυακός εκφοβισμός)” όσο και στην κατηγορία “Not cyberbullying (Όχι διαδκτυακός εκφοβισμός)” τα f1-scores που σημειώθηκαν είναι 57% και 52% αντίστοιχα. Σε αντίθεση με τις προηγούμενες 2 περιπτώσεις, το χαμηλότερο recall που σημειώνεται δεν αφορά την κατηγορία “Not cyberbullying (Όχι διαδικτυακός εκφοβισμός)” αλλά την κατηγορία “Religion cyberbullying (Θρησκευτικός διαδικτυακός εκφοβισμός)”. Μάλιστα, το recall που καταγράφεται εκεί είναι μόλις 42%, το χαμηλότερο από όλα μέχρι στιγμής. Πρακτικά, με βάση τον ορισμό της μετρικής recall, αυτό σημαίνει πως από όλα τα κείμενα που ανήκουν στην κατηγορία “Religion cyberbullying (Θρησκευτικός διαδικτυακός εκφοβισμός)” ο KNN αλγόριθμος ανιχνεύει σωστά μόλις το 42%, δηλαδή λιγότερο από 1 στα 2 κείμενα.

Μία ακόμη παρατήρηση σε σχέση με το classification report, είναι πως οι διαφορές μεταξύ των μετρικών precision και recall είναι αρκετά υψηλές σε όλες τις κατηγορίες/κλάσεις, κάτι που δεν συνέβη σε κανέναν από τους προηγούμενους 2 αλγορίθμους. Αυτό υποδεικνύει πως από τις προβλεπόμενες τιμές για κάθε κλάση, μεγάλο ποσοστό κειμένων ανήκει όντως στην κλάση, αλλά από τις πραγματικές τιμές που ανήκουν στην κλάση, μεγάλο μέρος δεν προβλέπεται σωστά, δηλαδή είναι μεγάλος ο παράγοντας FN.

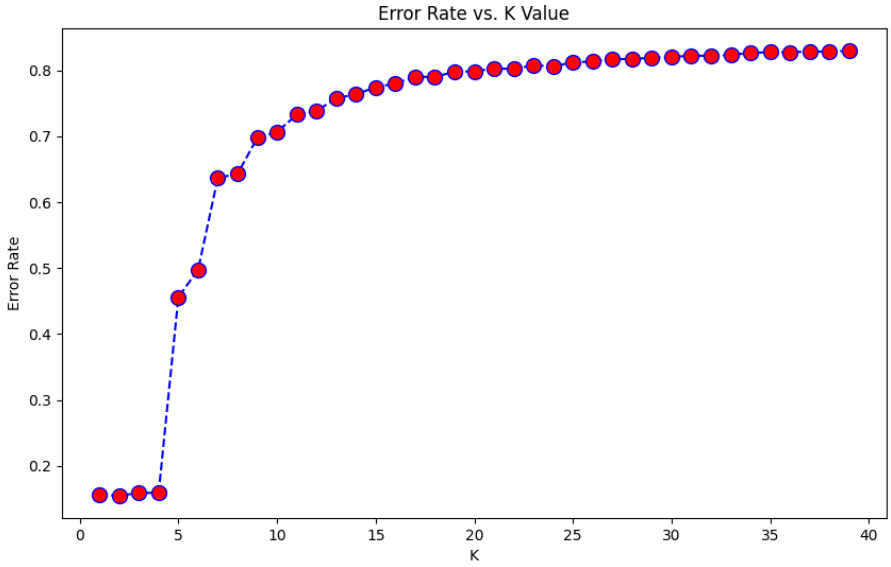
Αυτό επιβεβαιώνεται και από το confusion matrix που παρουσιάζεται παρακάτω. Οι τιμές που είναι FN για κάθε κλάση (δηλαδή όχι τα διαγώνια στοιχεία του πίνακα) είναι πολύ υψηλότερες σε σχέση με τους προηγούμενες 2 αλγορίθμους. Πιο συγκεκριμένα, όλες οι τιμές είναι διψήφιες ή τριψήφιες και δεν υπάρχει καμία μονοψήφια όπως πριν.



Εικόνα 83 Confusion matrix αλγορίθμου KNN για το SOSNet Twitter Dataset

5.5.2 Suspicious Tweets Dataset

Όπως πραγματοποιήθηκε διερεύνηση για την εύρεση του βέλτιστου k στο SOSNet Twitter Dataset, ακριβώς η ίδια διαδικασία ακολουθήθηκε και για το δεύτερο. Αυτήν τη φορά, το βέλτιστο k δεν είναι το 3 όπως στο SOSNet Twitter Dataset , αλλά k=1 με error rate= 0.1515358361774744.



Εικόνα 84 Γραφική παράσταση του παράγοντα k εν συναρτήσει του error rate (Suspicious Tweets Dataset)

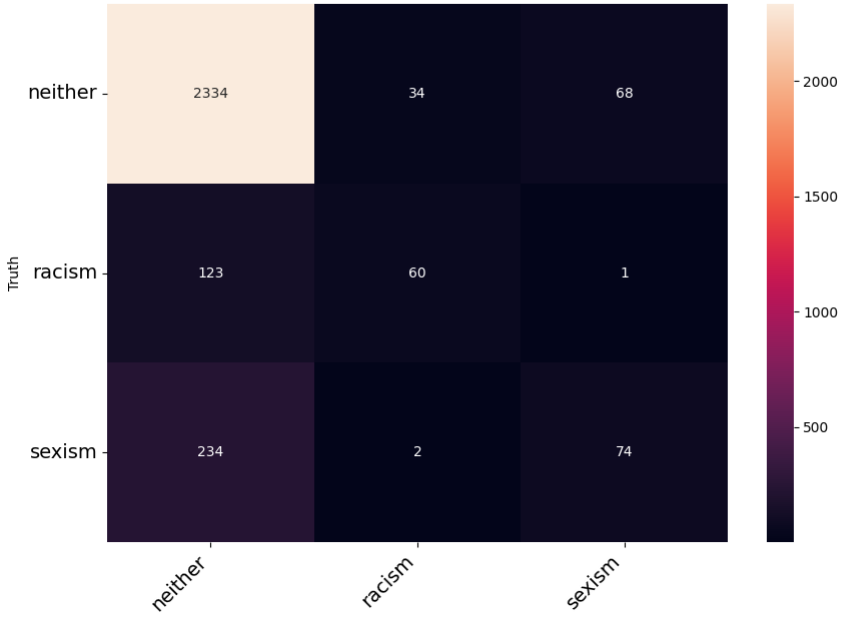
Επομένως, η εκπαίδευση έγινε με αυτήν την επιλογή του k, δηλαδή, ουσιαστικά, κάθε κείμενο κατηγοριοποιείται στην κλάση που ανήκει ο κοντυνότερος γείτονάς του. Τα αποτελέσματα του αλγορίθμου δίνονται παρακάτω.

Classification Report for K-nearest neighbours:

	precision	recall	f1-score	support
neither	0.87	0.96	0.91	2436
racism	0.62	0.33	0.43	184
sexism	0.52	0.24	0.33	310
accuracy			0.84	2930
macro avg	0.67	0.51	0.56	2930
weighted avg	0.82	0.84	0.82	2930

Εικόνα 85 Classification αλγορίθμου KNN για το Suspicious Tweets Dataset

Ο αλγόριθμος είχε συνολική ακρίβεια 84%, δηλαδή, 15% υψηλότερα σε σχέση με το προηγούμενο dataset. Ακόμη, παρατηρούνται υψηλές αποκλίσεις μεταξύ των μετρικών precision και recall και ιδίως στις κατηγορίες racism (ρατσισμός) και sexism (σεξισμός). Συγκεκριμένα, στην κατηγορία sexism (σεξισμός) το recall είναι 24%, παρά τη γενικά καλή επίδοση του αλγορίθμου. Ουσιαστικά, από τα κείμενα που ανήκουν στην κατηγορία sexism (σεξισμός) μόλις 1 στα 4 προβλέπεται σωστά.



Εικόνα 86 Confusion matrix αλγορίθμου KNN για το Suspicious Tweets Dataset

Ενδεικτικά, όπως φαίνεται και από το confusion matrix, η πλειοψηφία των κειμένων που ανήκουν στην κατηγορία racism cyberbullying (ρατσιστικός διαδικτυακός εκφοβισμός) προβλέπονται λανθασμένα ως neither (123), μόλις 1 προβλέπεται λανθασμένα ως sexism (σεξισμός) ενώ σωστά προβλέπονται τα 60 κείμενα.

5.6 Αποτελέσματα Decision Tree

5.6.1 SOSNet Twitter Dataset

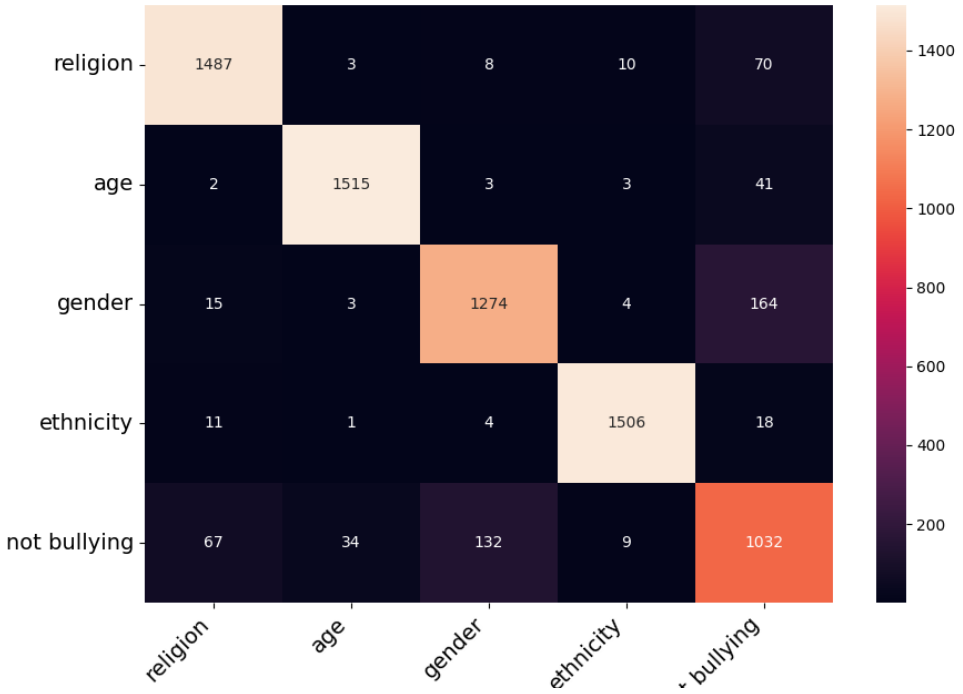
Classification Report for Decision Tree:

	precision	recall	f1-score	support
religion	0.94	0.94	0.94	1578
age	0.97	0.97	0.97	1564
gender	0.90	0.88	0.89	1460
ethnicity	0.99	0.98	0.98	1540
not bullying	0.78	0.82	0.80	1274
accuracy			0.92	7416
macro avg	0.92	0.92	0.92	7416
weighted avg	0.92	0.92	0.92	7416

Εικόνα 87 Classification report αλγορίθμου Decision Tree για το SOSNet Twitter Dataset

Στην περίπτωση του Decision Tree, τα αποτελέσματα παρουσιάζονται βελτιωμένα. Η συνολική ακρίβεια του μοντέλου είναι 92%, λίγο μικρότερη από αυτήν του SVM. Τα f1-scores είναι πολύ υψηλά, ιδιαίτερα στην κατηγορία “Ethnicity cyberbullying (Διαδικτυακός εκφοβισμός εθνικότητας)”, όπου και το recall είναι πάρα πολύ υψηλό. Το ίδιο ισχύει και για την κατηγορία “Age cyberbullying (Ηλικιακός διαδικτυακός εκφοβισμός)”. Αυτό που είναι αξιοσημείωτο είναι πως και στον αλγόριθμο SVM και πάλι αυτές οι 2 κατηγορίες

είχαν υψηλότερα recalls και f1-scores. Φαίνεται λοιπόν, πως η ανίχνευση tweets που σχετίζονται με αυτές τις 2 κατηγορίες διαδικτυακού εκφοβισμού είναι πιο εύκολο να υλοποιηθεί από τους υπό εξέταση αλγορίθμους. Όσον αφορά την κατηγορία “Not cyberbullying (Όχι διαδικτυακός εκφοβισμός)”, τα αποτελέσματα είναι και σε αυτόν τον αλγόριθμο χειρότερα σε σχέση με τις υπόλοιπες κατηγορίες, ωστόσο, κυμαίνονται κοντά στο 80% για όλες τις μετρικές σε αντίθεση με τον αλγόριθμο KNN, που η μετρική precision έφθασε στο πολύ χαμηλό 39%.



