ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάπτυξη συστήματος συγκέντρωσης, ανάλυσης και επιβεβαίωσης συμβάντων σε ειδήσεις και κοινωνικά δίκτυα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΕΥΣΤΑΘΙΟΥ ΤΖΑΘΑ

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2020

Η σελίδα αυτή είναι σκόπιμα λευκή.

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

Ανάπτυξη συστήματος συγκέντρωσης, ανάλυσης και επιβεβαίωσης συμβάντων σε ειδήσεις και κοινωνικά δίκτυα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΕΥΣΤΑΘΙΟΥ ΤΖΑΘΑ

Επιβλέπων :
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την Τετάρτη 11 Μαρτίου 2020.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτριος Ασκούνης Καθηγητής Ε.Μ.Π.	Ιωάννης Ψαρράς, Καθηγητής Ε.Μ.Π.	Χρυσόστομος Δούκας, Αναπληρωτής Καθηγητής Ε.Μ.Π

Αθήνα, Μάρτιος 2020

(Υπογραφή)

.....

ΕΥΣΤΑΘΙΟΣ ΤΖΑΘΑΣ

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

Περίληψη

Η εξάπλωση των ειδήσεων στις μέρες μας γίνεται με ταχύτατους ρυθμούς. Τα νέα ποτέ στην ιστορία δεν μεταδίδονταν τόσο γρήγορα και σε τέτοιο πλήθος από πλευράς πληροφορίας. Το διαδίκτυο είναι το μέσο αυτής της γιγαντιαίας ανταλλαγής πληροφορίων και τα κοινωνικά δίκτυα τα κέντρα μετάδοσης τους. Καθένας μπορεί να γίνει δίαυλος ενημέρωσης χωρίς όμως να ελέγχεται η αξιοπιστία του. Αυτήν την αδυναμία ελέγχου της ποιότητας των ειδήσεων χρησιμοποιούν κακόβουλοι χρήστες ώστε να προωθήσουν τα προσωπικά τους οικονομικά συμφέροντα ή να για αναπαράγουν αναλήθειες.

Ο σκοπός της διπλωματικής εργασίας είναι η ανάπτυξη μεθοδολογίας για την ανίχνευση, αναγνώριση και καταγραφή των clickbaits , μιας τακτικής χειραγώγησης του αναγνώστη για την μετάδοση και εξάπλωση ψευδών ειδήσεων. Αρχικό ζητούμενο αποτέλεσε η αυτοματοποιημένη αξιολόγηση των ειδήσεων ως προς την αξιοπιστία τους. Η μεθοδολογία που ακολουθήσαμε για την εύρεση και συλλογή παραπλανητικών τίτλων έγινε κατόπιν μελέτης των σύγχρονων τεχνολογικών ερευνών πάνω σε παρόμοια έργα της επιστήμης των δεδομένων. Στόχος ήταν η ταυτοποίηση των τίτλων ειδήσεων που ξεχωρίζουν από την πλειοψηφία των δεδομένων ως προς τα γλωσσικά τους χαρακτηριστικά. Στο πλαίσιο αυτό υλοποιήθηκε σύστημα ανάλυσης του γραπτού λόγου σε μια προσπάθεια να εξαχθούν νοήματα από τα γλωσσικά δεδομένα. Κατόπιν υλοποιήθηκαν μοντέλα αναδράσεων , για την εκτίμηση και ταξινόμηση των τίτλων σε κατηγορίες . Στα μοντέλα αυτά συμπεριλήφθηκαν Δέντρα Αποφάσεων , μηχανές SVM και νευρωνικά δίκτυα (RNN , MLP). Το σύστημα υλοποιήθηκε εξολοκλήρου σε γλώσσα Python και τα δεδομένα αλιεύθηκαν από το ίντερνετ και με τη χρήση διασύνδεσης προγραμματισμού εφαρμογών του κοινωνικού δικτύου Twitter.

Λέξεις Κλειδιά: κοινωνικά δίκτυα, ψευδείς ειδήσεις , clickbait , Python , NLP, νευρωνικά δίκτυα, ταξινόμηση , δέντρα αποφάσεων , SVM

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

In our days news spreads rapidly in every corner of the globe. It is unprecedented in human history for so much information, of any kind or form, to be able to reach as many people and as quickly as it does today. The internet has been the tool that has enabled this development. Moreover, over the past 15 years social networks have turbocharged the spread of information. On social networks anyone can reach an enormous amount of people and spread any kind of information he wishes without any verification of whether the information is credible or not. This lack of verification and quality control of information of news spread online allows malicious agents to spread misinformation so they can further their economic or political interests. The purpose of this thesis is the development of a methodology for detecting, identifying and recording clickbaits, a tactic that is often used for the manipulation of readers and the spreading of fake news.

The initial aim was the automation of the evaluation of the credibility of news. The methodology we used for finding and collecting misleading article titles resulted after studying recent research on similar topics of data science. The goal was the identification of news titles that stand out due to their linguistic characteristics. In this context, a system of analysis of written speech was implemented in an attempt to extract the meaning of linguistic data. Feedback models were then implemented to evaluate and classify the titles into categories. These models included Decision Trees, Support-Vector Machines (SVM in brief), and neural networks such as Recurrent Neural Network and Multilayer Perceptron (RNN and MLP in brief). The system was implemented entirely in Python and the data was captured from the Internet and using the Twitter application programming interface.

Keywords: social media, fake news , clickbait , Twitter , Python , NLP, neural networks, ταξινόμηση , δέντρα αποφάσεων , SVM

Η σελίδα αυτή είναι σκόπιμα λευκή

Περιεχόμενα

1. Εισαγωγή

1.1 Ψευδείς ειδήσεις και ιστορικότητα του φαινομένου

Τα ευρήματα μελετών που έχουν γίνει εδώ και τουλάχιστον τέσσερις δεκαετίες έχουν δείξει πως οι άνθρωποι δεν είναι και τόσο καλοί στην ανίχνευση εξαπάτησης από κάποιο κείμενο. Στην πραγματικότητα, με βάση μια ανάλυση πάνω από διακοσίων (200) πειραμάτων, οι απαντήσεις που δίνουμε βασιζόμενοι στην κρίση μας είναι μόλις 4/100 πιο σωστές από τις απαντήσεις που θα δίναμε εντελώς τυχαία. (Bond and DePaulo, 2006)[11]

Είναι άραγε οι ψευδείς ειδήσεις αθώες; Μήπως διακυβεύονται πολλά σε πολιτικό, κοινωνικό και οικονομικό επίπεδο πίσω από μια ψευδή είδηση;

Αυτή είναι πάντως η άποψη του έλληνας νομοθέτη, ο οποίος έχει ποινικοποιήσει τη διασπορά ψευδών ειδήσεων. Σύμφωνα, με το άρθρο 191 του Ποινικού Κώδικα, με την πλημμεληματική ποινή φυλάκισης τουλάχιστον τριών μηνών και χρηματική ποινή καταδικάζεται όποιος διασπείρει ψευδείς ειδήσεις ή φήμες ικανές να επιφέρουν ανησυχίες ή φόβο στους πολίτες ή να ταράξουν τη δημόσια πίστη ή να κλονίσουν την εμπιστοσύνη του κοινού στο εθνικό νόμισμα ή στις ένοπλες δυνάμεις της χώρας ή να επιφέρουν διαταραχή στις διεθνείς σχέσεις της χώρας.

Είναι γεγονός πως το φαινόμενο της παραπληροφόρησης μέσω διασποράς ψευδών ειδήσεων, γνωστό ως “fake news”, έχει λάβει μεγάλες διαστάσεις στον ψηφιακό κόσμο της υπερπληροφόρησης, όπου οι κοινωνοί της πληροφορίας έχουν περιορισμένο χρόνο να την αναλώσουν και ακόμη λιγότερο για να την επαληθεύσουν. Έτσι, οι λεζάντες αποκτούν μεγαλύτερη σημασία από την είδηση καθεαυτή, με την ανάγνωση εύπεπτων τίτλων να κυριαρχεί.

Αν και το φαινόμενο της παραπληροφόρησης δεν είναι νέο, ήδη ψήγματα ψευδών ειδήσεων εντοπίζουμε στον κλασσικό κίτρινο τύπο, οι ψευδείς ειδήσεις στον ψηφιακό κόσμο αποκτούν άλλη δυναμική, χάρις στην ευκολία και την ταχύτητα με την οποία διαδίδονται μέσω των διαφόρων ιστολογίων και κυρίως των μέσων κοινωνικής δικτύωσης. Άλλωστε πλέον, το διαδίκτυο αποτελεί πρωτογενή πηγή πληροφόρησης για το μεγαλύτερο μέρος του αναγνωστικού κοινού, παραμένοντας ωστόσο αχαρτογράφητο αναφορικά με την πηγή, την πορεία και το φιλτράρισμα των ειδήσεων. Παράλληλα, ο ρυθμός με τον οποίο οι διάφοροι ειδησεογραφικοί ιστότοποι τροφοδοτούνται με ειδήσεις -ψευδείς και μη-, οι οποίες

διακινούνται, καθιστά εξαιρετικά δύσκολο τον εντοπισμό όχι μόνο της παραπληροφόρησης, αλλά κυρίως της πηγής αυτής.

Όπως αναφέρει ο Donald Trumb σε ένα tweet του στις 13 Δεκεμβριου του 2018:

"Wow, more than 90% of Fake News Media coverage of me is negative, with numerous forced retractions of untrue stories. Hence my use of Social Media, the only way to get the truth out. Much of Mainstream Media has become a joke!"

"Είναι τρομερό , πάνω από το 90% των καλύψεων από τα Μίντια Ψευδών Ειδήσεων που έχουν να κάνουν με εμένα είναι αρνητικές, με πολλαπλές εξαναγκασμένες ανακλήσεις για την αναλήθεια των ιστοριών τους. Ως εκ τούτου η χρήση των μέσων κοινωνικής δικτύωσης είναι ο μόνος τρόπος να μεταδώσω την αλήθεια. Τα περισσότερα συμβατικά κανάλια είναι πλέον για γέλια."

Η παρουσίαση των μέσων κοινωνικής δικτύωσης από τον Πλανητάρχη ως θεματοφύλακες της διαφάνειας και της αλήθειας φαίνεται οξύμωρη, ειδικά αν σκεφτεί κανείς ότι η αναπαραγωγή των ειδήσεων σε αυτά γίνεται συνήθως χωρίς καμία αναφορά στην πηγή της πληροφορίας, με αποτέλεσμα ο συντάκτης να διαδραματίζει ήσσονος σημασίας ρόλο για τον έλεγχο της αξιοπιστίας της αυτής. Στον αντίποδα της συντακτικής ευθύνης και της δεοντολογίας του δημοσιογραφικού λειτουργήματος, ανευρίσκουμε το ποσοτικό κριτήριο των αναδημοσιεύσεων ως ενδείκτη αλήθειας. Σε κάθε περίπτωση πάντως, ανεξαρτήτως μέσου έντυπου ή ψηφιακού ο εντοπισμός και η κατηγοριοποίηση των ψευδών ειδήσεων θα πρέπει να γίνεται με κοινά μεθοδολογικά εργαλεία.

Σκοπός της παρούσας εργασίας είναι να παρουσιάσει μια συνολική και συνεκτική προσέγγιση του φαινομένου των ψευδών ειδήσεων, εισφέροντας τα απαραίτητα μεθοδολογικά εργαλεία προκειμένου να συστηματοποιηθεί και να διευκολυνθεί η κατηγοριοποίηση των ψευδών ειδήσεων.

1.2 Δομή Εργασίας

Αφού έγινε κατανοητή η σημασία του φαινομένου και οι επιπτώσεις που έχει σε κοινωνικό , οικονομικό και πολιτικό επίπεδο , στόχος μας έγινε η προσπάθεια αντιμετώπισης του και η άντληση συμπερασμάτων σχετικά με τις μεθόδους που χρησιμοποιούν οι επίδοξοι δημοσιογράφοι.

Αρχικά αναγάγαμε το φαινόμενο σε πρόβλημα συνηθισμένης περίπτωσης αναγνώρισης προτύπων ώστε να το αντιμετωπίσουμε με την βοήθεια αλγορίθμων μηχανικής μάθησης. Κατηγοριοποιήσαμε τους τίτλους σε δύο σύνολα , τους παραπλανητικούς (clickbaits) και τους μη παραπλανητικούς (not-clickbaits) με βάση την συλλογιστική πορεία που περιγράφεται στα παρακάτω κεφάλαια. Το επόμενο βήμα ήταν να συλλέξουμε και να αποθηκεύσουμε τα δεδομένα. Αυτό το κάναμε με χρήση κώδικα που μπορεί να βρεθεί αναρτημένος στην ηλεκτρονική βιβλιοθήκη (github) έχοντας ως στόχο λογαριασμούς στο Twitter από συγκεκριμένα ειδησεογραφικά κανάλια την περίοδο του Νοεμβρίου του 2018. Έπειτα συγχωνεύσαμε το δείγμα αυτό με κάποιο έτοιμο που συλλέξαμε στο ίντερνετ , τα στοιχεία του οποίου περιγράφουμε στην συνέχεια.

Αφού ολοκληρώσαμε την συλλογή , συγχώνευση και καθαρισμό των δειγμάτων, μετατρέψαμε τα δεδομένα μας σε πίνακες έτοιμους να αναλυθούν με χρήση βιβλιοθηκών της γλώσσας Python. Η επόμενες μετατροπές των δεδομένων μας έγιναν ώστε οι τίτλοι ή οι πίνακες χαρακτηριστικών, από απλό κείμενο να πάρουν διανυσματικές μορφές ώστε να μπορούμε να τους εισάγουμε με αλγορίθμους ταξινόμησης. Οι αλγόριθμοι ταξινόμησης στην πορεία κατηγοριοποίησαν τους υποψήφιους τίτλους στις επιθυμητές μας τάξεις (clickbait / not-clickabait).

Χρησιμοποιώντας διάφορους αλγορίθμους ταξινόμησης και νευρωνικά δίκτυα , πάνω σε διαφορετικές μορφοποιήσεις του δείγματος μας βγάλαμε συμπεράσματα με βάση την απόδοση τους και τους συγκρίναμε για το πως αποδίδει ο καθένας σε περιπτώσεις ασκήσεων Επεξεργασίας Φυσικής Γλώσσας , όπως στην περίπτωση μας.

Τα επόμενα κεφάλαια επεξηγούν την διαδικασία της έρευνας μας.

2 Θεωρητικό Υπόβαθρο

2.1 Κατηγορίες ψευδών ειδήσεων

Προτού μιλήσει κανείς για την διαδικασία ανίχνευσης της εξαπάτησης, θα ήταν χρήσιμο να προηγηθεί ο ορισμός των εννοιών που θα αποτελέσουν τις λέξεις κλειδιά στο πλαίσιο της παρούσας εργασίας:

Ως fake news [19] ορίζονται οι ειδήσεις, που είναι εσφαλμένες, μη απόλυτα έγκυρες και χωλαίνουν ως προς την απόδοση της αλήθειας , ο στόχος τους κυμαίνεται από το να παραπλανήσουν μέχρι και να διακωμωδήσουν, με τα όρια μεταξύ παραπλάνησης και διακωμώδησης να μην είναι απόλυτα ευκρινή. Τα fake news μπορεί να εξαπλώνονται από μικρούς λογαριασμούς εντός κάποιου microblog μέχρι και από μεγάλα ειδησεογραφικά κανάλια (ΕΡΤ - ΤοΚουλούρι) [14].

Ακολούθως, ένας άλλος κεντρικής σημασίας ορισμός που θα χρειαστεί να γνωρίζουμε είναι τα clickbaits. Ως click-baits ορίζουμε την στρατηγική συγγραφής τίτλων με βαρύγδουπο περιεχόμενο, το οποίο εντυπωσιάζει και διεγείρει το συναίσθημα των αναγνωστών, με αποτέλεσμα την αύξηση της επισκεψιμότητας ή αλλιώς των clicks, των ιστοτόπων που τις φιλοξενούν. Την στρατηγική αυτή ακολουθούν πολλά ειδησεογραφικά πρακτορεία, προκειμένου να προσελκύσουν περισσότερα κλικς στον ιστότοπο τους και συνακόλουθα κατά τον απολογισμό της επισκεψιμότητας τους να αποκομίσουν μεγαλύτερο κέρδος από τις διαφημίσεις. Οι τίτλοι εντυπωσιασμού έχουν συνήθως μικρή ή ακόμη και καμία σχέση με το περιεχόμενο του άρθρου, του οποίου αποτελούν επιστέγασμα. Σε άλλη εκδοχή τους, οι τίτλοι εντυπωσιασμού, διεγείρουν την περιέργεια του αναγνώστη, ο οποίος για να την ικανοποιήσει σπεύδει να επισκεφθεί το site που φιλοξενεί τη σχετική είδηση. Στην περίπτωση αυτή ο στρατηγικά δομημένος τίτλος έχει πετύχει το στόχο του, καθώς τα ερωτήματα που γέννησε, κρίθηκαν άξια διερεύνησης με αναφορά στο σύνολο της είδησης, η οποία ωστόσο στερείται ουσιώδους πληροφορίας.

Μπορούμε να κατατάξουμε τα fake news στις ακόλουθες τρεις επιμέρους μεγάλες κατηγορίες σύμφωνα και με την μελέτη των Verstraete, Mark and Bambauer, Derek E. and Yakowitz Bambauer [9].

- Σοβαρές Κατηγορίες ψευδών ειδήσεων: Αυτό το είδος αφορά κυρίως τις περιπτώσεις δόλιας δημοσιογραφίας, όπου ο ρεπόρτερ παρουσιάζει «κατασκευασμένες» ειδήσεις και ψευδείς κατηγορίες συνήθως για κάποιο δημόσιο πρόσωπο. Τέτοιου τύπου νέα χρησιμοποιούν click-bait επικεφαλίδες, έχουν θέματα που αφορούν την προσωπική ζωή κάποιου ανθρώπου ή κάποιο σκάνδαλο, καθώς επίσης ενδέχεται να εστιάζουν σε θέματα όπως οι συγκλονιστικές ιστορίες εγκλημάτων, η αστρολογία, τα κουτσομπολιά σχετικά με διασημότητες και οι ειδήσεις για τα πρόχειρα φαγητά.

- Hoaxes: Το hoaxing είναι ένας άλλος τύπος σκόπιμης κατασκευής ή παραποίησης ειδήσεων και εμφανίζεται συνήθως στα κοινωνικά δίκτυα. Η βασική διαφορά μεταξύ των Σοβαρών Κατηγοριών ψευδών ειδήσεων και των hoaxes, έγκειται στο ότι οι δεύτερες μπορεί να επικυρωθούν και να μεταδοθούν κατά λάθος ακόμα και από τα αναγνωρισμένα κανάλια ειδήσεων. Περαιτέρω, η διαφορά του hoax από μια απλή φάρσα διακρίνεται κατά Brunvand (1998) [13] με βάση τη φαινομενικά μεγάλη έκταση που λαμβάνει ,την πολυπλοκότητα του θέματος της, καθώς και το στοιχείο της παραπλάνησης, που περιλαμβάνει το οποίο και υπερβαίνει κάποιο χιουμοριστικό και αθώο κίνητρο, καθώς μπορεί να προκαλέσει υλικές ζημίες ή πνευματικές βλάβες στο θύμα.

- Κωμικές ή σατυρικές ειδήσεις: Υπάρχει διάκριση μεταξύ κατασκευασμένων ειδήσεων που εμπεριέχουν βαριούς ισχυρισμούς , από τις χιουμοριστικές ειδήσεις, καθώς στο πλαίσιο των δεύτερων οι αναγνώστες έχουν πλήρη επίγνωση της χιουμοριστικής πρόθεσης του συντάκτη τους και τις διαβάζουν όχι με σκοπό την πληροφόρησή τους, αλλά την τέρψη τους. Άλλωστε, υπάρχουν τεχνολογικά φίλτρα, τα οποία μπορούν να εντοπίσουν το χιούμορ, καθώς και τις πηγές προέλευσής τους, ώστε οι αναγνώστες να λάβουν σχετική προειδοποίηση . Σε αυτή την κατηγορία εντάσσονται και οι ειδήσεις σατυρικών ιστότοπων. Η παρωδία των ειδήσεων, αποτελεί συγκεκριμένο είδος ειδήσεων, το οποίο είναι προσφιλές σε πολλούς ιστοτόπους (π.χ. Το Κουλούρι [14]) . Το ειδησεογραφικό κείμενο με σκοπό την παρωδία, διατηρεί την τυπική και κρατούσα σύμφωνα με τους κανόνες της δημοσιογραφίας μορφή, πλην όμως περιστρέφεται γύρω από την ειρωνεία και το χιούμορ. Οι ιστότοποι που φιλοξενούν τέτοιου τύπου ειδήσεις, μιμούνται αξιόπιστες πηγές ειδήσεων και συχνά τυγχάνουν της ευρείας αποδοχής του αναγνωστικού κοινού, με αποτέλεσμα κάποιοι να πέφτουν θύματα φάρσας , όπως στην περίπτωση των δημοσιογράφων της ΕΡΤ.

