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

Ευφυής ανίχνευση γεγονότων και εκτίμηση αξιοπιστίας ειδήσεων στο Twitter

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΟΥ

Σταμάτιου Κατσαούνη Μολύβα

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

Αθήνα, Μάιος 2017

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

Ευφυής ανίχνευση γεγονότων και εκτίμηση αξιοπιστίας ειδήσεων στο Twitter

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Σταμάτιου Κατσαούνη Μολύβα

Επιβλέπων : Ανδρέας Γεώργιος Σταφυλοπάτης
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 25η Μαΐου 2017.

.....	.....	.....
Ανδρέας Γεώργιος Σταφυλοπάτης	Γεώργιος Στάμου	Παναγιώτης Τσανάκας
Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Μάιος 2017

.....

Κατσαούνης Μολύβας Σταμάτιος

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

Περίληψη

Τα νέα αποτελούν πληροφορίες για τρέχοντα γεγονότα. Πληροφορίες οι οποίες διαδίδονται από άνθρωπο σε άνθρωπο, με ή χωρίς γνώση για την πηγή προέλευσής τους. Η έμφυτη ανάγκη του ανθρώπου να μαθαίνει και να διαδίδει νέα τον οδήγησε από πολύ παλιά στη δημιουργία μέσων για την ανταλλαγή τους. Η πιο σύγχρονη και ταυτόχρονα ευρεία μορφή μέσου διάδοσης νέων είναι τα κοινωνικά δίκτυα. Μέσω αυτών, οι χρήστες τους διακινούν τεράστιο όγκο πληροφοριών, πολλές φορές αμφιβόλου αξιοπιστίας. Την αδυναμία τους αυτή έρχονται να εκμεταλλευτούν κακόβουλοι χρήστες αναπαράγοντας ψευδείς ειδήσεις με σκοπό την επιρροή της κοινής γνώμης. Σήμερα, περισσότερο από ποτέ, είναι υπαρκτή η ανάγκη προστασίας του αγαθού που λέγεται νέο. Ο σκοπό του προτεινόμενος συστήματος και η συμβολή του στην αντιμετώπιση τέτοιων καταστάσεων είναι διττός. Πρώτο ζητούμενο αποτελεί η αξιολόγηση των ειδήσεων ως προς την αξιοπιστία τους. Στο πλαίσιο αυτό υλοποιήθηκε σύστημα ταξινόμησης, κάνοντας χρήση μοντέλων Δέντρων Αποφάσεων και μηχανών SVM. Το σύστημα έχει την ικανότητα να εκτιμήσει ανά πάσα στιγμή και μάλιστα ιδιαίτερα αποτελεσματικά την αξιοπιστία της πηγής προέλευσης μίας δημοσίευσης. Δεύτερο ζητούμενο, συμπληρωματικό του πρώτου, είναι η τάχιστη αναγνώριση μίας είδησης ανάμεσα στη ροή πληροφοριών που διακινούνται ανά πάσα στιγμή μέσω των κοινωνικών δικτύων. Με άμεση εφαρμογή στο κοινωνικό δίκτυο Twitter, το σύστημα είναι ικανό να εντοπίσει νέες ειδήσεις σε ελάχιστο χρονικό διάστημα από την πρώτη εμφάνισή τους. Μέσω της παραπάνω διαδικασίας προσφέρει στον χρήστη άμεση ενημέρωση για γεγονότα που συμβαίνουν σε ζωντανό χρόνο χωρίς να προϋποθέτει προσπάθεια από μέρος του. Τα δύο συστατικά μέρη του συστήματος υλοποιήθηκαν εξολοκλήρου σε γλώσσα Python καθώς ενδείκνυται για ευφυή συστήματα που έχουν ως είσοδο δεδομένα σε μορφή κειμένου. Επιπλέον, υπάρχει πληθώρα σε σύγχρονα εργαλεία και βιβλιοθήκες για τη γλώσσα αυτή, τα οποία και χρησιμοποιήθηκαν προσφέροντας ποιοτικότερα αποτελέσματα στη μελέτη που πραγματοποιήθηκε. Ο πηγαίος κώδικας του συστήματος που συνοδεύει την παρούσα βρίσκεται διαθέσιμος σε δημόσιο αποθετήριο για χρήση, βελτίωση και επέκταση από όποιον ενδιαφερόμενο.

Λέξεις Κλειδιά: κοινωνικά δίκτυα, Twitter, ψευδείς ειδήσεις, ανάλυση δεδομένων, εντοπισμός ειδήσεων, Δέντρα Αποφάσεων, μηχανές SVM, scikit & NTLK Python

Abstract

News is information about current events; information which spreads from person to person, with or without knowledge of their source of origin. The inherent need of man to learn and spread news has long led him to create tools for exchanging news. The most modern and broad tool for exchanging news is social networks. Through them, users are exposed to a huge amount of information which credibility is often doubted. Malicious users often exploit that weak point of social networks by reproducing false news for the influence of public opinion. Today, more than ever, there is huge need of protecting the act of exchanging news. The contribution of the proposed system to dealing with such situations is twofold. First goal of the suggested system is the evaluation of news about their credibility. To achieve this, a classification system was implemented, using SVM Decision Models and Decision Trees. System has the ability to assess the reliability of the source of a post at most of the times and very effectively in particular. Second goal of our system, supplementary to the first, is the rapid recognition of news between the information streams exchanged through social networks. With immediate application to the Twitter social network, the system is capable of identifying news in a very short space of time since they first appeared. Through the above process, it offers users immediate information about events that occur in live time without any effort demanded from them. The two components of the system are fully implemented in Python language as it is suitable for intelligent systems dealing with text data as input. In addition, there are plenty of modern tools and libraries for this language, which were used to offer better results in the study. The source code of the system accompanying this thesis is available in a public repository for use, improvement, and extension by any interested party.

Key Words: social networks, Twitter, fake news, data analysis, first story detection, Decision Trees, SVM Decision Models, scikit & NTLK Python

Ευχαριστίες

Αφιερώνεται σε όλους όσους βοήθησαν με τον τρόπο τους στη συγγραφή της διπλωματικής αυτής.

Στην Ευρυδίκη, που με γεμίζει με τις ευχές της.

Στη Μαρία, που δεν με αφήνει στιγμή από τα μάτια της.

Στον Σπύρο, που με ενθάρρυνε στις δύσκολες στιγμές.

Στον Νίκο, που ανέχτηκε τις ιδιοτροπίες μου πολλάκις.

Στην Αλεξάνδρα, που ομορφαίνει τη ζωή μου.

Στους Αλέξιο, Γιάννο και Κωστή για τις στιγμές νηφαλιότητας που μοιραστήκαμε.

Στους Δημήτρη και Νίκο, που επιθυμούσαν διακαώς να τελειώσει.

Στη Νατάσα, που περιμένει τη σειρά της.

Στον Γιώργο, χωρίς τον οποίο δε θα ήταν τίποτα εφικτό.

Στους καθηγητές μου, που έβαλαν δύσκολα και με βοήθησαν να τα ξεπεράσω.

Σας ευχαριστώ όλους θερμά!

Πίνακας Περιεχομένων

Πίνακας Εικόνων

Πίνακας Γραφημάτων

Πίνακας Πινάκων

1 Εισαγωγή

Μέρα με τη μέρα όλο και περισσότεροι άνθρωποι χρησιμοποιούν τα κοινωνικά δίκτυα ως μέσο έκφρασης των προσωπικών τους απόψεων καθώς και για την ενημέρωσή τους για την επικαιρότητα. Η αμεσότητα της διάδοσης της πληροφορίας που προσφέρει το διαδίκτυο καθώς και η απήχηση που μπορεί να έχει μία πληροφορία που διαδίδεται σε αυτό συναρπάζει το ανθρώπινο είδος κάνοντάς το ολοένα πιο εξαρτώμενο από αυτό. Αφορμή για τη μελέτη που παρουσιάζεται στην εργασία αυτή αποτέλεσε το διαφαινόμενο αντίκτυπο που είχαν τα κοινωνικά δίκτυα στις πρόσφατες εκλογές ανάδειξης προέδρου των Ηνωμένων Πολιτειών της Αμερικής. Με αφορμή τις εκλογές αυτές η μαζική διασπορά ψευδών ειδήσεων και οι τρόποι πρόληψης από αυτήν αποτέλεσαν αντικείμενο πληθώρας συζητήσεων και διαμαχών. Προτού αναπτύξουμε τις βασικές λειτουργίες του συστήματος που προτείνουμε για τη διασφάλιση του αγαθού των ειδήσεων πραγματοποιούμε μία αναδρομή στις συνθήκες που διαμορφώνουν το περιβάλλον που έπρεπε να εργαστούμε.

1.1 Η σημασία των νέων στην ιστορία της ανθρωπότητας

Από πολύ νωρίς στην ιστορία του ανθρώπινου είδους, τα νέα και η διάδοσή τους κατείχαν σημαντική θέση στην κουλτούρα διάφορων πολιτισμών. Χαρακτηριστικό παράδειγμα αποτελούν οι φυλές της Αμερικής, οι οποίες όχι μόνο είχαν ειδικό χώρο για ανταλλαγή ιστοριών μεταξύ των ανθρώπων αλλά ακολουθούσαν την τακτική να ανακρίνουν τους ταξιδιώτες για την εκμάθηση νέων. Με τον τρόπο αυτό, τα σημαντικότερα νέα διαδίδονταν με μεγάλη ταχύτητα από στόμα σε στόμα σε πολύ μεγάλες γεωγραφικές εκτάσεις. (Stephens, 1988).

1.2 Από την τυπογραφία στο διαδίκτυο, από την ειδησεογραφία στα κοινωνικά δίκτυα

Καθώς η ανθρωπότητα εξελισσόταν η διάδοση ειδήσεων, αληθών και ψευδών διαρκώς και διαφοροποιούταν τόσο ως προς τον τρόπο όσο και ως προς τα μέσα. Εφημερίδες, ραδιόφωνο και τηλεόραση αποτελούν μέχρι και τις μέρες μας σημαντική πηγή για νέα γεγονότα. Η μεγαλύτερη επανάσταση όμως γίνεται με την εφεύρεση του διαδικτύου. Η ασύγκριτα μεγαλύτερη ταχύτητα που προσφέρει λόγω της φύσης του προσελκύει όλο και περισσότερους ανθρώπους να το επιλέγουν ως μέσο ενημέρωσής τους. Ένα ακόμα δυνατό στοιχείο του διαδικτύου είναι η καθολικότητα που προσφέρει, καθώς σε πραγματικό χρόνο, άνθρωποι από όλο τον πλανήτη μπορούν να επικοινωνήσουν χωρίς το παραμικρό εμπόδιο.

Το παραδοσιακό μοντέλο διάδοσης ειδήσεων μέσω του διαδικτύου αποτελούταν από δημοσιογραφικές ιστοσελίδες, τις οποίες μπορεί να επισκεφτεί κάποιος και να λάβει γνώση για ειδήσεις σε παγκόσμιο επίπεδο. Η ανάγκη όμως του ανθρώπου να συμμετέχει ο ίδιος στην είδηση, να μπορεί να την αναπαράγει και να τη διαδίδει δημιουργήσει νέες μορφές ανταλλαγής πληροφοριών. Την τελευταία δεκαετία ολοένα και περισσότεροι άνθρωποι γίνονται χρήστες κοινωνικών δικτύων. Τα κοινωνικά δίκτυα είναι ιστότοποι όπου χρήστες μπορούν είτε με πλήρη είτε με σχετική ελευθερία να διακινούν πληροφορίες μεταξύ τους. Η δυνατότητα διαμοίρασης περιεχομένου ταυτόχρονα με εκατομμύρια χρήστες και η ευκολία

επικοινωνίας κάνουν ιδιαίτερα θελκτικά τα κοινωνικά δίκτυα, με αποτέλεσμα να αυξάνουν διαρκώς το μέγεθος και τη δυναμικότητά τους.

1.3. Ψευδείς Ειδήσεις και χειραγώγηση στα κοινωνικά δίκτυα

Ο τεράστιος όγκος δεδομένων και χρηστών των κοινωνικών δικτύων δυσκολεύει κατά πολύ τον έλεγχο της πληροφορίας που διακινείται εντός τους. Το γεγονός αυτό δίνει τη δυνατότητα σε κακόβουλους χρήστες να κινούνται εντός των κοινωνικών δικτύων και να προσπαθούν να χειραγωγήσουν τους χρήστες τους. Σκοπός ενός κακόβουλου χρήστη είναι είτε να παραπλανήσει είτε να καθοδηγήσει την κοινή γνώμη για ίδιο όφελος. Για να το επιτύχει πολλές φορές προσποιείται προσωπικότητες κοινής αποδοχής και δημοσιεύει πληροφορίες τις οποίες αποδίδει αργότερα σε αυτές. Άλλη μορφή εξαπάτησης είναι η δημοσίευση ψευδών ειδήσεων μέσω λογαριασμών με εικονικά προσωπικά στοιχεία. Η εξάρτηση των σύγχρονων κοινωνιών από τα κοινωνικά δίκτυα ανάγει τη διασπορά ψευδών ειδήσεων σε θέμα μείζονος σημασίας, ικανό να προκαλέσει κινδύνους στη λειτουργία τους.

1.4 Συμβολή της τεχνολογίας στην προστασία των χρηστών του διαδικτύου

Η ανθρώπινη προσπάθεια δεν αρκεί από μόνη της στην καταπολέμηση φαινομένων παραπλάνησης και χειραγώγησης στα κοινωνικά δίκτυα. Ο όγκος των δεδομένων και η ανάγκη για άμεση ανταπόκριση οδηγούν τους ερευνητές στην υλοποίηση ευφυών συστημάτων ικανών να επεξεργάζονται ροές δεδομένων. Τέτοιου είδους συστήματα χρησιμοποιούνται τόσο για τον εντοπισμό ειδήσεων όσο και για την αξιολόγηση του διακινούμενου περιεχομένου. Προκειμένου τα συστήματα αυτά να είναι όσο το δυνατόν πιο αποδοτικά απαιτείται ανάπτυξη τεχνικών εξόρυξης και επεξεργασίας δεδομένων.

1.5. Εξόρυξη Δεδομένων και εξαγωγή συμπερασμάτων

Εξόρυξη δεδομένων είναι η διαδικασία μέσω της οποίας ένα σύστημα μπορεί και εξάγει νέα δεδομένα από ήδη υπάρχουσα. Πρόκειται για ένα συνδυασμό ευφυίας και στατιστικής που βοηθάει τον άνθρωπο να κατανοήσει καλύτερα τα δεδομένα με τα οποία αλληλοεπιδρά. Πολλές φορές μέσω της διαδικασίας εξόρυξης παράγεται νέα γνώση καθώς βοηθάει στην εξάλειψη των περιορισμών και δίνει τη δυνατότητα νέας προοπτικής των ίδιων δεδομένων. Ως διαδικασία είναι στενά συνδεδεμένη με τα συστήματα που μελετούν κοινωνικά δίκτυα ενώ είναι ένας κλάδος που διαρκώς εξελίσσεται. Στην περίπτωση της μελέτης που πραγματοποιήθηκε η εξόρυξη δεδομένων μέσα από δημοσιεύσεις του Twitter ήταν από τα πρώτα ζητούμενα που μας απασχόλησαν.

1.6 Εντοπισμός ταχέως διαδιδόμενων θεμάτων και αναγνώριση ψευδών ειδήσεων

Δύο από τα κυριότερα ζητήματα που ενδιαφέρουν τους ερευνητές των κοινωνικών δικτύων είναι ο εντοπισμός ταχέως διαδιδόμενων θεμάτων και η αναγνώριση των ψευδών ειδήσεων. Κάθε ένας από τους δύο αυτούς κλάδους είναι ιδιαίτερα ενεργός και διαρκώς προτείνονται διαφορετικές προσεγγίσεις και συστήματα που τις υλοποιούν. Στις ενότητες που ακολουθούν, γίνεται προσπάθεια καταγραφής των πιο σύγχρονων συστημάτων και για τους

δύο τομείς. Τα συστήματα αυτά και οι προσεγγίσεις που πραγματοποιούν αποτέλεσαν πηγή έμπνευσης για την εκπόνηση της μελέτης αυτής.

1.7 Προτεινόμενο Σύστημα

Το σύστημα που προτείνεται αποτελείται από δύο συστατικά μέρη. Το πρώτο, αφορά τον χαρακτηρισμό ειδήσεων ως προς το περιεχόμενό τους. Στο υποσύστημα αυτό αφότου συλλεχθούν δημοσιεύσεις του Twitter, παράγονται αριθμητικές τιμές ανάλογα με τα χαρακτηριστικά τους. Με τη βοήθεια μοντέλων ταξινομητών, ανάλογα τις τιμές για τα χαρακτηριστικά μίας δημοσίευσης, το σύστημα αποφαίνεται για την κατηγορία χρηστών στην οποία ανήκει ο δημιουργός της. Κατόπιν μελέτης επιλέχθηκαν τα χαρακτηριστικά και το μοντέλο εκείνο που μεγιστοποίησαν το ποσοστό επιτυχίας των προβλέψεων του συστήματος. Το σύστημα, στην καλύτερη περίπτωση, εμφανίζει ποσοστό επιτυχίας μεγαλύτερο του 75%, ταξινομώντας ορθά 3 στις 4 δημοσιεύσεις. Το δεύτερο υποσύστημα μελετά την ταχύτητα διάδοσης όρων ανά τακτά χρονικά παράθυρα κατά τη ροή δεδομένων του Twitter. Υλοποιώντας διάφορες τεχνικές, άλλες προτεινόμενες στη βιβλιογραφία και άλλες πρωτότυπες, επιδιώκει τον εντοπισμό ταχέως διαδιδόμενων όρων. Στην προσπάθεια αξιολόγησής του, καταγράψαμε το χρονικό παράθυρο στο οποίο εντόπισε σημαντικά γεγονότα της περιόδου μελέτης. Τα αποτελέσματα που επιτεύχθηκαν υποδεικνύουν μεγάλη επίδοση εμφανίζοντας μηδενικές αποκλίσεις από τις πρώτες εμφανίσεις των υπό μελέτη γεγονότων στη ροή των δημοσιεύσεων.

2 Υλοποίηση

Στην ενότητα αυτή καταγράφονται οι τεχνολογίες και τα πακέτα λογισμικού που χρησιμοποιήθηκαν για της ανάγκες της μελέτης. Περιλαμβάνονται πληροφορίες για κάθε μία από τις τεχνολογίες, οι λόγοι που οδήγησαν στη χρήση τους καθώς και ο τρόπος που χρησιμοποιήθηκαν για την υλοποίηση του συστήματος που προτείνει η μελέτη αυτή.

Η υλοποίηση του προτεινόμενου συστήματος έγινε εξολοκλήρου σε γλώσσα Python και περιλαμβάνει τόσο παραγωγή ιδίου κώδικα όσο χρήση μεθόδων διαθέσιμων πακέτων λογισμικού. Εκτός της μεγάλης εξοικείωσης με τη συγκεκριμένη γλώσσα, υπήρξαν αρκετοί λόγοι που οδήγησαν στη χρήση της.

Σημαντικό κριτήριο στην επιλογή της Python ήταν τα διαθέσιμα εργαλεία ανοικτού κώδικα που αφορούν επεξεργασία φυσικής γλώσσας και αλγορίθμους μηχανικής μάθησης. Τόσο το scikit-learn όσο και το NLTK προσφέρουν πολύ ισχυρά εργαλεία, τα οποία και χρησιμοποιήθηκαν σε μεγάλο βαθμό. Επιπλέον, μεγάλο μέρος του συστήματος αφορά επεξεργασία συμβολοσειρών και προσπέλαση σε λεξικά κλειδιών-τιμών. Η python είναι μία γλώσσα πολύ ισχυρή πάνω στον τομέα αυτό καθώς ενσωματώνει λ-λογισμό και διαθέτει ιδιαίτερα χρήσιμες και αποδοτικές δομές. Τέλος, τόσο το twitter API όσο και η MongoDB είναι ιδιαίτερα φιλικά ως προς τη Python μέσω των διαθέσιμων υλοποιήσεών τους στη γλώσσα αυτή.