Εικόνα 88 Confusion matrix αλγορίθμου Decision Tree για το SOSNet Twitter Dataset

Από το confusion matrix επίσης παρατηρείται ότι στις περισσότερες περιπτώσεις που υπήρχε FN για την κατηγορία “Not cyberbullying (Όχι διαδικτυακός εκφοβισμός)” το κείμενο προβλέφθηκε ως “Gender cyberbullying (Διαδικτυακός εκφοβισμός με βάση το φύλο)” (164 περιπτώσεις). Αυτό δικαιολογεί και το μικρότερο Precision αυτής της κατηγορίας σε σχέση με τις υπόλοιπες 3 μορφές διαδικτυακού εκφοβισμού. Βάσει του ορισμού της μετρικής (ενότητα Precision), με την άυξηση του FP αυξάνεται ο παρονομαστής, άρα, ως επακόλουθο, μειώνεται και το κλάσμα.

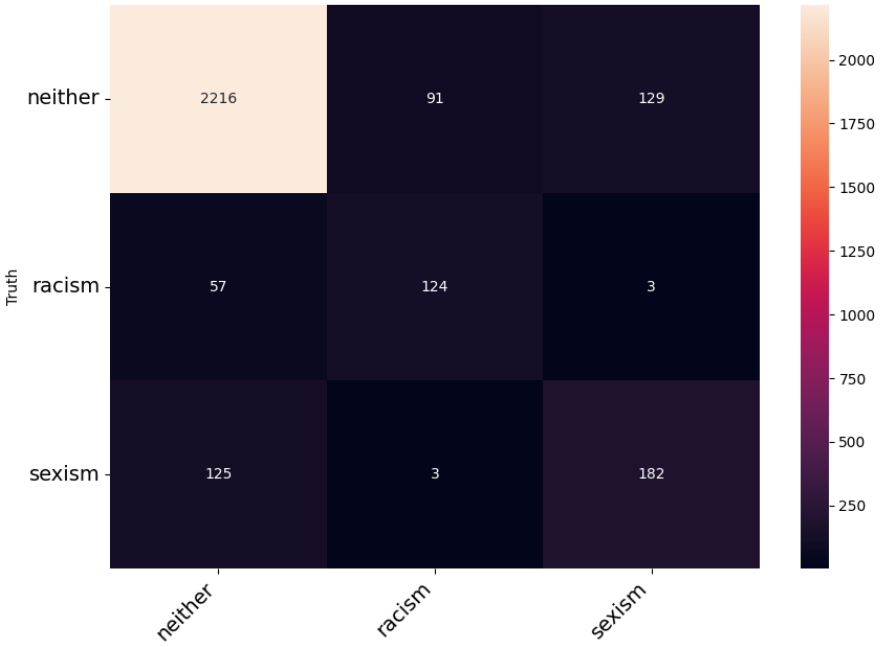
5.6.2 Suspicious Tweets Dataset

Classification Report for Decision Tree:

	precision	recall	f1-score	support
neither	0.92	0.91	0.92	2436
racism	0.57	0.67	0.62	184
sexism	0.58	0.59	0.58	310
accuracy			0.86	2930
macro avg	0.69	0.72	0.71	2930
weighted avg	0.87	0.86	0.86	2930

Εικόνα 89 Classification report αλγορίθμου Decision Tree για το Suspicious Tweets Dataset

Η εικόνα του classification report είναι βελτιωμένη, καθώς δεν παρατηρείται κάποιο εξαιρετικά χαμηλό score όπως συνέβη στον προηγούμενο αλγόριθμο. Επιπλέον, οι αποκλίσεις μεταξύ των μετρικών precision και recall είναι μικρές. Η συνολική ακρίβεια είναι 86%.



Εικόνα 90 Confusion matrix αλγορίθμου Decision Tree για το Suspicious Tweets Dataset

Αν συγκρίνουμε την εικόνα με το confusion matrix του προηγούμενου αλγορίθμου για την κατηγορία racism (ρατσισμός) υπάρχει σαφής βελτίωση. Τα 124 κείμενα προβλέπονται σωστά, ενώ 57 προβλέπονται ως neither (κανένα από τα δύο) και μόλις 3 ως sexism (σεξισμός). Αυτό δικαιολογεί και την αύξηση της μετρικής recall (πολύ χαμηλότερο FN).

5.7 Αποτελέσματα Random Forest

5.7.1 SOSNet Twitter Dataset

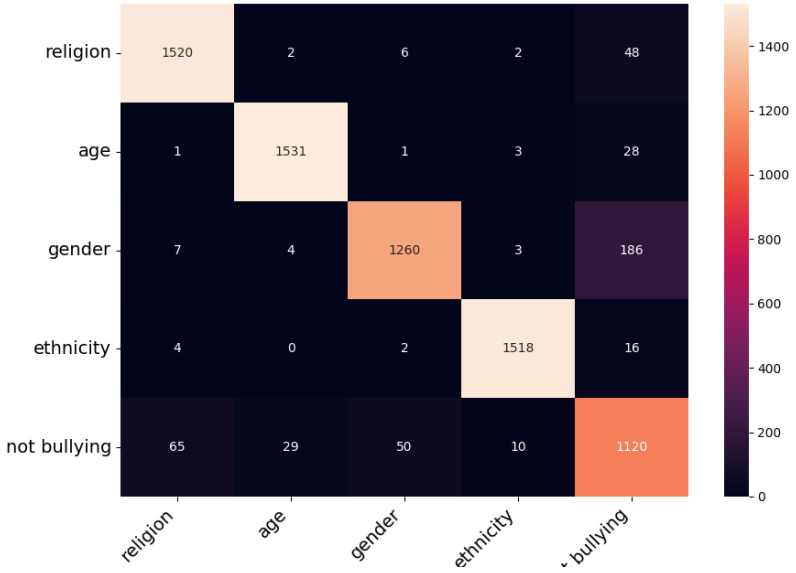
Classification Report for Random Forest:

	precision	recall	f1-score	support
religion	0.96	0.97	0.96	1578
age	0.98	0.98	0.98	1564
gender	0.95	0.86	0.91	1460
ethnicity	0.99	0.99	0.99	1540
not bullying	0.81	0.89	0.84	1274
accuracy			0.94	7416
macro avg	0.94	0.94	0.94	7416
weighted avg	0.94	0.94	0.94	7416

Εικόνα 91 Classification report αλγορίθμου Random Forest για το SOSNet Twitter Dataset

Τα αποτελέσματα του εν λόγω αλγορίθμου είναι πολύ εντυπωσιακά. Παρατηρείται ότι η συνολική ακρίβεια φθάνει το 94%. Επιπλέον, σε σχέση με τα decision trees υπάρχει βελτίωση, γεγονός που αναμενόταν και σύμφωνα με τη βιβλιογραφία [82].

Ιδιαίτερα στην κατηγορία “Ethnicity cyberbullying (Διαδικτυακός εκφοβισμός εθνικότητας)” και οι 3 μετρικές recall, precision και f1-score είναι 99%, κάτι που δεν είχε ξανασυναντηθεί σε κάποιον από τους προηγούμενους αλγορίθμους. Το χαμηλότερο recall αυτήν τη φορά σημειώνεται στην κατηγορία “Gender cyberbullying (Διαδικτυακός εκφοβισμός με βάση το φύλο)” με 86% και όχι στην κατηγορία “Not cyberbullying (Όχι διαδικτυακός εκφοβισμός)” όπως συνέβαινε μέχρι τώρα.



Εικόνα 92 Confusion matrix αλγορίθμου Random Forest για το SOSNet Twitter Dataset

Παρατηρώντας και το confusion matrix, επιβεβαιώνεται η εξαιρετική επίδοση του αλγορίθμου για την κατηγορία “Ethnicity cyberbullying (Διαδικτυακός εκφοβισμός εθνικότητας)”, αφού μόλις 18 κείμενα αυτής της κατηγορίας προβλέφθηκαν να ανήκουν σε κάποια άλλη.

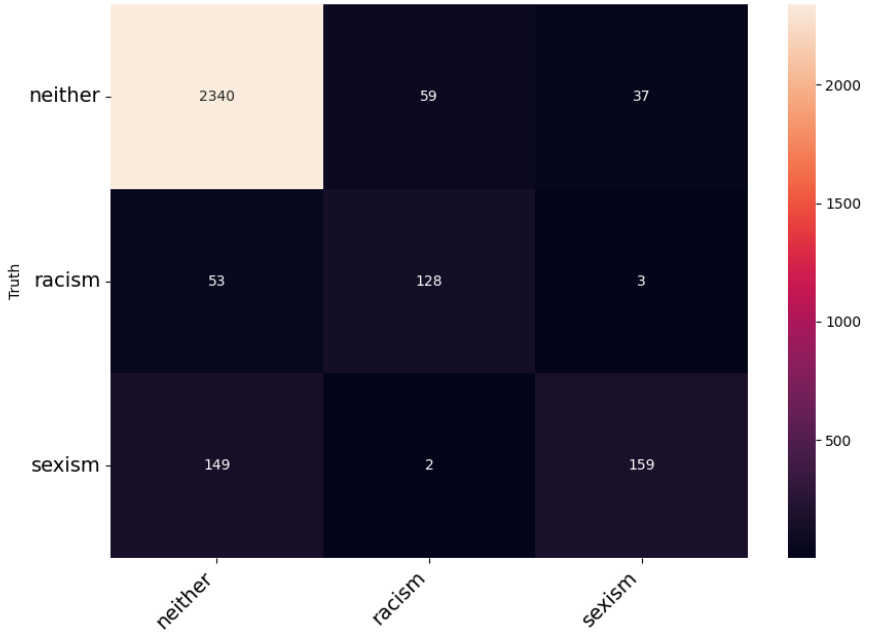
5.7.2 Suspicious Tweets Dataset

Classification Report for Random Forest:

	precision	recall	f1-score	support
neither	0.92	0.96	0.94	2436
racism	0.68	0.70	0.69	184
sexism	0.80	0.51	0.62	310
accuracy			0.90	2930
macro avg	0.80	0.72	0.75	2930
weighted avg	0.89	0.90	0.89	2930

Εικόνα 93 Classification report αλγορίθμου Random Forest για το Suspicious Tweets Dataset

Και σε αυτό το dataset, ο αλγόριθμος αποδίδει καλύτερα σε σύγκριση με το Decision Tree. Γενικώς, η απόδοση είναι πολύ καλή, με τη συνολική ακρίβεια να φθάνει το 90% και τα scores των κατηγοριών που σχετίζονται με εκφοβισμό και είναι unbalanced να ανεβαίνουν αρκετά. Ενδεικτικά, για την κατηγορία sexism (σεξισμός), το precision φθάνει το 80%, ποσοστό που δεν έχει παρατηρηθεί σε προηγούμενους αλγορίθμους για αυτήν την κλάση.