2.2 Ψευδείς Ειδήσεις και Clickbaits

Στην προσπάθεια μας για την αντιμετώπιση του φαινομένου και προτού να προβούμε στην ανάλυση της υλοποίησης αλλά και των εργαλείων είναι σημαντικό εδώ να συσχετίσουμε τα clickbaits με τα fake news . Τα clickbaits είναι το δολώματα που προσελκύουν τους αναγνώστες σε άρθρα αμφιβόλου περιεχόμενος, που απώτερο σκοπό δεν έχουν μόνο το μεγάλο αριθμό από «κλικς» , αλλά και την αναδημοσίευση των αναληθειών ώστε να υποστηριχτούν τα κρυφά συμφέροντα των ανθρώπων που τα δημιουργούν.

Αξίζει να παρατηρήσουμε για την σοβαρότητα του φαινομένου, πως ενέχει κινδύνους διαστρέβλωσης των ειδήσεων με πολλές κοινωνικές και οικονομικές συνέπιες. Παράδειγμα αποτελεί το hoax που εμφανίστηκε το 2008 όπου διέδιδε ως είδηση ότι ο CEO της Apple, Steve Jobs, υπέστη καρδιακό επεισόδιο με αποτέλεσμα η μετοχή της εταιρείας να πέσει κατά 10%.

Οι μέθοδοι ανίχνευσης ψεύτικων ειδήσεων και clickbaits σε τίτλους ειδήσεων μπορεί να γίνει από την εμφάνιση απλών λέξεων αλλά και μέσω πιο σύνθετων γλωσσικών και γραμματικών δομών. Καθώς και μέσω κάποιου υβριδικού μοντέλου που συμπεριλαμβάνει και τα δύο.

Χρησιμοποιώντας Support Vector Machines (SVM), and Naive Bayes classifiers βασισμένους στο να μετράνε λέξεις κλειδιά από κάποια παραδείγματα αλιευμένα για training αυτοματοποιήσαμε το σύστημα μας ώστε να ανιχνεύει παραπληροφόρηση. Βάζοντας υψηλά βάρη σε λέξεις που εμφανίζονταν σε πρωτότυπα παραδείγματα clickbaits και χαμηλά βάρη σε λέξεις που δεν εμφανίζονταν καθόλου. Μετατρέποντας αυτά τα βάρη σε πιθανότητες και χρησιμοποιώντας την εξίσωση του Bayes ή την μέθοδο ελαχίστων τετραγώνων φτιάξαμε φίλτρα πιθανοτήτων.

Για την προσέγγιση της συντακτικής δομής βασιστήκαμε και στην εργασία των Niall J. Conroy, VictoriaL.Rubin and Yimin Chen [1] όπου είχαν επιτυχή αποτελέσματα σε μελέτες ανίχνευσης εξαπάτησης χτίζοντας δομές προτάσεων προς-τα-εμπρός αναφορών. Συνδυάζοντας έτσι μοντέλα επεξεργασίας φυσικής γλώσσας και αλγορίθμων ταξινόμησης.

Η στρατηγική clickbaits ενδείκνυται να μελετηθεί κατά το πρότυπο της εργασίας των Jonas Nygaard Blom και Kenneth Reinecke Hansen [3], δηλαδή μέσω των προς-τα-εμπρός αναφορών, όπου η επιτυχία των τίτλων κρίνεται εκ του αποτελέσματος, δηλαδή με μετάθεση

προς τα εμπρός (στο εξής forward-reference), προς το περιεχόμενο της είδησης την οποία συνοδεύουν

Ειδικότερα, στην ορθόδοξη δημοσιογραφία, οι συντάκτες των ειδήσεων γράφουν τίτλους για τα άρθρα τους,οι οποίοι έχουν ένα αυτοτελές και αυθύπαρκτο περιεχόμενο, το οποίο δύναται να διαβαστεί και να κατανοηθεί και μεμονωμένα πέραν της είδησης που συνοδεύει. Στον αντίποδα, στο πλαίσιο της στρατηγικής clickbaits, οι δημοσιογράφοι δομούν σκόπιμα τους τίτλους τους με ελλείψεις απαραίτητων συντακτικών εκφράσεων και παράλληλα τους εμπλουτίζουν με υπονοούμενες αναφορές του κειμένου που ακολουθεί. Ειδικότερα, οι παραπάνω συγγραφείς κατέληξαν σε έξι διαφορετικές εκφάνσεις του φαινομένου forward-reference όσον αφορά στο συντακτικό και στη δομή των προτάσεων τίτλων των ειδήσεων, επισημαίνοντας πως αυτό παρατηρείται περισσότερο σε ειδήσεις ελαφρού περιεχομένου, με θεματικές που αφορούν για παράδειγμα το lifestyle και τα gadgets και λιγότερο στις περιπτώσεις ειδήσεων για σημαντικότερα νέα όπως καιρικά φαινόμενα , αθλητικά και πολιτική. Αντίστοιχα, το clickbaits, παρατηρείται εντονότερα στο πλαίσιο ιστοσελίδων ,οι οποίες στηρίζονται στις διαφημίσεις, παρά σε συνδρομητικές ειδησεογραφικές ιστοσελίδες.

Πιο αναλυτικά, οι έξι εκφάνσεις του forward reference είναι οι ακόλουθες.

- Χρήση δεικτικών αντωνυμιών (this, that etc.)
- Χρήση προσωπικών αντωνυμιών (he,she etc.)
- Χρήση επιρρημάτων (how, when, where etc.)
- Χρήση οριστικών και αόριστων άρθρων (the, a, an)
- Έλλειψη ουσιαστικών επιχειρημάτων
- Ερωτήματα, που ωθούν στην εύρεση των απαντήσεών τους στο πλήρες κείμενο

Συστημικά, υπάρχουν πολλοί τρόποι ανίχνευσης της υποκρυπτόμενης στρατηγικής των clickbaits, όπως είναι για παράδειγμα η κατασκευή εργαλείων παρακολούθησης της συμπεριφοράς των χρηστών. Μέσω μιας τέτοιας ποσοτικής μελέτης, ανεξάρτητης του περιεχομένου στο οποίο συνοδεύει ο τίτλος, μπορούν να βγουν χρήσιμα συμπεράσματα για το είδος του τίτλου που διάβασε ο χρήστης. Οι Yimin Chen , Niall J. Conroy , Victoria L. Rubin [1] για παράδειγμα, στην δική τους μελέτη, παρακολούθησαν την κυκλοφορία των ειδήσεων, καθώς και τον χρόνο που οι χρήστες παρέμειναν στις ειδησεογραφικές ιστοσελίδες αφού έκαναν το κλικ σε κάποιον σύνδεσμο που βρήκαν στα κοινωνικά δίκτυα. Με βάση αυτά τα στοιχεία, οι συγγραφείς εξήγαγαν συμπεράσματα, τα οποία φανερώνουν το ενδιαφέρον των χρηστών, αξιοποιώντας δύο δείκτες: αφενός τον χρόνο που αφιέρωσαν οι χρήστες για να

διαβάσουν τα άρθρα, τα οποία επισκέφτηκαν και αφετέρου τις περαιτέρω δημοσιοποιήσεις που έκαναν, αλλά και τα σχόλια τους.

Η χρήση Probabilistic Context Free γραμματικών ενδείκνυται να επιτευχθεί μια βαθύτερη ανάλυση του λόγου. Ειδικότερα, στη συντακτική ανάλυση, οι προτάσεις μετασχηματίζονται σε ένα σύνολο κανόνων ανασύνταξης (parse trees) για να περιγράψουν τη δομή της σύνταξης. Για παράδειγμα φράσεις με έμφαση στο ουσιαστικό (υποκείμενο ή αντικείμενο) και φράσεις με έμφαση στο κεντρικό ρήμα, οι οποίες με τη σειρά τους ξαναγράφονται από τα συντακτικά συστατικά μέρη τους. Το τελικό σύνολο των αναδημιουργημένων προτάσεων γράφεται σε ένα δένδρο που τα φύλλα ή οι ακμές του ταυτίζονται με κάποια πιθανότητα. Αυτή η μέθοδος, χρησιμοποιείται για τη διάκριση των κανόνων κατηγοριοποίησης (λεξικοποιημένοι, μη λεξικοποιημένοι, γονικοί κόμβοι, κ.λπ.). Χρήσιμα εργαλεία εδώ πάνω στο deep syntax είναι το Stanford Parser και το AutoSlog-TS syntax analyzer.

2.3 Επεξεργασία Φυσικής Γλώσσας

Στο πλαίσιο της παρούσας εργασίας, οι τρόποι που θα μας απασχολήσουν για την ανίχνευση φαινομένων clickbaits, αφορούν μια αυτοματοποιημένη διαδικασία που θα σχετίζεται με την εμφάνιση λέξεων-κλειδιών αλλά και πιο σύνθετων γλωσσικών και γραμματικών δομών.

Θέτοντας λοιπόν υψηλά βάρη σε λέξεις που εμφανίζονται σε πρωτότυπα παραδείγματα clickbaits και χαμηλά βάρη σε λέξεις που δεν εμφανίζονται καθόλου, και μετατρέποντας συνακόλουθα αυτά τα βάρη σε πιθανότητες, θα μπορέσουμε να εκπαιδεύσουμε τους Naive Bayes ταξινομητές μας [1].

Η απλούστερη μέθοδος, που μπορεί να χρησιμοποιηθεί είναι η αναπαράσταση κειμένων με το λεγόμενο "bag of words", η οποία θεωρεί κάθε λέξη ως ενιαία, εξίσου σημαντική μονάδα. Σε αυτή τη μέθοδο οι συχνότητες λέξεων ή οι συχνότητες των "n-grams" (δέσμες λέξεων) συγκεντρώνονται και αναλύονται για να αποκαλύψουν τις παραπλανήσεις. Με αυτό τον τρόπο παίρνουμε σύνολα συχνοτήτων για την αποκάλυψη γλωσσικών δεξιοτήτων εξαπάτησης.

Ένας άλλος ενδιαφέρων τρόπος προσέγγισης είναι μέσω του μοντέλου αναπαράστασης των λέξεων σε πίνακες-διανύσματα που θεμελίωσε ο T. Mikolov. Με χρήση του word2vec οι Junfeng Fu, Liang Liang, Xin Zhou, Jinkun Zheng [5] στην δική τους εργασία προσπάθησαν να φέρουν καλά αποτελέσματα ανίχνευσης των click-baits με τις ήδη υπάρχουσες βιβλιοθήκες με τίτλους,οι οποίοι δεν εκφέρονται μόνο στα Αγγλικά, ώστε να επιτευχθεί μια γενική αντιμετώπιση χωρίς γλωσσικούς φραγμούς.

Οι ανωτέρω συγγραφείς, κατάφεραν να φτιάξουν έναν ταξινομητή που να έχει αρκετά καλύτερη επίδοση σχετικά με άλλους ήδη γνωστούς βασικούς ταξινομήτες (SVM, Decision Tree και Random Forest).,συγκρίνοντας τον δικό τους τρόπο με των άλλων σε τέσσερις (4) διαφορετικές μετρικές. Έφτιαξαν έτσι το μοντέλο τους χρησιμοποιώντας ως εισόδους σύνολο προτάσεων από δυο datasets διαφορετικών γλωσσών (αγγλικά και κινέζικα) από ισομοιρασμένες clickbait και non-clickbait επικεφαλίδες, διοχετεύοντας τες σε ένα CNN (συνελικτικό νευρωνικό δίκτυο) χρησιμοποιώντας φίλτρα από την εργασία του Y. Kim [12] για ταξινόμηση προτάσεων. Η χρήση του πλαισίου CNN για ανίχνευση click-baits, έδειξε μετά τα εμπειρικά πειράματα των συγγραφέων ότι το μοντέλο μπορεί να καταγράψει καλύτερα το συντακτικό και τις σημασιολογικές σχέσεις μεταξύ των λέξεων σε παγκόσμιο επίπεδο, αφού λειτούργησε καλά σε όλες τις γλώσσες χωρίς να βασίζεται σε συγκεκριμένες γλωσσικά χαρακτηριστικά.

2.4 Ταξινόμηση

Για τη διαδικασία της ταξινόμησης θα γίνει numeric clustering, χρησιμοποιώντας αυτοματοποιημένες αριθμητικές διαδικασίες, εκπαιδεύοντας μηχανές SVM και τους Naïve Bayes αλγορίθμους, με εισόδους τα σύνολα των συχνοτήτων εμφάνισης λέξεων-κλειδιών. Όταν το μαθηματικό μοντέλο είναι επαρκώς εκπαιδευμένο από προ-κωδικοποιημένα παραδείγματα μπορεί να προβλέψει περιπτώσεις της μελλοντικής εξαπάτησης με βάση την αριθμητική ομαδοποίηση και τις αποστάσεις των σημείων πάνω σε καρτεσιανό επίπεδο.

Οι αλγόριθμοι Naïve Bayes χρησιμοποιήθηκαν ώστε να γίνουν ταξινομήσεις βασισμένες σε συσσωρευμένες αποδείξεις της συσχέτισης μεταξύ μιας δεδομένης μεταβλητής (π.χ. σύνταξη) και των άλλων μεταβλητών που ενυπάρχουν στο μοντέλο. Για την συναισθηματική ανάλυση η ταξινόμηση βασίζεται στην διαίσθηση ότι «οι απατεώνες» χρησιμοποιούν ακούσια συναισθηματική φόρτιση. Έτσι, εμφανίζονται συντακτικά μοτίβα (patterns), διαχωρίζοντας το συναίσθημα από τα επιχειρήματα, με βάση τα γεγονότα, συσχετίζοντας τον τρόπο εκμάθησης των τάξεων με βάση το στυλ επιχειρηματολογίας.

2.5 Ανάλυση Συναισθημάτων

Η ανίχνευση της πλάνης, της εξαπάτησης, της προπαγάνδας ή ακόμη του χιούμορ είναι μια λεπτή διαδικασία, που στην εργασία μας βασίζεται σε κείμενα συντακτών, οι οποίοι προσπαθούν να κρύψουν όσο καλύτερα γίνεται τους πραγματικούς σκοπούς τους. Τούτου δοθέντος μια αυτοματοποιημένη διαδικασία, η οποία π θα λειτουργεί έξυπνα, θα είναι αυτή που θα ανακαλύπτει τα πραγματικά συναισθήματα των αρθρογράφων, ώστε να μας οδηγήσει σε κατάλληλα συμπεράσματα. Ο κλάδος που ασχολείται με αυτήν ακριβώς την αποκωδικοποίηση των συναισθημάτων κάποιου ανθρώπου με βάση τον γραπτό του λόγο λέγεται συναισθηματική ανάλυση (Sentiment Analysis).

Στη μελέτη τους, οι Niall J. Conroy, Victoria L. Rubin, and YiminChen Automatic Deception Detection: Methods for Finding Fake News [2] επιχειρούν μία Συναισθηματική Ανάλυση μέσω του ελέγχου του βαθμού της συμβατότητας της προσωπικής εμπειρίας του αρθρογράφου σε σύγκριση με «προφιλ» εμπειριών, εξαγόμενων από σχετικά δείγματα. Στόχος εδώ είναι να αναγνωριστεί ο αρθρογράφος που εξαπατά πέφτοντας σε αντιφάσεις ή παραλήψεις γεγονότων. Το μοντέλο αυτό έχει καλή εφαρμογή, κυρίως σε ψεύτικα κρητικές προϊόντων , καθώς οι χρήστες των αληθινών κριτικών τείνουν να κάνουν παρόμοια σχόλια σχετικά με κάποιο χαρακτηριστικό του προϊόντος, ενώ οι κακοήθεις σχολιαστές γράφουν με αρκετά διαφορετικό ύφος. Έτσι τελικά, εξάγουν συμπεράσματα για την αξιοπιστία της άποψης με βάση τον βαθμό συμβατότητας με τα άλλα δείγματα. Όπως επισημαίνουν και οι συγγραφείς, αυτός ο τρόπος μπορεί να συμβάλει θετικά στην πρόβλεψη του ψεύδους, αλλά χρειάζεται πολλά δείγματα προφίλ για να λειτουργήσει αποδοτικά.

Οι παραπάνω συγγραφείς έκαναν ανάλυση συναισθημάτων σε διάφορους χρήστες του twitter μέσω ενός πρωτότυπου μοντέλου, που εστιάζει στην αποκάλυψη του είδους των απόψεων τους. Έτσι ταξινόμησαν τις απόψεις με βάση την αντικειμενικότητα τους ή το πόσο πολωμένες αυτές ήταν . Έκαναν συναισθηματική ανάλυση των χρηστών χωρίζοντας τα στοιχεία (features) που βρήκαν από άλλες μελέτες σε τρείς διαφορετικές σκοπιές:

- Το Polarity : η σκοπιά αυτή έχει να κάνει με το είδος της συναισθηματικής φόρτισης κάποιας συγκεκριμένης άποψης. Οι μέθοδοι που είναι προσανατολισμένες προς αυτήν την σκοπιά επιστρέφουν κάποια ποιοτική μεταβλητή που παίρνει τρείς πιθανές τιμές “θετική” ,”αρνητική ” και “ουδέτερη”.

- Το Strenght: αυτή η σκοπιά χρησιμοποιεί μεθόδους και πόρους που παρέχουν πληροφορίες με βάση τα επίπεδα φόρτισης. Τα features εδώ επιστρέφουν αριθμητικές τιμές σχετικές με την βαρύτητα αυτής της πόλωσης.

- Το Emotion: από αυτή τη σκοπιά οι συγγραφείς απέσπασαν πληροφορίες για την φύση του συναισθήματος και τη διάθεση των χρηστών. Οι μέθοδοι που χρησιμοποιήθηκαν επέστρεψαν ποιοτικές μεταβλητές που έχουν να κάνουν με τα συναισθήματα (βλ.Plutchik's emotion wheel.)

Οι συγγραφείς για την δική τους προσέγγιση χρησιμοποίησαν στοιχεία ανοιχτού κώδικα που έχουν χρησιμοποιηθεί και στο παρελθόν σε μελέτες. Τα στοιχεία αυτά τα ονόμασαν meta-level features. Κάποια από αυτά είναι το Opinion Finder Lexicon, το Bing Liu'sopinion Lexicon , το AFINN lexicon, το SentiWordNet lexicon, το Sentiment140 lexicon, το NRC-emotion , το SentiStrenght και το SenticNet. Είναι σημαντικό να παρατηρήσουμε εδώ, πως κάποια από αυτά τα λεξικά δημιουργήθηκαν πλήρως αυτοματοποιημένα, ενώ κάποια άλλα πήραν τιμές για τις λέξεις τους με βάση την ανθρώπινη κρίση με μεθόδους όπως το crowdsourcing. Δυο παραδείγματα είναι το SentiStrength που υπολογίζει θετικά και αρνητικά βάρη σε συναισθηματικό επίπεδο και το NRC word-emotion association lexicon, όπου ο δημιουργός του έβαλε ταμπέλες σε ένα πλήθος λέξεων σύμφωνα με τις κατηγορίες συναισθημάτων κατά Plutchik [7].

Ο τροχός συναισθημάτων κατα Plutchik [7], είναι μια σύνθεση από τέσσερα ζεύγη συναισθημάτων: χαρά-εμπιστοσύνη, θλίψη-θυμός, έκπληξη-φόβος, προσδοκία-απέχθεια. Με βάση αυτόν τον διαχωρισμό υπολογίστηκε η μοναδικότητα, δηλαδή το σύνολο των μοναδικών λέξεων που εμφανίζονται σε κάθε λεξικό ως προς το σύνολο των λέξεων που εμφανίζονται σε όλα τα λεξικά της ίδιας κατηγορίας. Επίσης υπολογίστηκε η ουδετερότητα , δηλαδή το σύνολο των λέξεων που έχουν ουδέτερη χροιά (δεν εμπεριέχουν δηλαδή κάποια σημαντική πληροφορία από συναισθηματικής απόψεως) ως προς το σύνολο των λέξεων του κάθε λεξικού. Τέλος, υπολογίστηκε ο βαθμός στον οποίο τα λεξικά που χρησιμοποιήθηκαν συμφωνούν μεταξύ τους, συγκρίνοντας τις λέξεις που έχουν από κοινού εμφάνιση.