2.1 Twitter API

Για τις ανάγκες της μελέτης έπρεπε να δημιουργηθεί ένα σύνολο δεδομένων αποτελούμενο από δημοσιεύσεις χρηστών σε κοινωνικά δίκτυα. Από την πληθώρα δεδομένων που προσφέρονται από τα μεγαλύτερα σε χρήση και αναγνώριση κοινωνικά δίκτυα επιλέχθηκε το Twitter ως πηγή για άντληση δεδομένων. Σημαντικό ρόλο στην επιλογή του έπαιξε το API που προσφέρει σε τρίτους για πρόσβαση στα δεδομένα που φιλοξενεί (Twitter Developers, 2013).

Το Twitter, μέσω του API του, προσφέρει σε διαπιστευμένους χρήστες πρόσβαση στα δεδομένα που φιλοξενεί. Τα δεδομένα αναπαριστώνται σε JSON και η δομή τους είναι προκαθορισμένη βάσει σύμβασης που ακολουθείται και περιγράφεται από το Twitter. Επιπλέον, μέσα στα δεδομένα εκτός της ίδιας της δημοσίευσης περιέχονται και μεταδεδομένα που έχει καταγράψει το Twitter όπως η τοποθεσία, ο αριθμός των αναδημοσιεύσεων κ.ο.κ.

Από το σύνολο του API έγινε χρήση δύο πολύ βασικών μεθόδων για ανάκτηση δεδομένων: GET friends/ids και GET statuses/user_timeline. Μέσω αυτών έγινε συλλογή αντικειμένων JSON, τα οποία αποθηκεύτηκαν για τη μετέπειτα μελέτη. Η πρώτη μέθοδος επιστρέφει τα μοναδικά αναγνωριστικά όλων των “φίλων” κάποιου χρήστη σε μορφή λίστας. Η δεύτερη μέθοδος, δεδομένου του αναγνωριστικού κάποιου χρήστη, επιστρέφει όλες τις δημοσιεύσεις που έχουν πραγματοποιηθεί μέσα από τον λογαριασμό του και βρίσκονται αναρτημένες στο χρονολόγιό του.

Η συλλογή δεδομένων πραγματοποιήθηκε με κλήσεις στα εν λόγω API services από δύο διαφορετικούς διαπιστευμένους λογαριασμούς λόγω των περιορισμών που θέτουν όσο αφορά τη χρήση τους. Πιο συγκεκριμένα, για λόγους ασφάλειας αλλά και απόδοσης της υποδομής του, το Twitter εφαρμόζει άνω όρια ως προς τη χρήση των υπηρεσιών που προσφέρει το API του. Ο πρώτος περιορισμός είναι ο αριθμός των κλήσεων που μπορούν να γίνουν μέσα σε διάστημα ενός τετάρτου της ώρας. Ο δεύτερος είναι ο μέγιστος αριθμός των δημοσιεύσεων που μπορούν να επιστραφούν ανά κλήση. Τελευταίος περιορισμός είναι ο αριθμός των δημοσιεύσεων που μπορεί να ανακτηθούν από το παρελθόν για κάθε έναν από τους χρήστες. Σε περίπτωση που τα όρια αυτά εξαντληθούν χρειάζεται ειδική πρόβλεψη ώστε το σύστημα να συνεχίσει απρόσκοπτα τη λειτουργία του.

Ο μηχανισμός που υλοποιήθηκε προσέφερε την αδιάκοπη τροφοδότηση του συστήματος με δημοσιεύσεις στοχευμένων λογαριασμών χρηστών και αποθήκευσής τους σε βάση δεδομένων. Ο μηχανισμός τέθηκε σε λειτουργία σε εικονική μηχανή server, φιλοξενούμενη στην υποδομή της υπηρεσίας ~okeanos. Υπήρξε ειδική μέριμνα κατά την εξάντληση των ορίων των υπηρεσιών του Twitter, εισάγοντας δεκαπεντάλεπτη καθυστέρηση και πραγματοποιώντας κυκλική ανάκτηση δημοσιεύσεων των χρηστών - στόχων. Ειδική μέριμνα υπήρξε και στην αποφυγή καταγραφής δημοσίευσης που είχε ήδη ανακτηθεί σε προηγούμενο έλεγχο.

2.2 MongoDB

Προκειμένου η ανάλυση των tweets να είναι πληρέστερη και να ξεπεραστούν οι περιορισμοί που εισάγονται από το twitter API, τα tweets που συλλέγονται πρέπει να αποθηκεύονται σε μία βάση δεδομένων.

Μία πρώτη προσέγγιση για την αποθήκευση είναι η χρήση παραδοσιακής σχεσιακής βάσης δεδομένων όπως είναι η mySQL (MySQL AB, 2002). Μία σχεσιακή βάση δεδομένων αποτελείται από συσχετιζόμενους πίνακες και τούπλες δεδομένων που αποθηκεύονται σε αυτούς. Προκειμένου να γίνει εφικτή η χρήση μίας τέτοιας βάσης δεδομένων θα πρέπει να υλοποιηθεί σχετικός μηχανισμός, ο οποίος θα επεξεργάζεται τα δεδομένα ενός tweet, θα εντοπίζει σε ποιες στήλες της βάσης δεδομένων αντιστοιχίζονται και θα καταγράφει τις τιμές τους στη βάση δεδομένων.

Το παραπάνω, αν και εφικτό, δεν αποτελεί καλή σχεδιαστική επιλογή για αποθήκευση δεδομένων προερχόμενων από το twitter API. Πιο συγκεκριμένα, ένα tweet, ή καλύτερα τα δεδομένα από τα οποία αποτελείται, επιστρέφονται σε μορφή JSON. Η JSON αναπαράσταση αν και απλή στον ορισμό της (λεξικό με μοναδικά κλειδιά που αντιστοιχίζονται σε τιμές), μπορεί να έχει ιδιαίτερα σύνθετη δομή. Στην περίπτωση των δεδομένων ενός tweet αυτό γίνεται πολύ εμφανές. Ένα tweet σε αναπαράσταση JSON πέραν του πλήθους το κλειδιών του, έχει πολλές κενές τιμές. Επιπλέον, είναι πιθανόν να βρίσκονται εμφολευμένα εντός του άλλα αντικείμενα σε αναπαράσταση JSON, κάτι που είναι πλήρως αποδεκτό από τον ορισμό της εν λόγω αναπαράστασης.

Για όλους αυτούς τους λόγους πρέπει να στραφούμε σε διαφορετική λύση από την αυστηρή δομή μίας σχεσιακής βάσης δεδομένων. Η βάση δεδομένων MongoDB (MongoDB Inc., 2016) και οι δυνατότητες που προσφέρει βοήθησαν στο να ξεπεραστούν όλα τα προβλήματα που προκύπτουν από τη μορφή των δεδομένων του προβλήματος που μελετήσαμε. Η MongoDB ως μία NoSQL βάση δεδομένων είναι απαλλαγμένη από πίνακες και σχέσεις μεταξύ αυτών. Επιπλέον, πρόκειται για μία document-based βάση δεδομένων που προσφέρει τη δυνατότητα αποθήκευσης αρχείων εντός της. Τα αρχεία αυτά έχουν μορφή BJSON, δηλαδή JSON σε δυαδική αναπαράσταση.

Η ιδιότητα της MongoDB να διαχειρίζεται BJSON αρχεία κάνει την αποθήκευση ενός tweet ιδιαίτερα απλή αφού αρκεί ένα INSERT για να αποθηκευτεί αυτούσιο, χωρίς να απαιτείται ο παραμικρός έλεγχος για την πληρότητα των δεδομένων του. Η ανάκτηση του tweet είναι εξίσου εύκολη καθώς η MongoDB το επιστρέφει ακριβώς όπως αποθηκεύτηκε. Η MongoDB επιτρέπει επίσης την ταχεία αναζήτηση εντός των κλειδιών ενός αποθηκευμένου tweet κάτι που την κάνει ιδανική για μελέτη δεδομένων προερχόμενων από το twitter API.

Φυσικά, η τάχιστη αποθήκευση, ανάκτηση και αναζήτηση έρχονται με κάποιο τίμημα. Η MongoDB προκειμένου να τα προσφέρει απαιτεί δημιουργία μεγάλων ευρετηρίων. Αυτό έχει ως αποτέλεσμα να καταλαμβάνει μεγάλο χώρο στον δίσκο, δημιουργώντας περιορισμούς καθώς τα δεδομένα που καλείται να διαχειριστεί κλιμακώνονται. Αυτό όμως δεν αποτελεί πρόβλημα στη μελέτη του προβλήματος. Το δυνατό σημείο της MongoDB είναι η αποθήκευση των δεδομένων όπως ακριβώς προσφέρονται από το twitter API. Σε ένα ολοκληρωμένο σύστημα μπορεί να χρησιμοποιηθεί για προσωρινή και όχι για μόνιμη

αποθήκευση δεδομένων. Μπορεί δηλαδή να διατηρεί τα δεδομένα των tweets που λαμβάνονται μέχρι αυτά να καταναλωθούν από το επόμενο component.

Στη μελέτη μας ο αποθηκευτικός χώρος ήταν επαρκής και αποφασίστηκε να χρησιμοποιηθεί εξολοκλήρου ως αποθηκευτικό μέσο, καθώς κρίθηκε αναγκαία η χρήση της άμεσης ανάκτησης και αναζήτησης στα δεδομένα που συλλέχθηκαν.

2.3 Scikit Learn

Προκειμένου να γίνει εφικτή η κατηγοριοποίηση των tweets κρίθηκε αναγκαία η χρήση ταξινομητών Δέντρων Αποφάσεων και Support Vector Machines. Η πολυπλοκότητα των μοντέλων αυτών έκανε ιδιαίτερα δύσκολη την υλοποίησή τους από μηδενική βάση κάτι που οδήγησε σε αναζήτηση έτοιμων υλοποιήσεων. Μεταξύ των διαθέσιμων πακέτων λογισμικού προτιμήθηκε το scikit-learn (Pedregosa *κ.ά.*, 2012).

Το scikit-learn, πρόκειται για ένα πακέτο γραμμένο σε Python, το οποίο 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 του.

Οι λόγοι που οδήγησαν στη χρήση του πακέτου scikit-learn είναι πολλαπλοί. Αρχικά, η χρήση του είναι ευρέως διαδεδομένη στους μελετητές μηχανικής μάθησης με πάνω από 17 χιλιάδες αστέρια στο δημόσιο αποθετήριο κώδικα github.com και πάνω από 5 χιλιάδες αναφορές της επιστημονικής δημοσίευσης που το συνοδεύει. Επιπλέον πρόκειται για ένα project με δημόσια προσβάσιμο κώδικα και συνεχή συντήρηση και επέκταση, γεγονός που μειώνει αισθητά τις πιθανότητες λάθους στην υλοποίηση των μοντέλων. Τέλος, έχει λάβει χρηματοδότηση και υποστήριξη από μεγάλους φορείς και ιδρύματα, κάτι που προσδίδει κύρος και αναγνώριση.

Για τις ανάγκες της μελέτης χρησιμοποιήθηκαν οι παρακάτω μέθοδοι του πακέτου scikit-learn:

- KFold: κατάτμηση των δεδομένων σε υποδείγματα
- F1_score: μετρική απόδοσης πρόβλεψης f1
- DecisionTreeClassifier: ταξινομητής Δέντρου Απόφασης
- SVC: ταξινομητής SVC, υποκατηγορίας ταξινομητών SVM

2.4 NLTK

Μεγάλο μέρος της μελέτης αφιερώθηκε στην επεξεργασία του φυσικού κειμένου των tweet που συλλέχθηκαν. Μέρος της επεξεργασίας ήταν η αναγνώριση μερών του λόγου, η εύρεση του λήμματος των όρων του κειμένου και η αφαίρεση εξαιρούμενων λέξεων. Όλα τα παραπάνω δεν θα μπορούσαν να πραγματοποιηθούν χωρίς τη βοήθεια του πακέτου NLTK (Bird και Loper, 2004).

Το πακέτο NLTK, εξίσου γραμμένου σε Python, είναι το πιο ευρέως διαδεδομένο πακέτο για επεξεργασία φυσικής γλώσσας. Από το 2001 που δημιουργήθηκε, συντηρείται και επεκτείνεται διαρκώς από τους δημιουργούς του. Επιπλέον είναι project ανοικτού κώδικα, δίνοντας έτσι τη δυνατότητα στον οποιοδήποτε το επιθυμεί να διορθώσει τυχόν σφάλματα ή να το επεκτείνει. Η ευρέως διαδεδομένη χρήση του και οι περίπου 200 συντελεστές του, το καθιστούν ιδανικό για χρήση.

Για τις ανάγκες της μελέτης χρησιμοποιήθηκαν οι παρακάτω μέθοδοι του πακέτου NLTK:

- WordNetLemmatizer: Εύρεση λήμματος όρου
- Pos_tag: Χαρακτηρισμός μέρους του λόγου
- RegexpTokenizer: Κατάτμηση συμβολοσειρές σε όρους
- SentimentIntensityAnalyzer: ανάλυση συναισθήματος πρότασης
- Stopwords: αφαίρεση εξαιρούμενων λέξεων

2.5 Matplotlib

Για τις ανάγκες οπτικοποίησης των αποτελεσμάτων της μελέτης αυτής δημιουργήθηκαν γραφήματα, μέρος των οποίων περιλαμβάνεται στην Ενότητα 4 (Αναγνώριση Γεγονότων και εξαγωγή θεμάτων από τη ροή δεδομένων του Twitter). 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.

Τα γραφήματα που δημιουργήθηκαν παρουσιάζουν το πλήθος και την ταχύτητα διάδοσης όρων που περιέχονται σε tweets κατά το πέρας χρονικών διαστημάτων.

3 Αναγνώριση ψευδών ειδήσεων στο Twitter

Το πρώτο μέρος του συστήματος που προτείνεται αφορά την αναγνώριση ψευδών ειδήσεων μέσα από τις δημοσιεύσεις χρηστών του Twitter. Σε ένα κόσμο που εξαρτάται όλο και περισσότερο από τη χρήση του διαδικτύου, η αξιολόγηση των πληροφοριών που φιλοξενούνται σε αυτό αποκτά όλο και πιο βαρύνουσα σημασία. Οι χρήστες του διαδικτύου βρίσκονται καθημερινά εκτεθειμένοι σε πληθώρα πληροφορίας αμφιβόλου πηγής και εγκυρότητας.

Επιδίωξη του προτεινόμενου συστήματος είναι η επιτυχής αναγνώριση μίας έγκυρης είδησης έναντι μίας ψευδής. Για να το επιτύχει, βασίζεται στα κοινά χαρακτηριστικά που προσδιορίζουν κάθε είδος είδησης και προσπαθεί να τα συγκεντρώσει και να τα διακρίνει.

Κάθε δημοσίευση προερχόμενη από το Twitter μετατρέπεται σε ένα διάνυσμα τιμών προερχόμενες από τα χαρακτηριστικά που τη διακρίνουν. Στη συνέχεια, ένα μοντέλο ταξινομητή τροφοδοτείται με τα διανύσματα των ειδήσεων και επιχειρεί να προβλέψει το είδος της κάθε μίας. Μέσω της πρόβλεψης που παράγει προειδοποιεί τον αναγνώστη της είδησης για το κατά πόσο αυτή είναι έγκυρη.

Στις παραγράφους που ακολουθούν, αφού αναφερθούν σύγχρονες προσεγγίσεις επί του θέματος αναγνώρισης ειδήσεων, παρουσιάζονται τα χαρακτηριστικά του συστήματος αναγνώρισης που προτείνεται. Τέλος, ακολουθούν τα αποτελέσματα της προσομοίωσης που πραγματοποιήθηκε από τα οποία μπορούν να βγουν συμπεράσματα ως προς την εγκυρότητα και την απόδοση του συστήματος.

3.1 Σύγχρονες μέθοδοι και προσεγγίσεις

Σκοπός της ενότητας αυτής είναι να παρουσιαστούν σύγχρονες μέθοδοι και προσεγγίσεις χαρακτηρισμού ειδήσεων στα κοινωνικά δίκτυα. Οι προσεγγίσεις που ακολουθούνται διαφοροποιούνται μεταξύ τους, έχουν όμως όλες ως κοινό σκοπό τον εντοπισμό ψευδών ειδήσεων. Πλην μίας, χρησιμοποιούν σαν δεδομένα εισόδου δημοσιεύσεις του Twitter, καθώς η μορφή του κοινωνικού δικτύου είναι τέτοια που προσφέρεται για το αντικείμενο της μελέτης.

Με αφορμή τον σεισμό που πλήττει τη Χιλή οι (Mendoza, Poblete και Castillo, 2010) μελετούν το φαινόμενο διασποράς ειδήσεων. Εκτός τη διασπορά ειδήσεων, τους προβληματίζει ιδιαίτερα και η διασπορά ψευδών ειδήσεων κατά τη διάρκεια ενός έκτακτου φαινομένου. Η μέθοδος που χρησιμοποιείται είναι καταρχάς ο εντοπισμός γνωστών ειδήσεων αλλά και ψευδών ειδήσεων μέσα από σχετικές λέξεις κλειδιά. Στη συνέχεια, τοποθετούνται ετικέτες ανάλογα με το αν ένας χρήστης επιβεβαιώνει ή απορρίπτει τη σχετική είδηση. Η τοποθέτηση ετικετών γίνεται με μη αυτόματο τρόπο για τα αρχικά γεγονότα και με αυτόματο για τις αναδημοσιεύσεις αυτών. Στη συνέχεια μελετώντας συχνότητα εμφάνισης όρων, συσχετίσεις μεταξύ τους αλλά και επιβεβαιώσεις χρηστών, η μελέτη αυτή καταδεικνύει διαφορετική συμπεριφορά ανά είδηση. Καταλήγει πως όταν η συντριπτική πλειοψηφία των χρηστών επιβεβαιώνουν ένα γεγονός τότε αυτό πρέπει να είναι αληθινό.

Οι (Poblete, Castillo και Mendoza, 2011) επανέρχονται με νέα τους έρευνα στην οποία προσπαθούν να καταγράψουν όλα τα χαρακτηριστικά εκείνα που καθιστούν μία δημοσίευση αληθινή είδηση ή όχι. Η νέα μελέτη, πάνω στην οποία στηρίχθηκε και η παρούσα προσέγγιση, προσπαθεί με αυτοματοποιημένο τρόπο να αποφανθεί για την εγκυρότητα ενός tweet, ή αλλιώς μίας δημοσίευσης χρήστη του Twitter. Προτού οδηγηθούν σε χρήση ταξινομητή πραγματοποιούν μία επιβλεπόμενη διαδικασία, απαραίτητη για την εκπόνηση της μελέτης. Αναζητούν βάσει συχνότητας εμφάνισης λέξεων αναδυόμενες ειδήσεις μέσα σε δημοσιεύσεις χρηστών. Στο πρώτο επίπεδο, ζητούν από μία ομάδα ανθρώπων να αποφανθεί αν ένα γεγονός παρουσιάζει κάποιο δημοσιογραφικό ενδιαφέρον. Σε δεύτερο επίπεδο για όλα τα γεγονότα που επιλέχθηκαν, μία δεύτερη ομάδα αποφαίνεται για την εγκυρότητά τους. Έχοντας καταλήξει στα δεδομένα που θα χρησιμοποιηθούν, εξάγουν από όλες τις σχετικές δημοσιεύσεις χρήσιμα χαρακτηριστικά. Με αυτά παράγουν διανύσματα χαρακτηριστικών τα οποία τροφοδοτούν σε μοντέλα ταξινομητών ώστε να καταλήξουν στην ταξινόμηση των γεγονότων σε αληθή ή ψευδή. Το τελικό σύστημα κάνει χρήση δέντρου

απόφασης J48 με μέθοδο παραγωγής gini και μέθοδο αξιολόγησης K-Fold Cross-Validation για k=3.