Εικόνα 94 Confusion matrix αλγορίθμου Random Forest για το Suspicious Tweets Dataset

Αυτό συμβαίνει διότι το FP αυτής της κατηγορίας είναι πλέον χαμηλό, αφού μόλις 37 κείμενα που πραγματικά είναι neither (κανένα από τα δύο) προβλέπονται ως sexism (σεξισμός) και μόλις 3 που σχετίζονται με racism (ρατσισμός) προβλέπονται λανθασμένα ως sexism (σεξισμός).

5.8 Συγκεντρωτικά Αποτελέσματα

Σε αυτήν την ενότητα, θα παρουσιαστούν πίνακες με τα συγκεντρωτικά αποτελέσματα όλων των αλγορίθμων στα 2 σύνολα δεδομένων, ώστε να είναι πιο εύκολο να πραγματοποιηθεί σύγκρισή τους.

5.8.1 SOSNet Twitter Dataset

Πίνακας 5 Συγκεντρωτικά αποτελέσματα αλγορίθμων στο SOSNet Twitter Dataset

	Naive Bayes	SVM	KNN (k=3)	Decision Tree	Random Forest
Συνολική Ακρίβεια	86%	93%	69%	92%	94%
f1-score Ethnicity cyberbullying (Διαδικτυακός εκφοβισμός εθνικότητας)	91%	98%	87%	98%	99%
f1-score Age cyberbullying (Ηλικιακός διαδικτυακός εκφοβισμός)	88%	97%	76%	97%	98%
f1-score Religion cyberbullying (Θρησκευτικός διαδικτυακός εκφοβισμός)	90%	95%	57%	94%	96%
f1-score Gender cyberbullying (Διαδικτυακός εκφοβισμός με βάση το φύλο)	87%	91%	77%	89%	91%
f1-score Not cyberbullying (Όχι διαδικτυακός εκφοβισμός)	62%	81%	52%	80%	84%

Η μεγαλύτερη συνολική ακρίβεια επετεύχθη με το Random Forest μοντέλο με 94% και η αμέσως επόμενη καλύτερη επίδοση είναι του SVM με 93%. Τρίτη καλύτερη επίδοση είναι αυτή του DT που ακολουθεί από κοντά με 92%. Οι 2 χειρότερες επιδόσεις είναι των Naive Bayes και KNN με 86% και 69% αντίστοιχα.

Όσον αφορά τα f1-scores στην κατηγορία Ethnicity cyberbullying (Διαδικτυακός εκφοβισμός εθνικότητας) και πάλι η κατάταξη είναι η ίδια, μόνο που αυτήν τη φορά στη δεύτερη θέση ισοβαθμούν οι SVM και DT με 98%. Επιπλέον, για τη συγκεκριμένη κατηγορία παρατηρείται πως όλα τα f1-scores κυμαίνονται αρκετά υψηλότερα από τη συνολική ακρίβεια, ακόμα και ο KNN που έχει γενικότερα μέτρια επίδοση. Συνεπώς, είναι ένα είδος cyberbullying που ανιχνεύεται πιο εύκολα σε σχέση με τα υπόλοιπα. Η ίδια εικόνα ισχύει και για το Age cyberbullying (Ηλικιακός διαδικτυακός εκφοβισμός), όπου και πάλι συναντώνται οι ίδιες ακριβώς ισοβαθμίες αλγορίθμων και τα f1-scores κυμαίνονται υψηλότερα από τη συνολική ακρίβεια.

Στην κατηγορία Religion cyberbullying (Θρησκευτικός διαδικτυακός εκφοβισμός), όλοι οι αλγόριθμοι αποδίδουν πολύ καλά με ποσοστά άνω του 90%, με μοναδική εξαίρεση τον KNN με ποσοστό 57%.

Στο Gender cyberbullying (Διαδικτυακός εκφοβισμός με βάση το φύλο) υπάρχει μία μικρή μείωση των f1-scores όλων των αλγορίθμων με όλους να αποδίδουν λίγο κάτω

από τη συνολική τους ακρίβεια με εξαίρεση τον Naive Bayes που βρίσκεται 1% επάνω και τον KNN που αποδίδει κατά 8% καλύτερα από τη συνολική του ακρίβεια αλλά εξακολουθεί να έχει το χαμηλότερο score.

Τέλος, στην κατηγορία Not cyberbullying (Όχι διαδικτυακός εκφοβισμός), το καλύτερο score που σημειώνεται είναι αυτό του Random Forest με 84%. Όλοι οι αλγόριθμοι σε αυτήν την κατηγορία αποδίδουν αρκετά χειρότερα από τη μέση επίδοσή τους με τελευταίους και πάλι τον Naive Bayes με 62% και τον KNN με μόλις 52%.

5.8.2 Suspicious Tweets Dataset

Πίνακας 6 Συγκεντρωτικά αποτελέσματα αλγορίθμων στο Suspicious Tweets Dataset

	Naive Bayes	SVM	KNN (k=1)	Decision Tree	Random Forest
Συνολική Ακρίβεια	79%	89%	84%	86%	90%
f1-score Neither (Κανένα από τα δύο)	87%	94%	91%	92%	94%
f1-score Racism (Ρατσισμός)	60%	62%	43%	62%	69%
f1-score Sexism (Σεξισμός)	53%	58%	33%	58%	62%

Τα scores στο Suspicious Tweets Dataset είναι μειωμένα λόγω των μη ισορροπημένων κλάσεων που χρησιμοποιούντα αυτήν τη φορά, θέτοντας ένα πιο ρεαλιστικό περιβάλλον πραγματικών συνθηκών [122]. Η μεγαλύτερη ακρίβεια σημειώνεται και πάλι από τον αλγόριθμο Random Forest με 90% και ακολουθεί ο SVM με 89%. Στην τρίτη θέση βρίσκεται ο DT, ενώ ο KNN δεν είναι πλέον ο τελευταίος σε απόδοση αλγόριθμος, αλλά ο Naïve Bayes. Η κατάταξη είναι ακριβώς η ίδια και για τα f1-scores στην κατηγορία Neither (Κανένα από τα δύο), με τη διαφορά ότι στην 1η θέση υπάρχει ισοβαθμία μεταξύ των αλγορίθμων SVM και Random Forest. Στις υπόλοιπες 2 κατηγορίες βρίσκεται στην πρώτη θέση ο αλγόριθμος Random Forest και ακολουθεί ισοβαθμία των SVM και DT. Τελευταίος αυτήν τη φορά είναι ο KNN που ειδικά στην κατηγορία Sexism (Σεξισμός) σημειώνει ιδιαίτερα χαμηλό score.

6. ΑΠΟΔΟΣΗ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΕΤΑΦΡΑΣΗ ΔΕΔΟΜΕΝΩΝ

Η ανάλυση των δεδομένων SOSNet Twitter DATASET και Suspicious Tweets DATASET δίνει την δυνατότητα μελέτης των συνηθέστερων μορφών διαδικτυακού εκφοβισμού (cyberbullying), τουλάχιστον για τις ΗΠΑ.

Η μετάφραση και απόδοση των συνηθέστερων εκφράσεων που σηματοδοτούν διαδικτυακό εκφοβισμό στα Ελληνικά ή σε μια άλλη φυσική γλώσσα δίνει την δυνατότητα της μελέτης και αξιολόγησης των δεδομένων αυτών σε ένα ευρύτερο, διεθνές κοινό. Ωστόσο, στην περίπτωση αυτή πρέπει να ληφθούν υπόψη, αφενός, τα μεταφραστικά λάθη και άλλες μεταφραστικές δυσκολίες που συχνά προκύπτουν από ευρέως διαθέσιμα συστήματα μηχανικής μετάφρασης όπως το GoogleTranslate και, αφετέρου, κοινωνιο-γλωσσολογικά χαρακτηριστικά και άλλοι κοινωνιο-πολιτισμικοί παράγοντες που είναι αναγκαίο να ληφθούν υπόψη για την ορθή απόδοση των δεδομένων αυτών σε άλλη γλώσσα. Σε αυτή την περίπτωση, η συμβολή των επαγγελματιών μεταφραστών είναι καθοριστικής σημασίας.

Εδώ θα παρουσιαστούν ενδεικτικά παραδείγματα μηχανικής μετάφρασης στοιχείων από τα δεδομένα SOSNet Twitter DATASET και Suspicious Tweets DATASET στα Ελληνικά.

Αρχικά, για τους συνηθέστερους όρους και εκφράσεις τόσο από το SOSNet Twitter DATASET όσο και από το Suspicious Tweets DATASET για κάθε κατηγορία διαδικτυακού εκφοβισμού ξεχωριστά, γίνεται μετάφρασή τους με τη βοήθεια του Google Translate.

Βάσει ενδεικτικών δεδομένων που επεξεργάστηκαν, παρατηρείται ότι η μηχανική μετάφραση είναι προβληματική σε όρους αρκετά συνηθισμένους στα μέσα κοινωνικής δικτύωσης – Social Media, ιδιαίτερα όταν εμφανίζονται εκτός του συνηθισμένου τους περικειμένου τους και συμφραζομένων (context), δεδομένου ότι σε Social Media όπως το Twitter, οι φράσεις πολύ συχνά είναι ελλειπτικές, με χαρακτηριστικά αποσπάσματος. Χαρακτηριστικά παραδείγματα αποτελούν οι λέξεις «Radical» (ριζοσπαστικός) και «White» (λευκός) που μεταφράστηκαν ως «Ριζικό» και «Άσπρο» αντίστοιχα.
Επιπλέον, παρατηρείται ότι η μηχανική μετάφραση είναι προβληματική σε περιπτώσεις σύνθετων λέξεων, ιδιαίτερα σύνθετων λέξεων που δεν αποτελούν όρους όπως, για

παράδειγμα, τις λέξεις «Girl high» (λύκειο θηλέων) και «Bully school» (σχολείο νταήδων), που μεταφράστηκαν ως «Κορίτσι ψηλά» και «Σχολείο νταής» αντίστοιχα, αλλά και συνηθισμένων σύνθετων λέξεων που αποτελούν καθημερινό λεξιλόγιο όπως «School girl» (μαθήτρια), που μεταφράστηκε ως «Σχολικό κορίτσι».