Από το Stanford Twitter Sentiment οι συγγραφείς χρησιμοποίησαν το σύνολο δεδομένων το οποίο έχει ένα tag για κάθε tweet με τιμή θετική, αρνητική ή ουδέτερη, που υποδηλώνει την χροιά της κάθε γνώμης. Από αυτά θεώρησαν ως υποκειμενικά τα τουιτς που είχαν θετική ή αρνητική χροιά, ενώ ως αντικειμενικά τα τουιτς που είχαν ουδέτερη χροιά. Εφάρμοσαν αυτή τη διαδικασία για να βγάλουν συμπεράσματα για την

αντικειμενικότητα/υποκειμενικότητα των απόψεων ενώ για την πόλωση δεν συμπεριέλαβαν καθόλου τα τουιτς με ουδέτερα tags. Έτσι, από κάθε τουιτ υπολόγισαν το information gain για κάθε feature από τις τρεις διαφορετικές σκοπιές με σκοπό τους να εξετάσουν πια features μειώνουν περισσότερο την εντροπία του δείγματος. Παρατήρησαν έτσι, ότι τα polarity-based features δίνουν καλύτερη πληροφορία αφού πετυχαίνουν τα υψηλότερα best splits στο δένδρο των αποφάσεων. Τέλος έκαναν ταξινόμηση χρησιμοποιώντας αλγορίθμους όπως CART, J48, NaïveBayes και SVMs [15], αλλάζοντας συνεχώς τις παραμέτρους του κάθε αλγορίθμου ώστε να πετύχουν το καλύτερο καλιμπράρισμα. Μια από τις παρατηρήσεις τους είναι πως ο αλγόριθμός τους δεν λειτουργούσε σωστά σε περιπτώσεις τουιτς, τα οποία είχαν ανάμεικτα θετικά και αρνητικά συναισθήματα, καθώς αυτά ταξινομούνταν ως ουδέτερα πολωμένα. Επίσης παρατήρησαν πως ο SVM ξεπερνά σε απόδοση τους άλλους αλγορίθμους συγκρίνοντας τα αποτελέσματα της F1.

3 *Σύγχρονες Μέθοδοι και Προσεγγίσεις*

Ο Ricky J. Sethi με την εργασία του με τίτλο Spotting Fake News: A Social Argumentation Framework for Scrutinizing Alternative Facts [4] στη δική του προσπάθεια να εντοπίσει fake news έφτιαξε μια πλατφόρμα που βασίζεται στην αλληλεπίδραση των χρηστών πάνω στην ταυτοποίηση μη αξιόπιστων πληροφοριών των alternatives facts όπου είναι δυναμικές πηγές hoaxes.Το μοντέλο του στηρίζεται στη δημιουργία ενός graph-theoretic framework, δηλαδή ενός γράφου που καταχωρεί κάθε νέο επιχείρημα που χρήζει εξέτασης. Ενώ το αλγοριθμικό κομμάτι έχει μικρό ρόλο στην εξαγωγή τελικών συμπερασμάτων και δεν βασίζεται σε κάποια εντελώς αυτοματοποιημένη διαδικασία, το μοντέλο αξίζει να εξεταστεί αφού η δημιουργία μίας κοινωνίας, όπου τα μέλη της θα είναι άνθρωποι που θα βοηθούν στην αντιμετώπιση του φαινομένου δοκιμάζεται όλο και περισσότερο στις μέρες μας.

- Για την δημιουργία του γράφου του ο Ricky J. Sethi [4], όρισε πως το επιχείρημα δομείται από Θέσεις (Stances), Ισχυρισμούς (Claims) και αποδεικτικά Στοιχεία (Evidence), όπου τα αποδεικτικά Στοιχεία και οι Ισχυρισμοί υποστηρίζονται από κάποιες διαδικτυακές πηγές, ενώ οι Θέσεις είναι αμοιβαία αποκλειόμενα ενδεχόμενα που αποτελούνται από στοιχεία ατομικής επιχειρηματολογίας. Ένας Ισχυρισμός μπορεί να υποστηρίζεται από κάποια πηγή ακόμα και όταν δεν υπάρχει κάποιο Στοιχείο συνδεδεμένο στον υπό-γράφο, αν συνδεθούν πολλά Αποδεικτικά Στοιχεία για τον συγκεκριμένο Ισχυρισμό τότε οι πηγές του Ισχυρισμού ενώνονται με τους κόμβους των Αποδεικτικών Στοιχείων.Η όλη πλατφόρμα βασίζεται στην διαδικτυακή κοινότητα, δηλαδή στα μέλη που την χρησιμοποιούν, τους χρήστες, (όπου ο καθένας έχει μια συνάρτηση σχετική με το πόσο ενεργός είναι και τα ratings που έχει πάρει από άλλους χρήστες, και αυτό παίζει ρόλο στα βάρη του γράφου που είναι σχετικά με τις Θέσεις του, τους responders που συνδέουν τις θέσεις με Ισχυρισμούς ή Στοιχεία και τους moderators.

- Σε μια άλλη εργασία με τίτλο AutomaticDeceptionDetection: Methods for Finding Fake News οι συγγραφείς Niall J. Conroy, Victoria L. Rubin, and Yimin Chen [2] σε μια άλλη προσέγγιση, πέρα από τη γνωστή πλέον γλωσσική ανάλυση, στην οποία ήδη αναφερθήκαμε παραπάνω, διαπίστωσαν πως ένας τρόπος αντιμετώπισης του φαινομένου είναι με βάση το network analysis. Πρόκειται για μία αρκετά εξειδικευμένη μέθοδο, μιας και αφορά τα fake news που έχουν ένα αληθινό μέρος (π.χ. ο Αλέξης Τσίπρας γεννήθηκε στην Αθήνα), αλλά συνδέονται με κάποια ψευδή στοιχεία (π.χ. ο Αλέξης Τσίπρας γεννήθηκε στην Αθήνα και έχει εβραϊκή καταγωγή), ώστε να δώσουν μεγαλύτερη βαρύτητα στην επιχειρηματολογία. Αυτή η

μέθοδος εξάγει συμπεράσματα για το ποσοστό αλήθειας με βάση τα structured data (δομημένα δεδομένα) που μπορεί να βρει κανείς από έμπιστες πηγές του διαδικτύου.

- Οι Zhiwei Jin, Juan Cao, Yu-Gang Jiang, Yong dong Zhang στην δική τους εργασία με τίτλο News Credibility Evaluation on Microblogwith a Hierarchical Propagation Model [6] αξιολογούν την αξιοπιστία των ειδήσεων χρησιμοποιώντας το μοντέλο του hierarchical propagation. Οι συγγραφείς εδώ συλλέξαν δύο (2) dataset από το Sina Weibo (το κινέζικο twitter). Το ένα dataset αφορούσε ειδήσεις αποκλειστικά για το θέμα "Flight MH370 LostContact", ενώ το άλλο που είχε σχεδόν ισάριθμα δείγματα αφορούσε γενικές ειδήσεις του 2013 ανεξαρτήτου θέματος. Για το πρώτο dataset, οι συγγραφείς χρησιμοποίησαν το rumor reporting service του SinaWeibo καθώς και αξιόπιστες δημοσιογραφικές πηγές για να κατηγοριοποιήσουν τις ειδήσεις σε fake ή όχι, ενώ για το δεύτερο χρησιμοποίησαν ειδήσεις από τα τοπ 10, fake news sites , με βάση λίστα ενός σχετικού κινέζικου πρακτορείου. Χρησιμοποίησαν δύο (2) datasets, καθώς διαχωρίζουν τα νέα που δημοσιεύονται στα microblogs ως topic-independent news και topic-relatednews.

Περαιτέρω, οι συγγραφείς έχτισαν το hierarchical credibility propagation χρησιμοποιώντας τρία (3) επίπεδα, ώστε να το αντιμετωπίσουν ως ένα γράφο χρησιμοποιώντας γνωστές μεθόδους βελτιστοποίησης. Στα επίπεδα τηρείται η ιεραρχία και με κορυφή της πυραμίδας το ίδιο συμβάν που μας ενδιαφέρει έχουμε:

1. Event: περιστατικό που συμβαίνει σε συγκεκριμένο τόπο και χρόνο. Το σύνολο των μηνυμάτων στο microblog που εμπεριέχουν συγκεκριμένες λέξεις κλειδιά για ένα συγκεκριμένο χρονικό διάστημα.

2. Sub-event: δευτερεύον συμβάν είναι το σύνολο των μηνυμάτων που αφορούν ένα συγκεκριμένο περιστατικό , αλλά αποκλίνουν από την κεντρική ιστορία του περιστατικού και εκφράζουν διαφορετικές σκοπιές, αμφιλεγόμενες απόψεις και άλλες παρασκηνιακές ιστορίες (τα εντοπίζουν με single-passincremental clustering) . Το επίπεδο αυτό υπάρχει ώστε να γίνει η συναισθηματική ανάλυση των δεδομένων.

3. Message: είναι το σύνολο των δεδομένων που εμπεριέχονται σε κάποιο post (κείμενο, hashtags , urllinksκ.λ.π.), καθώς και τα δεδομένα που μπορεί να συλλέξει κανείς από αυτό (ώρα δημοσίευσης, αριθμός αναδημοσιεύσεων, αριθμός σχολίων, αριθμός followers και followees του συγγραφέα). Το επίπεδο αυτό υπάρχει ώστε αφού γίνει η κατάλληλη ομαδοποίηση, να μειωθεί ο θόρυβος της πληροφορίας.

Την αξιοπιστία (Credibility), την διαχειρίζονται ως μια αριθμητική μεταβλητή με τιμές να ανήκουν στο [-1,1]. Η αξιοπιστία του συμβάντος είναι μια συνάρτηση της αξιοπιστίας των δευτερεύοντος συμβάντων ,τα οποία με την σειρά τους έχουν σχετική αξιοπιστία με αυτήν των μηνυμάτων που τα εκφράζουν, στα οποία τέλος έχουν δοθεί τιμές αξιοπιστίας με βάση την ταξινόμηση που τους έχει γίνει και τα χαρακτηριστικά τους.

Η διαδικασία για τη δημιουργία των sub-events γίνεται με χρήση του αλγορίθμου Single-pass incremental clustering όπου περνώντας ένα-ένα τα μηνύματα ως εισόδους, ο αλγόριθμος δημιουργεί σταδιακά τις διαφορετικές κατηγορίες με βάση την ομοιότητα τους αν αυτή ξεπερνά το threshold μιας default μεταβλητής. Αλλάζοντας αυτό το κατώφλι επηρεάζονται οι τελικές κατηγορίες,δηλαδή τα sub-events.

Με αυτόν τον τρόπο στους γράφους που δημιουργούνται, η κάθε είδηση μπορεί να αναπαρασταθεί με τρία διαφορετικά επίπεδα και να εξαχθούν συμπεράσματα για το κάθε ένα ξεχωριστά. Στη δημιουργία των γράφων σημαντικό ρόλο παίζουν και τα τέσσερα (4) είδη ακμών, αφενός οι δύο (2) που ενώνουν τους κόμβους των διαφορετικών επιπέδων και αφετέρου οι άλλες 2 που ενώνουν τους κόμβους των 2 πρώτων επιπέδων (messages,sub-events). Οι ακμές έχουν βάρη και οι συγγραφείς κάνουν την παραδοχή πριν τον υπολογισμό τους,ότι παρόμοια μηνύματα έχουν ίσες τιμές αξιοπιστίας.

Τα βάρη των ακμών μεταξύ κόμβου μηνύματος σε κόμβο μηνύματος είναι συμμετρικά και αφορούν την επιρροή του ενός μηνύματος στο άλλο. Υπολογίζουν την ομοιότητα των μηνυμάτων ανά δύο (2) χρησιμοποιώντας NLP γνωστές τεχνικές .
Τα βάρη των ακμών μεταξύ κόμβου sub-event σε κόμβο sub-event έχουν τα ίδια δεδομένα με το προηγούμενο είδος και υπολογίζονται, όχι με βάση την ομοιότητα, όπως πριν καθώς ως διαφορετικές κατηγορίες είναι λογικό να έχουν την μικρότερη δυνατή ομοιότητα αλλά με βάση την αναπαράσταση του centroid for a cluster που υπολογίζεται βάση του tfscore (μετρική δλδ που έχει να κάνει με την συχνότητα εμφάνισης λέξεων-κλειδιών ανά μήνυμα).
Τα βάρη των ακμών μεταξύ κόμβου μηνύματος σε κόμβο sub-event έχουν να κάνουν με το βαθμό που το κάθε μήνυμα επηρεάζει την κεντρική ιδέα της κατηγορίας. Όσο περισσότερο σχετίζεται το περιεχόμενο στον wordvector του μηνύματος με το περιεχόμενο στο centroid του sub-event τόσο μεγαλύτερο το βάρος της ακμής.

Τα βάρη των ακμών μεταξύ κόμβου sub-event σε κόμβο συμβάντος υπολογίζονται με παρόμοιο τρόπο με τους παραπάνω δύο χρησιμοποιώντας τα centroids του event και των sub-events καθώς και τον βαθμό επιρροής, και μια παράμετρο (λ) για να ισορροπήσουν το άθροισμα μεταξύ των δυο τρόπων.

Αφού κάνουν την ίδια διαδικασία για κάθε είδηση και δημιουργήσουν ένα ξεχωριστό γράφο για κάθε μια προχωρούν στην εκμάθηση του ταξινομητή τους. Η ταξινόμηση γίνεται στο επίπεδο των μηνυμάτων και τα αποτελέσματα της δίνουν μια τιμή αξιοπιστίας για καθένα ξεχωριστά , η τιμή της αξιοπιστίας του συμβάντος θα βγει τελικά από τις όλες συνδέσεις.
Στο paper αυτό κάνουν τέλος και ένα optimization χρησιμοποιώντας έναν αλγόριθμο ανατροφοδότησης με διαφορετικά είδη ταξινόμησης κάθε φορά , αλλά κεντρική ιδέα αντιμετώπισης αυτή των γράφων.

Ένα άλλο εργαλείο που χρησιμοποίησαν οι συγγραφείς στην μελέτη τους είναι η συμπεριφορά των καθαυτών χρηστών απέναντι στις ειδήσεις . Είναι μια ποσοτική μελέτη ανεξάρτητη περιεχομένου βάση της οποίας παρακολούθησαν την κυκλοφορία των ειδήσεων καθώς και τον χρόνο που οι χρήστες παρέμειναν στις ειδησεογραφικές ιστοσελίδες αφού έκαναν το κλικ σε κάποιον σύνδεσμο που βρήκαν στα κοινωνικά δίκτυα. Δυο είναι οι δείκτες που φανερώνουν το ενδιαφέρον των χρηστών ,πρώτον ο χρόνος που αφιέρωσαν για να διαβάσουν το άρθρο και δεύτερον οι δημοσιοποιήσεις που έκαναν καθώς και τα σχόλια τους.

Κάτι τέτοιο θα μπορούσε να διευρύνει την εργασία μας , με περισσότερα χαρακτηριστικά στοιχεία για την ανάπτυξη του συστήματος που προτείναμε εμείς.

4 Εύρεση και διαμόρφωση των Δεδομένων

4.1 Δεδομένα στο διαδίκτυο

Για την διαδικασία συλλογής των δεδομένων που θα βοηθούσαν την εκμάθηση των αλγορίθμων μας υπήρξαν πολλές σκέψεις. Μια από αυτές ήταν η εξαγορά έτοιμων dataset από ιδιωτικές εταιρείες όπως την Grimson Hexagon Database ή από εταιρίες που ο ενδιαφερόμενος ζητά να του φτιάξουν datasets με χρήση της διαδικασίας του πληθοπορισμού (crowdsourcing). Δηλαδή της πράξης της εξωτερικής ανάθεσης καθηκόντων, σε μια μεγάλη ομάδα εθελοντών ή μία κοινότητα, μέσω ανοικτής πρόσκλησης. Μια τέτοια πλατφόρμα είναι το MechanicalTurk της Amazon ή το Crowd Flower της www.figure-eight.com.

Η μέθοδος κατηγοριοποίησης με ανάθεση σε μεγάλο κοινό είναι η πιο συνήθης και αξιόπιστη διαδικασία. Ωστόσο κάτι τέτοιο δεν μπορούσε να πραγματοποιηθεί στο πλαίσιο της παρούσας εργασίας.

4.2 Πρώτο πακέτο δεδομένων

Ψάχνοντας στο διαδίκτυο και μετά από έλεγχο της ορθότητας πολλών δειγμάτων ,το dataset που πάρθηκε για τον έλεγχο των αλγορίθμων μας ήταν αυτό από το https://www.clickbait-challenge.org/ ενός workshop που πραγματοποιήθηκε το φθινόπωρο του 2017 στο Bauhaus-Universität Weimar της Γερμανίας. Όπου ακολουθώντας το σαιτ κανείς μπορεί να βρει και τα αποτελέσματα άλλων μελετητών πάνω στην ίδια πρόκληση.

Ο λόγος λοιπόν που χρησιμοποιήσαμε αυτό το δείγμα ήταν γιατί η μέθοδος κατηγοριοποίησης με ανάθεση σε μεγάλο κοινό είναι η πιο συνήθης και αξιόπιστη διαδικασία. Ωστόσο κάτι τέτοιο δεν μπορούσε να πραγματοποιηθεί στο πλαίσιο της παρούσας εργασίας.

Το συγκεκριμένο dataset όπως διαβάζουμε από το description του δείγματος περιέχει δυο αρχεία . Το ένα είναι τα instances και το άλλο το truthmap.

Το αρχείο instances είναι στην ουσία τα posts χρηστών που έχουν δημοσιευτεί στα κοινωνικά δίκτυα και έχει την εξής δομή:

1. id: ο μοναδικός αριθμός ταυτοποίησης του κάθε ποστ
2. postMedia: το αρχείο εικόνας που συνόδευε το ποστ (στην περίπτωση μας δεν μας ενδιαφέρει)
3. postText: το σχόλιο του χρήστη που έκανε τη δημοσίευση χωρίς τα λινκς
4. postTimestamp: η ημερομηνία και ώρα δημοσίευσης
5. targetCaptions: η λεζάντα που συνόδευε την φωτογραφία
6. targetDescription: το description tag του άρθρου
7. targetKeywords: το keyword tag του άρθρου
8. targetParagraphs: το κείμενο του άρθρου
9. targetTitle: ο τίτλος του άρθρου

Το αρχείο με τις τιμές αληθείας , περιέχει την κλάση κάθε τίτλου από τα instances(clickbait ή not-clickbait). Ο λόγος όμως που διαλέξαμε αυτό το δείγμα είναι ο εξής. Η κλάση του κάθε τίτλου έχει δοθεί στην κρίση τουλάχιστον 5 σχολιαστών με την διαδικασία του crowd-sourcing. Οι σχολιαστές έδωσαν μια τιμή από τις 4 πιθανές σχετικά με το κατά πόσο ο τίτλος που διάβαζαν ήταν προϊόν click-bait οι πιθανές επιλογές τους ήταν:

0.0 : για non-clicbait τίτλους
0.33 : για slightly clickbait τίτλους
0.66 : για considerably click-bait τίτλους
1 : για heavily click-bait τίτλους
ο πίνακας των κρίσεων υπάρχει στο στοιχείο truthJudgments για το κάθε id

Τα υπόλοιπα στοιχεία του αρχείου truth είναι οι εξής μετρήσεις που έγιναν σχετικά με τις κρίσεις που δόθηκαν:

Mean

mean example για totruth_Y[1]:

truthMean[1]= 0.0+0.6666667+0.0+0.33333334+0.05≈0.20.0+0.6666667+0.0+0.333 33334+0.05≈0.2

Median

median example για totruth_Y[0]:

0.0 , 0.0 , **0.0** , 0.33333334 , 0.6666667

η τιμή που χωρίζει το δείγμα τον 5 κρίσεων στα 2!

truthMeadian[0] = 0.0

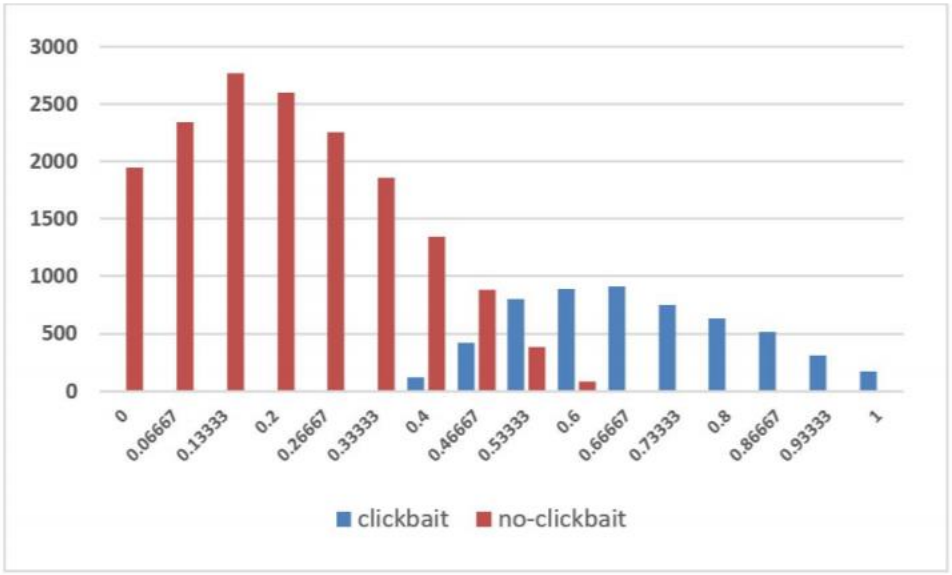
Mode

mode example γιατοtruth_Y[3]:

1.0 , 0.0 , **0.33333334** , **0.33333334** , 0.6666667

η τιμή που εμφανίζεται πιο συχνά στο δείγμα στην περίπτωση που όλες οι κρίσεις είναι διαφορετικές το mode είναι 0

truthMode[3] = 0.33333334



Σχήμα: Κατανομή των clickbait και non-clickbait τίτλων με βάση το Mean Judgment

4.3 Δεύτερο πακέτο δεδομένων

Μια άλλη ιδέα θα ήταν η άντληση δεδομένων από τα ίδια τα site κοινωνικών δικτύων , αφού δίνουν ελεύθερη πρόσβαση στον καθένα με χρήση των API τους. Τα θετικά σε αυτήν την περίπτωση είναι πολλά αφού πρώτων τα δεδομένα είναι ελεύθερα και δεύτερον αποτελούν τις

κύριες πηγές εξάπλωσης του φαινομένου που αναλύουμε . Παρ' όλα αυτά υπήρξε ένα σημαντικό εμπόδιο που έκανε την αυτοματοποιημένη δημιουργία των dataset από τα social media (για το πρώτο στάδιο της εκμάθησης) δύσκολη ώστε να μας είναι απαραίτητο και το δείγμα του πρώτου πακέτου . Το αρνητικό αυτό είναι το γεγονός ότι οι εταιρίες όπως το Twitter θέτουν σημαντικούς περιορισμούς τα τελευταία χρόνια στους χρήστες για το πλήθος των δεδομένων που κατεβάζουν μέσω του API, οπότε δεν θα ήταν δυνατή η αδιάκοπη άντληση δεδομένων ακόμα και με πολλαπλούς λογαριασμούς. Το εμπόδιο αυτό έρχεται σε συνδυασμό με τη δυσκολία στο να ξεχωρίσουμε τους λογαριασμούς που αναπαράγουν αληθείς από αυτούς που αναπαράγουν ψευδείς ειδήσεις .