Οι (Yang *κ.ά.*, 2012) αναγνωρίζουν τα αποτελέσματα της προηγούμενης έρευνας και επιθυμούν να την επεκτείνουν. Αντί του Twitter αποφασίζουν να συλλέξουν δεδομένα από αντίστοιχο κοινωνικό δίκτυο ευρέως διαδεδομένο στην Κίνα (Sina Weibo). Αφού συλλέξουν δεδομένα, διαφοροποιούμενοι ελάχιστα από την έρευνα του Castillo, χρησιμοποιούν ειδικούς για την ταξινόμηση των δημοσιεύσεων. Στη συνέχεια εισάγουν κάποια επιπλέον χαρακτηριστικά από τα προτεινόμενα στην έρευνα του Castillo και καταλήγουν σε διαφορετικό μοντέλο ταξινομητή. Ως ταξινομητής επιλέγεται μηχανή SVM με συνάρτηση πυρήνα γάμμα = 0.313 και μέθοδο αξιολόγησης K-Fold Cross-Validation για k=10. Αποδεικνύουν πως τα χαρακτηριστικά που εισάγουν και η αλλαγή στον ταξινομητή επιτυγχάνουν ακόμα καλύτερη επίδοση στην ήδη σημαντική της έρευνας του Castillo.

Την επόμενη χρονιά δημοσιεύεται μια εντελώς διαφορετική προσέγγιση για τον εντοπισμό ψευδών ειδήσεων στο Twitter. Στην έρευνα των (Jin *κ.ά.*, 2013) οι ειδήσεις του Twitter προσεγγίζονται με επιδημιολογικά μοντέλα. Η έρευνα προσπαθεί να καταδείξει πως η διασπορά ειδήσεων στο Twitter προσεγγίζει σε μεγάλο βαθμό τη διασπορά ασθενειών. Πιο συγκεκριμένα, η έρευνα μετατρέπει τον χώρο δεδομένων του Twitter σε ένα SEIZ μοντέλο. Ως ευπαθής (S) αναπαρίσταται ο χρήστης που ακόμα δεν έχει λάβει γνώση για την είδηση. Ως προσβεβλημένος αναπαρίσταται ο χρήστης που έχει αναρτήσει κάποια δημοσίευση σχετικά με την είδηση. Σκεπτικός (Z) χαρακτηρίζεται ο χρήστης που έχει λάβει γνώση για την είδηση αλλά δεν την έχει δημοσιεύσει. Τέλος, εκτεθειμένος (E) θεωρείται ένας χρήστης ο οποίος έχει λάβει γνώση για την είδηση μέσω κάποιας δημοσίευσης. Με την παραπάνω θεώρηση, βάσει της θεωρίας του επιδημιολογικού μοντέλου SEIZ και με στατιστική μελέτη του όγκου των δημοσιεύσεων που αφορούν γνωστά γεγονότα, η έρευνα υπολογίζει μία σχετική μετρική. Αποδεικνύεται πως ψευδή γεγονότα εμφανίζουν μηδενική τιμή της μετρικής αυτής, ενώ όσο πιο σημαντικό είναι ένα γεγονός τόσο μεγαλύτερη τιμή εμφανίζει στη μετρική αυτή.

Οι (Gupta, Kumaraguru και Castillo, 2014) προτείνουν στη συνέχεια ένα ολοκληρωμένο σύστημα αξιολόγησης ειδήσεων. Το σύστημα που προτείνεται (TweetCred) προσφέρει μία επέκταση στο Twitter, η οποία εμφανίζει μία βαθμολογία σχετικά με κάποια δημοσίευση. Η βαθμολογία δεν είναι τίποτα άλλο πέραν του αποτελέσματος εκτίμησης ενός ταξινομητή. Για τις ανάγκες του συστήματος, παρατηρούμε πως τα δεδομένα ταξινομούνται σε κατηγορίες και πάλι με τη βοήθεια ομάδας ανθρώπων. Μέσα από μία διαδικασία αξιολόγησης, συλλέχθηκαν και ταξινομήθηκαν δημοσιεύσεις σχετικά με κάποια συνταρακτικά γεγονότα του 2013. Χρησιμοποιώντας τα ως είσοδο, η έρευνα εντοπίζει ποιος ταξινομητής αποδίδει καλύτερα στην εκτίμηση της κλάσης τους. Οι ταξινομητές που συγκρίθηκαν ήταν οι Coordinate Ascent, AdaRank, RankBoost και SVM-rank ενώ κριτήριο επιλογής αποτέλεσε η επίδοση ως προς τη μετρική Normalized Discounted Cumulative Gain. Η μέθοδος αξιολόγησης που ακολουθήθηκε ήταν η K-Fold Cross-Validation για k=4. Βάσει μετρήσεων το μοντέλο SVM-rank κατέγραψε την καλύτερη επίδοση και επιλέχθηκε για το σύστημα ταξινόμησης δημοσιεύσεων. Αξίζει να σημειωθεί πως ο αριθμός των κλάσεων ήταν 7 και ήταν μία διαβάθμιση μεταξύ απόλυτα έγκυρης και πλήρως ψευδούς είδησης.

Το επόμενο έτος ο (Liu, 2015) προτείνει ένα νέο σύστημα, το οποίο έρχεται να χρησιμοποιήσει τις ιδέες όλων των προηγούμενων και να τις επεκτείνει. Το σύστημα αυτό προτείνει νέα χαρακτηριστικά τα οποία οδηγούν σε ακόμα καλύτερα αποτελέσματα ως προς την αξιοπιστία των δημοσιεύσεων του Twitter. Αρχικά, προσπαθεί να εντοπίσει φήμες μέσα από σχετικές ιστοσελίδες, τις οποίες στη συνέχεια προσπαθεί να αξιολογήσει. Μόλις εντοπίσει μία φήμη, συλλέγει τυχαία επιλεγμένα δεδομένα χρηστών του Twitter, από τα οποία διατηρεί μόνο τα σχετικά με αυτή. Στη συνέχεια, με τα δεδομένα που συλλέχθηκαν παράγει διανύσματα χαρακτηριστικών, στα οποία προσθέτει επιπλέον χαρακτηριστικά που έχει συλλέξει από προηγούμενα γεγονότα. Το σύστημα, τόσο με χρήση SVM όσο και δέντρων αποφάσεων J48 αποδίδει σαφέστατα καλύτερα από οποιοδήποτε άλλο.

Οι (Tolos, Tagarev και Georgiev, 2016) μέσα από την έρευνά τους έρχονται να αμφισβητήσουν τη σημασία κάποιων χαρακτηριστικών ως προς την αναγνώριση της αξιοπιστίας των δημοσιεύσεων του Twitter. Πιο συγκεκριμένα, κάνοντας χρήση των προτεινόμενων συστημάτων που στηρίζονται στα χαρακτηριστικά των δημοσιεύσεων μελετούν συγκεκριμένα γεγονότα (ψευδή και αληθή). Για τη μελέτη τους εφαρμόζουν τη μέθοδο leave-one-topic-out. Μελετούν δηλαδή την επιτυχία πρόβλεψης όταν απουσιάζουν από την εκμάθηση του μοντέλου οι δημοσιεύσεις που αφορούν ένα συγκεκριμένο γεγονός. Από τα αποτελέσματα που λαμβάνουν προσπαθούν να εντοπίσουν κατά πόσο ένα χαρακτηριστικό μπορεί να επηρεάσει την τελική εκτίμηση του συστήματος. Αποδεικνύουν πως για κάποια χαρακτηριστικά υπάρχει μεγάλη ανομοιογένεια ανάμεσα στα γεγονότα και έτσι δεν πρέπει να υπολογίζονται στην αξιολόγηση μίας δημοσίευσης. Τέλος καταλήγουν στον ελάχιστο αριθμό χαρακτηριστικών, τα οποία μπορούν να αποτελέσουν μία καλή βάση στον χαρακτηρισμό ενός γεγονότος όταν δεν υπάρχουν δεδομένα ώστε να γίνει ορθή εκμάθηση του μοντέλου προβλέψεων.

3.2 Αξιολόγηση και χαρακτηρισμός διαθέσιμων δεδομένων

Προκειμένου το σύστημα να μπορεί να κατηγοριοποιήσει μία είδηση ως προς την εγκυρότητά της θα πρέπει να έχει προηγηθεί ο σαφής ορισμός της κάθε μίας από τις κατηγορίες. Εν προκειμένω, οι κατηγορίες είναι οι αληθείς και οι ψευδείς ειδήσεις.

Από τη φύση του ένα σύστημα δε διαθέτει τη νοημοσύνη για να πραγματοποιήσει την κατηγοριοποίηση αυτή. Η συνηθέστερη αντιμετώπιση τέτοιων εργασιών είναι η ανάθεσή τους σε ανθρώπους, οι οποίοι μελετώντας τα δεδομένα, πραγματοποιούν την κατηγοριοποίηση αυτών βάσει της κρίσης τους. Παρά τους κινδύνους που ενέχει η υποκειμενικότητα του ανθρώπινου παράγοντα πρόκειται για την επικρατέστερη λύση για κατηγοριοποίηση αταξινόμητων δεδομένων. Φυσικά, ο χρόνος που απαιτείται για την επεξεργασία μεγάλου όγκου δεδομένων δημιουργεί χρηματικές απαιτήσεις από τα εμπλεκόμενα άτομα στη διαδικασία αυτή.

Εφόσον υπάρχει δυνατότητα χρηματοδότησης της μελέτης τότε η επεξεργασία των δεδομένων μπορεί να γίνει μέσω ηλεκτρονικών εφαρμογών, υλοποιημένων για τέτοιου είδους διεργασίες. Το πιο γνωστό από αυτά τα συστήματα είναι το Amazon Mechanical Turk (https://www.mturk.com). Σε αυτό μπορούν να τεθούν ερωτήσεις, οι απαντήσεις των οποίων αμείβονται σε όποιον τις προσφέρει. Αποτελεί ιδανική επιλογή καθώς δεν προϋποθέτει καμία δέσμευση μεταξύ των συμβαλλόμενων και μπορεί να συλλέξει απαντήσεις από αρκετά

διευρυμένο κοινό. Με αυτό τον τρόπο συμβάλει στο να ξεπεραστούν προκαταλήψεις που μπορεί να υπάρχουν μεταξύ ανθρώπων με παρόμοιο υπόβαθρο.

Στη μελέτη που πραγματοποιήσαμε απουσίαζε η δυνατότητα χρηματοδότησης. Συνεπώς έπρεπε να ξεπεραστεί το ζήτημα κατηγοριοποίησης ειδήσεων με διαφορετικό τρόπο. Προκειμένου να έχουμε προσεγγιστικά μία ταυτόσημη με το Mechanical Turk κατηγοριοποίηση έπρεπε να καταφύγουμε σε μία λύση που περιλαμβάνει εξίσου τον ανθρώπινο παράγοντα.

Ως προς τις αληθείς ειδήσεις πραγματοποιήθηκε η παραδοχή πως παγκόσμιοι δημοσιογραφικοί οργανισμοί μπορούν να χαρακτηριστούν χωρίς βλάβη της γενικότητας έγκυρες πηγές πληροφοριών. Στα πλαίσια αυτής της παραδοχής αναζητήθηκαν οι λογαριασμοί των οργανισμών αυτών στο Twitter και οι δημοσιεύσεις τους θεωρήθηκαν έγκυρες ειδήσεις. Κριτήριο επιλογής λογαριασμών αποτέλεσε η πιστοποίηση αυθεντικού λογαριασμού, η οποία παρέχεται από το Twitter καθώς και το πλήθος των ακολούθων τους. Στην ορολογία του Twitter ακόλουθος θεωρείται ο χρήστης που επιλέγει να εμφανίζονται στην αρχική σελίδα του οι δημοσιεύσεις του ακολουθούμενου και δηλώνει με κάποιο τρόπο την προτίμησή του στις δημοσιεύσεις του.

Ως προς τις ψευδείς ειδήσεις η επιλογή λογαριασμών δεν ήταν τόσο ξεκάθαρη διαδικασία. Αυτό συμβαίνει καθώς ο σκοπός ενός χρήστη που πραγματοποιεί διασπορά ψευδών ειδήσεων είναι να μην αποκαλυφθεί ποτέ η ταυτότητά του. Επίσης, πολλοί τέτοιοι λογαριασμοί προβαίνουν συχνά σε παραβίαση των όρων χρήσης του Twitter και απενεργοποιούνται. Το αποτέλεσμα είναι η εξεύρεση λογαριασμών που παράγουν ψευδείς ειδήσεις να είναι μία επίπονη διαδικασία. Αρχικά, θεωρήσαμε ως πηγές παραγωγής ψευδών ειδήσεων σατιρικούς λογαριασμούς, οι οποίοι αναπαράγουν ψευδείς σατιρικές ειδήσεις εις γνώση των ακολούθων τους. Παρότι δεν υπάρχει δόλος ως προς το περιεχόμενο των δημοσιεύσεών τους, αποτελούν καλό παράδειγμα ψευδών ειδήσεων καθώς η θεματολογία τους σχετίζεται με ειδήσεις της επικαιρότητας. Στη συνέχεια προστέθηκαν και λογαριασμοί χρηστών οι οποίοι έχουν αναφερθεί στο διαδίκτυο ως αναξιόπιστοι από πληθώρα χρηστών. Πηγή για τις αναφορές αποτέλεσε αφενός το σχετικό λήμμα της Wikipedia, αφετέρου σχετικές λίστες που ενημερώνονται διαρκώς από χρήστες του διαδικτύου. Η φερεγγυότητα των λιστών δεν μπορεί να αποδειχθεί επιστημονικά, όμως, αποτελεί μία πολύ καλή προσομοίωση της λύσης Mechanical Turk, όπου η φερεγγυότητα των απαντήσεων δεν είναι καθόλου δεδομένη.

Ακολουθούν σχετικές λίστες με τους λογαριασμούς κάθε κατηγορίας, οι οποίοι αποτέλεσαν πηγές για τα δεδομένα που συλλέχθηκαν.

Όνομα Λογαριασμού	**Αναγνωριστικό Λογαριασμού**
Telegraph News	@TelegraphNews
FRANCE 24 English	@France24_en
FRANCE 24	@FRANCE24
AFP news agency	@AFP
NBC	@nbc
Reuters Live	@ReutersLive
Reuters World	@ReutersWorld
dwnews	@dwnews
The Telegraph	@Telegraph
RT	@RT_com
euronews	@euronews
Telegraph World News	@TelegraphWorld
The Guardian	@guardian
CNN International	@cnni
CBS News	@CBSNews
NBC News	@NBCNews
Sky News	@SkyNews
Guardian news	@guardiannews
Sky News Newsdesk	@SkyNewsBreak
ABC News	@ABC
Breaking News	@BreakingNews
TIME	@TIME
Wall Street Journal	@WSJ
Washington Post	@washingtonpost
The Associated Press	@AP
The New York Times	@nytimes
CNN	@CNN
BBC News (World)	@BBCWorld
CNN Breaking News	@cnnbrk
Reuters Top News	@Reuters
BBC Breaking News	@BBCBreaking

Πίνακας 1: Αξιόπιστοι Λογαριασμοί

Όνομα Λογαριασμού	Αναγνωριστικό Λογαριασμού
ChristWire	@ChristWire
DerfMagazine.com	@DerfMagazine
369News	@369news
CAP News	@Capnews
The D.C. Clothesline	@DCClothesline
21st Century Wire	@21WIRE
Currentish.com	@TheCurrentish
Satira Tribune	@SatiraTribune
Breitbart News	@BreitbartNews
The Onion	@TheOnion
NewsBiscuit	@NewsBiscuit
DRUDGE REPORT	@DRUDGE_REPORT

Πίνακας 2: Αναξιόπιστοι Λογαριασμοί

3.3 Αξιόλογα χαρακτηριστικά των δημοσιεύσεων του Twitter

Μία δημοσίευση του Twitter εκτός του φυσικού κειμένου συνοδεύεται από επιπρόσθετες πληροφορίες οι οποίες διατίθενται εξίσου από το Twitter API. Οι πληροφορίες αυτές αφορούν τοποθεσία που έγινε η δημοσίευση, γλώσσα δημοσίευσης, πόσοι χρήστες πραγματοποίησαν αναδημοσίευση αυτής κ.ο.κ. Προκειμένου να επεξεργαστούμε τις δημοσιεύσεις και να τις μετατρέψουμε σε δεδομένα εισόδου του συστήματος έπρεπε να επιλέξουμε κατάλληλα τα χαρακτηριστικά που θα συμπεριλάβουμε στη μελέτη.

Για την επιλογή κατάλληλων χαρακτηριστικών η μελέτη στηρίχθηκε ιδιαίτερα στο έργο των (Poblete, Castillo και Mendoza, 2011) με τίτλο: “Information Credibility on Twitter”. Στο έργο τους μεταξύ των άλλων αναφέρονται τα χρήσιμα χαρακτηριστικά που περιέχονται σε μία δημοσίευση του Twitter. Κατά την επεξεργασία των δεδομένων καταγράφηκε ο προτεινόμενος πίνακας χαρακτηριστικών για κάθε μία από τις δημοσιεύσεις, ο οποίος επεκτάθηκε προσφέροντας βελτίωση στην ποιότητα των δεδομένων.

Ακολουθεί ο πίνακας με τα χρήσιμα στη μελέτη μας χαρακτηριστικά.