Σε πιο τεχνικό επίπεδο, παρατηρείται, επιπλέον, ότι μεταφραστικές δυσκολίες προκύπτουν από συντομογραφίες. Για παράδειγμα, ο όρος Rt δεν μεταφράστηκε διότι πρόκειται για αρχικά διαδικτυακής αργκό και αναπαριστά τον όρο Retweet. Ακόμη, δεν υπήρξε επιτυχής μετάφραση του όρου Moham. Πρόκειται για μία λέξη που χρησιμοποιούταν παλαιότερα στον Δυτικό κόσμο για να χαρακτηρίσει τους Μουσουλμάνους, ενώ δεν υιοθετήθηκε ποτέ από τους ίδιους τους Μουσουλμάνους. Αποτελεί συντομογραφία του όρου Mohammedan που μεταφράζεται επιτυχώς στα Ελληνικά ως Μωαμεθανός. Οι όροι Kat και Mkr δεν μεταφράστηκαν καθώς αντιπροσωπεύουν hashtags που χρησιμοποιήθηκαν για τη δημιουργία του dataset. Για παράδειγμα, το #MKR συνδέεται με το ριάλιτι My Kitchen Rules που χρησιμοποιήθηκε όπως έχει προαναφερθεί για τη συγκέντρωση tweets συνδεδεμένα με σεξισμό. Τέλος, επιτυχής μετάφραση στα Ελληνικά δεν υπήρχε ούτε για τον όρο ISIS.

SOSNet Twitter DATASET

6.1 Κατηγορία Ηλικία (Age) - Σχολείο

Πίνακας 7 Μηχανική μετάφραση συχνότερων λέξεων της κατηγορίας Ηλικιακός Διαδικτυακός εκφοβισμός του SOSNet Twitter Dataset

Όρος	Μηχανική Μετάφραση
School	Σχολείο
High	Υψηλός
Girl	Κορίτσι
Bully	Νταής
Bullied	Εκφοβίζεται
One	Ένας
People	Άνθρωποι
Now	Τώρα
Got	Πήρε
Kid	Παιδί

Πίνακας 8 Μηχανική μετάφραση των συχνότερων εκφράσεων 2 λέξεων της κατηγορίας Ηλικιακός διαδικτυακός εκφοβισμός του SOSNet Twitter Dataset

Έκφραση	Μηχανική Μετάφραση
High school	Λύκειο
Girl bullied	Κορίτσι που εκφοβίζεται
Bullied high	Εκφοβίζονται ψηλά
School bully	Σχολικός εκφοβιστής
Middle school	Γυμνάσιο
Girl high	Κορίτσι ψηλά
Got bullied	Δέχτηκε εκφοβισμό
Bully school	Σχολείο νταής
School bullied	Σχολικός εκφοβισμός
School girl	Σχολικό κορίτσι

6.2 Κατηγορία Εθνικότητα-Φυλή (Ethnicity)

Πίνακας 9 Μηχανική μετάφραση συχνότερων λέξεων της κατηγορίας Διαδικτυακός εκφοβισμός Εθνικότητας του SOSNet Twitter Dataset

Όρος	Μηχανική Μετάφραση
Nigger	Αράπης
Fuck	Γαμώ
Dumb	Χαζός
Black	Μαύρος
White	Άσπρο
People	Άνθρωποι
Obama	Ομπάμα
One	Ένας
Bitch	Σκύλα
Called	Που ονομάζεται

Πίνακας 10 Μηχανική μετάφραση των συχνότερων εκφράσεων 2 λέξεων της κατηγορίας Διαδικτυακός εκφοβισμός εθνικότητας του SOSNet Twitter Dataset

Έκφραση	Μηχανική Μετάφραση
Dumb nigger	Χαζός μαύρος
Dumb fuck	Χαζό σκατά
Fuck Obama	Γάμα τον Ομπάμα
Obama dumb	Ο Ομπάμα χαζός
Tayyoung fuck	-[Δεν μεταφράστηκε][1]
Fuck dumb	γαμώ χαζό
Black people	Μαύροι άνθρωποι
Fuck nigger	Γαμώ τον μαύρο
Nigger fuck	μαύρος γαμώ
Nigger dumb	μαύρος χαζός

6.3 Κατηγορία Θρησκεία (Religion)

Πίνακας 11 Μηχανική μετάφραση συχνότερων λέξεων της κατηγορίας Θρησκευτικός Διαδικτυακός εκφοβισμός του SOSNet Twitter Dataset

Όρος	Μηχανική Μετάφραση
Muslim	μουσουλμάνος
Idiot	Βλάκας
Christian	Χριστιανός
Terrorist	Τρομοκράτης
Right	σωστά
Islamic	Ισλαμική
Woman	Γυναίκα
Islam	Ισλάμ
Terrorism	Τρομοκρατία
Radical	Ριζικό

- [1] Η έκφραση «Tayyoung» είναι όνομα από το πεδίο της hip hop μουσικής

Πίνακας 12 Μηχανική μετάφραση των συχνότερων εκφράσεων 2 λέξεων της κατηγορίας Θρησκευτικός διαδικτυακός εκφοβισμός του SOSNet Twitter Dataset

Έκφραση	Μηχανική Μετάφραση
Islamic terrorism	Ισλαμική τρομοκρατία
Christian woman	Χριστιανή γυναίκα
Radical Islamic	Ριζοσπαστικό Ισλαμικό
Muslim idiot	Μουσουλμάνος ηλίθιος
Non muslim	Μη-μουσουλμάνος
Muslim country	μουσουλμανική χώρα
Radical Christian	Ριζοσπάστης Χριστιανός
Support radical	Στήριξη ριζική
Muslim terrorist	μουσουλμάνος τρομοκράτης
Good Christian	Καλός Χριστιανός

6.4 Κατηγορία Φύλο (Gender)

Πίνακας 13 Μηχανική μετάφραση συχνότερων λέξεων της κατηγορίας Διαδικτυακός εκφοβισμός με βάση το φύλο του SOSNet Twitter Dataset

Όρος	Μηχανική Μετάφραση
joke	Αστείο
Rape	Βιασμός
Gay	Ομοφυλόφιλος/γκέι
Woman	Γυναίκα
Bitch	Σκύλα
Female	Θηλυκός
Call	Κλήση
People	Άνθρωποι
Sexist	Σεξιστικό
Make	Φτιάχνω, κάνω

Πίνακας 14 Μηχανική μετάφραση των συχνότερων εκφράσεων 2 λέξεων της κατηγορίας Διαδικτυακός εκφοβισμός με βάση το φύλο του SOSNet Twitter Dataset

Έκφραση	Μηχανική Μετάφραση
Rape joke	Ανέκδοτο βιασμού
Gay joke	Γκέι αστείο
Joke gay	Ανέκδοτο γκέι
Gay rape	Ομοφυλοφιλικός βιασμός
Joke rape	Ανέκδοτο βιασμό
Rape gay	Βίασε γκέι
Call female	Καλέστε θηλυκό
Gay people	Ομοφυλόφιλοι άνθρωποι
Prison rape	Βιασμός στη φυλακή
Make rape	Κάντε βιασμό

Suspicious Tweets DATASET

6.5 Κατηγορία Ρατσισμός (Racism)

Πίνακας 15 Μηχανική μετάφραση των συχνότερων λέξεων της κατηγορίας Ρατσιστικός διαδικτυακός εκφοβισμός του Suspicious Tweets Dataset

Όρος	Μηχανική Μετάφραση

Islam	Ισλάμ
Muslim	μουσουλμάνος
Murder	Δολοφονία
Rt	-[Δεν μεταφράστηκε]
Moham	-[Δεν μεταφράστηκε]
Like	Αρέσει
Religion	Θρησκεία
Isis	-[Δεν μεταφράστηκε]
People	Άνθρωποι
jew	Εβραίος

6.6 Κατηγορία Σεξισμός (Sexism)

Πίνακας 16 Μηχανική μετάφραση των συχνότερων λέξεων της κατηγορίας Σεξιστικός διαδικτυακός εκφοβισμός του Suspicious Tweets Dataset

Όρος	Μηχανική Μετάφραση
Rt	-[Δεν μεταφράστηκε]
Sexist	Σεξιστικό
Mkr	-[Δεν μεταφράστηκε]
Women	γυναίκες
Girl	Κορίτσι
Kat	-[Δεν μεταφράστηκε]
Female	Θηλυκός
Call	Κλήση
Get	Παίρνω
Like	Αρέσει

Όσον αφορά την αναζήτηση και επεξεργασία δεδομένων για τα Ελληνικά, εδώ είναι απαραίτητο να επισημανθεί ότι υπάρχει η ανάγκη ορισμού συγκεκριμένων κριτηρίων βάσει των οποίων τα σύνολα δεδομένων από τα μέσα κοινωνικής δικτύωσης – Social Media ταξινομούνται ως διαδικτυακός εκφοβισμός (cyberbullying) και διαχωρίζονται από την όποια κοινωνική-πολιτική ή/και δημοσιογραφική τοποθέτηση. Αυτός ο διαχωρισμός είναι ιδιαίτερα προβληματικός, καθότι δεν είναι πάντα σαφές το όριο μεταξύ μιας - έστω και ακραίας- τοποθέτησης για κάποιο θέμα και του διαδικτυακού εκφοβισμού. Αν ληφθεί υπόψη ότι, τα τελευταία χρόνια και σε αρκετές περιπτώσεις, ο δημόσιος λόγος στην Ελλάδα είναι αρκετά οξύς και επιθετικός, ο διαχωρισμός μεταξύ ακραίας- τοποθέτησης για κάποιο θέμα και του διαδικτυακού εκφοβισμού δεν είναι πάντα ευδιάκριτος και απαιτείται η συμβολή ειδικών αναλυτών.

Κατά, συνέπεια, εδώ θα παραθέσουμε παραδείγματα περιπτώσεων οι οποίες είναι ανεξάρτητες από θέματα επικαιρότητας, μολονότι, από τα σύνολα δεδομένων που εξετάστηκαν, προκύπτουν οι εξής πέντε (5) κατηγορίες, οι οποίες, σε ένα μεγάλο βαθμό, αντιστοιχούν στις κατηγορίες διαδικτυακού εκφοβισμού (cyberbullying) που παρουσιάστηκαν για τα Αγγλικά σύνολα δεδομένων:

(1) Εκφοβισμός με βάση τις πολιτικές πεποιθήσεις
(2) Εκφοβισμός με βάση την εθνικότητα
(3) Εκφοβισμός με βάση την άποψη για τα δικαιώματά των φύλων
(4) Εκφοβισμός με βάση τη σωματική διάπλαση/εμφάνιση
(5) Εκφοβισμός με βάση την ηλικία

Από τις κατηγορίες «Εκφοβισμός με βάση τη σωματική διάπλαση/εμφάνιση» και «Εκφοβισμός με βάση την ηλικία», ενδεικτικά παραδείγματα ελληνικών σχολίων που ανακτήθηκαν από το Twitter και προβαίνουν σε διαδικτυακό εκφοβισμό είναι οι εκφράζεις «γριά»/ «κ@γρια», «βρ@γερε», «χοντρή», «άσχημη», «σίχαμα». Από τις κατηγορίες «Εκφοβισμός με βάση την εθνικότητα», ενδεικτικά παραδείγματα ελληνικών σχολίων που ανακτήθηκαν από το Twitter και προβαίνουν σε διαδικτυακό εκφοβισμό είναι οι εκφράζεις «αλλοδαπός» και ονόματα συγκεκριμένων εθνικοτήτων. Όσον αφορά τις κατηγορίες «Εκφοβισμός με βάση τις πολιτικές πεποιθήσεις» και «Εκφοβισμός με βάση την άποψη για τα δικαιώματά των φύλων», εδώ διαπιστώθηκε ότι δεν ήταν σαφής ο διαχωρισμός μεταξύ διαδικτυακού εκφοβισμού (cyberbullying) και της όποιας κοινωνικής-πολιτικής τοποθέτησης. Ενδεικτικά αναφέρουμε τις εκφράσεις «φασίστας», «αναρχοάπλυτος», «φεμιναζί», «βόδι», «σκουπίδι» και «σκουλίκι».