Η διαδικασία αυτή δεν είναι καθόλου ξεκάθαρη ειδικά στην εποχή μας που το φαινόμενο των fake news έχει πάρει τόσο μεγάλες διαστάσεις και οι εταιρείες λαμβάνουν σημαντικά μέτρα για την αντιμετώπιση του. Οι λογαριασμοί που αναπαράγουν fake news έχουν ως στόχο να κάνουν ζημία χωρίς να αποκαλυφθεί η ταυτότητα τους άρα και είναι λογαριασμοί σχεδόν άγνωστοι με λίγους ακόλουθους (followers) είτε κατάφεραν να γίνουν γνωστοί με αποτέλεσμα να τους γίνουν αναφορές από ανθρώπους που καταπολεμούν το φαινόμενο και να κλείσουν μιας και παραβιάζουν τους όρους χρήσης της ιστοσελίδας.

Η δική μας προσέγγιση, ήταν να φτιάξουμε ένα πρόγραμμα αυτόματης συλλογής δεδομένων με την χρήση των API που μας πρόσφερε το Twitter, όπου οι χρήστες εμπιστεύονται περισσότερο για την ενημέρωση τους. Ο περιορισμός του Twitter λοιπόν σε νούμερα είναι πως όσο κατεβάζεις με το API δεν μπορείς να ξεπεράσεις τον αριθμό των 200 tweets ανά λογαριασμό.

Τα κανάλια που επιλέξαμε δεν ήταν τυχαία αλλά βρέθηκαν από πηγές όπως το Wikipedia για την αξιοπιστία τους . Έτσι καταλήξαμε να μελετήσουμε 20 ειδησεογραφικά κανάλια , 10 για clickbait ειδήσεις και άλλα 10 για not-clickbait. Παίρνοντας 200 tweets απο το κάθε ένα σύμφωνα με τους περιορισμούς του API.

Αποφασίσαμε να έχουμε μοιρασμένο δείγμα και έτσι κατατλήξαμε στα παρακάτω κανάλια.

{Clickbait :}

UNILAD , BUZZFEED , PUREWOW, UPWORTHY , DIGG , FARK , CRACKED , MASHABLE , CLICKHOLE , HUFFIGTONPOST

{Not-Clickbait :}
WASHIGTONPOST , WALL STREET JOURNAL , BBC , TheEconomist , USATODAY , REUTERS , THE ASSOCIATED PRESS , POLITICO , BLOOMBERG , FINANCIAL TIMES.

Οπως είπαμε παραπάνω διαλέξαμε τα clickbait κανάλια με βάση πληροφορίες που πήραμε από το wikipedia, αλλά διασταυρώνοντας τις πληροφορίες και από άρθρα στο ιντερνέτ όπου οι χρήστες έχουν δηλώσει την δυσαρέσκεια τους ως προς την μη αξιοπιστία των παραπάνω σαιτς και κατ' επέκταση των λογαριασμών τους στο Twitter.

Για τα not-clicbait κανάλια η διαδικασία ήταν διαφορετική μιας και ήταν εύκολο να βρούμε αξιόπιστες πηγές και πρακτορεία ειδήσεων που εμπιστεύεται ο κόσμος. Κρατήσαμε 10 που με βάση την κρίση μας και την πολιτική που ακολουθούν είχαν ανομοιομορφίες μεταξύ τους. Κάποια μεταδίδουν καθαρά πολιτικές ειδήσεις , άλλα οικονομικές , κάποια είναι μεροληπτούν (biased) αφού γράφουν δημοσιογράφοι με «αριστερές» ιδεολογίες και κάποια που γράφουν με πιο «δεξιές» , παρόλα αυτά εξακολουθούν να είναι not-clickbait μιας και στην δική μας εργασία αυτό είναι το αντικείμενο που μελετάμε.
Με όλα τα παραπάνω να είναι γνωστά σε 3 διαφορετικές περιόδους του χρόνου αντλήσαμε 12.000 διαφορετικά tweets εκ των οποίων κρατήσαμε τα 4000 τυχαία. 200 από το κάθε ειδησεογραφικό κανάλι που στοχεύσαμε.

4.4 Αποθήκευση δεδομένων

Η διαδικασία της αποθήκευσης των δεδομένων έγινε με χρήση βάσης δεδομένων. Εδώ για λόγους ευκολίας χρησιμοποιήσαμε No-SQL βάση μιας και ο κώδικάς ήταν γραμμένος σε python αυτή που μας βόλεψε ήταν η MongoDB . Αφού την συνδέσαμε με το API του twitter δεν αποθηκεύσαμε μόνο τα δεδομένα αλλά τα περάσαμε και από ένα process data cleansing.
Αυτό γιατί χρειαζόμασταν απλούς τίτλους ειδήσεων και η ανομοιομορφία των tweets από το κάθε ειδησεογραφικό κανάλι μας εμπόδιζε να πάρουμε τα δεδομένα με την μορφή που θέλαμε. Βρίσκοντας ομοιότητες σε κάποια κανάλια αυτοματοποιήσαμε την διαδικασία με χρήση "φλαγκς" στον κώδικά μας. Παραδείγματα τέτοιας ανομοιομορφίας ήταν για παράδειγμα που βρισκόταν ο τίτλος της είδησης. Κάποια κανάλια είχαν τον τίτλο στο κείμενο του tweet μαζί

με κάποια άλλα σχόλια που αφαιρούσαμε ,σε κάποια άλλα κανάλια ο τίτλος ήταν "κρυμμένος" στον υπερσύνδεσμο.

Κατά την διάρκεια της διαδικασίας αποθήκευσης σημαντικό κομμάτι ήταν ο αλγόριθμος να απορρίπτει αναδημοσιεύσεις (retweets) των καναλιών από άλλους χρήστες μιας και θέλαμε μόνο το αρχικό περιεχόμενο (original content) των καναλιών που επιλέξαμε.
Στην διαδικασία του data cleansing επίσης απορρίψαμε τα emoji μιας και θεωρήσαμε πως είναι αντικείμενο άλλης μελέτης.

Τελικά στην βάση δεδομένων επιστρέφαμε ένα JSON file με keys για το κάθε tweet τα ακόλουθα:

- id
- tweet_@
- tweet_text
- tweet_created_at
- tag

5 Μοντέλα Επεξεργασίας

5.1 Bag of Words

Η πρώτη απόπειρα εξαγωγής αποτελεσμάτων έγινε με τη χρήση του **countVectorizer** της scikitlearn(https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html) . Είναι η απλή αντιμετώπιση των τίτλων από τις ειδήσεις μας ως «τσουβάλι» από λέξεις. Έτσι μετατρέποντας την συλλογή των τίτλων-κειμένων σε διανυσματικό πίνακα και κάνοντας μια ταξινόμηση με τον αλγόριθμο Naïve Bayes παίρνουμε μια χαμηλή απόδοση πρόβλεψης της τάξης του 85/100, ενώ με Random Forest αυξάνετε λίγο αλλά όχι σημαντικά σε απόδοση 89/100. Τέλος με χρήση του ταξινομητή MLP πήραμε απόδοση 91/100

5.2 Συχνότητα των όρων

Η δεύτερη απόπειρα για την καλύτερη ανάθεση βαρών σε κάθε όρο, χρησιμοποιήσαμε το **TF-IDF**(TermFrequency - InverseDocumentFrequency).
Όπως προσδίδει και το όνομά του, το tf-idf αποτελείται από 2 όρους. Ο πρώτος είναι το **TermFrequency (TF)**:

$$tf(i,d) = \frac{f(i,d)}{\Sigma_i \mathrm{f(i,d)}}$$

Όπου i ο όρος στο κείμενο d. Το tf είναι στην ουσία η συχνότητα με την οποία εμφανίζεται ο κάθε όρος στο κείμενο. Λέξεις με μεγάλη συχνότητα είναι σημαντικότερες για το κείμενο από ό,τι λέξεις με μικρή. Κάνοντας ταξινομήσεις με διάφορους αλγορίθμους η μεγαλύτερη απόδοση που πήραμε ήταν 90/100 αυτή του ταξηνομητή MLP.

5.3 Εξαγωγή Χαρακτηριστικών

Επειδή όμως το θέμα μας απαιτεί μια πιο έξυπνη επεξεργασία φυσικής γλώσσας των τίτλων ειδήσεων , η Τρίτη απόπειρα για εξαγωγή χαρακτηριστικών έγινε χτίζοντας έναν πίνακα για τον κάθε ένα με διάφορα χαρακτηριστικά. Για να το κάνουμε αυτό χρησιμοποιήσαμε τα εργαλεία του πακέτου NLTK και πιο συγκεκριμένα το tokenizer για την κατάτμηση του κειμένου σε όρους ,το stopwords για αφαίρεση εξαιρουμένων λέξεων, το Sentiment Intensity Analyzer για την ανάλυση συναισθήματος πρότασης και το Part of Speech Tagger για τον χαρακτηρισμό μέρους του λόγου.

Έτσι κάνοντας το preprocessing καταλήξαμε από τους τίτλους που δώσαμε στα έξης στοιχεία:

1. Word-count , το πλήθος των λέξεων του τίτλου ,
2. Avg-word-len ,το μέσο μήκος των λέξεων σε γράμματα
3. Max-word-len, το μέγεθος της μεγαλύτερης λέξης
4. Start-wn , αν ξεκινάει η πρόταση με αριθμό
5. Is-number , εύρεση αριθμών στον τίτλο
6. Start-ws , εύρεση ερωτηματικής αντωνυμίας στην αρχή του τίτλου
7. Start-the αν ξεκινάει με 'THE'
8. Qm-exist , εύρεση ερωτηματικών στον τίτλο
9. In-num , πλήθος τοπικών επιρρημάτων
10. Wrb-num , τα επιρρήματα 'πού', ' πότε' , 'πως' , 'πόσο'
11. Rp-num , πλήθος προθέσεων
12. Conjuctions, determiners and 'to'
13. Cc-num , απαρίθμηση συνδετικών λέξεων (συμπλεκτικών, διαζευκτικών, αντιθετικών κλπ.)
14. Cd-num , απαρίθμηση χρονολογιών και λοιπών αριθμητικών στοιχείων εκφρασμένων ολογράφως
15. Dt-num , επιρρήματα προσδιορισμού
16. To-num , πλήθος εμφάνισης της πρόθεσης 'προς'

Predeterminer and pronouns

17. Pdt-num , πλήθος ποσοτικών επιρρημάτων
18. Prp-num , πλήθος προσωπικών αντωνυμιών
19. Prpp-num , πλήθος κτητικών αντωνυμιών

Nouns

20. Nn-num , πλήθος ουσιαστικών στον ενικό
21. Nnp-num , πλήθος κύριων ονομάτων
22. Nns-num , πλήθος ουσιαστικών στον πληθυντικό

Adverbs

23. Rb-num , πλήθος επιρρημάτων
24. Rbr-num , πλήθος επιρρημάτων στον συγκριτικό βαθμό
25. Rbs-num , πλήθος επιρρημάτων στον υπερθετικό βαθμό

Adjectives

26. Jj-num , πλήθος επιθέτων
27. Jjr-num , πλήθος επιθέτων στον συγκριτικό βαθμό
28. Jjs-num , πλήθος επιθέτων στον υπερθετικό βαθμό

Verbs and particles

29. Vb-num , πλήθος ρημάτων
30. Vbd-num , πλήθος ρημάτων παρελθοντικού χρόνου
31. Vbg-num , πλήθος μετοχών ενεστώτα ή γερουνδίων
32. Vbn-num , πλήθος μετοχών αορίστου
33. Vbp-num , πλήθος ρημάτων που δεν είναι στο γ' ενικό
34. Vbz-num , πλήθος ρημάτων που είναι στο γ' ενικό

ΔύοΛεξικά

35. Cmn , πληθος συνηθισμένων λέξεων
36. Hpr , πλήθος υπερβολικών λέξεων

Vader

37. Neu , neutral
38. Neg , negative
39. Pos , positive
40. Com , compound

N-gram

41. Gram2-nnp
42. Gram2-vbz
43. Gram2-in
44. Gram2-in2
45. Gram2-in3
46. Gram2-nns
47. Gram2-jj
48. Gram2-prp
49. Gram2-teleia

Εφαρμόσαμε τις συναρτήσεις εξαγωγής των χαρακτηριστικών αυτών ώστε να μπουν ως στήλες στους δυο πίνακες που φτιάξαμε. Έναν για το σύνολο εκπαίδευσης και έναν για το

σύνολο τεσταρίσματος, με γραμμές τους τίτλους που αποσπάσαμε με τους τρόπους που περιγράψαμε νωρίτερα.

Στην συνέχεια φτιάξαμε έναν transformer ώστε να χρησιμοποιήσουμε δικά μας χαρακτηριστικά τα οποία βρήκαμε με την βοήθεια άλλων επιστημονικών δημοσιεύσεων και θα εξηγήσουμε στην πορεία. Ας δούμε όμως πρώτα τι είναι ένας μετασχηματιστής (transformer).

Όπως εξηγήσαμε παραπάνω τον τρόπο λειτουργίας του tfidfvectorizer που ήταν μια έτοιμη βιβλιοθήκη και έπαιρνε τους τίτλους των ειδήσεων και τις μετέτρεπε σε πίνακες ώστε να πάρεχει πληροφορία για το πόσο σημαντικές είναι οι λέξεις στο μεγαλύτερο μέρος του δείγματος, δηλαδή μετατροπή κειμένου σε αριθμητικές τιμές Κάθε έτοιμο vectorizer λοιπόν που μπορούμε να βρούμε σε βιβλιοθήκες τις Python είναι ένας μετασχηματιστής που μπορεί να μας φανεί χρήσιμος στην εξαγωγή συμπερασμάτων. Έτσι και ο δικό μας μετασχηματιστής έχοντας ως εισόδους καθαρά τους τίτλους ως κείμενο μετασχηματίζει και μας δίνει ως έξοδο πίνακες με τον καθέναν να έχει μια διαφορετική πληροφορία για τον κάθε τίτλο, ακριβώς όπως τον προγραμματίσαμε.
Χρησιμοποιώντας τις βιβλιοθήκες One-Hot-Encoder (https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html) και Feature-Union (https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.FeatureUnion.html) κάναμε fit και έπειτα transform τα δεδομένα μας τόσο για το training set όσο και για το testing set που είχαμε ήδη χωρίσει.

6 Αναπαράσταση κειμένου

6.1 Κανονικοποίηση κειμένου

Η Κανονικοποίηση κειμένου επιτελείται από ορισμένα βήματα, τα οποία παραλλάσσονται ανάλογα με την εφαρμογή, με σημαντικότερο την ανάλυση σε λεκτικές μονάδες (tokenization). Ο ορισμός του τι αποτελεί λεκτική μονάδα (token) μπορεί να εξαρτάται από το είδος των προς επεξεργασία κειμένων, αλλά στα πλαίσια της φυσικής γλώσσας πρόκειται για λέξεις. Ο συνήθης τρόπος με τον οποίο γίνεται η εν λόγω ανάλυση είναι με τη χρήση κανονικών εκφράσεων (regular expressions). Συγκεκριμένα, για την πραγματοποίηση της εργασίας χρησιμοποιήθηκε ο αλγόριθμος που παρέχεται από το Stanford και αναπτύχθηκε για το tokenization tweets στην εκπαίδευση των αναπαραστάσεων GloVe οι οποίες θα αναφερθούν παρακάτω. Ο προαναφερθείσας αλγόριθμος αντιμετωπίζει ορισμένες δυσκολίες οι οποίες οφείλονται σε χαρακτηριστικά που δεν εμφανίζονται σε τυπικά κείμενα, όπως για παράδειγμα τα λογοτεχνικά ή τα επιστημονικά. Πρόκειται για χαρακτηριστικά του άτυπου λόγου, δηλαδή ανορθόγραφες (εσκεμμένα ή μη) λέξεις, συντμήσεις, κτλ. καθώς και για χαρακτηριστικά που αφορούν το διαδίκτυο (π.χ. υπερσύνδεσμοι, emoticon) και άλλα που αφορούν την πλατφόρμα του tweeter (π.χ. αναφορές σε ονόματα χρηστών). Για την αναπαράσταση των τελευταίων χρησιμοποιούνται επιπλέον tokens πέραν αυτών που αντιστοιχούν σε λέξεις της αγγλικής γλώσσας. Παρακάτω παρατείθεται ο κώδικας σε Python και ακολουθεί μια σύντομη επεξήγηση του.

```
import regex as re

FLAGS = re.MULTILINE | re.DOTALL

def hashtag(text):
    text = text.group()
    hashtag_body = text[1:]
    if hashtag_body.isupper():
        result = "<hashtag> {} ".format(hashtag_body)
    else:
        result = " ".join(["<hashtag>"] + re.split(r"(?=[A-Z])", hashtag_body, flags=FLAGS))
    return result

def allcaps(text):
    text = text.group()
    return text.lower() + " <allcaps> "


def tokenize(text):
    # Different regex parts for smiley faces
    eyes = r"[8:=;]"
    nose = r"['`\-]?"

    # function so code less repetitive
    def re_sub(pattern, repl):
        return re.sub(pattern, repl, text, flags=FLAGS)

    text = re_sub(r"https?:\/\/\S+\b|www\.(\w+\.)+\S*", " <url> ")
    text = re_sub(r"@\w+", "<user>")
    text = re_sub(r"{}{}[)dD]+|[(dD]+{}{}".format(eyes, nose, nose, eyes), " <smile> ")
    text = re_sub(r"{}{}p+".format(eyes, nose), "<lolface>")
    text = re_sub(r"{}{}\(+|\)+{}{}".format(eyes, nose, nose, eyes), " <sadface> ")
    text = re_sub(r"{}{}[\/|l*]".format(eyes, nose), " <neutralface> ")
    text = re_sub(r"<3","<heart>")
    text = re_sub(r"/"," / ")
    text = re_sub(r"[-+]?[.\d]*[\d]+[:,.\d]*", " <number> ")
    text = re_sub(r"#\S+", hashtag)
    text = re_sub(r"([!?.]){2,}", r" \1 <repeat> ")
    text = re_sub(r"\b(\S*?)(.)\2{2,}\b", r"\1\2 <elong>")
    text = re_sub(r"([A-Z]){2,}", allcaps)

    return list(filter(None, text.lower().split()))
```

Ο παραπάνω κώδικας μετασχηματίζει το κείμενο μέσω μιας σειρά από regular expression substitutions τα οποία, με τη σειρά που εμφανίζονται στον κώδικα της συνάρτησης tokenize, αντικαθιστούν τα εξής:

1. Τους υπερσυνδέσμους με το token<url>
2. Τα tags χρηστών με το token<user>
3. Συνήθη emoticon με αντίστοιχα tokens
4. Το χαρακτήρα '/', περιβάλλοντάς τον με κενά ώστε να διαχωρίσει τυχόν λέξεις που βρίσκονται εκατέρωθεν του
5. Αριθμούς με το token<number>
6. Τα hashtags με το token<hashtag> ακολουθούμενο με το κείμενο του
7. Θαυμαστικά, ερωτηματικά και τελείες που επαναλαμβάνονται 2 ή παραπάνω φορές με το σημείο στίξης ακολουθούμενο από το token<repeat>
8. Λέξεις οι οποίες έχουν γραφεί με το τελευταίο τους γράμμα να επαναλαμβάνεται με τη λέξη χωρίς την επανάληψη ακολουθούμενη από το token<elong>
9. Λέξεις γραμμένες με όλα τους τα γράμματα κεφαλαία με το τη λέξη ακολουθούμενη από το token<allcaps>

Τέλος, τα κεφαλαία γράμματα αντικαθίστανται με μικρά και το κείμενο χωρίζεται στο κενό δίνοντας μια λίστα από tokens.
Προφανώς ο αλγόριθμος δεν αντιμετωπίζει ζητήματα που αφορούν την λανθασμένη ορθογραφία. Κάτι τέτοιο θα ήταν αρκετά δυσκολότερο και επίσης στην προκείμενη περίπτωση δεν θα είχε νόημα. Ο λόγος για αυτό (και ο λόγος που επιλέχθηκε ο αλγόριθμος) είναι ότι τα tokens στο λεξικό των αναπαραστάσεων GloVe έχουν προκύψει από αυτόν το αλγόριθμο. Παρ' όλα αυτά υπάρχει ακόμα η περίπτωση να εμφανιστούν tokens τα οποία δεν υπάρχουν στο λεξικό, στην οποία περίπτωση η αναπαράσταση του token είναι μια ειδική για άγνωστες λέξεις. Η παραπάνω είναι μια συνήθης πρακτική για την αντιμετώπιση λέξεων εκτός λεξιλογίου.