α/α	Χαρακτηριστικό	Περιγραφή	Τιμές
1	Μέγεθος δημοσίευσης σε χαρακτήρες	Το μέγεθος του φυσικού κειμένου της δημοσίευσης σε πλήθος χαρακτήρων	1 έως N
2	Μέγεθος δημοσίευσης σε λέξεις	Το μέγεθος του φυσικού κειμένου της δημοσίευσης σε πλήθος λέξεων	1 έως N
3	Περιέχει ερωτηματικό "?"	Το φυσικό κείμενο της δημοσίευσης περιέχει τουλάχιστον ένα ερωτηματικό	0 ή 1
4	Περιέχει θαυμαστικό "!"	Το φυσικό κείμενο της δημοσίευσης περιέχει τουλάχιστον ένα θαυμαστικό	0 ή 1
5	Περιέχει πολλαπλά ερωτηματικά ή θαυμαστικά	Το φυσικό κείμενο της δημοσίευσης περιέχει περισσότερα του ενός ερωτηματικά ή θαυμαστικά	0 ή 1
6	Περιέχει σύμβολο χαμογελαστού προσώπου	Το φυσικό κείμενο της δημοσίευσης περιέχει happy emoticon	0 ή 1
7	Περιέχει σύμβολο σκυθρωπού προσώπου	Το φυσικό κείμενο της δημοσίευσης περιέχει frown emoticon	0 ή 1
8	Πλήθος συμβόλων συναισθήματος (emoticons & emojis)	Το πλήθος των emojis & emoticons που περιέχονται στο φυσικό κείμενο της δημοσίευσης	0 έως N
9	Περιέχει αντωνυμία α' προσώπου	Το φυσικό κείμενο της δημοσίευσης περιέχει τουλάχιστον μία αντωνυμία α' προσώπου	0 ή 1
10	Περιέχει αντωνυμία β' προσώπου	Το φυσικό κείμενο της δημοσίευσης περιέχει τουλάχιστον μία αντωνυμία β' προσώπου	0 ή 1
11	Περιέχει αντωνυμία γ' προσώπου	Το φυσικό κείμενο της δημοσίευσης περιέχει τουλάχιστον μία αντωνυμία γ' προσώπου	0 ή 1
12	Πλήθος κεφαλαίων γραμμάτων	Το πλήθος των κεφαλαίων χαρακτήρων που περιέχονται στο φυσικό κείμενο της δημοσίευσης	0 έως N
13	Αριθμός υπερσυνδέσμων Δημοσίευσης	Το πλήθος των urls που αναφέρονται εντός της δημοσίευσης	0 έως N
14	Περιέχει αναφορά στις 10 διασημότερες ιστοσελίδες	Τουλάχιστον ένα από τα url της δημοσίευσης προέρχεται από τις 10 διασημότερες ιστοσελίδες	0 ή 1
15	Περιέχει αναφορά στις 100 διασημότερες ιστοσελίδες	Τουλάχιστον ένα από τα url της δημοσίευσης προέρχεται από τις 100 διασημότερες ιστοσελίδες	0 ή 1
16	Περιέχει αναφορά στις 1000 διασημότερες ιστοσελίδες	Τουλάχιστον ένα από τα url της δημοσίευσης προέρχεται από τις 1000 διασημότερες ιστοσελίδες	0 ή 1
17	Περιέχει αναφορά σε χρήστη του Twitter	Τουλάχιστον μία εμφάνιση αναφοράς χρήστη "@<όνομα_χρήστη>"	0 ή 1
18	Περιέχει Hashtag	Τουλάχιστον μία εμφάνιση hashtag "#<συμβολοσειρά>"	0 ή 1
19	Περιέχει σύμβολο μετοχής "$"	Τουλάχιστον μία εμφάνιση μετοχής "$<όνομα_μετοχής>"	0 ή 1

20	Αποτελεί αναδημοσίευση	Πρόκειται για αναδημοσίευση προηγούμενης δημοσίευσης του Twitter	0 ή 1
21	Μέρα της εβδομάδας που δημοσιεύτηκε	Μία από τις μέρες της εβδομάδας (Δευ, Τρι, Τετ, Πεμ, Παρ, Σαβ, Κυρ)	1,2,3,4,5, 6 ή 7
22	Θετικές Λέξεις Δημοσίευσης	Βαθμολογία ως προς τις θετικές συναισθηματικά λέξεις	-1.0 έως 1.0
23	Αρνητικές Λέξεις Δημοσίευσης	Βαθμολογία ως προς τις αρνητικές συναισθηματικά λέξεις	-1.0 έως 1.0
24	Ουδέτερες Λέξεις Δημοσίευσης	Βαθμολογία ως προς τις ουδέτερες συναισθηματικά λέξεις	-1.0 έως 1.0
25	Συναίσθημα Δημοσίευσης	Βαθμολογία ως προς το συνολικό συναίσθημα	-1.0 έως 1.0

Πίνακας 3: Χρήσιμα Χαρακτηριστικά Δημοσιεύσεων

3.3.2 Ποσοτικοποίηση Χαρακτηριστικών

Προκειμένου να μελετήσουμε τα χαρακτηριστικά των δημοσιεύσεων έπρεπε αυτά να ποσοτικοποιηθούν. Για την απεικόνιση των χαρακτηριστικών σε αριθμητικές τιμές ακολουθήθηκαν οι παρακάτω παραδοχές:

- Για χαρακτηριστικά που αφορούν πλήθος επιλέξαμε την ίδια τιμή του πλήθους
- Για χαρακτηριστικά που αφορούν έλεγχο εμφάνισης επιλέξαμε την αρίθμηση 0 εάν το χαρακτηριστικό δεν εμφανίζεται και 1 εάν το χαρακτηριστικό εμφανίζεται.
- Για τις ημέρες της εβδομάδας ακολουθήσαμε την παρακάτω 1 προς 1 αντιστοιχία:
  - 1 για τη Δευτέρα
  - 2 για την Τρίτη
  - 3 για την Τετάρτη
  - 4 για την Πέμπτη
  - 5 για την Παρασκευή
  - 6 για το Σάββατο
  - 7 για την Κυριακή
- Για τις μετρικές που αφορούν τη συναισθηματική ανάλυση της δημοσίευσης καταγράψαμε την τιμή που επιστρέφεται από τη μέθοδο SentimentIntensityAnalyzer και είναι μεταξύ των τιμών -1.0 και 1.0

3.3.3 Κανονικοποίηση Χαρακτηριστικών

Αφότου τα χαρακτηριστικά αναπαραστάθηκαν αριθμητικά ακολούθησε κανονικοποίηση των τιμών τους. Ο λόγος της κανονικοποίησης είναι η φύση των ταξινομητών και το πως καταλήγουν στην παραγωγή εκτιμήσεων. Πιο συγκεκριμένα, οι ταξινομητές χρησιμοποιούν την ευκλείδεια απόσταση μεταξύ δύο σημείων. Πολύ μεγάλες αποκλίσεις μεταξύ των τιμών κάποιου χαρακτηριστικού μπορεί να οδηγήσουν σε επικυριαρχία του έναντι των άλλων χαρακτηριστικών. Η ανάγκη όλα τα χαρακτηριστικά να συνεισφέρουν εξίσου στην εκτίμηση του ταξινομητή χρήσει την κανονικοποίηση αναγκαία διαδικασία.

Για τις ανάγκες τις κανονικοποίησης εφαρμόστηκε η μέθοδος της αλλαγής κλίμακας του εύρους των χαρακτηριστικών. Κατόπιν κανονικοποίησης όλες οι τιμές των χαρακτηριστικών βρίσκονταν εντός του εύρους 0.0 έως 1.0 ή -1.0 έως 1.0, ανάλογα με τις αρχικές τιμές των χαρακτηριστικών. Η αλλαγή κλίμακας πραγματοποιείται με χρήση της κάτωθι φόρμουλας:

$$\hat{x} = \frac{x - min(x)}{max(x) - min(x)} \text{ εάν } max(x) \neq min(x) \text{ αλλιώς } \hat{x} = 0.5$$

Όπου $min(x)$ είναι η ελάχιστη τιμή ενός χαρακτηριστικού στο σύνολο των δημοσιεύσεων και $max(x)$ η μέγιστη τιμή αυτού.

3.3.4 Ορισμός Κλάσεων

Προκειμένου να τροφοδοτήσουμε τα δεδομένα σε μορφή διανυσμάτων χαρακτηριστικών σε ταξινομητές έπρεπε να ορίσουμε τις κλάσεις στις οποίες ανήκε το κάθε διάστημα.

Από τα δεδομένα της μελέτης, οι κλάσεις του προβλήματος ήταν δύο: οι αξιόπιστοι λογαριασμοί και οι αναξιόπιστοι λογαριασμοί. Η αριθμητική αναπαράσταση που χρησιμοποιήσαμε ήταν η απόδοση της τιμής 0 στην κλάση των αξιόπιστων λογαριασμών και της τιμής 1 στην κλάση των αναξιόπιστων λογαριασμών.

Τέλος καταγράψαμε για κάθε ένα διάνυσμα χαρακτηριστικών την κλάση στην οποία ανήκει ανάλογα της ομάδας λογαριασμών που ανήκει ο χρήστης που πραγματοποίησε τη δημοσίευση, με την οποία αυτό αντιστοιχίζεται.

Έχοντας απεικονίσει τα δεδομένα σε μορφή κατάλληλη για χρήση ταξινομητών, προχωρήσαμε στην εύρεση του καταλληλότερου μοντέλου για εκτίμηση της κλάσης που ανήκει ένα διάνυσμα χαρακτηριστικών. Με άλλα λόγια, της εκτίμησης για την αξιοπιστία της πηγής προέλευσης μίας νέας δημοσίευσης του Twitter.

3.4 Ταξινομητές - Δέντρα Αποφάσεων

Τα δέντρα αποφάσεων αποτελούν αναπαραστάσεις προβλημάτων και αποφάσεων επί των προβλημάτων σε μορφή γράφου. Ως ρίζα νοείται η αρχική κατάσταση ενός προβλήματος. Οι ενδιάμεσοι κόμβοι αναπαριστούν έναν έλεγχο ως προς κάποια παράμετρο του προβλήματος. Οι ακμές αντιπροσωπεύουν το αποτέλεσμα του ελέγχου, ενώ τα φύλλα αντιπροσωπεύουν τελικές αποφάσεις. Έτσι, μία απόφαση που λαμβάνεται σε κάποιο φύλλο του δέντρου προκύπτει από ελέγχους και αποτελέσματα αυτών που πραγματοποιούνται από την αρχική κατάσταση του προβλήματος. (Safavian και Landgrebe, 1991)

Εξαιτίας της μορφής των δέντρων αποφάσεων, αυτά βρίσκουν πληθώρα εφαρμογών σε διάφορους κλάδους. Ο λόγος είναι πως κατά έναν τρόπο προσομοιώνουν τη διαδικασία που ακολουθεί ο άνθρωπος στη λήψη αποφάσεων. Κατά αυτόν τον τρόπο τα αποτελέσματα που παράγουν μπορούν εύκολα να κατανοηθούν καθώς και να επαληθευτούν. Σημαντική είναι επίσης και η πολυπλοκότητα υλοποίησής τους, η οποία εξαιτίας της δεντρικής μορφής παραμένει λογαριθμική.

Τα δέντρα αποφάσεων αποτελούν επίσης μοντέλο που μπορεί να χρησιμοποιηθεί ως ταξινομητής κλάσεων για διανύσματα (ή σύνολα χαρακτηριστικών). Το μόνο που χρειάζεται να γίνει είναι κάποιες παραδοχές ως προς το τι αντιπροσωπεύουν τα συστατικά τους μέρη. Πιο συγκεκριμένα, σε έναν ταξινομητή δέντρο απόφασης τα φύλλα αντιπροσωπεύουν τις διάφορες κλάσεις του συνόλου των δεδομένων. Οι ενδιάμεσοι κόμβοι αποτελούνται από ελέγχους πάνω σε τιμές των χαρακτηριστικών, οι οποίοι οδηγούν στην τελική εκτίμηση ως προς την εκτίμηση της κλάσης ενός συνόλου χαρακτηριστικών.

Για την αποδοτικότερη εκτίμηση της κλάσης ενός συνόλου χαρακτηριστικών, το δέντρο απόφασης πρέπει να περιέχει όσο το δυνατόν πιο αποδοτικούς ελέγχους στους ενδιάμεσους κόμβους. Για να επιτευχθεί κάτι τέτοιο το δέντρο απόφασης περνάει από ένα στάδιο εκμάθησης. Στο στάδιο αυτό το δέντρο σχηματίζεται τμηματικά βάσει κάποιων μεθόδων. Τα δέντρα που μελετήσαμε χρησιμοποιούν τις μεθόδους Entropy & Information Gain και Gini.

3.4.1 Μέθοδος Entropy & Information Gain

Η μέθοδος αυτή στηρίζεται σε δύο μετρικές για την κατασκευή κατάλληλου δέντρου εκμάθησης για τα υπό μελέτη σύνολα χαρακτηριστικών. Η πρώτη (entropy) έχει να κάνει με την τυχαιότητα ως προς της τιμές ενός χαρακτηριστικού. Λαμβάνει τιμή 0 εάν το χαρακτηριστικό έχει την ίδια τιμή για όλα τα δεδομένα, ενώ έχει τιμή 1 όταν οι τιμές του χαρακτηριστικού ισομοιράζονται μεταξύ δύο συγκεκριμένων τιμών. Η μετρική information gain υπολογίζει τη μείωση που επέρχεται στην τυχαιότητα με την αφαίρεση κάποιου χαρακτηριστικού.

Το δέντρο που κατασκευάζεται είναι δυαδικό και τοποθετεί στη ρίζα το χαρακτηριστικό με τη μεγαλύτερη τιμή στη μετρική information gain. Παιδιά του μπορεί να είναι είτε κάποιος έλεγχος πάνω σε κάποιο άλλο χαρακτηριστικό είτε κάποιο φύλλο. Στα φύλλα τοποθετούνται χαρακτηριστικά που παρουσιάζουν μηδενική τιμή στη μετρική entropy.

Εφόσον έχουν τοποθετηθεί όλα τα χαρακτηριστικά σε κάποιον ενδιάμεσο κόμβο είτε σε φύλλα θα πρέπει να γίνει έλεγχος ότι όλα τα μονοπάτια του δέντρου οδηγούν σε φύλλο. Εφόσον κάτι τέτοιο δεν είναι εφικτό θα πρέπει να πραγματοποιηθεί διαχωρισμός κάποιου χαρακτηριστικού στα δύο. Σκοπός του διαχωρισμού είναι να μειώσει τη μετρική entropy του συγκεκριμένου χαρακτηριστικού, αφού διαχωριζόμενο σε περισσότερα υποσύνολα, η τυχαιότητα των τιμών του μειώνεται. Η διαδικασία επαναλαμβάνεται μέχρις ότου επιτευχθεί ένα επιθυμητό βάθος δέντρου είτε όλοι οι εσωτερικοί κόμβοι έχουν τιμή 0 στην μετρική entropy.

3.4.2 Μέθοδος Gini

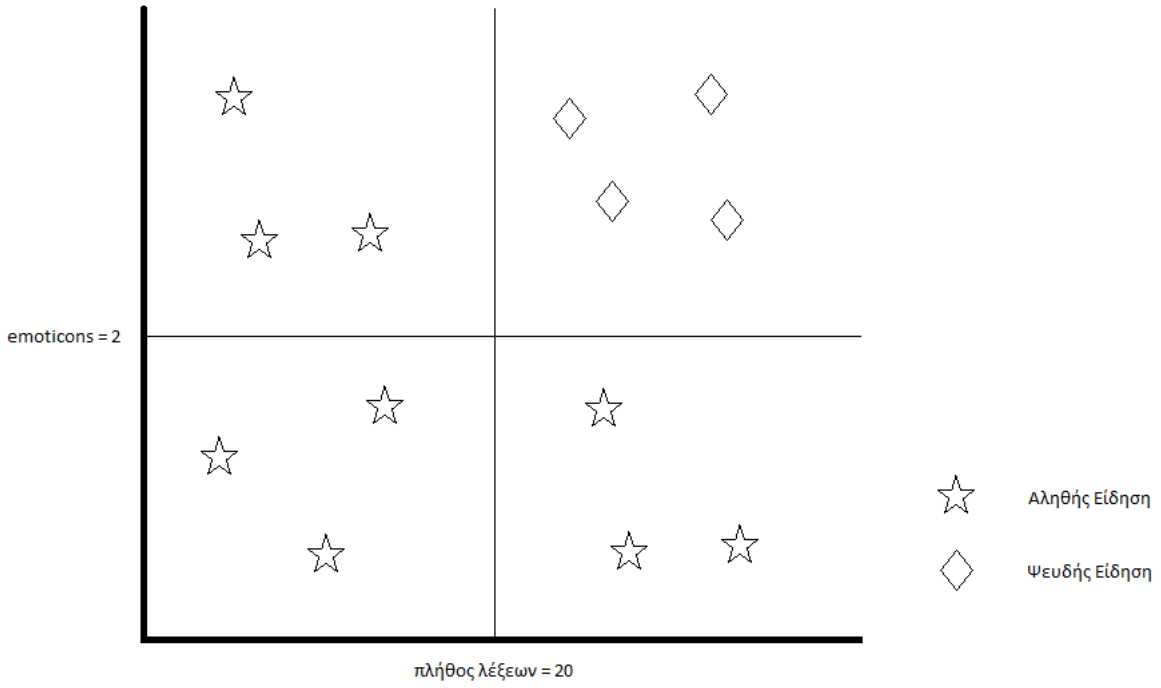
Η μέθοδος κατασκευής δέντρου απόφασης Gini στηρίζεται στην ομώνυμη μετρική. Η μετρική αυτή δείχνει τη συχνότητα λάθους αναγνώρισης ενός τυχαία επιλεγμένου στοιχείου. Όπως είναι προφανές επιλέγουμε τα χαρακτηριστικά από το μικρότερο προς το μεγαλύτερο ως προς την τιμή της μετρικής αυτής για την κατασκευή του δέντρου. Στη ρίζα τοποθετείται το χαρακτηριστικό με τη μικρότερη τιμή Gini, ακολουθεί το χαρακτηριστικό με τη δεύτερη μικρότερη κ.ο.κ.

Και στις δύο μεθόδους το τελικό αποτέλεσμα είναι ένα δέντρο απόφασης το οποίο δεδομένου σταδίου εκμάθησης, μπορεί να εκτιμήσει για ένα σύνολο χαρακτηριστικών την κλάση στην οποία ανήκουν. Προκειμένου να το επιτύχει θα πρέπει, τέλος να συσχετίσει κάθε φύλλο του με μία κλάση. Για να το επιτύχει, ο ταξινομητής δέντρου απόφασης πραγματοποιεί μία απλή διαδικασία. Μετά το πέρας της διαδικασίας εκμάθησης, για κάθε ένα φύλλο, καταγράφει την κλάση της πλειοψηφίας των συνόλων χαρακτηριστικών που οδηγήθηκαν στο φύλλο αυτό.

Μόλις ολοκληρωθεί η κατασκευή του δέντρου απόφασης, ο ταξινομητής θεωρείται εκπαιδευμένος και μπορεί να χρησιμοποιηθεί για την εκτίμηση της κλάσης νέων διανυσμάτων χαρακτηριστικών, τα οποία θα λάβει ως είσοδο. Για κάθε μία εκτίμηση, ο ταξινομητής ενημερώνει στην έξοδό του για την κλάση που θεωρεί πως ανήκει το εκάστοτε διάνυσμα.

3.4.3 Παράδειγμα κατασκευής Δέντρου Απόφασης

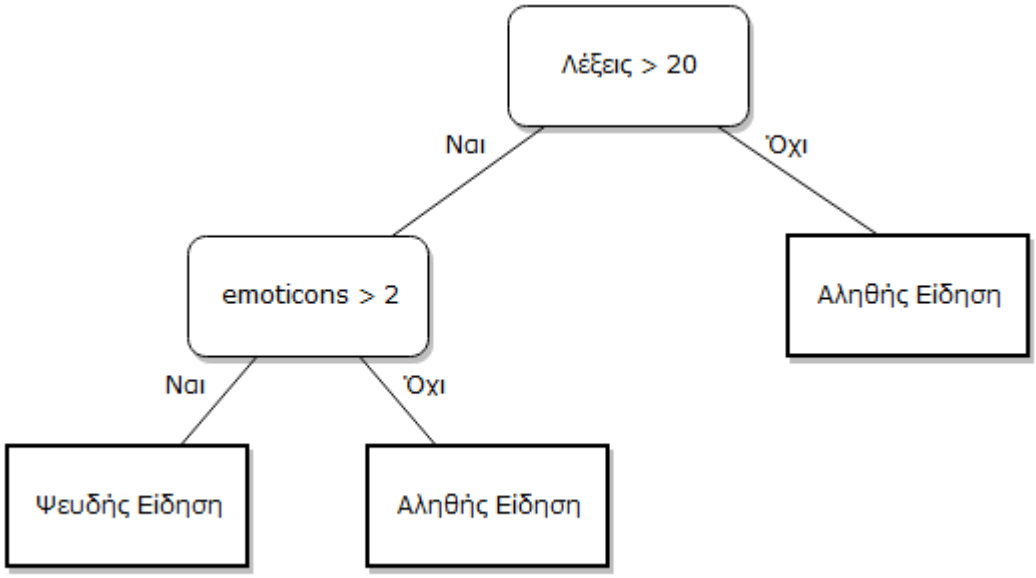
Προκειμένου να γίνει καλύτερα κατανοητός ο τρόπος με τον οποίο κατασκευάζεται ένας ταξινομητής δέντρου απόφασης κατά το στάδιο εκμάθησης, κρίνεται σκόπιμο να παρουσιαστεί ένα σχετικό παράδειγμα. Σε αυτό το απλό παράδειγμα έχουμε διανύσματα δύο χαρακτηριστικών: πλήθος λέξεων και αριθμό emoticons. Μία κατανομή των σημείων που τα αντιπροσωπεύουν είναι αυτή της παρακάτω εικόνας.