Επειδή και στις έξι κατηγορίες είτε γίνεται αναφορά σε δημόσια πρόσωπα είτε πρόκειται για ζητήματα της πρόσφατης επικαιρότητας σε δημόσιο διάλογο και αντιπαράθεση, τα tweets που εξετάστηκαν και από τα οποία προκύπτουν τα ενδεικτικά παραδείγματα δεν παρατίθενται εδώ.

Επιπλέον, επιχειρήθηκε να βρεθούν σύνολα δεδομένων αντίστοιχου μεγέθους με τα δεδομένα που επεξεργάστηκαν και αναλύθηκαν, τα οποία να προβαίνουν σε αντίστοιχες κατηγοριοποιήσεις στην Ελληνική γλώσσα, έτσι ώστε να εντοπιστούν κάποιοι ενδεικτικοί όροι διαδικτυακού εκφοβισμού στα Ελληνικά και παράλληλα να προκύψουν νέες κατηγορίες διαδικτυακού εκφοβισμού ως κίνητρο για κάποια μελλοντική δουλειά στον τομέα αυτόν. Ένα πολύ ενδιαφέρον σύνολο δεδομένων που βρέθηκε είναι το Offensive Language Identification in Greek(OLID) (Pitenis et al., LREC 2020). Το συγκεκριμένο σύνολο δεδομένων - dataset διαχωρίζει σχόλια από το ελληνικό twitter σε Offensive (Προσβλητικά) και Not offensive (μη-προσβλητικά). Δεν είναι ακριβώς αυτό που ήταν επιθυμητό ως επέκταση, ωστόσο η μελέτη αυτού του συνόλου δεδομένων αποτέλεσε έμπνευση και για την προσαρμογή της αναζήτησής μας στο Twitter με σκοπό εύρεση σχολίων διαδικτυακού εκφοβισμού. Όπως και στο SOSNet Twitter Dataset και το Suspicious Tweets Dataset, το OLID περιέχει tweets από hashtags τηλεοπτικών εκπομπών της Ελληνικής τηλεόρασης όπως #GNTM, #SurvivorGR κ.λ.π. Επομένως, η μελέτη των κειμένων του εν λόγω dataset κατηύθυνε τη μετέπειτα αναζήτησή για εύρεση σχολίων διαδικτυακού εκφοβισμού.

7. ΣΥΜΠΕΡΑΣΜΑΤΑ

Σε αυτήν την ενότητα θα πραγματοποιηθεί απαρίθμηση των συμπερασμάτων που εξήχθησαν από την τρέχουσα έρευνα τόσο σε επίπεδο απόδοσης των αλγορίθμων όσο και γενικά συμπεράσματα βασισμένα στα δεδομένα που μελετήθηκαν. Τα συμπεράσματα είναι τα ακόλουθα:

1. Όσα κείμενα συνδέονταν με διαδικτυακό εκφοβισμό ήταν αρκετά μεγαλύτερα σε έκταση συγκριτικά με αυτά που κατατάσσονταν ως μη-εκφοβιστικά. Συγκεκριμένα, στο SOSNet Twitter Dataset τα tweets της κατηγορίας Όχι διαδικτυακός εκφοβισμός/ Not cyberbullying είχαν μέσο μήκος 7.78 λέξεις με την αμέσως επόμενη κατηγορία να είναι ο Διαδικτυακός εκφοβισμός με βάση το φύλο/Gender cyberbullying με 13.19 λέξεις δηλαδή ποσοστό αύξησης 69%. Όλες οι υπόλοιπες κατηγορίες διαδικτυακού εκφοβισμού είχαν ακόμα μεγαλύτερο μέσο αριθμό λέξεων. Αυτό το συμπέρασμα είναι δυνατό να φανεί χρήσιμο για την εύκολη αναγνώριση μη-εκφοβιστικών κειμένων.

2. Όλοι οι αλγόριθμοι ανίχνευσαν με ευκολία τον Διαδικτυακό εκφοβισμό εθνικότητας/Ethnicity cyberbullying. Αυτό το συμπέρασμα μπορεί να εξηγηθεί βάσει και της μηχανικής μετάφρασης που πραγματοποιήθηκε στα ελληνικά για τις πιο συνηθισμένες εκφράσεις των κειμένων της εν λόγω κατηγορίας. Όπως διαπιστώθηκε κατά τη μετάφραση, στην κατηγορία αυτήν παρατηρούνται με διαφορά οι πιο υβριστικές εκφράσεις σε σύγκριση με όλες τις υπόλοιπες κατηγορίες. Συνεπώς, ήταν πιο εύκολο για τους αλγορίθμους να εντοπίσουν τις εκφράσεις μίσους και συνεπώς τον εκφοβισμό με βάση την εθνικότητα.

3. Όλοι οι αλγόριθμοι είχαν μεγαλύτερη δυσκολία στον εντοπισμό του Διαδικτυακού εκφοβισμού με βάση το φύλο/Gender cyberbullying στο SOSNet Twitter Dataset και του Σεξισμού/Sexism στο Suspicious Tweets Dataset που αποτελούν δύο κατηγορίες με παρόμοιο περιεχόμενο. Οι συνηθέστερες εκφράσεις αυτών των κατηγοριών δεν είναι υβριστικές ούτε φανερώνουν μίσος. Επομένως, είναι χρήσιμο να επιχειρηθεί μελλοντικά βελτίωση σε αυτές τις κατηγορίες.

4. Μέσω των δεδομένων που χρησιμοποιήθηκαν στην τρέχουσα έρευνα επιβεβαιώνεται πως ,όπως αναφέρεται και στην πλειοψηφία των βιβλιογραφικών αναφορών, ο διαδικτυακός εκφοβισμός κορυφώνεται κατά την εφηβική ηλικία. Το συμπέρασμα αυτό επιβεβαιώθηκε και στην παρούσα διπλωματική, καθώς οι λέξεις που κυριάρχησαν στα κείμενα του SOSNet Twitter Dataset στην κατηγορία Ηλικιακός διαδικτυακός εκφοβισμός/Age cyberbullying αφορούσαν το σχολικό πλαίσιο (π.χ. ‘High school’ , ‘school’ κ.λ.π)

5. Όσον αφορά το φύλο και τον σεξουαλικό προσανατολισμό του ατόμου, διαπιστώθηκε πως οι μη-ετεροφυλόφιλοι άνδρες έχουν σαφώς μεγαλύτερες πιθανότητες να υπάρξουν θύματα διαδικτυακού εκφοβισμού. Επιπλέον, ενώ τα ομοφυλόφιλα άτομα είναι πιο συχνά θύματα διαδικτυακού εκφοβισμού σε σχέση με τα ετεροφυλόφιλα, οι ομοφυλόφιλοι άνδρες πέφτουν συχνότερα θύματα σε σύγκριση με τις ομοφυλόφιλες γυναίκες. Οι περισσότερες έρευνες που συνδέουν το διαδικτυακό εκφοβισμό με το χαρακτηριστικό του φύλου δίχως να λαμβάνουν υπόψη τον σεξουαλικό προσανατολισμό υποστηρίζουν πως οι γυναίκες είναι πιο συχνά θύματα διαδικτυακού εκφοβισμού. Τα δεδομένα των κατηγοριών Σεξισμός/Sexism του Suspicious Tweets Dataset και Διαδικτυακός εκφοβισμός με βάση το φύλο/Gender Cyberbullying του SOSNet Twitter Dataset επιβεβαιώνουν αυτούς τους ισχυρισμούς καθώς οι λέξεις που κυριαρχούν στις κατηγορίες αυτές είναι ‘gay (ομοφυλόφιλος)’, ‘woman (γυναίκα)’,’female (θηλυκός)’

υπογραμμίζοντας και την εγκυρότητα των δεδομένων. Επιπλέον έρευνα θα μπορούσε να πραγματοποιηθεί για τα τρανσέξουαλ άτομα καθώς επειδή τα χρησιμοποιηθέντα δείγματα μαθητών είναι σαφώς μικρότερα από τις άλλες ομάδες, δεν μπορούν να διεξαχθούν ασφαλή συμπεράσματα.

6. Τόσο στο SOSNet Twitter Dataset όσο και στο Suspicious Tweets Dataset, οι αλγόριθμοι μηχανικής μάθησης που απέδωσαν καλύτερα ήταν οι Random Forest και SVM. Οι 2 αλγόριθμοι είχαν πολύ κοντινή απόδοση με τον Random Forest να υπερτερεί κατά 1% και στα 2 υπό εξέταση σύνολα δεδομένων. Η καλή επίδοση του SVM ήταν αναμενόμενη βάσει βιβλιογραφίας. Ωστόσο, εξάγεται το συμπέρασμα πως και ο Random Forest είναι χρήσιμος σε δεδομένα που προέρχονται από κοινωνικά δίκτυα,όπως στην παρούσα έρευνα, διότι πρόκειται για δεδομένα υψηλών διαστάσεων (high dimensional data) και η εκπαίδευση γίνεται με υποσύνολα των δεδομένων. Σε έρευνα με μέσα κοινωνικής δικτύωσης και ανάλυση συναισθήματος για ανίχνευση του άγχους [130] και πάλι ο Random Forest υπερτερούσε του SVM.

7. Βάσει της απόδοσης των αλγορίθμων, οι 2 αλγόριθμοι που έχουν χειρότερη απόδοση είναι οι KNN και Naïve Bayes. Συγκεκριμένα, στο SOSNet Twitter Dataset, ο KNN αποδίδει πολύ χειρότερα από οποιονδήποτε άλλον αλγόριθμο ενώ στο Suspicious Tweets Dataset ο χειρότερος αλγόριθμος είναι ο Naïve Bayes. Συμπεραίνεται λοιπόν πως σε μη-ισορροπημένα δεδομένα όπως αυτά του Suspicious Tweets Dataset, τα οποία προσομοιώνουν την πραγματική εικόνα του Διαδικτύου, ο Naïve Bayes δεν είναι ικανός να ανιχνεύσει με επιτυχία το Διαδικτυακό εκφοβισμό.

8. Στο Suspicious Tweets Dataset, οι αλγόριθμοι συναντούν μεγαλύτερη δυσκολία στο να ξεχωρίσουν ποια tweets σχετίζονται με διαδικτυακό εκφοβισμό και ποια όχι παρά να ανιχνεύσουν σωστά ποιο είναι το είδος του εκφοβισμού (ρατσιστικός ή σεξιστικός). Συνεπώς, σε πολύ μεγάλο βαθμό προβλέπεται σωστά το είδος και θα μπορούσε να χρησιμοποιηθεί σε περιπτώσεις εκφοβισμού για αυτόματη κατηγοριοποίησή εκφοβιστικών tweets, αλλά χρειάζεται βελτίωση (πιθανώς εκπαίδευση σε περισσότερα δεδομένα των ίδιων κατηγοριών) ώστε να βελτιωθεί και η αυτόματη ανίχνευση του διαδικτυακού εκφοβισμού. Στο SOSNet Twitter Dataset που είναι μεγαλύτερο σύνολο δεδομένων και με ισορροπημένες κλάσεις, δεν τίθεται αυτό το ζήτημα καθώς ανιχνεύεται και ο διαδικτυακός εκφοβισμός και το είδος του με αρκετά μεγάλη ακρίβεια.