Καταλήγοντας, αξίζει να αναφερθεί ότι αν και δεν χρησιμοποιήθηκαν στην εργασία ένα άλλο σύνηθες στάδιο της προεπεξεργασίας κειμένου είναι το stemming, δηλαδή η αντικατάσταση καταλήξεων των λέξεων ώστε να βρίσκονται σε μία κανονική μορφή (π.χ. running→run).

6.2 Διανυσματική σημασιολογία

Ο όρος διανυσματική σημασιολογία αναφέρεται σε διάφορες τεχνικές αναπαράστασης λέξεων ως διανύσματα τα οποία κωδικοποιούν μερικά από τα σημασιολογικά χαρακτηριστικά τους. Η σημασία της παραπάνω πρότασης θα γίνει εμφανής στη συνέχεια. Συγκεκριμένα, οι παραπάνω τεχνικές βασίζονται στην παρατήρηση ότι όμοιες σημασιολογικά λέξεις χρησιμοποιούνται σε παρόμοια συμφραζόμενα. Η έννοια της ομοιότητας θα μπορούσε να οριστεί με διάφορους τρόπους, για παράδειγμα θα μπορούσε να οριστεί ως συνωνυμία. Παρ' όλα αυτά η ομοιότητα την οποία προσπαθούν να καταγράψουν οι εν λόγω τεχνικές είναι γενικότερη και περισσότερο διαισθητική (άλλωστε δύο αντώνυμες λέξεις μπορούν να θεωρηθούν όμοιες). Ένας ακριβής ορισμός μάλλον θα πήγαζε από την προαναφερθείσα παρατήρηση, δηλαδή ο βαθμός ομοιότητας δύο λέξεων εξαρτάται από το κατά πόσο χρησιμοποιούνται σε παρόμοια συμφραζόμενα. Για παράδειγμα οι λέξεις «γάτα» και «σκύλος» είναι αρκετά όμοιες σύμφωνα με αυτήν την άποψη.

6.2.1 Ομοιότητα διανυσμάτων

Η ομοιότητα μεταξύ δύο λέξεων-διανυσμάτων μετράτε με την απόσταση συνημίτονου:

$$similatity(a,b) = \frac{a \cdot b}{\|a\|\|b\|}$$

Το εσωτερικό γινόμενο θα μπορούσε να χρησιμοποιηθεί από μόνο του ως μέτρο ομοιότητας. Παρ' όλα αυτά σε ορισμένες τεχνικές αναπαράστασης λέξεων το μέτρο των διανυσμάτων εξαρτάται από τη συχνότητα εμφάνισης της λέξης και συνεπώς η ομοιότητα της με κάποια άλλη λέξη επηρεάζεται από αυτό, γεγονός το οποίο είναι ανεπιθύμητο εφόσον ένα μέτρο ομοιότητας θα έπρεπε να είναι ανεξάρτητο από τη συχνότητα των λέξεων.

6.2.2 Κανονικότητα στη δομή των αναπαραστάσεων

Παρόλο που τα περισσότερα μοντέλα εκπαιδεύονται ώστε οι σημασιολογικά παραπλήσιες λέξεις να έχουν μικρή απόσταση συνημίτονου, παρατηρήθηκε ότι η δομή των αναπαραστάσεων που προκύπτουν αποτυπώνει περεταίρω σημασιολογική πληροφορία με τη μορφή γραμμικών συσχετίσεων. Πιο συγκεκριμένα αυτό που ανακαλύφθηκε είναι πως το κοντινότερο (σύμφωνα με την απόσταση συνημίτονου) διάνυσμα σε αυτό που προκύπτει αν από την εξής πράξη:

$$vector(\text{"Athens"}) - vector(\text{"Greece"}) + vector(\text{"France"})$$

είναι το διάνυσμα για την λέξη Παρίσι. Γενικότερα, ερωτήσεις για σχέσεις αναλογίας της μορφής «το α ειναι για το β, ότι το γ για το _» μπορούν να απαντηθούν βρίσκοντας το κοντινότερο διάνυσμα στο
$vector(\beta) - vector(\alpha) + vector(\gamma)$. Για την αξιολόγηση της ποιότητας των αναπαραστάσεων ως προς την παραπάνω ιδιότητα δημιουργήθηκε ένα dataset με τέτοιες ερωτήσεις δύο ειδών: α) συντακτικές, π.χ. big :bigger, small : smaller, και β) σημασιολογικές, π.χ. ζεύγη χωρών με τις πρωτεύουσές τους όπως το παραπάνω παράδειγμα.
Το παραπάνω παρατηρήθηκε αρχικά σε μοντέλα RNN (recurrent neural network) και οδήγησε στην ανάπτυξη των μοντέλων CBOW (continuous bag of words) και SkipGram, τα οποία είναι απλούστερα στην αρχιτεκτονική τους με σκοπό να διατηρούν τις παραπάνω γραμμικές σχέσεις.

6.2.3 Πίνακας συνεμφάνισης

Το GloVe προκειμένου να καταγράψει την πληροφορία για τα συμφραζόμενα μιας λέξης χρησιμοποιεί τον πίνακα συνεμφάνισης. Πρόκειται για έναν πίνακα X διαστάσεων |V| ×|V|, όπου |V| το μέγεθος του λεξικού, στον οποίο το στοιχείοX_{ij} είναι ίσο με τον αριθμό των φορών τον οποίων η j-οστή λέξη υπάρχει εντός ενός παραθύρου μεγέθους W εκατέρωθεν της i-οστής λέξης στο χρησιμοποιούμενο dataset. Ο πίνακας αυτός είναι αραιός και συμμετρικός.

Διανυσματικές αναπαραστάσεις μπορούν να προκύψουν απευθείας από τον παραπάνω πίνακα χρησιμοποιώντας τις γραμμές ή στήλες του. Τέτοια διανύσματα, όμως, είναι αραιά και μεγάλης διάστασης (|V|), το οποίο καθιστά την εκπαίδευση μοντέλων αργή. Ακόμα, δεν κωδικοποιούν επιτυχημένα τις σημασιολογικές σχέσεις των λέξεων.

6.2.4 Το μοντέλο Glove

Οι δημιουργοί του GloVe αφορμήθηκαν από την προαναφερθείσα κανονικότητα της δομής των διανυσματικών αναπαραστάσεων ώστε να σχεδιάσουν ένα μοντέλο το οποίο εξηγεί τον τρόπο με τον οποίο εμφανίζονται. Επίσης προσπάθησαν να συνδυάσουν τα πλεονεκτήματα των δύο βασικών οικογενειών μοντέλων, οι οποίες είναι η παραγοντοποίηση ενός πίνακα συχνότητας συνεμφάνισης (π.χ. LSA) και οι μέθοδοι «ρηχού» παραθύρου (π.χ. CBOW), οι οποίες εκπαιδεύουν μοντέλα με online, στοχαστκό τρόπο. Το πλεονέκτημα των πρώτων μοντέλων είναι ότι χρησιμοποιούν συνολικές στατιστικές πληροφορίες των δεδομένων, ενώ των δεύτερων ότι καταφέρουν να αποτυπώσουν καλύτερα τις γραμμικές σχέσεις των λέξεων.

Το μοντέλο βασίζεται στη διαίσθηση ότι ο λόγος $\frac{P_{ik}}{P_{jk}}$αποτελεί ένα δείκτη διάκρισης των λέξεων i, j όσον αφορά τη λέξη k, δηλαδή για λέξεις k οι οποίες σχετίζονται με την i, αλλά όχι με την j ο λόγος είναι μεγάλος, για το αντίθετο ο λόγος είναι μικρός, ενώ για λέξεις που σχετίζονται το ίδιο και με τις δύο ο λόγος είναι περίπου 1. Στον παραπάνω συμβολισμό P_{ij} είναι η εκτίμηση μέγιστηςπιθανοφάνειας της πιθανότητας η λέξη j να εμφανιστεί στα συμφραζόμενα της λέξης i και συνεπώς ισούται με $\frac{X_{ij}}{X_i}$, όπουX_{ij} το στοιχείο ij του πίνακα συνεμφάνισης και X_i ο αριθμός των φορών που η λέξη i εμφανίζεται στο dataset.

Λόγω του παραπάνω αναζητείται μία συνάρτηση τέτοια ώστε:

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}},$$

όπου w και $\overline{w}$ δύο διαφορετικές αναπαραστάσεις. Έπειτα με την επιβολή ορισμένων απαιτήσεων όπως είναι η γραμμική δομή την οποία είναι επιθυμητό να κωδικοποιήσει το μοντέλο, καθώς και η συμμετρία τόσο του πίνακα X όσο και μεταξύ των λέξεων που αποτελούν συμφραζόμενα και των λέξεων που θεωρούνται κέντρα του παραθύρου οι συγγραφείς καταλήγουν στην ακόλουθη εξίσωση:

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = log(X_{ik}),$$

όπου τα b_i, $\overline{b_k}$ είναι βαθμωτές ποσότητες.
Από την παραπάνω εξίσωση προκύπτει η συνάρτηση κόστους για την εκπαίδευση.

$$\sum_{i,j}^{|V|} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - log(X_{ij}))^2$$

όπου το f μία συνάρτηση βάρους, ώστε τα στοιχεία του πίνακα X με μικρή τιμή, τα οποία συνεπώς δεν παρέχουν πολύ πληροφορία, να μην υπερτιμώνται, η οποία είναι η εξής:

$$f(x) = \begin{cases} \left(\frac{x}{x_{max}}\right)^{\frac{3}{4}} & \text{αν } x < x_{max} \\ 1 & \text{διαφορετικά} \end{cases}$$

Η τελική αναπαράσταση της λέξης i προκύπτει από το άθροισμα $w_i + \overline{w_i}$ καθώς η χρήση των δύο ξεχωριστών αναπαραστάσων βοηθάει στην αποφυγή του overfitting.

6.2.5 Μεταβαίνοντας από τις λεκτικές μονάδες στο κείμενο

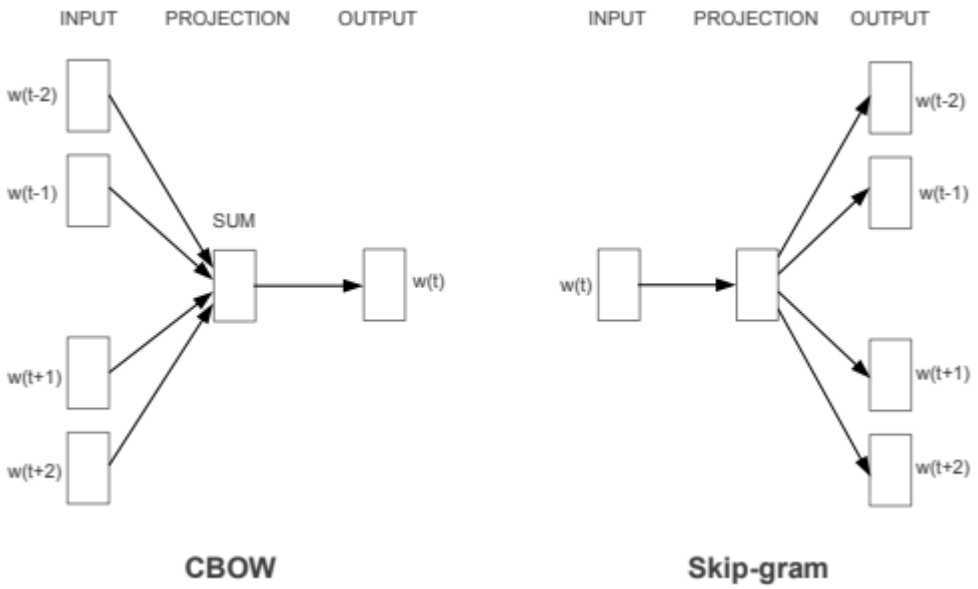
Μέχρι στιγμής έχει γίνει συζήτηση για την αναπαράσταση των λέξεων. Παρ' όλα αυτά, εφόσον ο τελικός στόχος είναι η ταξινόμηση tweets, χρειάζεται να συνδυάσουμε τις ξεχωριστές αναπαραστάσεις των λέξεων των του κειμένου σε μία ενιαία.

Ο πιο απλός τρόπος να γίνει αυτό είναι αθροίζοντας, ενδεχομένως με κάποια βάρη (όπως tf-idf), τα επιμέρους διανύσματα δημιουργώντας μια bag-of-words αναπαράσταση. Αυτή η τεχνική έχει την προφανή αδυναμία δεν λαμβάνει υπόψη τη σειρά των λέξεων στο κείμενο.

Ο τρόπος που χρησιμοποιήθηκε για την εργασία και ο οποίος είναι ο πιο συνήθης για προβλήματα μάθησης τα οποία περιλαμβάνουν ακολουθίες είναι η χρήση ενός RNN. Συγκεκριμένα χρησιμοποιήθηκε ένα εμφίδρομο LSTM, δηλαδή δύο LSTM που επεξεργάζονται το κείμενο με αντίθετη φορά, και η αναπαράσταση του κειμένου είναι η συνένωση (concatenation) των τελικών εξόδων των δύο.

6.3 Μοντέλα CBOWκαι Skip-Gram

Έχοντας παρατηρήσει της προαναφερθείσες γραμμικές σχέσεις στις αναπαραστάσεις που προέκυψαν από τις ενδιάμεσες αναπαραστάσεις κάποιων νευρωνικών στοιχείων οι Mikolov et al.[17] θέλησαν να δημιουργήσουν αποδοτικότερα μοντέλα τα οποία τις δημιουργούν. Δύο είναι τα μοντέλα στα οποία κατέληξαν. Το continuous bag-of-words (εφεξής CBOW) και το continuous skip-gram (εφεξής SG). Η λειτουργία τους είναι αντίστροφη: το πρώτο χρησιμοποιεί τα συμφραζόμενα για να προβλέψει τη λέξη, ενώ το δεύτερο χρησιμοποιεί τη λέξη για να προβλέψει τα συμφραζόμενα.



Σχήμα: Η αρχιτεκτονική των δύο δικτύων. Εικόνα από εργασία των Mikolov et al

Ο τρόπος εκπαίδευσης και των δύο μοντέλων είναι παρόμοιος οπότε θα αναλύσουμε συνοπτικά μόνο το δεύτερο.

Αν συμβολίσουμε με T τον αριθμό των λέξεων στο σύνολο εκπαίδευσης και ως c το μισό τον αριθμό των λέξεων που θέλουμε να προβλέψουμε εκατέρωθεν της λέξης εισόδου (άρα συνολικά 2c λέξεις) τότε η συνάρτηση κόστους που θέλουμε να ελαχιστοποιήσουμε είναι η:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

Η πιθανότητα $p(w_{t+j} \mid w_t)$ ορίζεται ως:

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^{\top} v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left({v'_w}^{\top} v_{w_I}\right)}$$

όπου w_o και w_i είναι διαφορετικές αναπαραστάσεις της λέξης ανάλογα με το αν είναι είσοδος ή έξοδος του δικτύου και W το μέγεθος του λεξικού. Η πιθανότητα αυτή είναι προφανώς το softamax της ομοιότητας των λέξεων-διανυσμάτων.

Το παραπάνω softmax είναι υπολογιστικά ακριβό λόγω του μεγέθους του λεξικού, οπότε μία τεχνική που χρησιμοποιήθηκε για την υπέρβαση αυτού του εμποδίου είναι το Negative Sampling. Σύμφωνα με αυτήν για κάθε θετικό δείγμα, δηλαδή για κάθε λέξη που ανήκει στα συμφραζόμενα της λέξης εισόδου επιλέγουμε ένα αρνητικό παράδειγμα, δηλαδή μια λέξη η οποία επιλέγεται τυχαία. Κατά αυτόν τον τρόπο προκύπτει η εξής συνάρτηση κόστους:

$$\log \sigma({v'_{w_O}}^{\top} v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^{\top} v_{w_I}) \right]$$

Προφανώς η μέση τιμή που υπάρχει στον παραπάνω τύπο προσεγγίζεται από τη δειγματοληψία που γίνεται. Από τη μορφή της συνάρτησης είναι φανερό ότι η μάθηση μπορεί να γίνει με online τρόπο.

Υπάρχουν διάφορες επιλογές για την κατανομή πιθανότητας βάσει της οποίας γίνεται η δειγματοληψία των αρνητικών δειγμάτων. Μία από αυτές είναι ακόλουθη:

$$P_\alpha(w) = \frac{count(w)^\alpha}{\sum_{w'} count(w')^\alpha}$$

όπου count(w) ο αριθμός εμφανίσεων της λέξης w στο σύνολο εκπαίδευσης και α μια σταθερά η οποία συνήθως τίθεται ίση με 0.75.

7 Μέθοδοι Αξιολόγησης

7.1 Ταξινομητές

Η αξιολόγηση των ταξινομητών γίνεται πάντα σε δεδομένα που δεν έχουν δει κατά την εκπαίδευση έτσι ώστε να αξιολογήσουμε τη δυνατότητα γενίκευσής τους. Συνεπώς, προτού φτιάξουμε το μοντέλο κάθε ταξινομητή χωρίζουμε τα δεδομένα μας τυχαία σε ένα σύνολο εκπαίδευσης (train set) και ένα σύνολο ελέγχου (test set).

Στην περίπτωση μας διαχωρίσαμε το dataset με αναλογία 70-30. Ακολουθήσαμε το ίδια διαδικασία για όλες τις περιπτώσεις ταξινόμησης. Αρχικά χρησιμοποιούμε το train set για να εκτιμούμε και να βελτιώνουμε το μοντέλο του ταξινομητή κατά την ανάπτυξή του. Χρησιμοποιούμε μετά το test set για να αξιολογήσουμε στατιστικά την απόδοση του μοντέλου μας.

Η scikitklearn έχει τη συνάρτηση train_test_split() που ανακατεύει τυχαία τα δείγματα και τα διαχωρίζει σε train και test με βάση το ποσοστό που της δώσαμε.

Στην συνέχεια θα περιγράψουμε τον τρόπο λειτουργίας και χαρακτηριστικά των ταξινομητών που χρησιμοποιήθηκαν.

7.1.1 Gaussian Naïve Bayes

Η βασική ιδέα λειτουργίας του ταξινομητή είναι ο γνωστός νόμος του Bayes και η (naïve) υπόθεση ότι τα χαρακτηριστικά είναι όλα ανεξάρτητα μεταξύ τους (δεν ισχύει γενικά, αλλά ο ταξινομητής είναι πρακτικά καλός σε πολλές περιπτώσεις). Παράδειγμα: θα βρέξει σήμερα? Naive Bayes: "Θα το προβλέψω με βάση το παρελθόν θεωρώντας ότι τα χαρακτηριστικά θερμοκρασία, νεφοκάλυψη και ατμοσφαιρική πίεση είναι όλα ανεξάρτητα μεταξύ τους".