Εικόνα 1: Διανύσματα χαρακτηριστικών

Κάνοντας χρήση του κριτηρίου entropy υποβάλλουμε το δέντρο απόφασης σε διαδικασία εκμάθησης. Το σύστημα διαπιστώνει πως μπορεί να δημιουργήσει ένα δυαδικό δέντρο δύο επιπέδων. Αυτό συμβαίνει καθώς αρκούν δύο συνθήκες προκειμένου να μπορεί να απαντήσει με σαφήνεια για την κλάση ενός νέου διανύσματος. Αρκεί να ελέγξει αν το πλήθος

λέξεων δεν ξεπερνά τις 20 και εάν το πλήθος των emoticons δεν ξεπερνάει τα δύο. Χρησιμοποιώντας τους δύο παραπάνω ελέγχους σε ενδιάμεσους κόμβους και τοποθετώντας τις κλάσεις στα φύλλα προκύπτει το δέντρο της παρακάτω εικόνας.



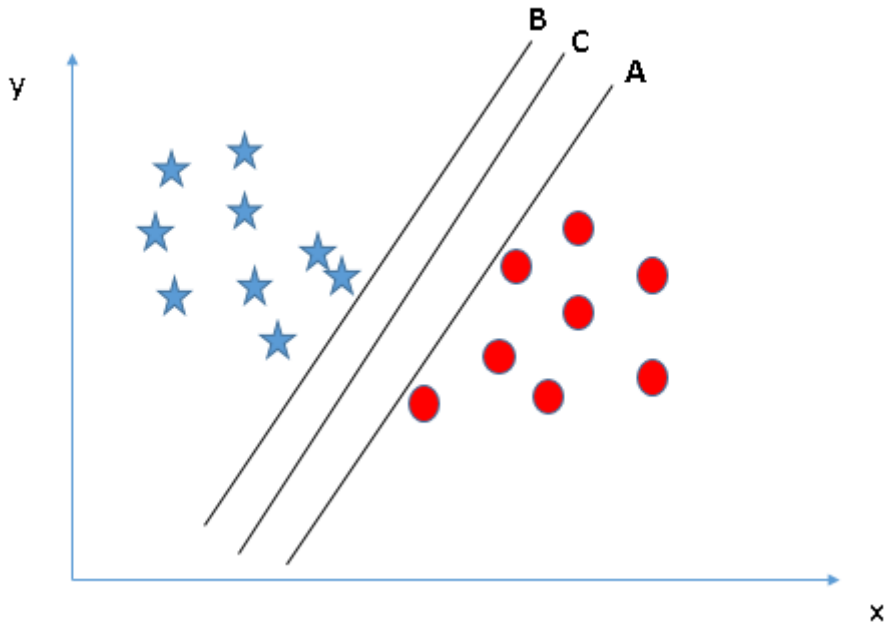
Εικόνα 2: Δέντρο απόφασης

Έχοντας ολοκληρώσει τη διαδικασία εκμάθησης με τα διαθέσιμα δεδομένα, το δέντρο απόφασης μπορεί οποιαδήποτε στιγμή να αποφανθεί για την κλάση ενός αγνώστου διανύσματος χαρακτηριστικών. Για να το επιτύχει, αρκεί να εκκινήσει από τη ρίζα και να οδηγείται στον επόμενο κόμβο ανάλογα το αποτέλεσμα του ελέγχου που συναντά στον προηγούμενο. Οποιαδήποτε διαδρομή ελέγχων και αν ακολουθήσει θα καταλήξει σε μία από τις δύο κλάσεις του προβλήματος: αληθής ή ψευδής είδηση.

3.5 Ταξινομητές - Support Vector Machines (SVM)

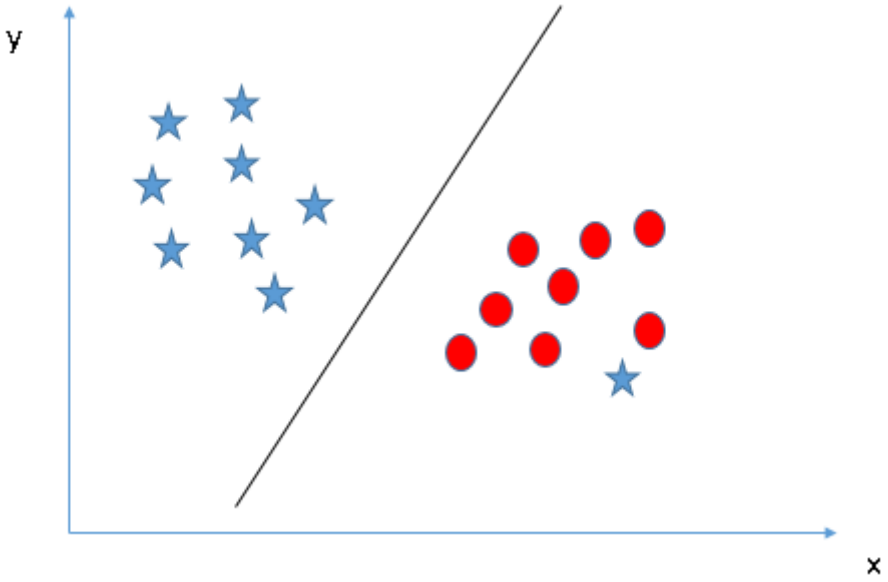
Οι μηχανές διανυσμάτων υποστήριξης ή αλλιώς SVM αποτελούν ένα από τα σημαντικότερα μοντέλα ταξινομητών στον χώρο της μηχανικής μάθησης. Στηρίζονται στη γραφική απεικόνιση των διάφορων στοιχείων και τον διαχωρισμό αυτών ανάλογα την κλάση τους. Στόχος μίας μηχανής διανυσμάτων υποστήριξης είναι η εύρεση ενός υπερεπιπέδου, το οποίο έχει διπλή ιδιότητα. Αφενός διαχωρίσει πλήρως ή σχεδόν πλήρως τα σημεία που αντιπροσωπεύουν τα στοιχεία των κλάσεων του προβλήματος. Αφετέρου, από όλα τα πιθανά υπερεπίπεδα είναι εκείνο το οποίο μεγιστοποιεί την απόσταση όλων των σημείων του προβλήματος από το ίδιο το υπερεπίπεδο.

Στην πιο απλή μορφή της μία μηχανή SVM επιτυγχάνει τον γραμμικό διαχωρισμό των στοιχείων των διαφόρων κλάσεων καθώς το υπερεπίπεδο που σχηματίζεται είναι γραμμικό ως προς μία από τις διαστάσεις του. Αξίζει να σημειωθεί πως η απεικόνιση των στοιχείων γίνεται σε πολλές διαστάσεις, καθώς κάθε χαρακτηριστικό του μπορεί να αντιπροσωπεύει μία διάσταση.



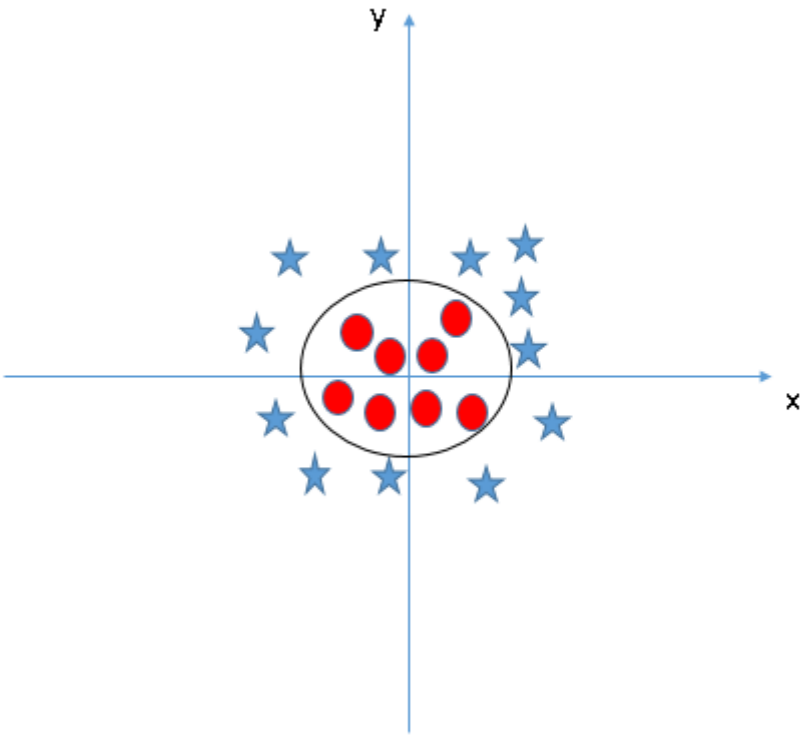
Εικόνα 3: Επιλογή υπερεπιπέδου

Στην παραπάνω εικόνα μπορούμε να δούμε μία τυπική απεικόνιση ενός προβλήματος, όπου τα σημεία αντιπροσωπεύουν την απεικόνιση των δεδομένων διανυσμάτων χαρακτηριστικών. Παρατηρούμε πως όλα τα σημεία ανήκουν σε μία εκ των δύο κλάσεων του προβλήματος. Επίσης παρατηρούμε τρία πιθανά υπερεπίπεδα ανάμεσα στα οποία καλείται μία γραμμική μηχανή SVM να επιλέξει. Η επιλογή είναι σχετικά εύκολη καθώς παρατηρούμε πως όλα τα σημεία απέχουν το μέγιστο δυνατό από το υπερεπίπεδο C, το οποίο τελικά και επιλέγεται.



Εικόνα 4: Διαχείριση outlier

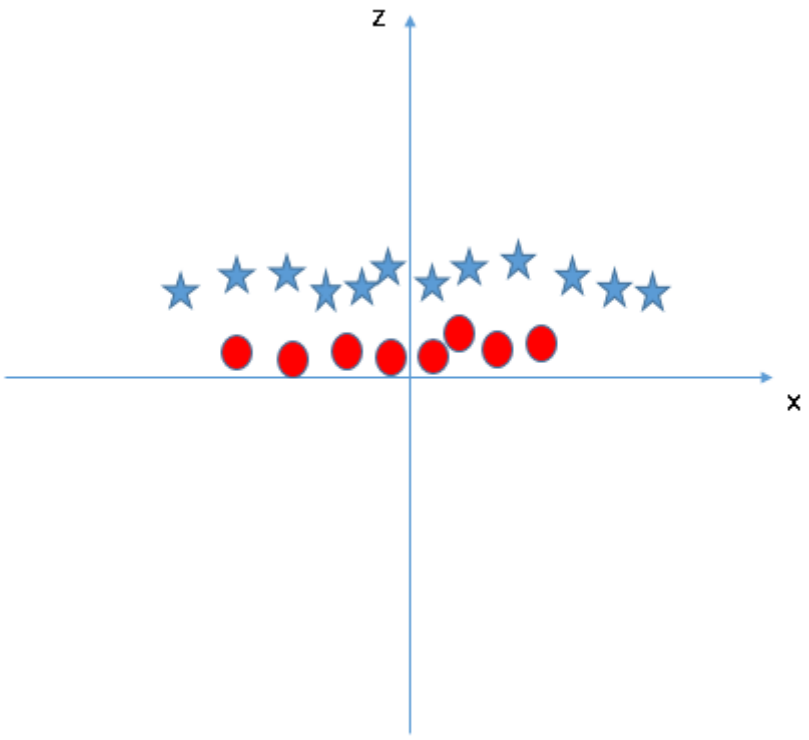
Στην παραπάνω εικόνα παρατηρούμε το εξής παράδοξο. Ενώ όλα τα σημεία της μπλε κλάσης βρίσκονται συγκεντρωμένα σε κάποιο χώρο, υπάρχει ένα το οποίο βρίσκεται σε χώρο που βρίσκονται στοιχεία άλλης κλάσης. Στοιχεία σαν και αυτό θεωρούνται ανωμαλίες “outliers” και μία μηχανή SVM καλείται να τα αντιμετωπίσει. Ο τρόπος που αντιμετωπίζονται τα outliers από τις SVM είναι η παράλειψή τους κατά την επιλογή του βέλτιστου υπερεπιπέδου, κάτι το οποίο γίνεται αντιληπτό και από το υπερεπίπεδο που έχει επιλεχθεί. Τα outliers δεν είναι κάτι ασυνήθιστο, καθώς τα δεδομένα αντιπροσωπεύουν μετρήσεις από την πραγματική ζωή.



Εικόνα 5: Αδυναμία εύρεσης γραμμικού υπερεπιπέδου

Σε καταστάσεις όπως της παραπάνω εικόνας μία γραμμική μηχανή SVM είναι αδύνατο να επιλέξει κάποιο γραμμικό υπερεπίπεδο για να διαχωρίσει πλήρως τα στοιχεία διαφορετικών κλάσεων. Παρατηρούμε όμως πως έχει επιλεγεί στη θέση του ένα κυκλικό υπερεπίπεδο, το οποίο παρουσιάζει και τις δύο επιθυμητές ιδιότητες που χρειάζεται ένα παραγόμενο υπερεπίπεδο από μηχανές SVM. Για να είναι σε θέση μία μηχανή SVM να παράξει τέτοια υπερεπίπεδα αφενός δεν αρκεί να είναι γραμμική, αφετέρου πρέπει να εφαρμόσει κάποια επιπλέον μέθοδο. Η παραπάνω μηχανή SVM ανήκει στην κατηγορία μη γραμμικών SVM.

Οι μη γραμμικές SVM προκειμένου να δημιουργήσουν το κατάλληλο υπερεπίπεδο χρησιμοποιούν τεχνικές μετατροπής των συντεταγμένων των σημείων του προβλήματος. Πιο συγκεκριμένα χρησιμοποιούν μετασχηματισμούς των σημείων σε άλλες διαστάσεις στις οποίες η εύρεση κατάλληλου υπερεπιπέδου είναι εφικτή.



Εικόνα 6: Μετασχηματισμός συντεταγμένων για εύρεση υπερεπιπέδου

Στην παραπάνω εικόνα είναι εμφανής ο μετασχηματισμός που έγινε σε z συντεταγμένες των σημείων του προηγούμενου προβλήματος. Όπως μπορεί να γίνει αντιληπτό πλέον είναι πολύ εύκολη η επιλογή κατάλληλου υπερεπιπέδου που να διαχωρίζει πλήρως τα σημεία των δύο κλάσεων.

Η παραπάνω μετατροπή είναι πολύ χρονοβόρα και αυξάνει αισθητά τις διαστάσεις του προβλήματος και άρα την πολυπλοκότητα του αλγορίθμου εύρεσης υπερεπιπέδου. Στην πραγματικότητα οι μηχανές SVM χρησιμοποιούν κάποιες ειδικές συναρτήσεις που μπορούν να βρίσκουν το ιδανικό υπερεπίπεδο χωρίς να πραγματοποιούν τη μετατροπή. Οι συναρτήσεις αυτές λέγονται συναρτήσεις πυρήνα “kernel functions” και χρησιμοποιούν τεχνάσματα εσωτερικού γινομένου μεταξύ σημείων. Για τα πειράματά μας χρησιμοποιήσαμε τη συνάρτηση πυρήνα RBF (Radial Basis Function), η οποία είναι από τις ευρύτερα χρησιμοποιούμενες.

3.5.1 Συνάρτηση Πυρήνα RBF

Η συνάρτηση RBF ή αλλιώς Γκαουσιανή πρόκειται για ένα τέχνασμα που χρησιμοποιείται από τις μηχανές SVM όταν πρέπει να γίνει εύρεση υπερεπιπέδου σε χώρο που δεν διαχωρίζεται γραμμικά. Όπως αναφέρθηκε παραπάνω σε μη γραμμικά διαχωρίσιμους χώρους χρειάζεται μετατροπή των συντεταγμένων σε άλλο σύστημα για να βρεθεί κατάλληλο υπερεπίπεδο. Έτσι εφαρμόζεται η παρακάτω φόρμουλα:

$$g(z) = \sum_{j=1}^{k} w_j \, \varphi_j(x)$$

Η παραπάνω φόρμουλα δεν είναι τίποτα άλλο από το σύνολο όλων των συναρτήσεων μετασχηματισμών των συναρτήσεων των k διαστάσεων του αρχικού χώρου x στον νέο χώρο z. Φυσικά για τις ανάγκες της εύρεσης υπερεπιπέδου είναι απαραίτητος ο υπολογισμός αποστάσεων μεταξύ σημείων. Εφόσον θέλουμε να υπολογίσουμε τις αποστάσεις στον νέο χώρο, τότε ο υπολογισμός κοστίζει ιδιαίτερα καθώς όχι μόνο αυξάνονται οι διαστάσεις του προβλήματος, αλλά και η ίδια η μετατροπή έχει μεγάλο υπολογιστικό κόστος.

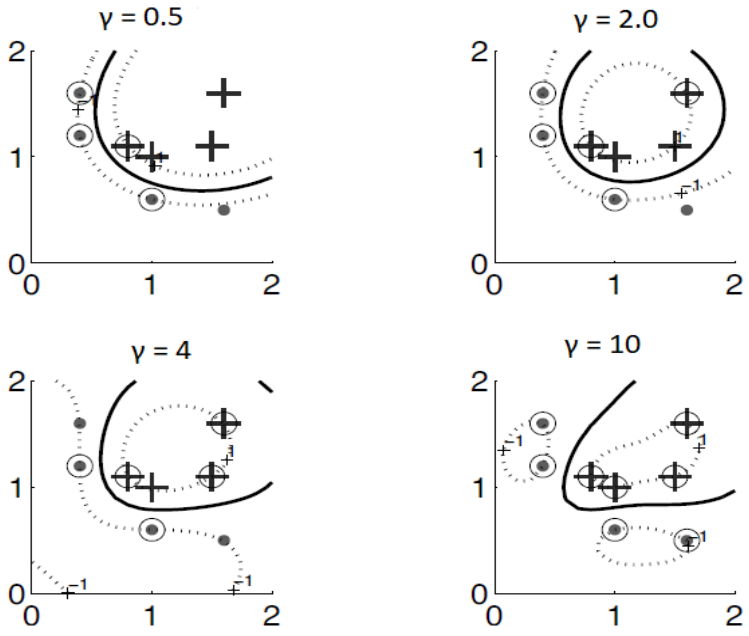
Προκειμένου να αποφευχθεί το κόστος αυτό μπορεί να υπολογιστεί ένα ισοδύναμο πρόβλημα. Μέσω του τρικ πυρήνα (Alpaydin, 2010), οι αποστάσεις μπορούν να υπολογιστούν στον αρχικό χώρο χωρίς να γίνει η παραπάνω μετατροπή. Η συνάρτηση πυρήνα RBF είναι μία από τις πολλές κατηγορίες τέτοιων συναρτήσεων. Μέσω της RBF αντί να υπολογίζουμε αποστάσεις στον χώρο z κάνουμε χρήση της παρακάτω φόρμουλας:

$$K(x^t) = exp[-\gamma \|x^t - x\|^2 \|]$$

Όπως παρατηρούμε, η παράμετρος γ (γάμμα) καθορίζει την τελική επιλογή του κατάλληλου υπερεπιπέδου. Η παράμετρος αυτή καθορίζεται από τον χρήστη. Η συνάρτηση συνολικά παριστάνει εποπτικά μία σφαίρα με κέντρο το σημείο x^t και ακτίνα s, η οποία σχετίζεται με την παράμετρο γ μέσω της παρακάτω σχέσης:

$$\gamma = \frac{1}{s^2}$$

Για τον σχηματισμό του κατάλληλου υπερεπιπέδου, η επιφάνειά του καθορίζεται από τα σημεία x^t τα οποία θα επιλεγούν και άρα διαφέρει ανάλογα της ακτίνας της σφαίρας που προκύπτει από τη συνάρτηση RBF. Στα παρακάτω σχήματα βλέπουμε διαφοροποιήσεις ανάμεσα σε υπερεπίπεδα που σχηματίζονται για διάφορες τιμές της παραμέτρου γ (γάμμα). Παρατηρούμε πως όσο μικραίνει το γάμμα τόσο πιο λείες επιφάνειες παράγονται.