Πίνακας 17 Συγκεντρωτικός πίνακας συμπερασμάτων

Συμπέρασμα	Συνοπτική παρουσίαση
1	Τα κείμενα που δεν σχετίζονται με Διαδικτυακό εκφοβισμό έχουν σημαντικά μικρότερη έκταση από αυτά που σχετίζονται.
2	Ο Διαδικτυακός εκφοβισμός Εθνικότητας/ Ethnicity cyberbullying ανιχνεύτηκε πιο εύκολα λόγω των εκφράσεων μίσουςυβριστικού λόγου.
3	Ο Διαδικτυακός εκφοβισμός με βάση το φύλο/ Gender cyberbullying και ο Σεξισμός/Sexism ανιχνεύονται πιο δύσκολα από τους αλγορίθμους.

4	Οι συχνότερες εκφράσεις του Ηλικιακού Διαδικτυακού εκφοβισμού υποδεικνύουν ότι ο διαδικτυακός εκφοβισμός λαμβάνει χώρα κυρίως κατά την εφηβική ηλικία.
5	Ο διαδικτυακός εκφοβισμός με βάση το φύλο έχει ως θύματα περισσότερο τις γυναίκες.
6	Ο διαδικτυακός εκφοβισμός με βάση τη σεξουαλική προτίμηση έχει ως θύματα κατά κύριο λόγο ομοφυλόφιλα άτομα
7	Ο διαδικτυακός εκφοβισμός με βάση συνδυαστικά το φύλο και τη σεξουαλική προτίμηση έχει κυρίως ως θύματα ομοφυλόφιλους άνδρες και τρανς άτομα (χρειάζεται μεγαλύτερο δείγμα για ασφαλέστερα συμπεράσματα)
8	Ο καλύτερος αλγόριθμος και στα 2 datasets ήταν ο Random Forest διότι τα δεδομένα κοινωνικής δικτύωσης είναι υψηλής διάστασης.
9	Ο δεύτερος καλύτερος και για τα 2 datasets ήταν ο SVM με μόλις 1% διαφορά.
10	Στο Suspicious Tweets Dataset, χειρότερος αλγόριθμος ήταν ο Naive Bayes. Αυτός ο αλγόριθμος δεν είναι κατάλληλος για εφαρμογή σε μη-ισορροπημένες κλάσεις δηλαδή σε πραγματικά δεδομένα που προέρχονται από κοινωνικά δίκτυα.
11	Τα είδη διαδικτυακού εκφοβισμού δεν συγχέονται μεταξύ τους σε κανένα από τα 2 datasets.
12	Στο Suspicious Tweets Dataset χρειάζεται βελτίωση ο διαχωρισμός των εκφοβιστικών από τα μη-εκφοβιστικά tweets (εμπλουτισμός του συνόλου δεδομένων).

Βάσει των αποτελεσμάτων, οι αλγόριθμοι που απέδωσαν καλύτερα είναι οι Random Forest και SVM, με τον πρώτο να υπερτερεί ελάχιστα. Η συνεισφορά της παρούσας ανάλυσης είναι η ανάπτυξη μίας αυτοματοποιημένης διαδικασίας για την ανίχνευση κειμένων σχετικών με το διαδικτυακό εκφοβισμό και η περαιτέρω κατηγοριοποίησή τους ανάλογα με το είδος του εκφοβισμού με υψηλή ακρίβεια, απλά με αλγορίθμους μηχανικής μάθησης και χωρίς τη χρήση νευρωνικού δικτύου.

Οι μελλοντικοί στόχοι που τίθενται είναι : (1) Η εύρεση και δημιουργία Ελληνικού συνόλου δεδομένων με κατηγορίες όπως διατυπώθηκαν στο κεφάλαιο ΑΠΟΔΟΣΗ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΕΤΑΦΡΑΣΗ ΔΕΔΟΜΕΝΩΝ. Έπειτα, εφαρμογή της αντίστοιχης μεθοδολογίας προκειμένου να εξετασθεί αν γενικεύεται και σε δεδομένα άλλων γλωσσών (2) εξέταση πιθανής βελτίωσης των αποτελεσμάτων με χρήση νευρωνικών δικτύων και (3) δημιουργία λεξιλογίου διαδικτυακού εκφοβισμού συγκεκριμένου είδους με βάση τα ευρήματα που παρουσιάστηκαν στις ενότητες Συννεφόλεξα Suspicious Tweets Dataset

και Συννεφόλεξα SOSNet Twitter Dataset και διάθεσή του στην επιστημονική κοινότητα προκειμένου να συμβάλουμε σε μελλοντικές έρευνες με χαρακτηριστικά περιεχομένου. για συγκεκριμένα είδη εκφοβισμού (π.χ ηλικιακού).

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
Train dataset	Σύνολο Δεδομένων Εκπαίδευσης
Test dataset	Σύνολο Δεδομένων Ελέγχου
Split ratio	Αναλογία Διαχωρισμού
Supervised	Εποπτευόμενος
K-Nearest Neighbors	Κ-Πλησιέστεροι Γείτονες
Root Node	Κόμβος Ρίζα
Internal Node	Εσωτερικός Κόμβος
Leaf Node	Κόμβος Φύλλο
Precision	Ακρίβεια
Recall	Ανάκληση
Classification Report	Έκθεση ταξινόμησης
Confusion Matrix	Πίνακας σύγχυσης
Balanced classes	Ισορροπημένες κλάσεις
Resampling	Αναδειγματοληψία
Undersampling	Υποδειγματοληψία
Oversampling	Υπερδειγματοληψία
Cluster	Συστάδα

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

NLP	Natural Language Processing
NLTK	Natural Language Toolkit
IR	Information Retrieval
FS	Feature Selection
TF	Term Frequency
IDF	Inverse Document Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
DT	Decision Tree
BERT	Bidirectional Encoder From Transformers
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
DFA	Deterministic Finite Automata
CBOW	Continuous Bag Of Words
ML	Machine Learning
DNN	Deep Neural Network

ΑΝΑΦΟΡΕΣ

[1] Mehendale, N., Rajpara, K., Shah, K., & Phadtare, C. (2022). A Review on Cyberbullying Detection Using Machine Learning. In SSRN Electronic Journal. Elsevier BV. https://doi.org/10.2139/ssrn.4116153 .

[2] Gurusamy, Vairaprakash & Kannan, Subbu. (2014). Preprocessing Techniques for Text Mining.https://www.researchgate.net/publication/273127322_Preprocessing_Techniques_for_Text_Mining

[3] Mohan, Vijayarani. (2015). Preprocessing Techniques for Text Mining - An Overview.https://www.researchgate.net/publication/339529230_Preprocessing_Techniques_for_Text_Mining_-_An_Overview

[4] Palomino, M. A., & Aider, F. (2022). Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis. In Applied Sciences (Vol. 12, Issue 17, p. 8765). MDPI AG. https://doi.org/10.3390/app12178765

[5] Medar, R., Rajpurohit, V. S., & Rashmi, B. (2017). Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning. In 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA). 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA). IEEE. https://doi.org/10.1109/iccubea.2017.8463779

[6] Pintas, J. T., Fernandes, L. A. F., & Garcia, A. C. B. (2021). Feature selection methods for text classification: a systematic literature review. In Artificial Intelligence Review (Vol. 54, Issue 8, pp. 6149–6200). Springer Science and Business Media LLC. https://doi.org/10.1007/s10462-021-09970-6

[7] Jurafsky, D., & Martin, J. H. (2013). *Speech and language processing: Pearson new international edition* (2nd ed.). Pearson Education.https://web.stanford.edu/~jurafsky/slp3/

[8] Basu, A., Walters, C., & Shepherd, M. (2003). Support vector machines for text categorization. In 36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the. 36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the. IEEE. https://doi.org/10.1109/hicss.2003.1174243

[9] Rennie, Jason & Rifkin, Ryan & Memo, Ai. (2002). Improving Multiclass Text Classification with the Support VectorMachine.https://www.researchgate.net/publication/2522390_Improving_Multiclass_Text_Classification_with_the_Support_Vector_Machine

[10] Nikhath, A. K., Subrahmanyam, K., & Vasavi, R. (2016). Building a k-nearest neighbor classifier for text categorization. *International Journal of Computer Science and Information Technologies*, *7*(1), 254-256.

[11] M.U. Noormanshah, W., N.E. Nohuddin, P., & Zainol, Z. (2018). Document Categorization Using Decision Tree: Preliminary Study. In International Journal of Engineering & Technology (Vol. 7, Issue 4.34, p. 437). Science Publishing Corporation. https://doi.org/10.14419/ijet.v7i4.34.26907

[12] Ali, J., Ahmad, N., & Khan, R. (2012). Random Forest and Decision Trees. *International Journal of Computer Science Issues*, *9*(5), 272–277.

[13] Desai, A., Kalaskar, S., Kumbhar, O., & Dhumal, R. (2021). Cyber Bullying Detection on Social Media using Machine Learning. In M. D. Patil & V. A. Vyawahare (Eds.), ITM Web of Conferences (Vol. 40, p. 03038). EDP Sciences. https://doi.org/10.1051/itmconf/20214003038

[14] Carvalho, A. M. de, & Prati, R. C. (2018). Improving kNN classification under Unbalanced Data. A New Geometric Oversampling Approach. In 2018 International Joint Conference on Neural Networks (IJCNN). 2018 International Joint Conference on Neural Networks (IJCNN). IEEE. https://doi.org/10.1109/ijcnn.2018.8489411

[15] Ladani, D. J., & Desai, N. P. (2020). Stopword Identification and Removal Techniques on TC and IR applications: A Survey. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE. https://doi.org/10.1109/icaccs48705.2020.9074166

[16] K., J., & R., J. (2016). Stop-Word Removal Algorithm and its Implementation for Sanskrit Language. In International Journal of Computer Applications (Vol. 150, Issue 2, pp. 15–17). Foundation of Computer Science. https://doi.org/10.5120/ijca2016911462

[17] V. Jha, N. Manjunath, P. D. Shenoy and K. R. Venugopal, "HSRA: Hindi stopword removal algorithm," 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), 2016, pp. 1-5, doi: 10.1109/MicroCom.2016.7522593.

[18] R. M. Rakholia and J. R. Saini, "Lexical classes based stop words categorization for Gujarati language," 2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA) (Fall), 2016, pp. 1-5, doi: 10.1109/ICACCAF.2016.7749005.

[19] Rakholia, R. M., & Saini, J. R. (2017). A Rule-Based Approach to Identify Stop Words for Gujarati Language. In Advances in Intelligent Systems and Computing (pp. 797–806). Springer Singapore. https://doi.org/10.1007/978-981-10-3153-3_79

[20] S. V. S. Gunasekara and P. S. Haddela, "Context aware stopwords for Sinhala Text classification," 2018 National Information Technology Conference (NITC), 2018, pp. 1-6, doi: 10.1109/NITC.2018.8550073.

[21] Sharma, D. (2012). Stemming Algorithms: A Comparative Study and their Analysis. In International Journal of Applied Information Systems (Vol. 4, Issue 3, pp. 7–12). Foundation of Computer Science. https://doi.org/10.5120/ijais12-450655

[22] Porter, M. F. (1980). An algorithm for suffix stripping. In Program (Vol. 14, Issue 3, pp. 130–137). Emerald. https://doi.org/10.1108/eb046814

[23] Lovins, J.B. (1968). Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics, 11*, 22-31.

[24] Paice, C. D. (1990). Another stemmer. In ACM SIGIR Forum (Vol. 24, Issue 3, pp. 56–61). Association for Computing Machinery (ACM). https://doi.org/10.1145/101306.101310

[25] Jivani, A.G. (2011). A Comparative Study of Stemming Algorithms Ms .