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Με δεδομένα μια μεταβλητή κατηγορίας (κλάσης) X και ένα εξαρτώμενο διάνυσμα χαρακτηριστικών $C1$ μέχρι Cn, ισχύει:

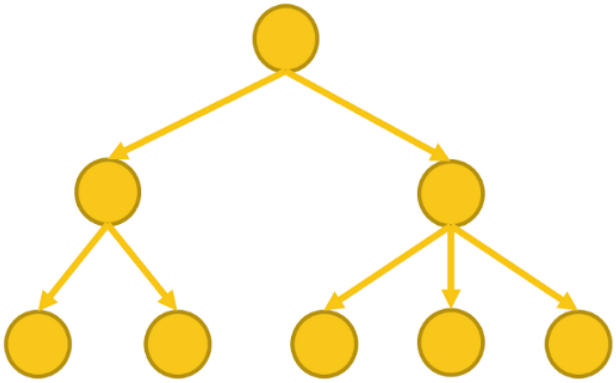
$$P(\mathbf{X} \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots \times P(x_n \mid C_i)$$

Τέλος, η πιθανότητα κάθε κλάσης που δίνεται σε μια περίπτωση (παράδειγμα δοκιμής) υπολογίζεται χρησιμοποιώντας το Θεώρημα Bayes.

$$P(c_i \mid x) = \frac{P(x \mid c_i) * P(c_i)}{\sum_j P(x \mid c_j) * P(c_j)}$$

7.1.2 Random Forest

Τόσο για παλινδρόμηση όσο και για ταξινόμηση τα δέντρα αποφάσεων αποτελούν ένα τύπο μοντέλου αρκετά συχνό. Τα δέντρα αποφάσεων αποτελούνται από απαντήσεις σε διαδοχικές ερωτήσεις που αποτελούν τα «κλαδιά» των δέντρων. Κάθε συγκεκριμένη διαδρομή αυτού έχει μια δεδομένη απάντηση, χρησιμοποιώντας ισχυρισμούς, αν συμβαίνει κάποιο γεγονός δίνει ένα συγκεκριμένο αποτέλεσμα ενώ ένα διαφορετικό γεγονός μπορεί να δώσει διαφορετικό αποτέλεσμα-απάντηση. Η δομή ενός δέντρου αποφάσεων φαίνεται στη παρακάτω εικόνα:



Σχήμα : Δομή μονού Δέντρου Αποφάσεων

Μια έννοια που αξίζει να αναφερθεί όσον αφορά τα δέντρα αποφάσεων είναι το βάθος του δέντρου. Τ βάθος του δέντρου αντιπροσωπεύει πόσες ερωτήσεις καλούνται πριν φτάσουμε στην προβλεπόμενη ταξινόμησή μας. Παρακάτω ακολουθούν κάποια πλεονεκτήματα και μειονεκτήματα της χρήσης δέντρων αποφάσεων [20]:

Πλεονεκτήματα:

1. Εύκολη ερμηνεία και απεικόνιση.

2. Οι εσωτερικές λειτουργίες μπορούν να παρατηρηθούν.

3. Μπορεί να χειριστεί και αριθμητικά και κατηγορικά δεδομένα.

4. Εφαρμόζεται σε μεγάλα σύνολα δεδομένων.

5. Είναι εξαιρετικά γρήγορα.

Μειονεκτήματα:

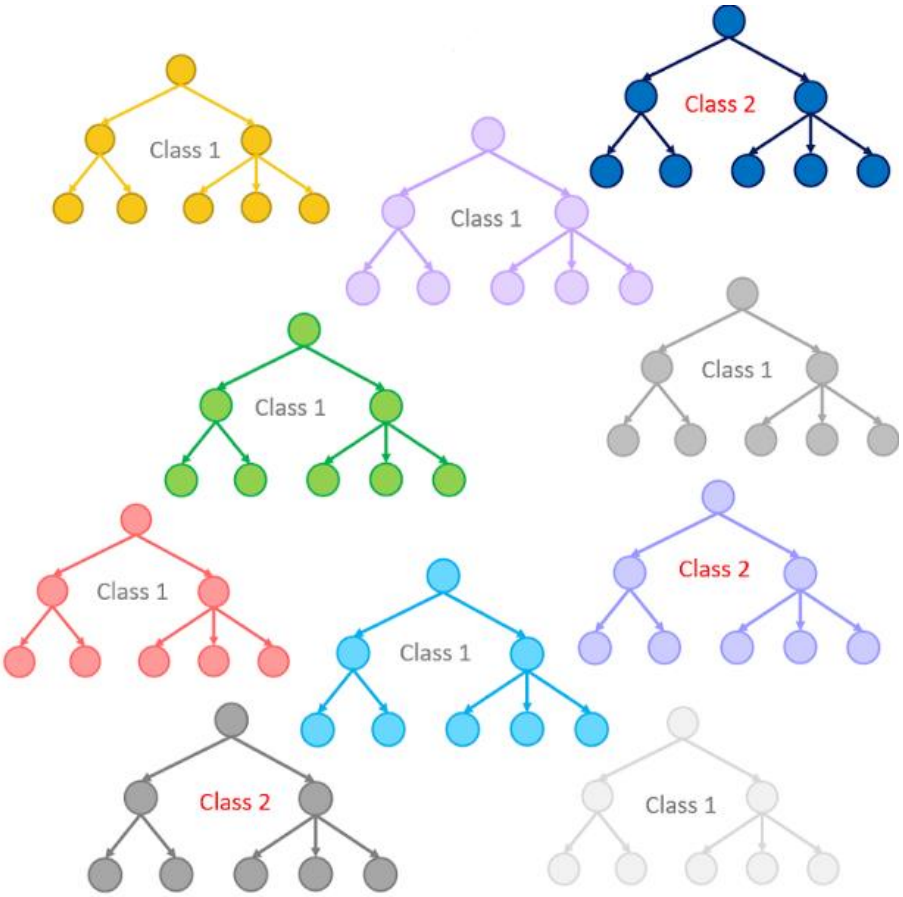
1. Για τη δημιουργία δέντρων αποφάσεων απαιτούνται αλγόριθμοι ικανοί να καθορίσουν μια βέλτιστη επιλογή σε κάθε κόμβο. Ένας δημοφιλής αλγόριθμος είναι ο αλγόριθμος του Hunt, ο οποίος εντοπίζει τη βέλτιστη απόφαση σε κάθε κλαδί – βήμα χωρίς όμως να λαμβάνει υπόψη το παγκόσμιο βέλτιστο. Δηλαδή, επιλέγοντας το καλύτερο αποτέλεσμα σε ένα δεδομένο βήμα δεν διασφαλίζει ότι θα κατευθύνεται προς τη διαδρομή που θα οδηγήσει στη βέλτιστη απόφαση στον τελικό κόμβο του δέντρου.

2. Τα δέντρα απόφασης είναι επιρρεπή σε υπερφόρτωση, ειδικά όταν ένα δέντρο έχει μεγάλο βάθος. Αυτό οφείλεται στην μεγάλη εξειδίκευση που οδηγεί τελικά σε μικρότερο δείγμα γεγονότων που πληρούν τις προηγούμενες παραδοχές. Αυτό το μικρό δείγμα θα μπορούσε να οδηγήσει σε αβάσιμα συμπεράσματα. Αυτό το μειονέκτημα μπορεί να αντιμετωπιστεί με τη ρύθμιση του βάθους του δέντρου όμως θα υπάρχει πάντα το σφάλμα της μεροληψίας.

Για να ελαχιστοποιηθούν τα σφάλματα λόγω μεροληψίας και διακύμανσης χρησιμοποιούνται τα τυχαία δάση. Ένα τυχαίο δάσος είναι απλώς μια συλλογή δέντρων αποφάσεων των οποίων τα αποτελέσματα συγκεντρώνονται σε ένα τελικό αποτέλεσμα. Η ικανότητά τους να περιορίζουν την υπερφόρτωση χωρίς να αυξάνεται σημαντικά το σφάλμα λόγω της μεροληψίας είναι ο λόγος για τον οποίο είναι τόσο ισχυρά μοντέλα.

Με τη χρήση τυχαίων δασών τα σφάλματα διακύμανσης ελαχιστοποιούνται με την εκπαίδευση σε διαφορετικά δείγματα δεδομένων. Ένας δεύτερος τρόπος είναι η χρήση ενός τυχαίου υποσυνόλου χαρακτηριστικών. Αυτό σημαίνει ότι αν έχουμε 30 χαρακτηριστικά, τα

τυχαία δάση θα χρησιμοποιούν μόνο ένα συγκεκριμένο αριθμό αυτών των χαρακτηριστικών σε κάθε μοντέλο, για παράδειγμα πέντε. Δυστυχώς, έχουμε παραλείψει 25 χαρακτηριστικά που θα μπορούσαν να είναι χρήσιμα. Αλλά όπως αναφέρθηκε, ένα τυχαίο δάσος είναι μια συλλογή από δέντρα αποφάσεων. Έτσι, σε κάθε δέντρο μπορούμε να χρησιμοποιήσουμε πέντε τυχαία χαρακτηριστικά. Εάν χρησιμοποιούμε πολλά δέντρα στο δάσος μας, τελικά πολλά ή όλα τα χαρακτηριστικά μας θα έχουν συμπεριληφθεί. Αυτή η συμπερίληψη πολλών χαρακτηριστικών θα συμβάλει στον περιορισμό των σφαλμάτων. Εάν τα χαρακτηριστικά δεν επιλέχθηκαν τυχαία, τα δέντρα βάσης στο δάσος μας θα μπορούσαν να έχουν υψηλή συσχέτιση. Αυτό συμβαίνει επειδή μερικά χαρακτηριστικά γνωρίσματα θα μπορούσαν να είναι ιδιαίτερα προβλέψιμα και έτσι, τα ίδια χαρακτηριστικά θα επιλέγονταν σε πολλά από τα βασικά δέντρα. Αν πολλά από αυτά τα δέντρα περιλάμβαναν τα ίδια χαρακτηριστικά, δεν θα μπορούσαμε να καταπολεμήσουμε το σφάλμα λόγω διακύμανσης. Καταλήγοντας, τα τυχαία δάση είναι μια ισχυρή τεχνική μοντελοποίησης και αρκετά ισχυρότερη από ένα ενιαίο δέντρο αποφάσεων.



Σχήμα : Δομή Τυχαίου Δάσους

7.1.3 Support Vector Machines and Logistic Regression

Οι μηχανές διανυσμάτων υποστήριξης (SVM) βασίζονται στον διαχωρισμό των στοιχείων με βάση την κλάση τους . Ο διαχωρισμός αυτός γίνεται με την χρήση ενός υπερεπιπέδου σε έναν χώρο Ν διαστάσεων (όπου Ν ο αριθμός των στοιχείων διαχωρισμού). Αυτό το υπερεπίπεδο αφενός προσπαθεί να χωρίσει τα σημεία ανάλογα με την κλάση τους , αφετέρου ψάχνει και διαλέγει το βέλτιστο υπερεπίπεδο για το οποίο μεγιστοποιούνται οι αποστάσεις όλων των σημείων από αυτό.

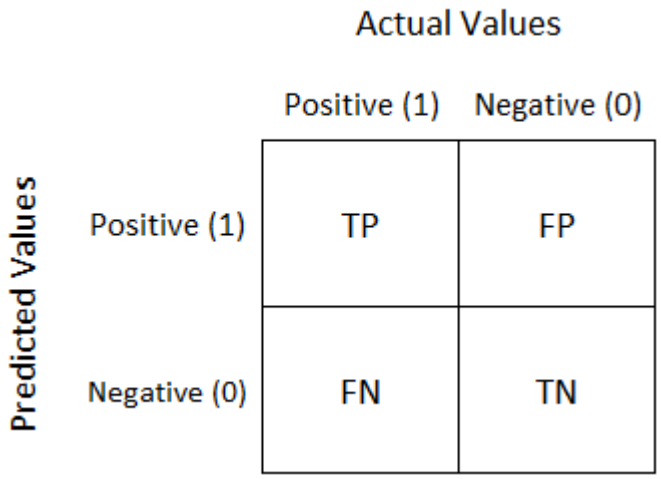
Οι παραπάνω μηχανές είναι από τα σημαντικότερα εργαλεία ταξινόμησης στον χώρο της μηχανικής μάθησής. Όπως επίσης και οι αλγόριθμοι λογικής παλινδρόμησης, όπου περιγράφουν την σχέση μεταξύ των μεταβλητών εισόδου. Οι μηχανές SVM με γραμμικό πυρήνα και τα μοντέλα γραμμικής παλινδρόμησης έχουν παρόμοια απόδοση , αλλά ανάλογα με τα χαρακτηριστικά εισόδου κάποια μοντέλο μπορεί να γίνει πιο αποτελεσματικό από το άλλο. Σε ασκήσεις ταξινόμησης συνιστάτε να χρησιμοποιούνται και τα δύο.

Η διαφορά μεταξύ Λογικής και Γραμμικής παλινδρόμησης [21] είναι ότι η λογική παλινδρόμηση δίνει έχει ως έξοδο διακριτό αποτέλεσμα αντίθετα με την Γραμμική που δίνει συνεχές αποτέλεσμα.

Χρησιμοποιήσαμε Λογική αντί για Γραμμική παλινδρόμηση μιας και τα δεδομένα μας χωρίζονται σε 2 τάξεις, αυτό τα καθιστά γραμμικά διαχωρίσιμα. Το αποτέλεσμα της παλινδρόμησης πρέπει να είναι διακριτό και όχι συνεχές μιας και ένα γραμμικό μοντέλο με την έξοδο του να κυμαίνεται από - άπειρο έως το +άπειρο θα δυσκολευτεί να ερμηνεύσει την τιμή απόκρισης μεταξύ του 0 και του 1.

7.2 Μετρικές απόδοσης

Θα χρησιμοποιήσουμε τις μετρικές accuracy, precision, recall και F1 για να εξάγουμε συμπεράσματα για την αποδοτικότητα της ταξινόμησης. Αυτές οι μετρικές βασίζονται στον πίνακα σύγχυσης:



Διακρίνουμε 4 περιπτώσεις Tp ,Tn , Fp , Fn όπου true positive είναι οι περιπτώσεις θετικής πρόβλεψης με αληθείς συνθήκες ,truenegative (δεν επηρεάζει) θετικής πρόβλεψης με ψευδείς συνθήκες, false positive λάθος πρόβλεψης με ψευδείς συνθήκες και false negative λάθος πρόβλεψης με αληθείς συνθήκες

Έτσι έχουμε τους παρακάτω τύπους για τις μετρικές μας

$$\text{Precision} = \frac{\text{Tp}}{Tp+Fp}$$

$$\text{Recall} = \frac{\text{Tp}}{Tp+Fn}$$

$$\text{Και } \text{F1} = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$$

7.3 Εκτίμηση

Η εκτίμηση της παραπάνω διαδικασίας έγινε με χρήση του αλγορίθμου cross-validation. Εδώ θα περιγράψουμε τι είναι το cross-validation. Η διασταυρωμένη επικύρωση είναι μια στατιστική μέθοδος που χρησιμοποιείται για την εκτίμηση της ικανότητας των μοντέλων μηχανικής μάθησης.

Χρησιμοποιείται για να συγκρίνει και να επιλέξει το μοντέλο για το ένα δεδομένο πρόβλημα πρόβλεψης μοντέλων. Εφαρμόζεται εύκολα και έχει ως αποτέλεσμα εκτιμήσεις δεξιοτήτων που γενικά έχουν χαμηλότερη απόκλιση από άλλες μεθόδους.
Το k-fold cross-validation , είναι η διαδικασία που χρησιμοποιήσαμε εμείς και θα εξηγήσουμε στην πορεία.

Η διασταυρούμενη επικύρωση είναι μια διαδικασία επαναδειγματοληψίας που χρησιμοποιείται για την αξιολόγηση μοντέλων μηχανικής μάθησης σε ένα περιορισμένο δείγμα δεδομένων.

Η διαδικασία έχει μια μόνο παράμετρο που ονομάζεται k που αναφέρεται στον αριθμό των ομάδων στις οποίες πρόκειται να χωριστεί ένα δεδομένο δείγμα δεδομένων. Ως εκ τούτου, η διαδικασία συχνά ονομάζεται k-fold cross-validation. Όταν επιλέγεται μια συγκεκριμένη τιμή για το k, μπορεί να χρησιμοποιηθεί στη θέση του k στην αναφορά στο μοντέλο, όπως το k = 10 να γίνει 10-πλάσιο cross-validation.

Με το cross-validation χρησιμοποιούμε ένα περιορισμένο δείγμα για να εκτιμήσουμε τον τρόπο με τον οποίο το μοντέλο αναμένεται να εκτελέσει γενικά όταν χρησιμοποιείται για να κάνει προβλέψεις για δεδομένα που δεν χρησιμοποιήθηκαν κατά την εκπαίδευση του μοντέλου.

Είναι μια δημοφιλής μέθοδος επειδή είναι απλή στην κατανόησή της και επειδή γενικά έχει ως αποτέλεσμα μια λιγότερο μεροληπτική ή λιγότερο αισιόδοξη εκτίμηση της ικανότητας του μοντέλου από άλλες μεθόδους.
Η γενική διαδικασία έχει ως εξής:

1. Ανακατεύει το σύνολο δεδομένων τυχαία.
2. Διαχωρίστε το σύνολο δεδομένων σε ομάδες k
3. Για κάθε μοναδική ομάδα:
    i. Διαχωρίζει την ομάδα ως ομάδα αναμονής ή ομάδα τεστ
    ii. Παίρνει τις υπόλοιπες ομάδες ως σύνολο δεδομένων εκπαίδευσης
    iii. Κάνει φιτ το μοντέλο (όπως περιγράψαμε πιο πάνω) και το εκπαιδεύει
    iv. Βγάζει το σκορ αξιολόγησης από την ομάδα τεστ και στην συνέχεια επαναλαμβάνει την διαδικασία
4. Συγκεντρώνει τέλος την ικανότητα του μοντέλου χρησιμοποιώντας το δείγμα βαθμολογίας αξιολόγησης μοντέλου.

Είναι σημαντικό ότι κάθε παρατήρηση στο δείγμα δεδομένων να ανατίθεται σε μια μεμονωμένη ομάδα και παραμένει σε αυτήν την ομάδα για όλη τη διάρκεια της διαδικασίας.

Αυτό σημαίνει ότι σε κάθε δείγμα δίνεται η ευκαιρία να χρησιμοποιηθεί ως ομάδα αναμονής και να χρησιμοποιηθεί για την εκπαίδευση του μοντέλου k-1 φορές.

8 Εργαλεία

8.1 Twitter API

Τα δεδομένα που χρησιμοποιήθηκαν συλλέχθηκαν με τη χρήση της διασύνδεσης προγραμματισμού εφαρμογών του Twitter [18]. Το Twitter , όπως και τα περισσότερα μέσα κοινωνικής δικτύωσης σήμερα, είναι εύκολα προσιτά σε μεγάλο τμήμα του πληθυσμού. Στα πλαίσια αυτής της διπλωματικής εργασίας δημιουργήθηκε ένας λογαριασμός στην εφαρμογή αυτή προκειμένου να συλλεχθούν δεδομένα από νέες δημοσιεύσεις – tweets (περίπου 60), 20 εκ των οποίων αποτελούσαν clickbaits, τα οποία στη συνέχεια αποθηκεύθηκαν σε μια βάση δεδομένων Mongo DB.

Η Διασύνδεση Προγραμματισμού Εφαρμογών του Twitter (αγγλικός όρος : Application Programming Interface - A.P.I. εν συντομία) αποτελεί τον τρόπο με τον οποίο τα δεδομένα που προκύπτουν από τις διάφορες δημοσιεύσεις στην εφαρμογή είναι προσιτά και μπορούν να χρησιμοποιηθούν για διάφορες αναλύσεις από εταιρείες, προγραμματιστές αλλά και γενικότερα χρήστες αυτής. Επιπλέον μέσω του Twitter A.P.I. οι χρήστες έχουν τη δυνατότητα να διαχειρίζονται τις μη δημόσιες πληροφορίες του λογαριασμού τους, όπως τα προσωπικά μηνύματα (Direct Messages).

Μέσω της εγγραφής στην εφαρμογή του Twitter δύναται απευθείας η πρόσβαση στον χρήστη όσον αφορά τις δημόσιες πληροφορίες της εφαρμογής. Στα πλαίσια της εργασίας δεν χρειάστηκε κάποια άδεια για περαιτέρω δικαιώματα στα δεδομένα της εφαρμογής.

Οι πληροφορίες στις οποίες αποκτήσαμε πρόσβαση μέσω του Twitter A.P.I. [18] χωρίζονται στις παρακάτω κατηγορίες:

- Λογαριασμοί και χρήστες: διαχείριση ρυθμίσεων ενός λογαριασμού, διαγραφή ή αποκλεισμός χρηστών, διαχείριση χρηστών και οπαδών, αναζήτηση πληροφοριών σχετικά με τη δραστηριότητα συγκεκριμένων λογαριασμών είναι αρκετές από τις δυνατότητες που παρέχει η διεπαφή προγραμματισμού της εφαρμογής.
- Tweets και απαντήσεις: Δημόσια Tweets και απαντήσεις σε αυτά είναι διαθέσιμες στους προγραμματιστές. Διευκολύνεται η αναζήτηση δημοσιεύσεων με συγκεκριμένες λέξεις-κλειδιά ή ζητώντας ένα δείγμα Tweets από συγκεκριμένους λογαριασμούς. Αυτή η κατηγορία δυνατοτήτων αποτέλεσε σημαντικό εργαλείο για την

ολοκλήρωση αυτής της εργασίας και συγκεκριμένα για τον εντοπισμό, την κατανόηση και την αντιμετώπιση της παραπληροφόρησης στις διάφορες δημοσιεύσεις.