Εικόνα 7: Συνάρτηση RBF για τιμές του γ

3.6 Επιλογή καταλληλότερου μοντέλου

Προκειμένου να καταλήξουμε ως προς το πιο κατάλληλο μοντέλο για την ταξινόμηση των δημοσιεύσεων του Twitter πραγματοποιήσαμε μετρήσεις για διάφορες παραμετροποιήσεις των παραπάνω ταξινομητών. Στην ενότητα αυτή περιγράφονται τόσο τα δεδομένα τα οποία χρησιμοποιήθηκαν όσο και οι υπερπαράμετροι που τέθηκαν στα υπό μελέτη μοντέλα. Εφόσον καταλήξαμε στις υπερπαράμετρους με τις οποίες τα μοντέλα ταξινομητών αποδίδουν καλύτερα, πραγματοποιήσαμε τη μεταξύ του σύγκριση. Τα αποτελέσματα αυτής που ήταν και τα τελικά παρουσιάζονται στην επόμενη ενότητα.

3.6.1 Επιλογή δεδομένων και αριθμητική απεικόνιση

Έχοντας αντλήσει ικανό πλήθος δεδομένων από δημοσιεύσεις του Twitter είχαμε τη δυνατότητα να συγκρίνουμε τις δύο οικογένειες μοντέλων. Έχοντας ξεκινήσει την καταγραφή από τον Νοέμβρη του 2016 έπρεπε να αποφανθούμε ποια δεδομένα θα συμπεριλάβουμε στην αξιολόγηση των ταξινομητών. Προκειμένου να υπάρξει όσο το δυνατόν μεγαλύτερη τυχαιότητα στην επιλογή αποφασίσαμε την παρακάτω διαδικασία.

Αρχικά, μετατρέψαμε όλα τα διαθέσιμα δεδομένα σε διανύσματα χαρακτηριστικών βάσει των κανόνων που αναφέρθηκαν στην παράγραφο 3.3.2. Επιπλέον πραγματοποιήσαμε κανονικοποίηση των τιμών των χαρακτηριστικών για τους λόγους που αναφέρθηκαν στην παράγραφο 3.3.3. Για λόγους ταχύτητας στην επεξεργασία αποφασίστηκε τα διανύσματα των χαρακτηριστικών να αποθηκευτούν προτού χρησιμοποιηθούν από τους ταξινομητές. Με αυτό τον τρόπο αποφεύχθηκε η προσθήκη επιπλέον πολυπλοκότητας στην ήδη υπάρχουσα των ταξινομητών κατά την εκτέλεσή τους.

Επόμενο βήμα ήταν η επιλογή κατάλληλων διανυσμάτων χαρακτηριστικών από τα διαθέσιμα. Λαμβάνοντας υπόψιν την πολυπλοκότητα των μηχανών SVM παράλληλα με τη μειωμένη διαθεσιμότητα δημοσιεύσεων μη έμπιστων λογαριασμών αποφασίστηκε το συνολικό πλήθος διανυσμάτων να είναι τα 26 χιλιάδες. Από αυτά, τα 13 χιλιάδες διανύσματα προέρχονταν από δημοσιεύσεις έμπιστων λογαριασμών και 13 χιλιάδες από μη έμπιστους. Για τα τελικά διανύσματα, πραγματοποιήθηκε τυχαία επιλογή από τα συνολικά διαθέσιμα διανύσματα για κάθε μία κλάση.

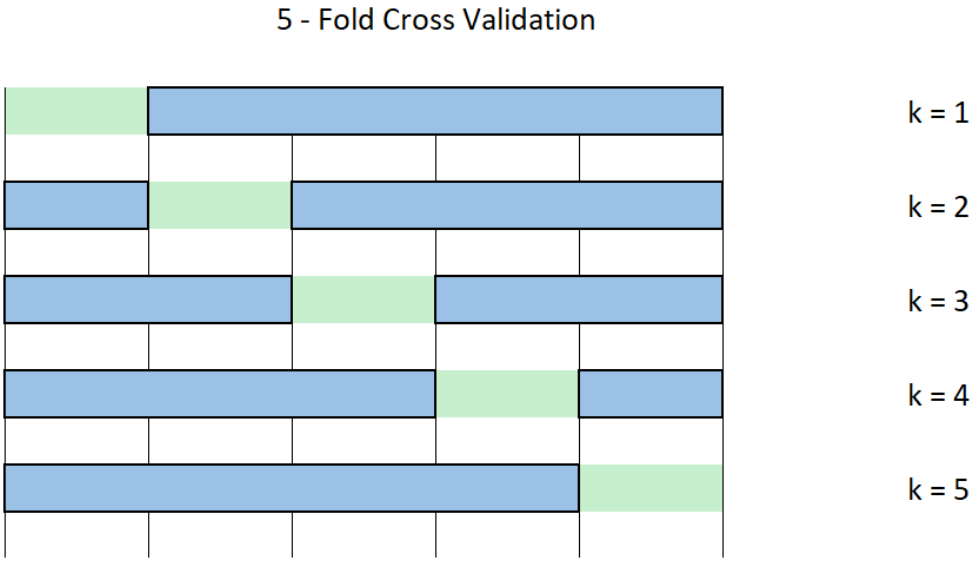
Τέλος, προκειμένου να μην υπάρχει πιθανή συσχέτιση μεταξύ δύο συνεχόμενων διανυσμάτων πραγματοποιήθηκε τυχαίο ανακάτεμα αυτών. Ολοκληρώνοντας την παραπάνω διαδικασία, χρησιμοποιήσαμε τα δεδομένα ως είσοδο για διάφορες παραμετροποιήσεις των δύο οικογενειών ταξινομητών.

3.6.2 Μέθοδος Αξιολόγησης

Προκειμένου να καταλήξουμε στο αποδοτικότερο μοντέλο ταξινομητή χρειάστηκε να τα συγκρίνουμε ως προς κάποια μετρική. Για τις μεθόδους αξιολόγησης που χρησιμοποιήσαμε ακολουθήσαμε τον οδηγό των (Hastie, Tibshirani και Friedman, 2009). Εφόσον ο χρόνος εκτέλεσης δεν αποτελεί πρόβλημα ο καλύτερος τρόπος αξιολόγησης ενός μοντέλου είναι η μέθοδος leave-one out. Στη μέθοδο αυτή για το σύνολο των διαθέσιμων δεδομένων θεωρούμε κάθε φορά ένα από αυτά άγνωστης κλάσης. Με τα υπόλοιπα πραγματοποιούμε εκμάθηση του ταξινομητή και στη συνέχεια ζητάμε από τον ταξινομητή να προβλέψει την κλάση του διανύσματος που διαχωρίσαμε. Πραγματοποιώντας το παραπάνω για κάθε ένα

από τα διανύσματα μπορούμε να συγκεντρώσουμε το ποσοστό επιτυχίας για κάθε ένα από τα μοντέλα ταξινομητών.

Αν και αποτελεί την πιο εξαντλητική μέθοδο για επιλογή του καλύτερου ταξινομητή πρόκειται για μία διαδικασία υψηλής πολυπλοκότητας και ιδιαίτερα χρονοβόρα. Αντί της παραπάνω μεθόδου πραγματοποιήθηκε μία ευρέως διαδεδομένη μέθοδος, προσεγγιστική της leave-one out. Η μέθοδος αυτή ονομάζεται K-Fold Cross-Validation. Η ιδέα της μεθόδου αυτής ακολουθεί αυτή της leave-one out καθώς και πάλι θεωρείται μέρος των δεδομένων άγνωστο. Πιο συγκεκριμένα, η μέθοδος χρησιμοποιεί μέρος των δεδομένων για εκμάθηση του μοντέλου και τα υπόλοιπα για αξιολόγηση αυτού. Η διαδικασία που ακολουθείται είναι η εξής. Αρχικά, τα δεδομένα του προβλήματος (διανύσματα χαρακτηριστικών) διαχωρίζονται σε K το πλήθος υποσύνολα ίσου μεγέθους. Για παράδειγμα, ένας διαχωρισμός για K=5 είναι ο παρακάτω:



Εικόνα 8: Επαναλήψεις K-fold cross validation

Στη συνέχεια, για κάθε ένα από τα k μικρότερα υποσύνολα (πράσινο χρώμα) ζητάμε από το μοντέλο να προβλέψει τις κλάσεις των διανυσμάτων που περιέχονται σε αυτό. Η εκμάθηση του μοντέλου πραγματοποιείται με τα k - 1 υπόλοιπα υποσύνολα (μπλε χρώμα). Η παραπάνω διαδικασία πραγματοποιείται K φορές για όλα τα υποσύνολα του συνόλου των δεδομένων. Για κάθε μία από τις K επαναλήψεις, συγκεντρώνεται η επίδοση του μοντέλου στην πρόβλεψη των κλάσεων των διανυσμάτων. Στο τέλος συνυπολογίζεται ο μέσος όρος των επιδόσεων και προκύπτει η τελική επίδοση του μοντέλου.

Όπως είναι φυσικό, θα πρέπει να διασφαλιστεί πως η διαχώριση πρέπει να είναι ακριβώς η ίδια για όλα τα υπό μελέτη μοντέλα, καθώς διαφορετική διαχώριση παράγει και διαφορετικά αποτελέσματα. Επιπλέον πρέπει να διασφαλίζεται η τυχαιότητα στην επιλογή των k υποσυνόλων των δεδομένων. Όσο αφορά την τυχαιότητα, αυτή διασφαλίστηκε με το τυχαίο ανακάτεμα που πραγματοποιήθηκε στα επιλεγμένα δεδομένα. Αναφορικά με την κοινή διαχώριση όλων των υπό μελέτη μοντέλων διασφαλίστηκε λόγω της διαδικασίας που ακολουθήθηκε. Πιο συγκεκριμένα, αντί να μελετηθούν όλα τα μοντέλα ξεχωριστά, αυτό έγινε

παράλληλα, με κοινή διαχώριση δεδομένων. Με αυτό τον τρόπο δόθηκε η δυνατότητα της απόλυτης σύγκρισης των μοντέλων ως προς την απόδοσή τους.

Για τη μέτρηση της απόδοσης χρησιμοποιήσαμε τη μετρική F1. Η μετρική αυτή πρόκειται ουσιαστικά για μία συνεκτίμηση δύο επιμέρους μετρικών και είναι ευρέως διαδεδομένη στη μέτρηση της απόδοσης μοντέλων. Οι μετρικές που συνεκτιμώμενες σχηματίζουν την παραπάνω μετρική είναι η ακρίβεια (precision) και η ανάκληση (recall). Όσο αφορά το precision πρόκειται για την ικανότητα ενός ταξινομητή να μην αποδίδει θετική τιμή σε μία πρόβλεψη ενώ αυτή είναι αρνητική στην πραγματικότητα. Όσο αφορά το recall πρόκειται για την ικανότητα ενός ταξινομητή να μην αποδίδει αρνητική τιμή σε κάποια πρόβλεψη όταν αυτή δεν είναι αρνητική στην πραγματικότητα. Ακολουθούν οι φόρμουλες από τις οποίες παράγονται οι τρεις αυτές μετρικές:

3.6.2.1 Μετρική Precision

$$precision = \frac{T_p}{T_p + F_p}$$

3.6.2.2 Μετρική Recall

$$recall = \frac{T_p}{T_p + F_n}$$

3.6.2.3 Μετρική F1 Score

$$f1score = \frac{2 * precision * recall}{precision + recall}$$

Στις παραπάνω φόρμουλες αναφέρονται οι μεταβλητές T_p, F_p, F_n. Οι μεταβλητές αυτές αναφέρονται στις περιπτώσεις true positive, false positive, false negative. Ο παρακάτω πίνακας, γνωστός και ως πίνακας σύγχυσης, παρουσιάσει τη σημασίας τους.

		Πρόβλεψη Συνθήκης	
		Θετική Πρόβλεψη	**Αρνητική Πρόβλεψη**
Αληθής συνθήκη	**Θετική Συνθήκη**	True Positive (Tp)	False Negative (Fn)
	Αρνητική Συνθήκη	False Positive (Fp)	True Negative (Tn)

Πίνακας 4: Πίνακας σύγχυσης

Βάσει του πίνακα σύγχυσης η μεταβλητή T_p υποδεικνύει τις περιπτώσεις όταν μία συνθήκη είναι στην πραγματικότητα αληθής και η πρόβλεψη για αυτήν ήταν θετική. Η μεταβλητή F_n αντιπροσωπεύει τις περιπτώσεις που υπάρχει αδυναμία πρόβλεψης αληθών συνθηκών. Η μεταβλητή F_p υποδεικνύει περιπτώσεις θετικής πρόβλεψης σε αρνητικές συνθήκες. Τέλος υπάρχει και η επιτυχής πρόβλεψης αρνητικών συνθηκών, μεταβλητή T_n, η οποία δεν επηρεάζει τις μετρικές που καταγράφουμε, αναφέρεται όμως για λόγους πληρότητας.

Όπως μπορεί να γίνει αντιληπτό, όσο μεγαλύτερη τιμή πετυχαίνει ένα μοντέλο ως προς τη μετρική F1 τόσο αποδοτικότερο είναι ως προς την ακρίβεια των προβλέψεων που παράγει. Σκοπός της μελέτης ήταν η καταγραφή του αποδοτικότερου δέντρου απόφασης καθώς και της αποδοτικότερης μηχανής SVM ως προς τη μετρική F1. Στη συνέχεια πραγματοποιήθηκε εκ νέου σύγκριση μεταξύ τους προκειμένου να καταλήξουμε στο συνολικά αποδοτικότερο μοντέλο για την επιτυχή ταξινόμηση των δεδομένων του προβλήματος που μελετήθηκε.

3.6.3 Επιλογή υπερπαραμέτρων Δέντρου Αποφάσεων

Όπως αναφέρθηκε και στην παράγραφο 3.4 υπάρχουν δύο διαφορετικοί τρόποι με τους οποίους μπορεί να κατασκευαστεί ένα δέντρο απόφασης κατά το στάδιο εκμάθησης. Για τις ανάγκες της μελέτης έπρεπε να δοκιμαστούν και οι δύο προτού καταλήξουμε στον πιο αποδοτικό. Μία ακόμα παράμετρος η οποία μπορεί να καθορίσει την απόδοση ενός ταξινομητή δέντρου αποφάσεων είναι το βάθος του δέντρου. Αφενός, ένα πολύ γενικό δέντρο και άρα μικρού βάθους, δεν μπορεί να προβλέψει με σημαντική απόδοση την κλάση ενός διανύσματος χαρακτηριστικών. Αυτό συμβαίνει καθώς το δέντρο δεν έχει τη δυνατότητα μέσω ικανοποιητικών ελέγχων να διαχωρίσει επαρκώς τα δεδομένα. Αφετέρου, ένα πολύ βαθύ δέντρο πολλές φορές μπορεί να δώσει την ψευδαίσθηση ότι είναι αποδοτικό, σε περιπτώσεις που εμφανίζεται το φαινόμενο “overfitting”.

3.6.3.1 Φαινόμενο Overfitting

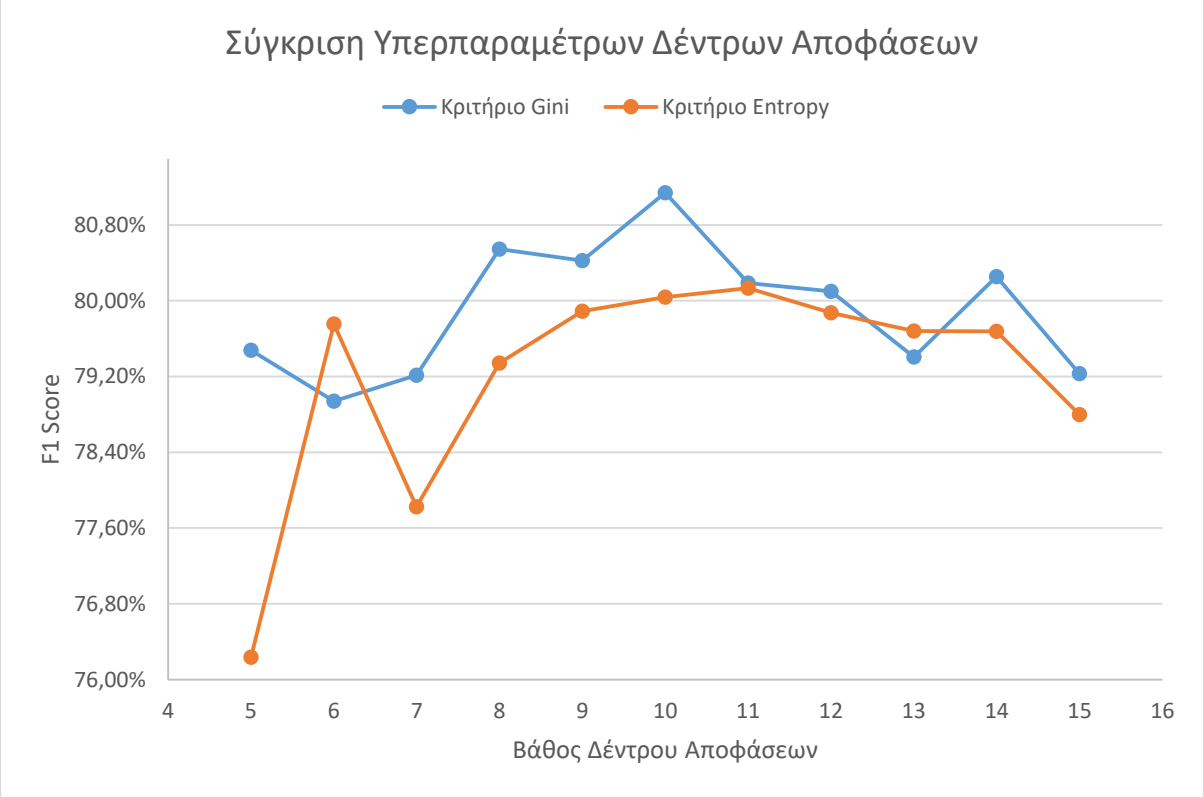
Είναι το φαινόμενο που συμβαίνει όταν ο ταξινομητής γίνεται πολύ εξειδικευμένος στα δεδομένα τα οποία λαμβάνει στο στάδιο εκμάθησης. Με άλλα λόγια, χρησιμοποιεί πολλές μεταβλητές για την παραγωγή πρόβλεψης, οι οποίες όμως προέκυψαν για τα συγκεκριμένα δεδομένα. Όταν του ζητηθεί να κάνει πρόβλεψη για διαφορετικά δεδομένα, τότε οι μεταβλητές αυτές έχουν μεγάλες διακυμάνσεις. Το αποτέλεσμα είναι να δημιουργείται αρκετός θόρυβος και άρα να φθίνει σε μεγάλο βαθμό η απόδοση του ταξινομητή.