[26] Melucci, M., & Orio, N. (2003). A novel method for stemmer generation based on hidden markov models. In Proceedings of the twelfth international conference on Information and knowledge management - CIKM '03. the twelfth international conference. ACM Press. https://doi.org/10.1145/956863.956889

[27] Plisson, J., Lavrac, N., & Mladenic, D. (2004). A Rule based Approach to Word Lemmatization.

[28] Krovetz, R. (2000). Viewing morphology as an inference process. In Artificial Intelligence (Vol. 118, Issues 1–2, pp. 277–294). Elsevier BV. https://doi.org/10.1016/s0004-3702(99)00101-0

[29] Porter, M.F. (2001). Snowball: A language for stemming algorithms.

[30] Kumar, V. (2014). Feature Selection: A literature Review. In The Smart Computing Review (Vol. 4, Issue 3). The Korea Academia-Industrial Cooperation Society. https://doi.org/10.6029/smartcr.2014.03.007

[31] Kozhevnikov, V. A., & Pankratova, E. S. (2020). RESEARCH OF THE TEXT DATA VECTORIZATION AND CLASSIFICATION ALGORITHMS OF MACHINE LEARNING. In Theoretical & Applied Science (Vol. 85, Issue 05, pp. 574–585). International Academy of Theoretical and Applied Sciences. https://doi.org/10.15863/tas.2020.05.85.106
[32] Learn. (n.d.). Retrieved October 30, 2022, from https://scikit-learn.org/stable/
[33] Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., & Gauvain, J.-L. (n.d.). Neural Probabilistic Language Models. In Innovations in Machine Learning (pp. 137–186). Springer-Verlag. https://doi.org/10.1007/3-540-33486-6_6
[34] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1301.3781
[35] Bhaya, Wesam. (2017). Review of Data Preprocessing Techniques in Data Mining. Journal of Engineering and Applied Sciences. 12. 4102-4107. https://medwelljournals.com/abstract/?doi=jeasci.2017.4102.4107
[36] Sumter, S. R., Baumgartner, S. E., Valkenburg, P. M., & Peter, J. (2012). Developmental Trajectories of Peer Victimization: Off-line and Online Experiences During Adolescence. In Journal of Adolescent Health (Vol. 50, Issue 6, pp. 607–613). Elsevier BV. https://doi.org/10.1016/j.jadohealth.2011.10.251
[37] Kowalski, R. M., Limber, S., & Agatston, P. W. (2012). *Cyberbullying: Bullying in the Digital age*. Wiley-Blackwell.
[38] Williams, K. R., & Guerra, N. G. (2007). Prevalence and Predictors of Internet Bullying. In Journal of Adolescent Health (Vol. 41, Issue 6, pp. S14–S21). Elsevier BV. https://doi.org/10.1016/j.jadohealth.2007.08.018
[39] Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: its nature and impact in secondary school pupils. In Journal of Child Psychology and Psychiatry (Vol. 49, Issue 4, pp. 376–385). Wiley. https://doi.org/10.1111/j.1469-7610.2007.01846.x
[40] Cohen-Almagor, R. (2018). Social responsibility on the Internet: Addressing the challenge of cyberbullying. In Aggression and Violent Behavior (Vol. 39, pp. 42–52). Elsevier BV. https://doi.org/10.1016/j.avb.2018.01.001
[41] Hymel, S., & Swearer, S. M. (2015). Four decades of research on school bullying: An introduction. In American Psychologist (Vol. 70, Issue 4, pp. 293–299). American Psychological Association (APA). https://doi.org/10.1037/a0038928
[42] Rivers, I., & Smith, P. K. (1994). Types of bullying behaviour and their correlates. In Aggressive Behavior (Vol. 20, Issue 5, pp. 359–368). Wiley. https://doi.org/10.1002/1098-2337(1994)20:5<359::aid-ab2480200503>3.0.co;2-j
[43] Olweus, D., & Limber, S. P. (2010, November). What do we know about bullying: Information from the Olweus Bullying Questionnaire. Paper presented at the meeting of the International Bullying Prevention Association, Seattle, WA.
[44] Dilmac, B. (2009). Psychological needs as a predictor of cyber bullying: A preliminary report on college students. Educational Sciences: Theory and Practice, 9, 1307–1325
[45] Kowalski, R. M., & Limber, S. P. (2007). Electronic bullying among middle school students. *The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine*, *41*(6 Suppl 1), S22-30. https://doi.org/10.1016/j.jadohealth.2007.08.017
[46] Hinduja, S., & Patchin, J. W. (2008). Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant Behavior*, *29*(2), 129–156. https://doi.org/10.1080/01639620701457816
[47] Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology*, *49*(2), 147–154. https://doi.org/10.1111/j.1467-9450.2007.00611.x
[48] Li, Q. (2006). Cyberbullying in schools: A research of gender differences. *School Psychology International*, *27*(2), 157–170. https://doi.org/10.1177/0143034306064547
[49] Sourander, A., Brunstein Klomek, A., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen, M., Ristkari, T., & Helenius, H. (2010). Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. *Archives of General Psychiatry*, *67*(7), 720. https://doi.org/10.1001/archgenpsychiatry.2010.79
[50] Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, *140*(4), 1073–1137. https://doi.org/10.1037/a0035618
[51] Black, W. W., Fedewa, A. L., & Gonzalez, K. A. (2012). Effects of “Safe School” Programs and Policies on the Social Climate for Sexual-Minority Youth: A Review of the Literature. In Journal of LGBT Youth (Vol. 9, Issue 4, pp. 321–339). Informa UK Limited. https://doi.org/10.1080/19361653.2012.714343
[52] Zych, I., Ortega-Ruiz, R., & Del Rey, R. (2015). Systematic review of theoretical studies on bullying and cyberbullying: Facts, knowledge, prevention, and intervention. In Aggression and Violent Behavior (Vol. 23, pp. 1–21). Elsevier BV. https://doi.org/10.1016/j.avb.2015.10.001

[53] Hinduja S, Patchin JW. Bullying, Cyberbullying, and LGBTQ Students. 2020. https://cyberbullying.org/bullying-cyberbullying-lgbtq
[54] Cénat, J. M., Blais, M., Hébert, M., Lavoie, F., & Guerrier, M. (2015). Correlates of bullying in Quebec high school students: The vulnerability of sexual-minority youth. In Journal of Affective Disorders (Vol. 183, pp. 315–321). Elsevier BV. https://doi.org/10.1016/j.jad.2015.05.011
[55] Duarte, C., Pittman, S. K., Thorsen, M. M., Cunningham, R. M., & Ranney, M. L. (2018). Correlation of Minority Status, Cyberbullying, and Mental Health: A Cross-Sectional Study of 1031 Adolescents. In Journal of Child & Adolescent Trauma (Vol. 11, Issue 1, pp. 39–48). Springer Science and Business Media LLC. https://doi.org/10.1007/s40653-018-0201-4
[56] Llorent, V. J., Ortega-Ruiz, R., & Zych, I. (2016). Bullying and Cyberbullying in Minorities: Are They More Vulnerable than the Majority Group? In Frontiers in Psychology (Vol. 7). Frontiers Media SA. https://doi.org/10.3389/fpsyg.2016.01507
[57] Camodeca, M., & Goossens, F. A. (2005). Aggression, social cognitions, anger and sadness in bullies and victims. In Journal of Child Psychology and Psychiatry (Vol. 46, Issue 2, pp. 186–197). Wiley. https://doi.org/10.1111/j.1469-7610.2004.00347.x
[58] Gerlsma, C., & Lugtmeyer, V. (2016). Offense Type as Determinant of Revenge and Forgiveness After Victimization: Adolescents' Responses to Injustice and Aggression. In Journal of School Violence (Vol. 17, Issue 1, pp. 16–27). Informa UK Limited. https://doi.org/10.1080/15388220.2016.1193741
[59] Edwards, L., Kontostathis, A. E., & Fisher, C. (2016). Cyberbullying, Race/Ethnicity and Mental Health Outcomes: A Review of the Literature. In Media and Communication (Vol. 4, Issue 3, pp. 71–78). Cogitatio. https://doi.org/10.17645/mac.v4i3.525
[60] Hinduja, S. (n.d.). *Cyberbullying statistics 2021: Age, gender, sexual orientation, and Race*. Cyberbullying Research Center. Retrieved November 17, 2022, from https://cyberbullying.org/cyberbullying-statistics-age-gender-sexual-orientation-race?fbclid=IwAR2RHDoskNXKvaCLXjkxPETcJaC3Ks5-voH1pnbT4NBWbmM2Rn4-E999VGA
[61] Pontes, N. M. H., Ayres, C. G., Lewandowski, C., & Pontes, M. C. F. (2018). Trends in bullying victimization by gender among U.S. high school students. In Research in Nursing & Health (Vol. 41, Issue 3, pp. 243–251). Wiley. https://doi.org/10.1002/nur.21868
[62] Hinduja, S. (no date) *Bullying because of religion: Our latest findings and best practices*, *Cyberbullying Research Center*. Available at: https://cyberbullying.org/bullying-and-religion (Accessed: November 18, 2022).
[63] LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. (1763). In Philosophical Transactions of the Royal Society of London (Vol. 53, pp. 370–418). The Royal Society. https://doi.org/10.1098/rstl.1763.0053
[64] Mosteller, F., & Wallace, D. L. (1963). Inference in an Authorship Problem. In Journal of the American Statistical Association (Vol. 58, Issue 302, p. 275). JSTOR. https://doi.org/10.2307/2283270
[65] Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer New York. https://doi.org/10.1007/978-1-4757-2440-0
[66] Kwok, James. (2000). Automated Text Categorization Using Support Vector Machine.
[67] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, 137–142. https://doi.org/10.1007/bfb0026683
[68] Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In Proceedings of the seventh international conference on Information and knowledge management - CIKM '98. the seventh international conference. ACM Press. https://doi.org/10.1145/288627.288651
[69] Zhao, Y., & Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. In Advances in Space Research (Vol. 41, Issue 12, pp. 1955–1959). Elsevier BV. https://doi.org/10.1016/j.asr.2007.07.020
[70] Decision Trees. (2009). In Classification Methods for Remotely Sensed Data, Second Edition (pp. 183–220). CRC Press. https://doi.org/10.1201/9781420090741.ch6
[71] Rafael Olivas, „Decision Trees – A primer for Decision-making Professionals", 2007
[72] Maimon, O. Z., & Rokach, L. (2005). *Data mining and knowledge discovery handbook* (O. Maimon & L. Rokach, Eds.). Springer.
[73] Nicolosi, N. (2008). Feature Selection Methods for Text Classification.
[74] Bashiri, H. & Oroumchian, Farhad & Moeini, Ali. (2005). Persian Email Classification Based on Rocchio and K-Nearest Neighbor Approach. Farhad Oroumchian.
[75] Alhutaish, Roiss & Omar, Nazlia. (2015). Arabic Text Classification Using K-Nearest Neighbour Algorithm. International Arab Journal of Information Technology. 12. 190-195.
[76] Zhongguo, Y., Hongqi, L., Liping, Z., Qiang, L., & Ali, S. (2017). A case based method to predict optimal k value for k-NN algorithm. In Journal of Intelligent & Fuzzy Systems (Vol. 33, Issue 1, pp. 55–65). IOS Press. https://doi.org/10.3233/jifs-161062