- Άμεσα μηνύματα: Επιπλέον υπάρχει η δυνατότητα πρόσβασης σε ιδιωτικές συνομιλίες των χρηστών που έχουν ρητά δώσει άδεια σε μια συγκεκριμένη εφαρμογή. Αξίζει να σημειωθεί ότι οι ιδιωτικές συνομιλίες των χρηστών (Direct Messages) δεν διατίθενται προς πώληση. Αντίθετα η πρόσβαση σε αυτά απαιτεί συγκεκριμένες άδειες και η πρόσβαση είναι περιορισμένη. Δεν χρησιμοποιήθηκαν στα πλαίσια αυτής της εργασίας.
- Διαφημίσεις: Παρέχεται μια σουίτα APIs για να μπορούν οι προγραμματιστές και οι εκάστοτε χρήστες της εφαρμογής να δημιουργήσουν και να διαχειριστούν αυτόματα διαφημιστικές καμπάνιες στο Twitter. Τα δημόσια Tweets μπορούν να χρησιμοποιηθούν για τον εντοπισμό θεμάτων και ενδιαφερόντων και να παράσχουν στις επιχειρήσεις εργαλεία για την εκτέλεση διαφημιστικών καμπανιών για να προσεγγίσουν το διαφορετικό κοινό στο Twitter.
- Εργαλεία εκδότη και Software Development Kit (S.D.K.): Τέλος παρέχονται εργαλεία για προγραμματιστές λογισμικού και εκδότες για να ενσωματώνονται τα χρονοδιαγράμματα του Twitter, τα κουμπιά κοινής χρήσης και άλλα περιεχόμενα του Twitter σε ιστοσελίδες. Αυτά τα εργαλεία επιτρέπουν τη χρήση των διαφόρων δημοσιεύσεων σε ιστοσελίδες πέραν του Twitter και διευκολύνουν τη κοινοποίηση πληροφοριών και άρθρων από και σε διάφορους ιστότοπους.

8.2 MongoDB

Η MongoDB είναι ένα σύστημα διαχείρισης βάσεων δεδομένων βασισμένο σε έγγραφα ανοιχτού κώδικα (DBMS) με ευέλικτα σχήματα.

Το έγγραφα αποθηκεύονται ως ένα σύνολο ζευγών κλειδιών / τιμών και έχουν δυναμικό σχήμα. Τα δυναμικά σχήματα σημαίνουν ότι τα έγγραφα μιας συλλογής δεν έχουν απαραιτήτως το ίδιο σύνολο πεδίων και δομών. Σημαίνει επίσης ότι τα κοινά πεδία στις συλλογές εγγράφων μπορούν να περιέχουν διαφορετικούς τύπους δεδομένων.

Η MongoDB χρησιμοποιεί έγγραφα μορφής JSON που είναι αποθηκευμένα στη δυαδική μορφή BSON. Χάρη στο πρωτόκολλο GridFS, ηMongoDB έχει τη δυνατότητα αποθήκευσης και ανάκτησης αρχείων. Με γνώμονα τα έγγραφα, η MongoDB δεν είναι

σχεσιακή βάση δεδομένων. Άρα δεν διαχειρίζεται των έλεγχο λειτουργιών όπως οι συναλλαγές. Η ατομικότητα είναι εγγυημένη μόνο σε επίπεδο εγγράφου, οπότε η μερική ενημέρωση του εγγράφου δεν μπορεί να συμβεί. Επίσης, δεν υπάρχει έννοια "απομόνωσης": οποιαδήποτε δεδομένα που διαβάζονται από έναν χρήστη μπορούν ταυτόχρονα να αλλάξουν από άλλο χρήστη, πράγμα αρκετά χρήσιμο στην περίπτωση μας μιας και κάναμε αρκετούς μετασχηματισμούς και πειραματισμούς στην μορφή των τουίτς για την ανάλυση κατά την διάρκεια της συλλογής των δεομένων όσο και μετά την ανάκτηση τους ,διατηρώντας ενεργές διαφορετικές εικόνες.

Κάποια από τα πλεονεκτήματα της MongoDB είναι ότι διατηρεί μια ιεραρχική δομή δεδομένων, έχει ευέλικτη ανάπτυξη, η αποθήκευση γίνεται σε έγγραφα (με τη μορφοποίηση JSON) , υποστηρίζει δυναμικά querieskαι έχει ανεκτικότητα σφαλμάτων και επεκτασιμότητας: με ασύγχρονη αναπαραγωγή, με σετ αντιγράφων και μια κατανεμημένη βάση δεδομένων συνδεδεμένη στους κόμβους . Υποστηρίζει επίσης την αναζήτηση πλήρους κειμένου για μορφολογική ανάλυση και την οριζόντια επεκτασιμότητα μέσω της διαμέρισης.
Το πιο βασικό πλεονέκτημα της είναι όμως ότι είναι συμβατή με έναν μεγάλο αριθμό προγραμμάτων οδήγησης και βιβλιοθηκών. Οι οδηγοί MongoDB χρησιμοποιούνται για τη σύνδεση εφαρμογών πελάτη και τη βάση δεδομένων.
Έτσι ήταν πολύ εύκολο να συνδέσουμε και να ενσωματώσουμε τα προγράμματα της Python έτσι ώστε μπορούν να λειτουργήσουν βλέποντας τη βάση δεδομένων MongoDB.
Με χρήση της βιβλιοθήκης Pymongo μάλιστα μπορούμε να αναζητήσουμε και δεδομένα μέσω της Pythonκατευθείαν από την βάση.

8.3 Pandas

Η Pandas είναι μια βιβλιοθήκη εργαλείων η οποία εξυπηρέτησε στην εξερεύνηση, το φόρτωμα , την επεξεργασία ,τον μετασχηματισμό, την ανάλυση των δεδομένων αλλά και στην μοντελοποιήση του δείγματος . Ώστε τελικά να φτάσει σε μια μορφή έτοιμο για χρήση από τους ταξινομητές αλλά και για να δουλέψουν πιο αποδοτικά άλλες βιβλιοθήκες όπως αυτές της ανάλυσης συναισθημάτων , ώστε να βγουν συμπεράσματα για το πόσο μεροληπτικοί είναι οι τίτλοι των ειδήσεων.

8.4 Scikit learn

Μια ακόμη βιβλιοθήκη που χρησιμοποιήθηκε για την κατηγοριοποίηση των τίτλων ειδήσεων είναι αυτή της scikit learn. Η οποία είναι γραμμένη σεPythonκαι αποτελεί επέκταση της επιστημονικής βιβλιοθήκης SciPy .Περιλαμβάνει διάφορες υλοποιήσεις μοντέλων μηχανικής μάθησης μεταξύ των οποίων είναι τα μοντέλα που χρησιμοποιήθηκαν στη μελέτη αυτή. Επιπλέον, διανέμεται κάτω από άδεια BSD κάτι που επέτρεψε την απρόσκοπτη χρήση της. Για τις ανάγκες της ταξινόμησης χρησιμοποιήθηκαν πολύπλοκοι ταξινομητές ,των οποίων η υλοποίηση των αντίστοιχων αλγορίθμων θα ήταν ιδιαίτερα δύσκολη , οπότε προτιμήθηκε η εύρεση των έτοιμων πακέτων της συγκεκριμένης βιβλιοθήκης.

Σύμφωνα με πολλές εργασίες και αντίστοιχες ασκήσεις τεχνητής νοημοσύνης (Pedregosa *κ.ά.*, 2012) η scikitlearnείναι η πιο χρήσιμη βιβλιοθήκη για εκμάθηση μηχανών σε Python. Περιέχει πολλά αποτελεσματικά εργαλεία για μηχανική μάθηση και στατιστική μοντελοποίηση, συμπεριλαμβανομένης της ταξινόμησης, της παλινδρόμησης, της ομαδοποίησης και της μείωσης των διαστάσεων, δηλαδή εργαλείων που χρησιμοποιήσαμε και στην εργασία μας και έχουμε περιγράψει παραπάνω.

8.5 Matplotlib

Για τις ανάγκες οπτικοποίησης των αποτελεσμάτων της μελέτης αυτής δημιουργήθηκαν γραφήματα, μέρος των οποίων περιλαμβάνεται στην ενότητα της ποιοτικής αξιολόγησης . Για την παραγωγή των γραφημάτων έγινε χρήση του πακέτου Matplotlib (Hunter, 2007).

Το πακέτο αυτό είναι υλοποιημένο σε Python και συντηρείται και επεκτείνεται από το 2003, όταν και αποδόθηκε πρώτη φορά για χρήση. Είναι ευρέως διαδεδομένο πακέτο και ανοικτό σε χρήση με δυνατότητες παραγωγής πληθώρας γραφημάτων.

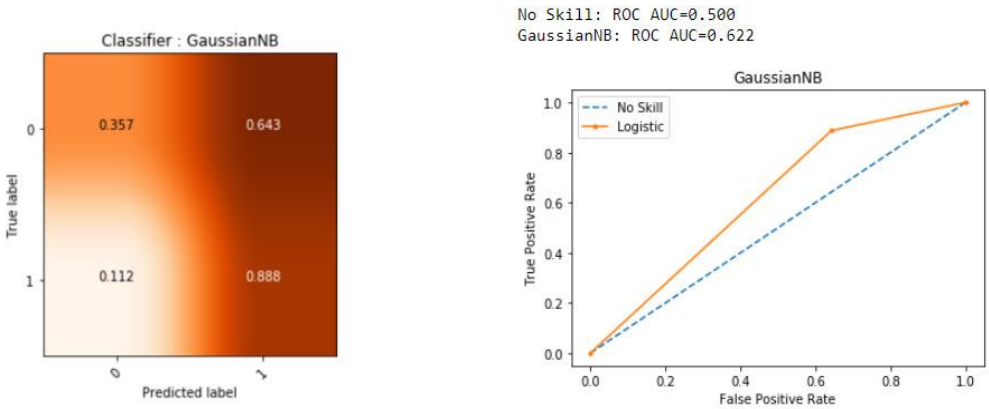
Ένα από τα μεγάλα οφέλη της απεικόνισης με την χρήση της matplotlib είναι ότι μας επιτρέπει την οπτική πρόσβαση σε μεγάλες ποσότητες δεδομένων για την εξαγωγή συμπερασμάτων με βάση εύκολα στην κατανόηση γραφήματα.

9 Ποιοτική Αξιολόγηση

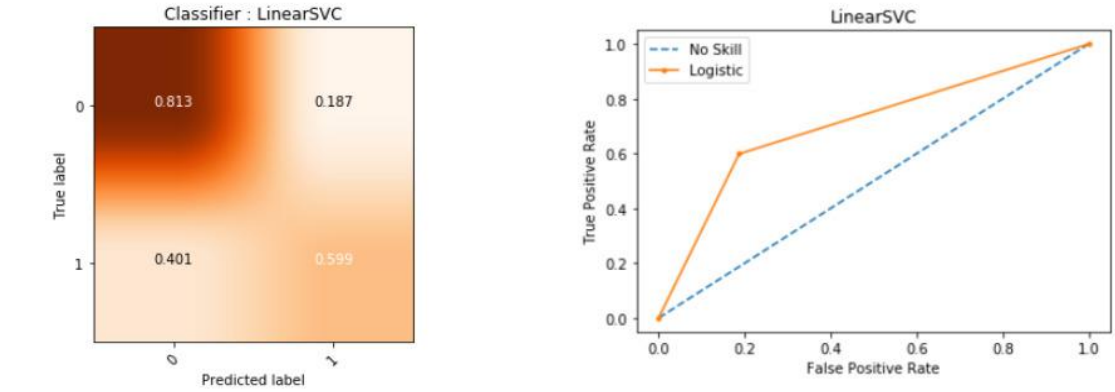
9.1 Χαρακτηριστικές Καμπύλες Λειτουργίας

Οι χαρακτηριστικές καμπύλες λειτουργίας ROC (Receiver Operation Charachteristic) είναι γραφήματα που δείχνουν την απόδοση των μοντέλων ταξινόμησης σε διάφορες ρυθμίσεις κατωφλίων. Ένα πολύ χρήσιμο χαρακτηριστικό τους είναι το μέτρο AUC (Area Under Curve) δηλαδή η επιφάνεια κάτω από την καμπύλη. Αυτό το μέτρο αντιπροσωπεύει τον βαθμό διαχωρισμού της ταξινόμησης. Όσο μεγαλύτερη είναι αυτή η επιφάνεια τόσο ισχυρότερο το μοντέλο να προβλέψει σωστά δηλαδή να διακρίνει πιο ξεκάθαρα τις περιπτώσεις clickbait από τις not-clickbait. Οι καμπύλες αυτές ορίζουν δυο παραμέτρους και σχεδιάζονται με βάση τον πίνακα σύγχυσης που εξηγήσαμε πιο πάνω .

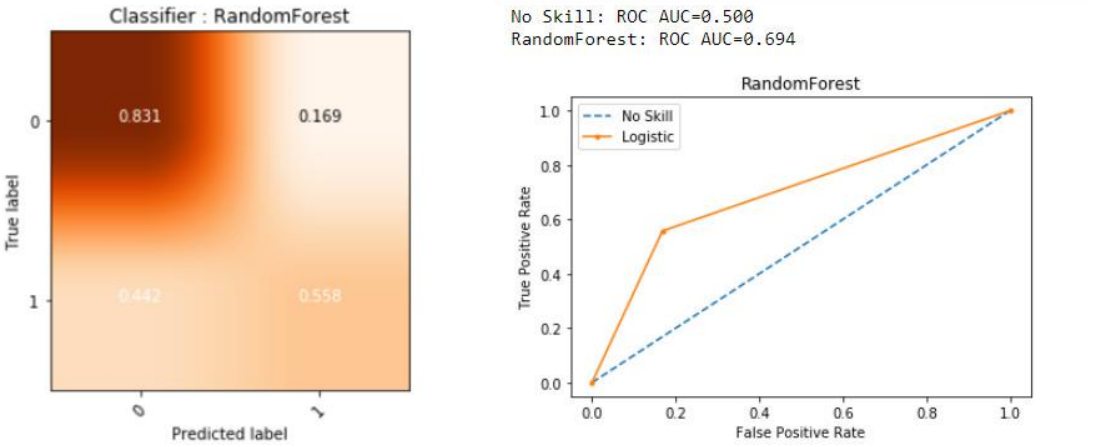
Παρακάτω βλέπουμε τις καμπύλες όπως τις σχεδιάσαμε για όλες τις περιπτώσεις ταξινομήσεων που υλοποιήσαμε μαζί με τους αντίστοιχους πίνακες σύγχυσης.



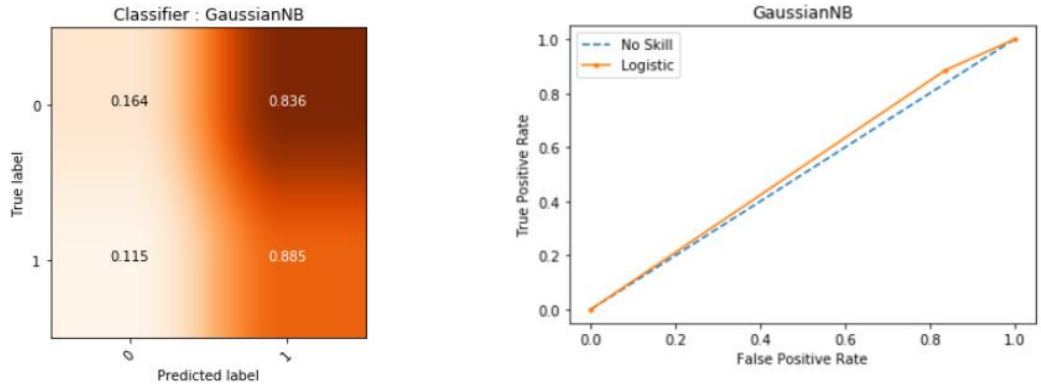
Εικόνα 1. BagOfWords / GaussianNB



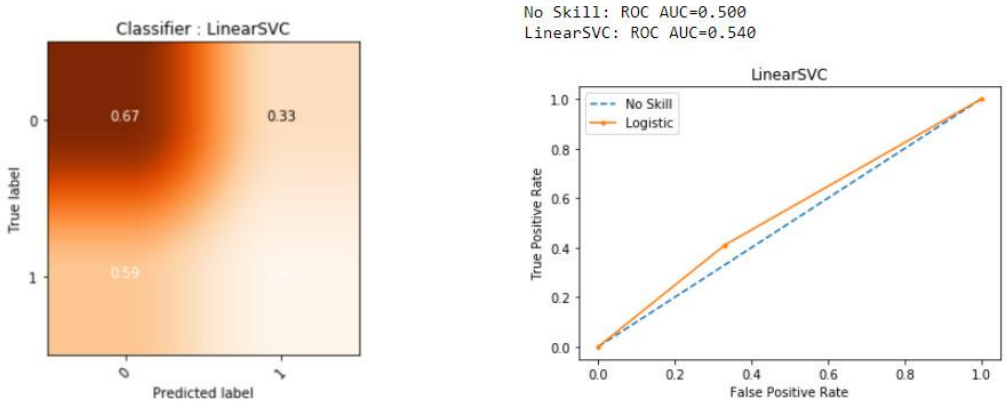
Εικόνα 2. BagOfWords / Linear SVC



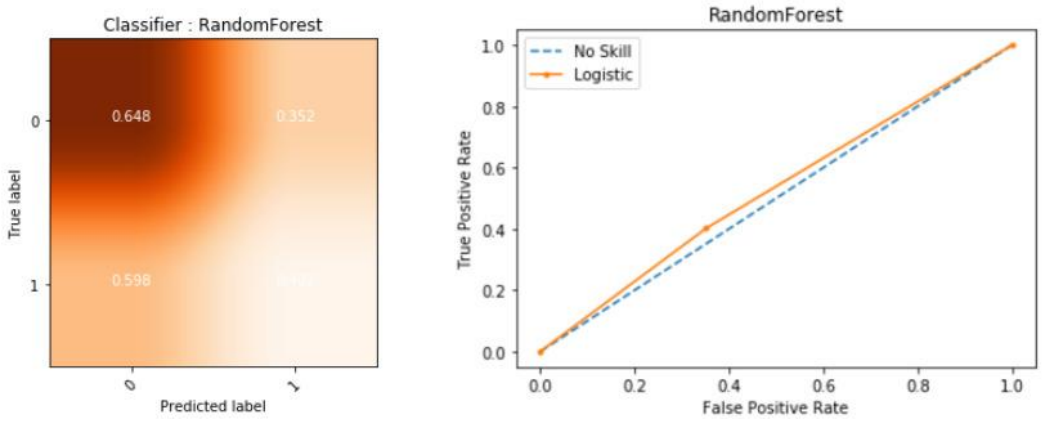
Εικόνα 3. BagOfWords / Random Forest



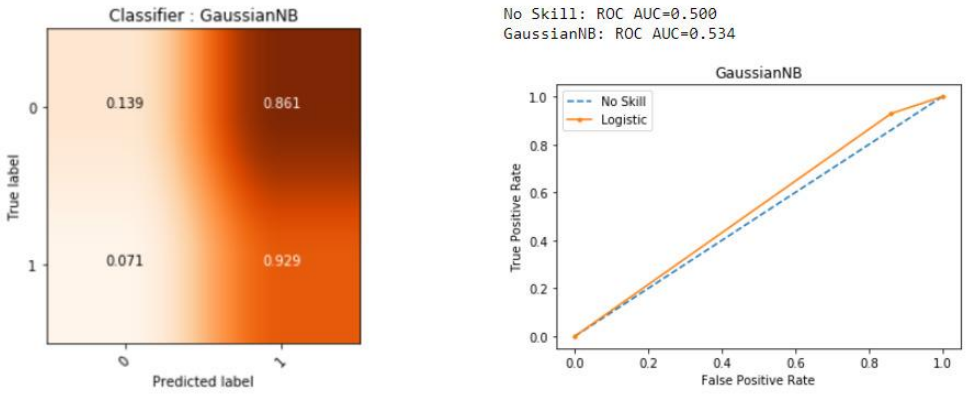
Εικόνα 4. Συχνότητα Όρων / GaussianNB



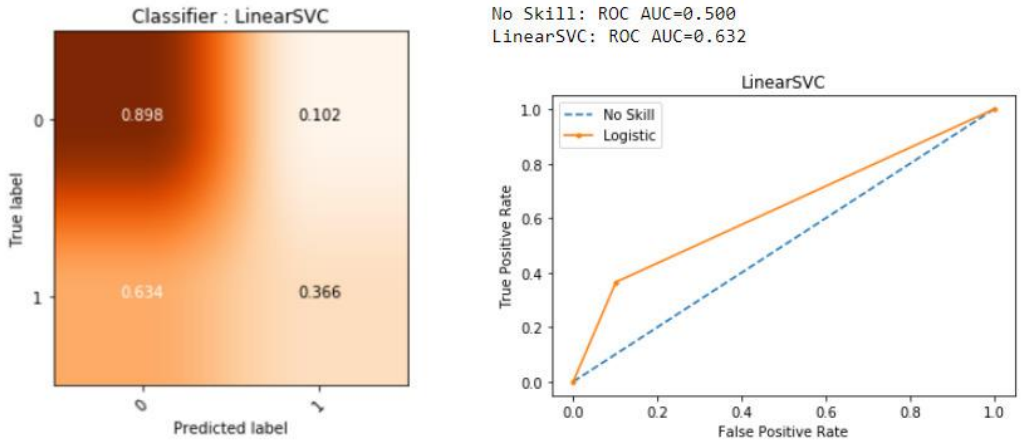
Εικόνα 5. Συχνότητα Όρων / Linear SVC



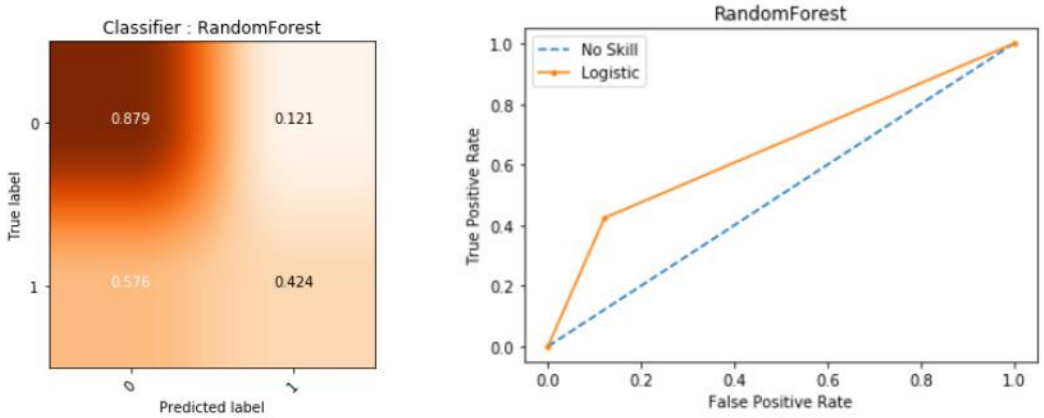
Εικόνα 6. Συχνότητα Όρων / Random Forest