3.6.3.2 Σύγκριση Δέντρων Αποφάσεων

Για τις ανάγκες τις μελέτης, δημιουργήσαμε μοντέλα δέντρων αποφάσεων τροποποιώντας τις δύο παραμέτρους (βάθος, κριτήριο δημιουργίας) και καταγράψαμε την απόδοσή τους. Για την υπερπαράμετρο βάθος μελετήσαμε ένα εύρος τιμών από 5 έως 15. Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα που λάβαμε επί των διανυσμάτων χαρακτηριστικών που χρησιμοποιήσαμε. Με πράσινο χρώμα σημειώνεται το αποδοτικότερο μοντέλο. Στη συνέχεια ακολουθεί γράφημα όπου παρουσιάζονται οπτικά οι επιδόσεις των μοντέλων που μελετήθηκαν.

Βάθος Δέντρου	Κριτήριο Επιλογής	Μετρική F1	Κριτήριο Επιλογής	Μετρική F1
5	gini	79.476%	entropy	76.235%
6	gini	78.941%	entropy	79.753%
7	gini	79.213%	entropy	77.824%
8	gini	80.546%	entropy	79.342%
9	gini	80.424%	entropy	79.890%
10	gini	81.139%	entropy	80.040%
11	gini	80.187%	entropy	80.134%
12	gini	80.098%	entropy	79.872%
13	gini	79.406%	entropy	79.679%
14	gini	80.254%	entropy	79.678%
15	gini	79.231%	entropy	78.798%

Πίνακας 5: Αποτελέσματα Δέντρων Αποφάσεων



Γράφημα 1: Επιδόσεις Δέντρων Αποφάσεων

Όπως φαίνεται και μέσω του γραφήματος η μέθοδος Gini πλην μίας περίπτωσης είναι διαρκώς καλύτερη της μεθόδου entropy. Μία πολύ ενδιαφέρουσα παρατήρηση είναι η επιβεβαίωση των όσων αναμέναμε για τις διάφορες τιμές του βάθους του δέντρου. Παρατηρούμε πως οι καλύτερες τιμές λαμβάνονται γύρω από το μεσαίο βάθος που είναι το 10. Αριστερά και δεξιά του παρατηρούμε και για τις δύο μεθόδους πως η απόδοση φθίνει.

Καταλαβαίνει κανείς πόσο σημαντική είναι η επιλογή του βέλτιστου βάθους να μην υποπίπτει σε overfitting ή υπεργενίκευση.

3.6.4 Επιλογή υπερπαραμέτρων Support Vector Machine

Για την εύρεση του καταλληλότερου μοντέλου ταξινομητή μηχανής SVM μας ενδιέφεραν δύο υπερπαράμετροι. Αυτές ήταν η C και η γάμμα. Κάθε μία καθορίζει σε μεγάλο βαθμό το τελικό υπερεπίπεδο που δημιουργείται από το μοντέλο. Το υπερεπίπεδο με τη σειρά του αποτελεί το κριτήριο του μοντέλου προκειμένου να εκτιμήσει την κλάση ενός διανύσματος χαρακτηριστικών.

Η υπερπαράμετρος γάμμα καθορίζει το πόσο πολύ επηρεάζεται το τελικό αποτέλεσμα από ένα μοναδικό διάνυσμα χαρακτηριστικών. Χαμηλή τιμή της παραμέτρου γάμμα συνεπάγεται μακρινή συσχέτιση με το διάνυσμα, ενώ υψηλή τιμή κοντινή συσχέτιση. Στην πράξη η τιμή της παραμέτρου σχετίζεται με το αντίστροφο της ακτίνας επιρροής των διανυσμάτων που επιλέγεται από το μοντέλο για τον καθορισμό του υπερεπιπέδου διαχωρισμού.

Η υπερπαράμετρος C επηρεάζει το κατά πόσο εκτιμώνται σωστά οι κλάσεις των διανυσμάτων που χρησιμοποιούνται κατά την εκμάθηση. Επιπλέον καθορίζει τη μορφή της επιφάνειας του υπερεπιπέδου διαχωρισμού. Τα δύο αυτά χαρακτηριστικά του μοντέλου είναι αντιστρόφως ανάλογα, αφού αύξηση του ενός συνεπάγεται μείωση του άλλου. Η τιμή της υπερπαραμέτρου C είναι αυτή που καθορίζει τον λόγο που χαρακτηρίζεται τη μεταξύ τους σχέση. Χαμηλή τιμή της υπερπαραμέτρου C θα έχει ως αποτέλεσμα η επιφάνεια του επιπέδου διαχωρισμού να είναι “λεία”. Αντίστροφα, υψηλή τιμή της C έχει ως στόχο την επιτυχή ταξινόμηση όλων των διανυσμάτων που χρησιμοποιούνται στη διαδικασία εκμάθησης. Αυτό συμβαίνει καθώς δίνει στη μηχανή SVM το περιθώριο να επιλέξει περισσότερα διανύσματα εντός της ακτίνας επιρροής για την κατασκευή του υπερεπιπέδου.

3.6.4.1 Αναζήτηση πλέγματος

Η επιλογή των κατάλληλων υπερπαραμέτρων μίας μηχανής SVM είναι μία σύνθετη διαδικασία, ενώ υπάρχουν πολλές προσεγγίσεις για το πως αυτή θα πραγματοποιηθεί. Η μελέτη που πραγματοποιήσαμε στηρίχθηκε στη μέθοδο της αναζήτησης πλέγματος, όπως περιγράφεται στο έργο των (Chih-Wei Hsu, Chih-Chung Chang, 2008). Η αναζήτηση πλέγματος, πρόκειται για ένα χώρο δύο διαστάσεων, τον οποίο καταλαμβάνουν σημεία με τιμές για τις παραμέτρους (C, γάμμα). Η τιμή που λαμβάνει κάθε σημείο είναι συνήθως μία μετρική επίδοσης του μοντέλου. Στην περίπτωση της μελέτης, έγινε χρήση της μετρικής F1 score, όπως αυτή περιγράφεται στην παράγραφο 3.6.2.

Για το εύρος των τιμών που μελετήθηκαν ακολουθήσαμε την προτεινόμενη μέθοδο των εκθετικά αυξανόμενων ακολουθιών. Πιο συγκεκριμένα, δοκιμάσαμε μοντέλα των οποίων οι υπερπαράμετροι C, γάμμα είχαν τιμές στο εύρος -5 έως 15 και -15 έως 5 αντίστοιχα. Οι παραπάνω κλίμακες είναι σε λογαριθμική κλίμακα βάσης 2. Για κάθε ένα συνδυασμό που προέκυψε για τις δύο υπερπαραμέτρους καταγράψαμε τη μετρική F1 score που πέτυχε η μηχανή SVM. Όπως περιγράφηκε και στην παράγραφο 3.6.2 η μέθοδος που ακολουθήσαμε για την εύρεση της επίδοσης του μοντέλου ήταν η K-Fold Cross-Validation για k = 5.

Στη συνέχεια παρουσιάζεται ο χάρτης heatmap για τις τιμές F1 score των συνδυασμών C και γάμμα. Στον χάρτη αυτό η κλίμακα ξεκινά από το πράσινο όπου βρίσκονται οι πιο χαμηλές τιμές και ολοκληρώνεται στο κόκκινο, χρώμα που πετυχαίνει η υψηλότερη τιμή F1 score.

gamma \ C	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	**12**	13	14	15
-15	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.623	0.705	0.705	0.705	0.705	0.738	0.772	0.779	0.778	0.783	0.785	0.784	0.784	0.784
-14	0.000	0.000	0.000	0.000	0.000	0.000	0.623	0.705	0.705	0.705	0.705	0.738	0.772	0.780	0.779	0.783	0.786	0.785	0.784	0.784	0.786
-13	0.000	0.000	0.000	0.000	0.000	0.621	0.705	0.705	0.705	0.705	0.738	0.772	0.779	0.779	0.784	0.785	0.786	0.786	0.786	0.791	0.794
-12	0.000	0.000	0.000	0.000	0.620	0.705	0.705	0.705	0.705	0.738	0.772	0.780	0.779	0.784	0.786	0.788	0.788	0.788	0.795	0.796	0.799
-11	0.000	0.000	0.000	0.618	0.705	0.705	0.705	0.705	0.738	0.773	0.780	0.780	0.785	0.788	0.789	0.790	0.794	0.797	0.800	0.802	0.802
-10	0.000	0.000	0.616	0.705	0.705	0.705	0.705	0.738	0.773	0.781	0.780	0.786	0.791	0.796	0.794	0.797	0.802	0.802	0.801	0.802	0.803
-9	0.000	0.610	0.705	0.705	0.705	0.705	0.738	0.773	0.782	0.782	0.789	0.796	0.796	0.798	0.801	0.802	0.803	0.802	0.803	0.808	0.806
-8	0.597	0.705	0.705	0.705	0.705	0.738	0.774	0.785	0.787	0.792	0.797	0.799	0.802	0.802	0.803	0.803	0.803	0.808	0.807	0.806	0.807
-7	0.705	0.705	0.705	0.705	0.738	0.776	0.789	0.790	0.795	0.801	0.803	0.805	0.803	0.804	0.804	0.807	0.808	0.806	0.807	0.810	0.812
-6	0.705	0.705	0.705	0.740	0.779	0.792	0.796	0.797	0.801	0.803	0.804	0.805	0.804	0.806	0.808	0.806	0.808	0.810	0.813	0.811	0.810
-5	0.705	0.705	0.744	0.782	0.794	0.798	0.798	0.801	0.804	0.805	0.807	0.807	0.808	0.806	0.806	0.812	0.814	0.811	0.809	0.810	0.812
-4	0.706	0.750	0.783	0.793	0.800	0.804	0.802	0.805	0.807	0.807	0.807	0.805	0.808	0.813	0.812	0.811	0.811	**0.815**	0.813	0.812	0.812
-3	0.763	0.781	0.792	0.801	0.803	0.804	0.807	0.809	0.806	0.804	0.809	0.811	0.811	0.812	0.813	0.811	0.813	0.812	0.810	0.810	0.810
-2	0.782	0.789	0.798	0.803	0.805	0.807	0.807	0.804	0.807	0.809	0.812	0.812	0.813	0.812	0.812	0.809	0.809	0.811	0.808	0.808	0.803
-1	0.788	0.794	0.803	0.803	0.804	0.803	0.803	0.805	0.809	0.812	0.811	0.811	0.810	0.810	0.807	0.805	0.802	0.799	0.799	0.796	0.792
0	0.790	0.797	0.799	0.802	0.803	0.806	0.807	0.805	0.806	0.805	0.805	0.804	0.804	0.801	0.799	0.798	0.796	0.793	0.793	0.790	0.787
1	0.788	0.795	0.797	0.801	0.799	0.799	0.802	0.801	0.803	0.803	0.800	0.796	0.796	0.795	0.792	0.792	0.788	0.788	0.785	0.783	0.786
2	0.784	0.790	0.792	0.797	0.796	0.791	0.795	0.799	0.799	0.794	0.792	0.793	0.796	0.793	0.789	0.786	0.786	0.779	0.780	0.774	0.775
3	0.771	0.787	0.792	0.793	0.796	0.794	0.793	0.797	0.789	0.790	0.785	0.783	0.784	0.782	0.774	0.773	0.770	0.773	0.773	0.770	0.766
4	0.764	0.775	0.784	0.788	0.790	0.789	0.787	0.787	0.787	0.785	0.780	0.778	0.768	0.765	0.766	0.765	0.763	0.760	0.757	0.754	0.753
5	0.750	0.765	0.776	0.779	0.783	0.780	0.781	0.773	0.773	0.770	0.765	0.765	0.762	0.760	0.760	0.755	0.752	0.750	0.751	0.751	0.752

Γράφημα 2: Χάρτης Heatmap

Όπως μπορούμε να δούμε και από το παραπάνω, καθώς το C αυξάνεται φαίνεται να πετυχαίνουμε καλύτερες επιδόσεις. Όπως αναμέναμε, υψηλό C συνεπάγεται καλύτερες εκτιμήσεις. Βέβαια, από ένα σημείο και μετά παρατηρούμε πως η αύξηση του C δεν επηρεάζει την επίδοση αφού εμφανίζεται κατώφλι κοντά στην τιμή C = 11.

Αναφορικά με το γάμμα, παρατηρούμε πως πετυχαίνει καλύτερη επίδοση για τιμές μεταξύ -6 και -1. Φαίνεται λοιπόν πως για να πετύχουμε καλύτερη επίδοση η επιφάνεια του υπερεπιπέδου δεν πρέπει να είναι πολύ λεία καθώς επίσης ούτε πολύ τραχιά. Με άλλα λόγια, πρέπει να αποτελείται μεν από ικανό αριθμό βοηθητικών διανυσμάτων χωρίς όμως ο αριθμός αυτός να είναι υπερβολικά μεγάλος.

Τέλος, με έντονη γραμματοσειρά παρατηρούμε την καλύτερη επίδοση που πετύχαμε για τη μηχανή SVM. Αυτή είναι η 81,54 % και επιτυγχάνεται για C = 12 και γάμμα = -4. Φυσικά, η τοποθέτηση της βέλτιστη τιμή κοντά στον συνδυασμό αυτό ήταν αναμενόμενη. Ο λόγος είναι πως βρίσκεται ανάμεσα στο βέλτιστο διάστημα της υπερπαραμέτρου γάμμα καθώς επίσης ιδιαίτερα κοντά στο κατώφλι της υπερπαραμέτρου C.

3.7 Τελική επιλογή μοντέλου

Έχοντας καταλήξει στο αποδοτικότερο δέντρο απόφασης και την αποδοτικότερη μηχανή SVM πρέπει να καταλήξουμε στην επιλογή του καταλληλότερου για τη μελέτη των δημοσιεύσεων του Twitter. Παρότι η μηχανή SVM απέδωσε καλύτερα ως προς το δέντρο απόφασης (81,54% έναντι 81,15%) αποφασίστηκε να πραγματοποιηθεί επιπλέον διερεύνηση ανάμεσα στα δύο μοντέλα. Για τον σκοπό αυτό, τα δύο επικρατέστερα μοντέλα δοκιμάστηκαν σε νέα κατάτμηση των δεδομένων διανυσμάτων χαρακτηριστικών. Χρησιμοποιώντας και πάλι τη μέθοδο K-Fold Cross-Validation για k = 5 μετρήσαμε το F1 score που επετεύχθη από κάθε μοντέλο.

Ο πίνακας που ακολουθεί παρουσιάζει τα αποτελέσματα που λάβαμε από τη νέα μελέτη. Επιπλέον της μετρικής F1 score, συμπεριλάβαμε και αυτούσιες τις μετρικές precision και recall προκειμένου να έχουμε μία καλύτερη εικόνα για το εκάστοτε μοντέλο.

3.7.1 Τελικά αποτελέσματα

	Δέντρο Απόφασης	Μηχανή SVM
Ακρίβεια (precision)	82.20%	83.69%
Ανάκληση (recall)	78.47%	79.03%
F1 score	80.27%	81.29%

Πίνακας 6: Σύγκριση επικρατέστερων μοντέλων

Όπως φαίνεται από τα αποτελέσματα του παραπάνω πίνακα και τα δύο μοντέλα πετυχαίνουν εξαιρετικά αποτελέσματα, δεδομένου του προβλήματος που καλούνται να επιλύσουν. Το ποσοστό επιτυχία κοντά στο 80% είναι ιδιαίτερα ενθαρρυντικό για χρήση του συστήματος, καθώς μεταφράζεται σε επιτυχή αναγνώριση αξιοπιστίας για 8 στις 10 δημοσιεύσεις χρηστών του Twitter. Φυσικά, η μηχανή SVM νικά κατά κράτος το δέντρο απόφασης σε οποιαδήποτε από τις υπό μελέτη μετρικές, κάτι που οδήγησε στην επιλογή της

ως το αποδοτικότερο μοντέλο για την εκτίμηση της αξιοπιστία των δημοσιεύσεων του Twitter.

Αξίζει να σημειωθεί πως εξαιτίας της πολυπλοκότητας της υλοποίησης των μηχανών SVM, η διαδικασία εκμάθησης μπορεί να διαρκέσει πολύ περισσότερο από ότι στα δέντρα αποφάσεων. Κάτι τέτοιο όμως δεν αποτελεί πρόβλημα καθώς η εκμάθηση προηγείται της χρήσης του ταξινομητή σε δεδομένα πραγματικού χρόνου. Τροφοδοτώντας τον ταξινομητή με δημοσιεύσεις που έχουν ήδη συλλεχθεί από το Twitter, μπορεί κανείς να πραγματοποιήσει τη διαδικασία εκμάθησης χωρίς χρονικούς περιορισμούς. Μόλις ολοκληρωθεί η εκμάθηση, ο ταξινομητής μπορεί να τεθεί σε χρήση με δημοσιεύσεις πραγματικού χρόνου, αφού μπορεί με ταχύτητα και ακρίβεια να πραγματοποιήσει τις εκτιμήσεις του για την αξιοπιστία των δημιουργών τους.

4 Αναγνώριση Γεγονότων και εξαγωγή θεμάτων από τη ροή δεδομένων του Twitter

Σημαντικό συστατικό μέρος του προτεινόμενου συστήματος αποτελεί η αναγνώριση γεγονότων μέσα στη ροή δημοσιεύσεων. Για την τάχιστη αναγνώριση υλοποιήθηκε ειδικός μηχανισμός, ακολουθώντας σε μεγάλο βαθμό το σύστημα TopicSketch των (Xie *κ.ά.*, 2016), όπως αυτό περιγράφεται από τους δημιουργούς του. Ο μηχανισμός που προτείνεται αποτελεί τροποποιημένη έκδοση του εν λόγω συστήματος και εισαγάγει βελτιώσεις ως προς την επεξεργασία και αναγνώριση γεγονότων και θεμάτων προερχόμενων από δημοσιεύσεις.

4.1 Σύγχρονες μέθοδοι και προσεγγίσεις

Σκοπός της ενότητας αυτής είναι να παρουσιαστούν σύγχρονες μέθοδοι και προσεγγίσεις ως προς τον ταχύ εντοπισμό νέων ειδήσεων στο Twitter. Οι μέθοδοι που παρουσιάζονται δεν είναι οι μοναδικές προτεινόμενες στη βιβλιογραφία, καθώς ο τομέας είναι ιδιαίτερα ενεργός. Στην ενότητα αυτή γίνεται προσπάθεια να καταγραφούν οι δημοφιλέστερες, οι οποίες εμφανίζονται να πετυχαίνουν τα καλύτερα αποτελέσματα.

Οι (Petrović, Osborne και Lavrenko, 2010) προτείνουν ένα σύστημα εντοπισμού ειδήσεων μέσα σε ροή γεγονότων με εφαρμογή στο Twitter. Το σύστημα που παρουσιάζουν έχει ως σκοπό την εύρεση δημοσιεύσεων - γειτόνων. Θέλοντας να πετύχουν καλή επίδοση τόσο στη μνήμη όσο και στον χρόνο εκτέλεσης, για τον εντοπισμό δύο γειτόνων χρησιμοποιούν και επεκτείνουν τη μέθοδο Locality Sensitive Hashing. Μέσω της τροποποιημένης έκδοσης που προτείνουν, συγκεντρώνουν τις δημοσιεύσεις ταξινομημένες σε νήματα γεγονότων. Φυσικά μία δημοσίευση μπορεί να ανήκει σε παραπάνω από ένα νήματα. Έχοντας ως προϋπόθεση την καλή απόδοση εφαρμόζουν περιορισμούς και στη δημιουργία νημάτων, προκειμένου να διατηρούν τον χώρο του προβλήματος σταθερό. Μία νέα είδηση στο σύστημα που προτείνεται ισοδυναμεί με ένα νήμα γεγονότων που μεγαλώνει γρηγορότερα από τα υπόλοιπα. Σε κάθε επανάληψη εκτέλεσης του αλγορίθμου, το σύστημα εμφανίζει ως ειδήσεις τα νήματα που πετυχαίνουν τους υψηλότερους ρυθμούς ανάπτυξης.