[77] Kang, P., & Cho, S. (2008). Locally linear reconstruction for instance-based learning. In Pattern Recognition (Vol. 41, Issue 11, pp. 3507–3518). Elsevier BV. https://doi.org/10.1016/j.patcog.2008.04.009
[78] Meesad, P., & Hengpraprohm, K. (2008). Combination of KNN-Based Feature Selection and KNNBased Missing-Value Imputation of Microarray Data. In 2008 3rd International Conference on Innovative Computing Information and Control. 2008 3rd International Conference on Innovative Computing Information and Control. IEEE. https://doi.org/10.1109/icicic.2008.635
[79] Lall, U., & Sharma, A. (1996). A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. In Water Resources Research (Vol. 32, Issue 3, pp. 679–693). American Geophysical Union (AGU). https://doi.org/10.1029/95wr02966
[80] Liu, H., Zhang, S., Zhao, J., Zhao, X., & Mo, Y. (2010). A New Classification Algorithm Using Mutual Nearest Neighbors. In 2010 Ninth International Conference on Grid and Cloud Computing. 2010 9th International Conference on Grid and Cloud Computing (GCC 2010). IEEE. https://doi.org/10.1109/gcc.2010.23
[81] Wang, J., Neskovic, P., & Cooper, L. N. (2006). Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence. In Pattern Recognition (Vol. 39, Issue 3, pp. 417–423). Elsevier BV. https://doi.org/10.1016/j.patcog.2005.08.009
[82] Breiman, L. (2001). In Machine Learning (Vol. 45, Issue 1, pp. 5–32). Springer Science and Business Media LLC. https://doi.org/10.1023/a:1010933404324
[83] *Random forests Leo Breiman and Adele Cutler*. Random forests - classification description. (n.d.). Retrieved November 22, 2022, from https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#prox
[84] Introduction to Decision Trees and Random Forests, Ned Horning; American Museum of Natural History's
[85] Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., & Alsaadi, F. E. (2020). Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. In Applied Soft Computing (Vol. 86, p. 105836). Elsevier BV. https://doi.org/10.1016/j.asoc.2019.105836
[86] Yang, J., Liu, Y., Zhu, X., Liu, Z., & Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. In Information Processing & Management (Vol. 48, Issue 4, pp. 741–754). Elsevier BV. https://doi.org/10.1016/j.ipm.2011.12.005
[87] Yang, Y., & Pedersen, J.O. (1997). A Comparative Study on Feature Selection in Text Categorization. International Conference on Machine Learning.
[88] Aghdam, M. H., Ghasem-Aghaee, N., & Basiri, M. E. (2009). Text feature selection using ant colony optimization. In Expert Systems with Applications (Vol. 36, Issue 3, pp. 6843–6853). Elsevier BV. https://doi.org/10.1016/j.eswa.2008.08.022
[89] Somol, P., & Novovičová, J. (2010). Evaluating Stability and Comparing Output of Feature Selectors that Optimize Feature Subset Cardinality. In IEEE Transactions on Pattern Analysis and Machine Intelligence (Vol. 32, Issue 11, pp. 1921–1939). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tpami.2010.34
[90] Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. In Knowledge-Based Systems (Vol. 36, pp. 226–235). Elsevier BV. https://doi.org/10.1016/j.knosys.2012.06.005
[91] Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. In Expert Systems with Applications (Vol. 36, Issue 3, pp. 5432–5435). Elsevier BV. https://doi.org/10.1016/j.eswa.2008.06.054
[92] Uysal, A. K. (2016). An improved global feature selection scheme for text classification. In Expert Systems with Applications (Vol. 43, pp. 82–92). Elsevier BV. https://doi.org/10.1016/j.eswa.2015.08.050
[93] Nigam, K., Mccallum, A. K., Thrun, S., & Mitchell, T. (2000). In Machine Learning (Vol. 39, Issue 2/3, pp. 103–134). Springer Science and Business Media LLC. https://doi.org/10.1023/a:1007692713085
[94] *Understanding the classification report through sklearn* (2018) *Muthukrishnan*. Available at: https://muthu.co/understanding-the-classification-report-in-sklearn/ (Accessed: December 12, 2022).
[95] Karimi, Zohreh. (2021). Confusion Matrix.
[96] Prakash, Sherly. (2021). A SURVEY ON CYBERBULLYING DETECTION.
[97] Statistic Brain. (2019, August 15). *Cyberbullying / bullying statistics*. Statistic Brain. Retrieved January 1, 2023, from https://www.statisticbrain.com/cyber-bullying-statistics/
[98] Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. In ACM Transactions on Interactive Intelligent Systems (Vol. 2, Issue 3, pp. 1–30). Association for Computing Machinery (ACM). https://doi.org/10.1145/2362394.2362400
[99] Van Hee, Cynthia & Lefever, Els & Verhoeven, Ben & Mennes, Julie & Desmet, Bart & Pauw, Guy & Daelemans, Walter & Hoste, Véronique. (2015). Automatic detection and prevention of cyberbullying.

[100] Van Hee, Cynthia & Lefever, Els & Verhoeven, Ben & Mennes, Julie & Desmet, Bart & Pauw, Guy & Daelemans, Walter & Hoste, Véronique. (2015). Detection and fine-grained classification of cyberbullying events.
[101] Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishr, S. (2015). Prediction of Cyberbullying Incidents on the Instagram Social Network (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1508.06257
[102] Zhao, R., Zhou, A., & Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. In Proceedings of the 17th International Conference on Distributed Computing and Networking. ICDCN '16: 17th International Conference on Distributed Computing and Networking. ACM. https://doi.org/10.1145/2833312.2849567
[103] Salawu, S., He, Y., & Lumsden, J. (2020). Approaches to Automated Detection of Cyberbullying: A Survey. In IEEE Transactions on Affective Computing (Vol. 11, Issue 1, pp. 3–24). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/taffc.2017.2761757
[104] Edwards, April & Leatherman, Amanda. (2009). ChatCoder: Toward the Tracking and Categorization of Internet Predators. 3.
[105] P.N. Tan,F. Chen, A. Jain, "Information assurance: Detection of web spam attacks in social media." Proceedings of Army Science Conference, Orland, Florida. 2010
[106] Yin, Dawei & Xue, Zhenzhen & Hong, Liangjie & Davison, Brian & Edwards, April & Edwards, Lynne. (2009). Detection of harassment on Web 2.0.
[107] CHISHOLM, J. F. (2006). Cyberspace Violence against Girls and Adolescent Females. In Annals of the New York Academy of Sciences (Vol. 1087, Issue 1, pp. 74–89). Wiley. https://doi.org/10.1196/annals.1385.022
[108] Bart Desmet and Véronique Hoste. 2014. Recognising suicidal messages in Dutch social media. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 830–835, Reykjavik, Iceland. European Language Resources Association (ELRA).
[109] Dinakar, K., Reichart, R., & Lieberman, H. (2021). Modeling the Detection of Textual Cyberbullying. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 5, Issue 3, pp. 11–17). Association for the Advancement of Artificial Intelligence (AAAI). https://doi.org/10.1609/icwsm.v5i3.14209
[110] Nahar, V., Li, X., & Pang, C. (2013). An Effective Approach for Cyberbullying Detection.
[111] Islam, M. M., Uddin, M. A., Islam, L., Akter, A., Sharmin, S., & Acharjee, U. K. (2020). Cyberbullying Detection on Social Networks Using Machine Learning Approaches. In 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). IEEE. https://doi.org/10.1109/csde50874.2020.9411601
[112] Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using Machine Learning to Detect Cyberbullying. In 2011 10th International Conference on Machine Learning and Applications and Workshops. 2011 Tenth International Conference on Machine Learning and Applications (ICMLA 2011). IEEE. https://doi.org/10.1109/icmla.2011.152
[113] Al-Ajlan, M. A., & Ykhlef, M. (2018). Deep Learning Algorithm for Cyberbullying Detection. In International Journal of Advanced Computer Science and Applications (Vol. 9, Issue 9). The Science and Information Organization. https://doi.org/10.14569/ijacsa.2018.090927
[114] Agrawal, S., & Awekar, A. (2018). Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms. In Lecture Notes in Computer Science (pp. 141–153). Springer International Publishing. https://doi.org/10.1007/978-3-319-76941-7_11
[115] Zhang, X., Tong, J., Vishwamitra, N., Whittaker, E., Mazer, J. P., Kowalski, R., Hu, H., Luo, F., Macbeth, J., & Dillon, E. (2016). Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE. https://doi.org/10.1109/icmla.2016.0132
[116] Dadvar, M., Trieschnigg, D., & de Jong, F. (2014). Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies. In Advances in Artificial Intelligence (pp. 275–281). Springer International Publishing. https://doi.org/10.1007/978-3-319-06483-3_25
[117] Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. In H. Suleman (Ed.), PLOS ONE (Vol. 13, Issue 10, p. e0203794). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0203794
[118] Wang, J., Fu, K., & Lu, C.-T. (2020). SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection. In 2020 IEEE International Conference on Big Data (Big Data). 2020 IEEE International Conference on Big Data (Big Data). IEEE. https://doi.org/10.1109/bigdata50022.2020.9378065

[119] Chatzakou, D., Leontiadis, I., Blackburn, J., Cristofaro, E. D., Stringhini, G., Vakali, A., & Kourtellis, N. (2019). Detecting Cyberbullying and Cyberaggression in Social Media. In ACM Transactions on the Web (Vol. 13, Issue 3, pp. 1–51). Association for Computing Machinery (ACM). https://doi.org/10.1145/3343484

[120] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1703.04009

[121] Bretschneider, Uwe & Wöhner, Thomas & Peters, Ralf. (2014). Detecting Online Harassment in Social Networks.

[122] Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Proceedings of the NAACL Student Research Workshop. Proceedings of the NAACL Student Research Workshop. Association for Computational Linguistics. https://doi.org/10.18653/v1/n16-2013

[123] Waseem, Z. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In Proceedings of the First Workshop on NLP and Computational Social Science. Proceedings of the First Workshop on NLP and Computational Social Science. Association for Computational Linguistics. https://doi.org/10.18653/v1/w16-5618

[124] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from Bullying Traces in Social Media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 656–666, Montréal, Canada. Association for Computational Linguistics

[125] Albright, M. (2021, August 12). *Suspicious tweets*. Kaggle. Retrieved January 4, 2023, from https://www.kaggle.com/datasets/munkialbright/suspicious-tweets?datasetId=1527522&sortBy=dateRun&tab=bookmarked

[126] Zaidi, S. A. R. (2021, February 17). *Suspicious tweets*. Kaggle. Retrieved January 4, 2023, from https://www.kaggle.com/datasets/syedabbasraza/suspicious-tweets

[127] *Using Twitter as a data source: An overview of Social Media Research Tools (2015)*. Impact of Social Sciences. Retrieved January 7, 2023, from https://blogs.lse.ac.uk/impactofsocialsciences/2015/07/10/social-media-research-tools-overview/

[128] Raskauskas, J., & Stoltz, A. D. (2007). Involvement in traditional and electronic bullying among adolescents. In Developmental Psychology (Vol. 43, Issue 3, pp. 564–575). American Psychological Association (APA). https://doi.org/10.1037/0012-1649.43.3.564

[129] Hong, J. S., Davis, J. P., Sterzing, P. R., Yoon, J., Choi, S., & Smith, D. C. (2014). A conceptual framework for understanding the association between school bullying victimization and substance misuse. In American Journal of Orthopsychiatry (Vol. 84, Issue 6, pp. 696–710). American Psychological Association (APA). https://doi.org/10.1037/ort0000036

[130] Saifullah S, Fauziah Y, Aribowo AS. Comparison of Machine Learning for Sentiment Analysis in Detecting Anxiety Based on Social Media Data. arXiv; 2021.