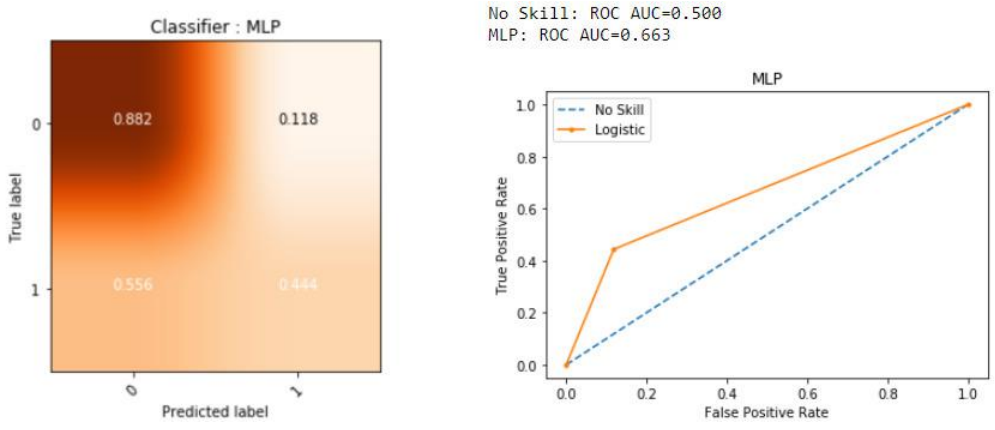
Εικόνα 7. Feature Extraction / GaussianNB



Εικόνα 8. Feature Extraction / Linear SVC



Εικόνα 9. Feature Extraction /Random Forest



Εικόνα 10. FeatureExtraction / MLP

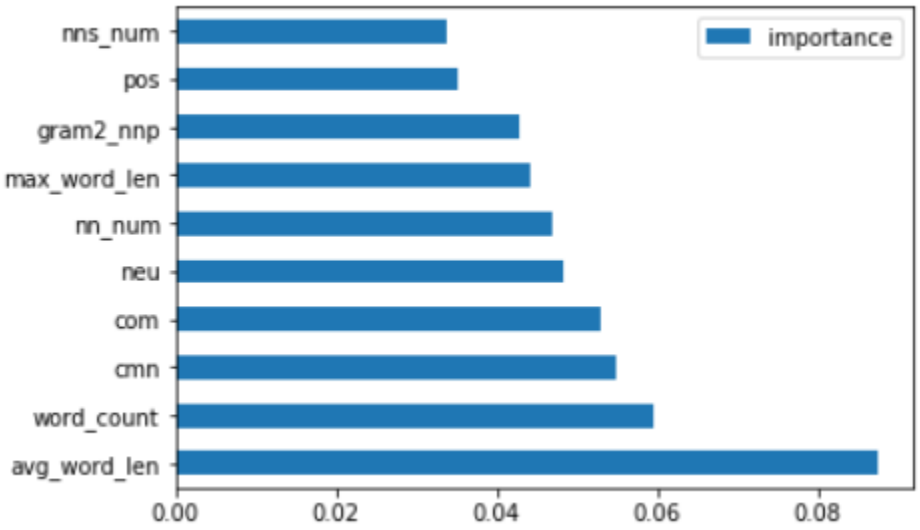
9.2 ***Σημασία*** των Χαρακτηριστικών

Ένα από τα πιο συνήθη προβλήματα της ταξινόμησης με χρήση πολλών χαρακτηριστικών είναι η αναγνώριση αυτών που έχουν άμεση συνάφεια μεταξύ τους , είναι λιγότερο σημαντικά ή και καθόλου σχετικά στο μοντέλο , με σκοπό την κατάργηση τους.

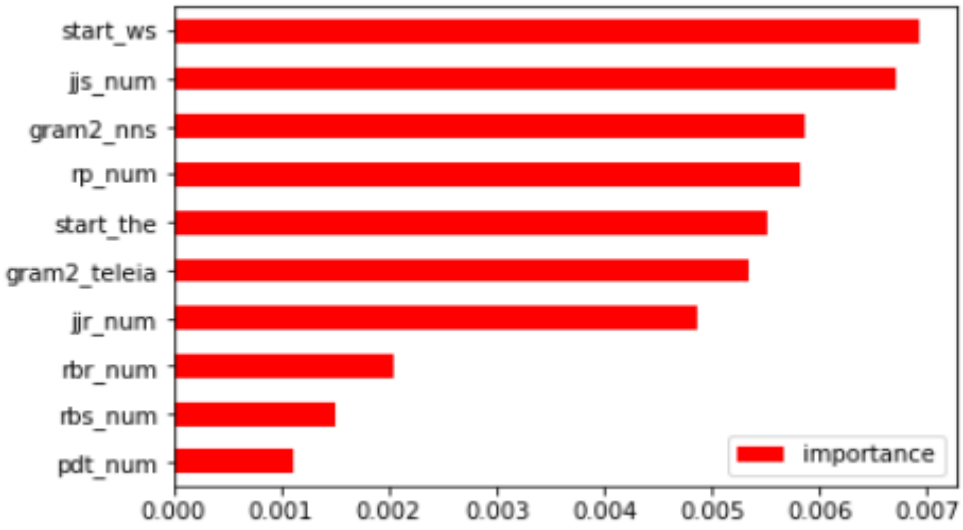
Η διαδικασία λέγεται Επιλογή Χαρακτηριστικών (Feature Selection) και είναι απαραίτητη διεργασία σε κάθε πρόβλημα Μηχανικής Μάθησης. Γίνεται και στην αρχή αλλά και στο τέλος του σχεδιασμού ενός μοντέλου , για την βελτίωση της απόδοσης του.

Μια τεχνική που μπορούμε να εφαρμόσουμε για την επιλογή χαρακτηριστικών στην περίπτωσή μας είναι να εξετάσουμε την σημαντικότητα των χαρακτηριστικών με χρήση του αντίστοιχου αλγορίθμου της βιβλιοθήκης Scikit-learn. Ο αλγόριθμος αυτός λειτουργεί αποδοτικά σε ταξινομητές που χρησιμοποιούν δένδρα αποφάσεων.

Τα δέντρα αποφάσεων εκπαιδεύονται με τέτοιον τρόπο ώστε να διαιρούν καλύτερα το σύνολο δεδομένων σε όλο και μικρότερα υποσύνολα για να προβλέψουν την σωστή κλάση. Οι συνθήκες παρουσιάζονται ως ‘φύλλα’ (κόμβοι) στο δένδρο και τα πιθανά αποτελέσματα ως ‘κλάδοι’ (άκρα). Αυτή η διαδικασία διαίρεσης συνεχίζεται μέχρις ότου να μην μπορεί να επιτευχθεί μεγαλύτερο ‘κέρδος’ ή μέχρι να ικανοποιηθεί κάποιος προκαθορισμένος κανόνας, π.χ. να φθάσει το δένδρο στο μέγιστο βάθος του.



Εικόνα 11. Χαρακτηριστικά μεγάλης σπουδαιότητας



Εικόνα 12. Χαρακτηριστικά χαμηλής σπουδαιότητας

9.3 *Όριο Διακύμανσης*

Τα χαρακτηριστικά γνωρίσματα των δεδομένων που επιλέγουμε είναι πολύ σημαντικά μιας και η εκπαίδευση στηρίζεται σε αυτά και έχουν μεγάλη επίδραση στην ακρίβεια του μοντέλου. Αυτά τα χαρακτηριστικά που δεν έχουν σχέση ή έχουν μερική σχέση με τις κατηγορίες των τάξεων μπορούν να επηρεάσουν ακόμα και αρνητικά την απόδοση.

Το μοντέλο πρέπει να εκπαιδεύεται πάντα μόνο με τα σχετικά και απαραίτητα χαρακτηριστικά. Πέρα από την βελτίωση της απόδοση, η διαδικασία της επιλογής είναι αναγκαία ούτως ώστε να μειώνουμε το φαινόμενο του overfittingto οποίο εξηγήσαμε και παραπάνω , αφού τα πλεονάζοντα χαρακτηριστικά δίνουν περισσότερες ‘ευκαιρίες’ δημιουργίας θορύβου κατά την λήψη αποφάσεων .Τέλος είναι σημαντικό να καταργήσουμε τα πλεονάζοντα χαρακτηριστικά αφού αυτά συμβάλλουν αρνητικά στον χρόνο της εκπαίδευσης, ώστε ο κώδικας να είναι πιο αποτελεσματικός.

Μια τεχνική επιλογής χαρακτηριστικών είναι αυτή που περιγράψαμε παραπάνω , δηλαδή της σημαντικότητας των χαρακτηριστικών. Μια δεύτερη τεχνική είναι αυτή του ορίου διακύμανσης.

Για να εφαρμόσουμε αυτή την τεχνική στα δικά μας δεδομένα θα χρησιμοποιήσουμε την βιβλιοθήκη της sklearn Variance Threshold (https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html) όπου επιλέγει και αφαίρει κατόπιν εφαρμογής μάσκας που φτιάξαμε εμείς τα χαρακτηριστικά που έχουν μικρή διακύμανση συγκριτικά με το δείγμα.

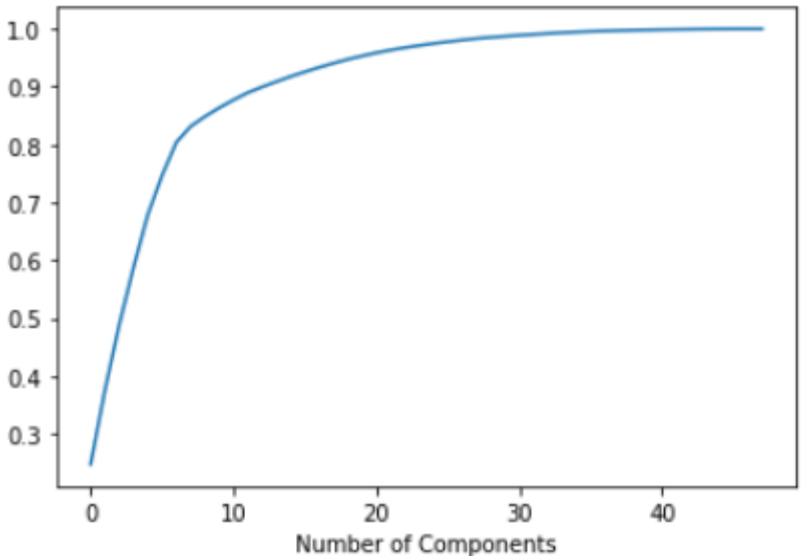
Τα αποτελέσματα που βγάλαμε από εφαρμογή αυτής της διαδικασίας είναι πως για τα δεδομένα μας το χαρακτηριστικό starts_theto οποίο εκφράζει το πόσο συχνά ξεκινάει ο τίτλος μιας είδησης με το άρθρο Theέχει μικρή σημασία στο τελικό αποτέλεσμα μιας και εμφανίζεται εξίσου συχνά σε περιπτώσεις clickbaitόσο και μη clickbait.

9.4 Ανάλυση Κύριων Συνιστωσών

Η ανάλυση σε κύριες συνιστώσες (PCA) είναι η ευρέως διαδεδομένη μέθοδος μείωσης της διαστατικότητας. Θα εξηγήσουμε περιγραφικά τις αρχές της μεθόδου. Αρχικά υπολογίζουμε τον πίνακα συσχέτισης (covariance matrix) των μεταβλητών που έχουμε στα δεδομένα. Από αυτόν τον πίνακα βρίσκουμε τις γραμμικώς συσχετισμένες μεταβλητές και βρίσκοντας τα ιδιοδιανύσματα του πίνακα μπορούμε να μετατρέψουμε τον πίνακα με έναν ορθογώνιο μετασχηματισμό και να βρούμε την βάση του νέου πίνακα. Αυτή η βάση του χώρου αποτελεί ένα νέο σύνολο μεταβλητών που είναι γραμμικά ασυσχέτιστες και ονομάζονται κύριες συνιστώσες.

Παρακάτω σχεδιάσαμε την γραφική παράσταση όπου παρουσιάζει την απόδοση του μοντέλου μας , με την χρήση και των 49 χαρακτηριστικών που εξήγαμε.

Όπως φαίνεται μετά τα πρώτα 35 χαρακτηριστικά το μοντέλο μας δεν προσφέρει σημαντική βελτίωση στην απόδοση. Κατ’αυτόν τον τρόπο και με χρηση του ορίου απόκλισης , αφαιρέσαμε κάποια εκ των χαρακτηριστικών και το αποτέλεσμα ήταν ίδιο , όπως καταγράψαμε με μια ακόμα ταξινόμηση.



Εικόνα 13. Ανάλυση Κύριων Συνιστωσών

10 Επίλογος

Ολοκληρώνοντας το ερευνητικό κομμάτι της παραπάνω μελέτης αξίζει να αναφερθούμε στην σημασία των αποτελεσμάτων που επιτεύχθηκαν με τη χρήση του συστήματος που υλοποιήθηκε καθώς και για τις μελλοντικές επεκτάσεις και βελτιώσεις που θα επέφεραν καλύτερες αποδόσεις στις ταξινομήσεις μας.

Ανακεφαλαιώνοντας , υλοποιήσαμε το παραπάνω σύστημα με σκοπό την αντιμετώπιση του φαινομένου των παραπλανητικών τίτλων ειδήσεων που μεταδίδονται από τα ειδησεογραφικά κανάλια στα κοινωνικά δίκτυα. Συνδέσαμε αυτό το κοινωνικό φαινόμενο με το επιστημονικό πεδίο της αναγνώρισης προτύπων και το αναγάγαμε σε μια πρόκληση μηχανικής μάθησης . Αρχικά προτάθηκαν αλγόριθμοι ταξινόμησης τους οποίους και περιγράψαμε παραπάνω , όπου δεν είχαν τα επιθυμητά αποτελέσματα. Στην συνέχεια χρησιμοποιήσαμε νευρωνικό δίκτυο με κατευθυνόμενο γράφο κατά μήκος χρονικής ακολουθίας, όπου κάναμε συνολική αποτίμηση χρησιμοποιώντας τις ίδιες μετρικές για την εκτίμηση της απόδοσης των μοντέλων (πίνακας σύγχυσης, πιστότητα , κλπ.) και πήραμε καλύτερα αποτελέσματα.

Το προτεινόμενο μοντέλο μετατρέπει τους τίτλους ειδήσεων σε πίνακες διανυσμάτων και κάνει την κατηγοριοποίηση των τάξεων κάνοντας ανάλυση των γλωσσικών χαρακτηριστικών τους. Μια άλλη μέθοδος που θα μπορούσε να εφαρμοστεί σαν προέκταση του παραπάνω μοντέλου θα ήταν και η ανάλυση των ίδιων των άρθρων που παραπέμπουν οι τίτλοι του δείγματός μας, η συνέπεια τους ως προς το περιεχόμενο , καθώς και η μελέτη των ψηφιακών αποτυπωμάτων των χρηστών σχετικά με το ενδιαφέρον που δείχνουν πριν και μετά το κλικ στον αμφιλεγόμενο υπερσύνδεσμο.

11 Βιβλιογραφία

[1] Chen, Y., Conroy, N. J., &Rubin, V. L. (2015). News in an Online World: The Need for an "Automatic Crap Detector". In *The Proceedings of the Association for Information Science and Technology Annual Meeting (ASIST2015)*, Nov. 6-10, St. Louis.

[2] Conroy, N. J., Chen, Y., & Rubin, V. L. (2015). Automatic Deception Detection: Methods for Finding Fake News. In *The Proceedings of the Association for Information Science and Technology Annual Meeting (ASIST2015)*, Nov. 6-10, St. Louis.

[3] Blom, Jonas & Hansen, Kenneth. (2015). *Click bait: Forward-reference as lure in online news headlines*. Journal of Pragmatics. 76. 10.1016/j.pragma.2014.11.010.

[4] Sethi, Ricky. (2017). *Spotting Fake News: A Social Argumentation Framework for Scrutinizing Alternative Facts.* 10.1109/ICWS.2017.108.

[5] J. Fu, L. Liang, X. Zhou and J. Zheng, "A Convolutional Neural Network for Clickbait Detection," 2017 4th International Conference on Information Science and Control Engineering (ICISCE), Changsha, 2017, pp. 6-10.

[6] Z. Jin, J. Cao, Y. Jiang and Y. Zhang, "News Credibility Evaluation on Microblog with a Hierarchical Propagation Model," *2014 IEEE International Conference on Data Mining*, Shenzhen, 2014, pp. 230-239.

[7] Plutchik, R. (2001). The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist, 89*(4), 344-350. Retrieved March 2, 2020, from www.jstor.org/stable/27857503

[8] Chen, Yimin & Conroy, Nadia & Rubin, Victoria. (2015). *Misleading Online Content: Recognizing Clickbait as "False News".* 10.1145/2823465.2823467.

[9] Verstraete, Mark and Bambauer, Derek E. and Yakowitz Bambauer, Jane R., *Identifying and Countering Fake News (August 1, 2017).* Arizona Legal Studies Discussion Paper No. 17-15.
SSRN: https://ssrn.com/abstract=3007971 or http://dx.doi.org/10.2139/ssrn.3007971

[10] Carlos Castillo, Marcelo Mendoza, Bárbara Poblete. Predicting Information Credibility in Time-Sensitive Social Media. *Internet Research (IR)* 23(5):560-588, Aug 2013. Emerald Group Publishing Limited. ISSN 1066-2243

[11] Bond, Charles & DePaulo, Bella. (2006). Accuracy of Deception Judgments. Personality and social psychology review: an official journal of the Society for Personality and Social Psychology, Inc. 10. 214-34. 10.1207/s15327957pspr1003_2.

[12] Kim, Yoon. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 10.3115/v1/D14-1181.

[13] Brunvand, Jan H. (2001). Encyclopedia of Urban Legends. W. W. Norton & Company. p. 194. ISBN 1-57607-076-X.

[14] Το Κουλούρι http://www.tokoulouri.com/society/toilet_paper/

[15] School of CTI DePaul University (2005) Classification via Decision Trees in WEKA http://facweb.cs.depaul.edu/mobasher/classes/ect584/weka/classify.html

[16] Omidvar, Amin et al. *"Using Neural Network for Identifying Clickbaits in Online News Media." SIMBig* (2018).

[17] T. Mikolov, A. Deoras, D. Povey, L. Burget and J. Černocký, "Strategies for training large scale neural network language models," *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, Waikoloa, HI, 2011, pp. 196-201.

[18] Twitter's Help Center : API Guidelines https://help.twitter.com/en/rules-and-policies/twitter-api

[19] Tandoc, Edson & Lim, Zheng & Ling, Rich. (2017*). Defining "Fake News": A typology of scholarly definitions*. Digital Journalism. 1-17. 10.1080/21670811.2017.1360143.

[20] Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). *Random Forests and Decision Trees.* International Journal of Computer Science Issues(IJCSI). 9.

[21] A Comprehensive Study of Linear vs Logistic Regresson to refresh the Basics https://towardsdatascience.com/a-comprehensive-study-of-linear-vs-logistic-regression-to-refresh-the-basics-7e526c1d3ebe