Οι (Osborne, Petrovic και McCreadie, 2012) προτείνουν τη βελτίωση του συστήματος που είχαν αναπτύξει το 2010. Παρατηρούν πως η ομαδοποίηση των δημοσιεύσεων σε νήματα και η εμφάνιση των ταχέως αναπτυσσόμενων περιέχει πολλές φορές θόρυβο στα

αποτελέσματα. Πιο συγκεκριμένα, εκτός από εμφάνιση πραγματικών ειδήσεων μπορεί να περιέχει και ανούσιες πληροφορίες που ανταλλάσσουν οι χρήστες του Twitter μεταξύ τους. Στα πλαίσια της εμφάνισης ποιοτικότερων αποτελεσμάτων οι ερευνητές αναζητούν μία μέθοδο φιλτραρίσματος. Διαπιστώνουν πως όταν ένα σημαντικό γεγονός συμβαίνει, εκτός από τάχιστη αύξηση του σχετικού νήματος, το αντίστοιχο λήμμα της γνωστής ηλεκτρονικής εγκυκλοπαίδειας Wikipedia παρουσιάζει κίνηση. Όπου κίνηση εννοείται ο αριθμός των προβολών του λήμματος, ο οποίος είναι αφύσικα μεγάλος σε σχέση με τον μέσο αριθμό προβολών για το λήμμα αυτό. Διατηρώντας χρονικά παράθυρα με τις προβολές των λημμάτων της Wikipedia, οι ερευνητές καταγράφουν τις περιπτώσεις που κάποιο λήμμα παρουσιάζει ανωμαλίες (outliers). Επιπλέον, γίνεται μία προσπάθεια συσχέτισης των δημοσιεύσεων – αντιπροσώπων των νημάτων με τα λήμματα αυτά. Το βελτιωμένο σύστημα, προτού παρουσιάσει τα ταχύτερα αποτελέσματα πραγματοποιεί το παραπάνω φιλτράρισμα, εξαλείφοντας έτσι τον θόρυβο που υπήρχε χωρίς αυτό.

Οι (Petrović *κ.ά.*, 2012) προσθέτουν ακόμα μία βελτίωση στο σύστημά τους. Διαπιστώνουν πως οι πολλές διαφορετικές εκφράσεις για την περιγραφή του ίδιου γεγονότος δημιουργούν εσφαλμένα αποτελέσματα στην αναγνώριση ειδήσεων. Πιο συγκεκριμένα, μία δημοσίευση αποτελεί τη γραπτή αποτύπωση του ανθρώπινου λόγου. Ως φυσικό επακόλουθο, οι εκφράσεις που περιλαμβάνονται στις δημοσιεύσεις εντοπίζονται σε διάφορες παραλλαγές, τόσο σε επίπεδο λέξεων όσο και σε επίπεδο ολόκληρων φράσεων. Κάνοντας χρήση της βιβλιοθήκης Wordnet (Fellbaum, 1998), δημιουργούν μία αυτοματοποιημένη διαδικασία, η οποία συσχετίζει τις παραλλαγές ίδιων εκφράσεων μεταξύ των κειμένων. Μέσω της διαδικασίας αυτής οι εκφράσεις θεωρούνται πλέον ως ταυτόσημοι όροι. Μία ακόμα προσθήκη στο σύστημα αποτελεί και η θεώρηση των hashtags ως λεκτικών όρων των δημοσιεύσεων, έχοντας αφαιρέσει από αυτά τον ειδικό χαρακτήρα #. Σύμφωνα με τη μελέτη τους οι παραπάνω προσθήκες βελτιώνουν αισθητά τις επιδόσεις του συστήματος που πρότειναν το 2010.

Οι (Benhardus και Kalita, 2013) προτείνουν τεχνικές για εντοπισμό όρων οι οποίοι είναι δημοφιλείς ανάμεσα στη ροή δεδομένων του Twitter. Αν και δεν γίνεται προσπάθεια πρώτου εντοπισμού των όρων αυτών, η έρευνά τους είναι ιδιαίτερα σημαντική καθώς αποτελεί βάση για εντοπισμό ειδήσεων. Πιο συγκεκριμένα, προτείνεται η λεξικογραφική μελέτη μίας δημοσίευσης του Twitter ως μέθοδος για εντοπισμό ειδήσεων. Αρχικά, εισάγεται η έννοια του όγκου (ή συχνότητας) όρων, όπου καταγράφεται το πλήθος εμφανίσεώς τους. Στη συνέχεια εισάγεται η αντίστροφη συχνότητα δημοσίευσης που δείχνει το πλήθος των δημοσιεύσεων που περιλαμβάνουν έναν όρο. Σε περαιτέρω βελτίωση του παραπάνω προτείνεται από τους ερευνητές ο ορισμός της κανονικοποιημένης συχνότητας όρου, που αναφέρει τη συχνότητα εμφάνισης του όρους ως προς το σύνολο των όρων όλων των δημοσιεύσεων. Προτού πραγματοποιήσουν μετρήσεις, οι ερευνητές προσπαθούν να αφαιρέσουν όρους οι οποίοι δημιουργούν θόρυβο στα αποτελέσματα. Αφαιρούν εξαιρούμενες λέξεις, hashtags, αναφορές σε χρήστες και υπερσυνδέσμους. Επιπλέον ορίζουν εντροπία ενός όρου, η οποία καθορίζει αν ένας όρος αποτελεί θόρυβο ή όχι. Τέλος πραγματοποιούν πειράματα τόσο με μοναδικούς όρους όσο και με ζευγάρια συνεμφανιζόμενων όρων. Με βάση τα παραπάνω κριτήρια οι ερευνητές παρουσιάζουν εξαιρετικά αποτελέσματα ως προς τη δημοφιλία όρων, χωρίς όμως να μπορούν να αναδείξουν την έναρξη αυτής.

Το 2016 προτείνεται το σύστημα TopicSketch των (Xie *κ.ά.*, 2016). Ουσιαστικά πρόκειται για αναθεώρηση του αντίστοιχου συστήματος που είχε προταθεί από τους ίδιους το 2013. Το σύστημα αυτό αποσκοπεί στην τάχιστη εύρεση μίας νέα είδησης ανάμεσα σε ροή δημοσιεύσεων του Twitter. Κάνοντας χρήση λεξικογραφικών μεθόδων συγκεντρώνει σε πρώτο στάδιο τον όγκο των ζευγαριών και των τριπλετών συνεμφανιζόμενων όρων. Αφού τα συγκεντρώσει, ανά χρονικά διαστήματα ορίζει και υπολογίζει τις μετρικές ταχύτητα και επιτάχυνση διάδοσης όρων για όλους τους όρους που συλλέγει. Για κάθε έναν από τους όρους διατηρεί ιστορικό, το οποίο ενημερώνει διαρκώς με τις νέες τιμές. Όταν εντοπίζει κάποια ανωμαλία στην επιτάχυνση ενός όρου ως προς το ιστορικό του, τότε διερευνά την πιθανότητα εμφάνισης νέας είδησης. Εφόσον αποφανθεί για την ύπαρξη νέων ειδήσεων ή αλλιώς για όρους που ξεπερνούν ένα προκαθορισμένο όριο επιτάχυνσης, το σύστημα τις συγκεντρώνει και τις παρουσιάζει ως αποτελέσματα. Προκειμένου να είναι ιδιαίτερα αποδοτικό εφαρμόζει επιπλέον της διαδικασίας σημαντικές βελτιστοποιήσεις. Αρχικά πραγματοποιεί φιλτράρισμα ως προς τους υπό μελέτη όρους. Αφαιρεί τις εξαιρούμενες και τις σπάνιες λέξεις, ειδήσεις με πολύ μικρό αριθμό δημοσιεύσεων καθώς και ειδήσεις με μεγάλη τυχαιότητα στους όρους των δημοσιεύσεών τους. Τέλος, προκειμένου να πετύχει αποδοτική χρήση του χώρου αποθήκευσης, χρησιμοποιεί τεχνικές κατακερματισμού για τον τρέχον πίνακα όρων.

4.2 Πλήθος εμφανίσεων και ταχύτητα διάδοσης όρων σε παράθυρα χρόνου

Το σύστημα που αναπτύχθηκε μελετά το πλήθος εμφανίσεων όρων καθώς και τη ταχύτητα διάδοσή τους μέσα σε ισομήκη χρονικά παράθυρα. Για να το επιτύχει ορίζει τις μετρικές Όγκος Όρου και Ταχύτητα Όρου και οι οποίες περιγράφονται παρακάτω.

4.2.1 Όγκος Όρου

Ο όγκος ενός όρου W_i τη χρονική στιγμή t είναι το πλήθος εμφάνισης του όρου αυτού από την αρχή του χρόνου έως το πέρας του τρέχοντος χρονικού παραθύρου και ορίζεται ως εξής:

$$W_i(t) = \sum_j w_{ij}(t) + W_i(t-1)$$

όπου w_{ij} είναι το πλήθος εμφάνισης του όρου w_i στη δημοσίευση j, η οποία δημοσιεύτηκε στο χρονικό διάστημα από t έως $t-1$.

4.2.2 Ταχύτητα Διάδοσης Όρου

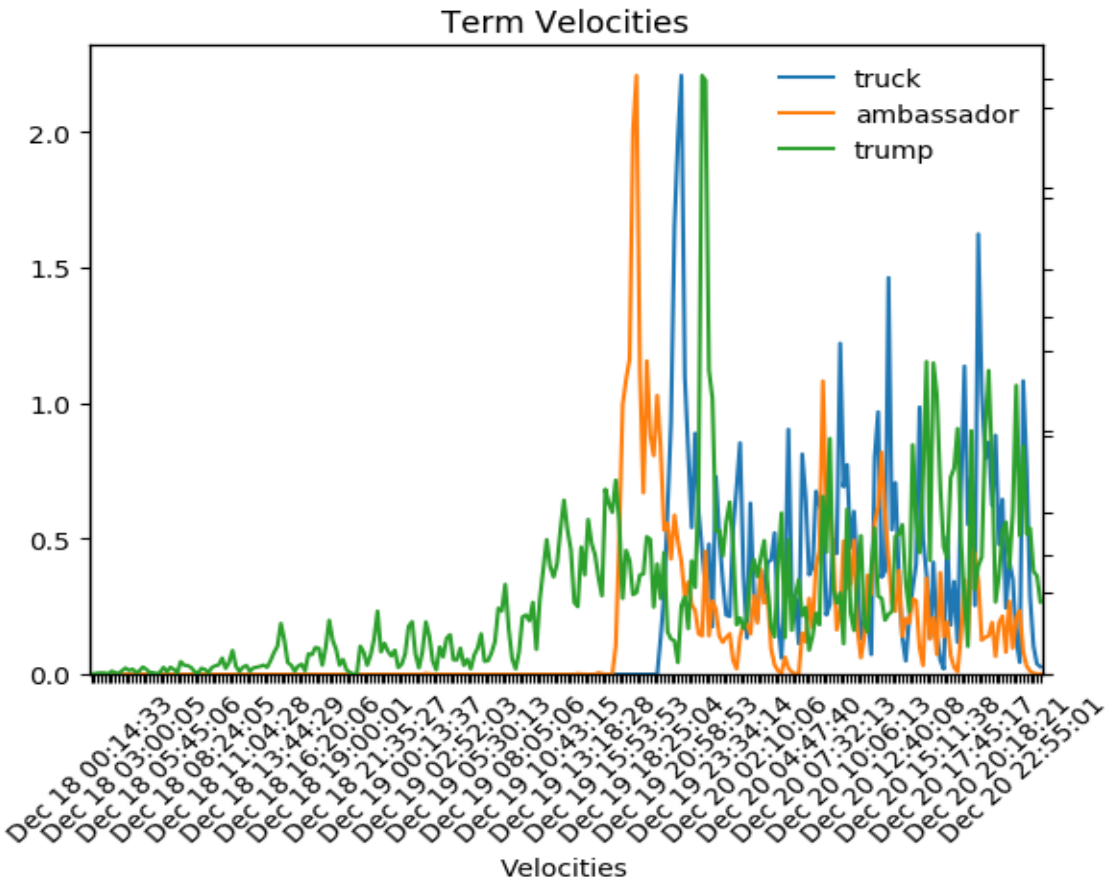
Η ταχύτητα ενός όρου v_i ορίζεται ως εξής:

$$v_i(t) = \sum_{t_i \le t} \frac{W_i(t) \cdot exp\left(\frac{t_i - t}{\Delta T}\right)}{\Delta T}$$

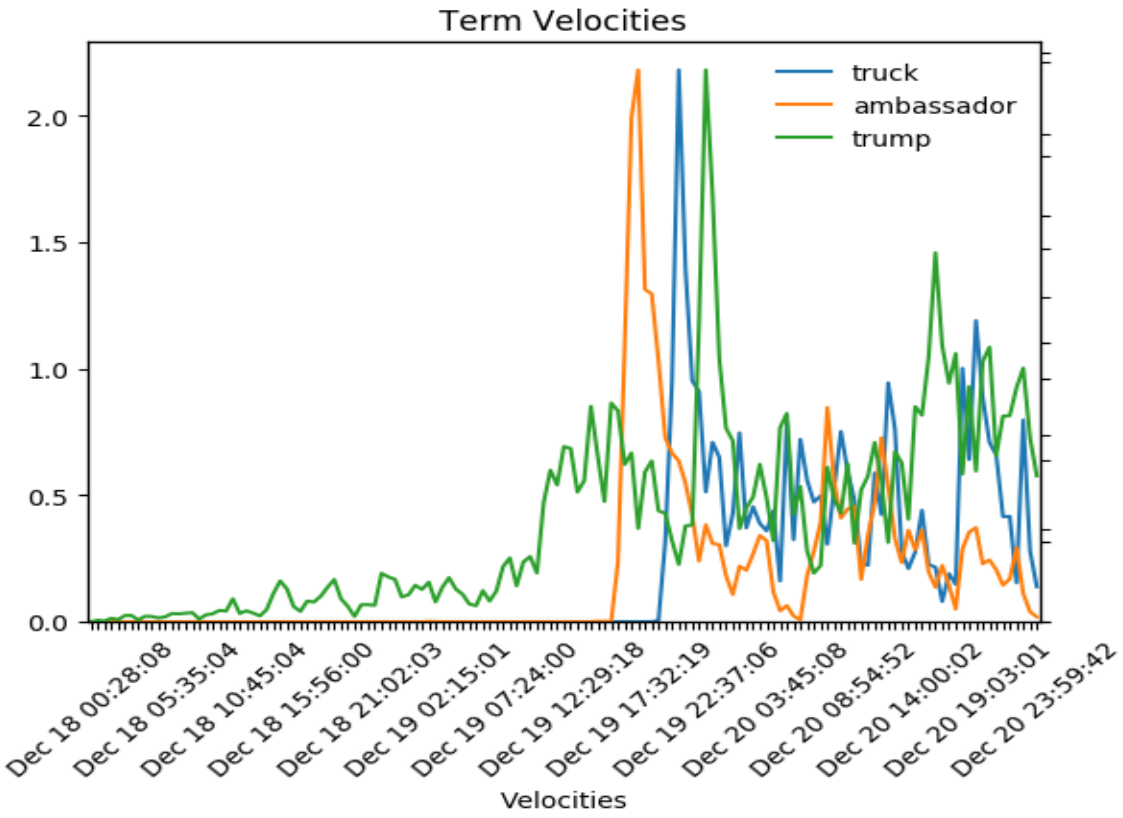
Όπου ΔT το μέγεθος του χρονικού παραθύρου σε δευτερόλεπτα, $W_i(t)$ ο όγκος του όρου w_i και t_i η χρονική στιγμή της δημοσίευσης που περιλαμβάνει τον όρο αυτό.

4.2.3 Χρονικό Παράθυρο

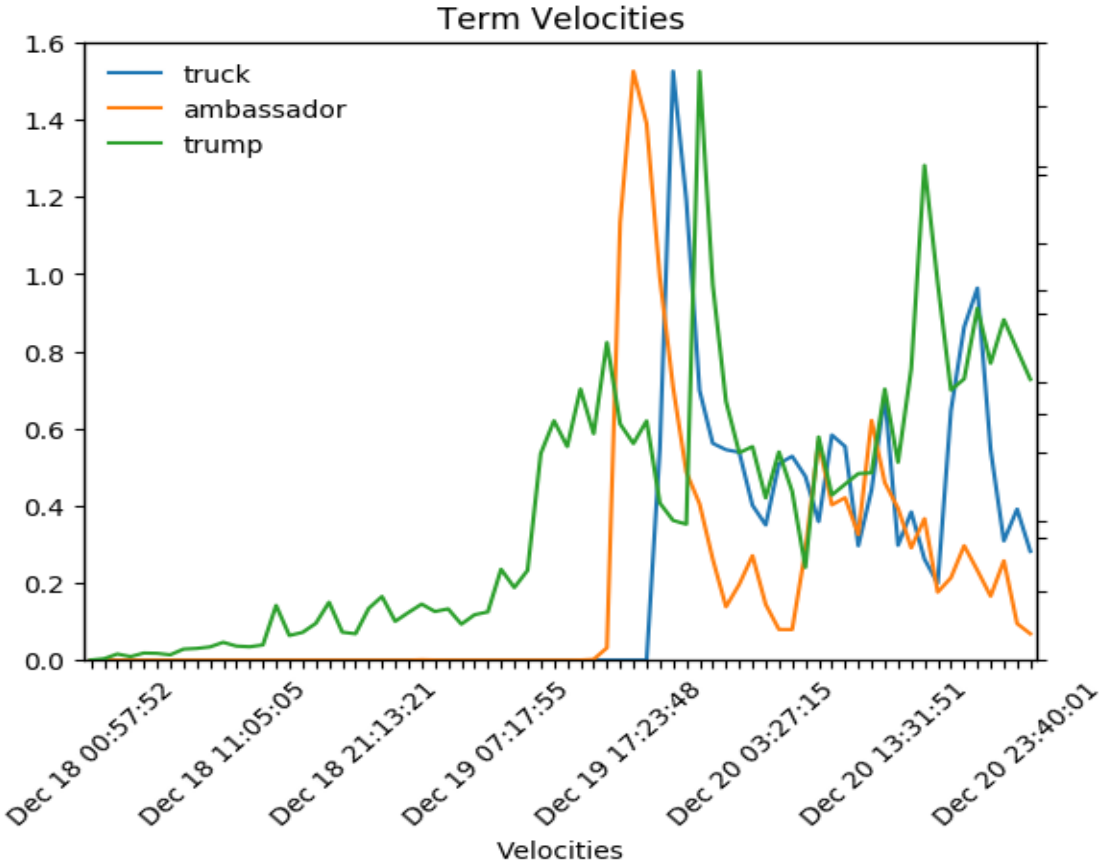
Το χρονικό παράθυρο είναι μία παράμετρος που καθορίζει ως ένα βαθμό τα παραγόμενα αποτελέσματα. Προκειμένου να επιλεχθεί η κατάλληλη τιμή της παραμέτρου πραγματοποιήθηκαν μετρήσεις, οι οποίες παρουσιάζονται στη συνέχεια.



Γράφημα 3: Χρονικό παράθυρο 15 λεπτών



Γράφημα 4: Χρονικό παράθυρο 30 λεπτών

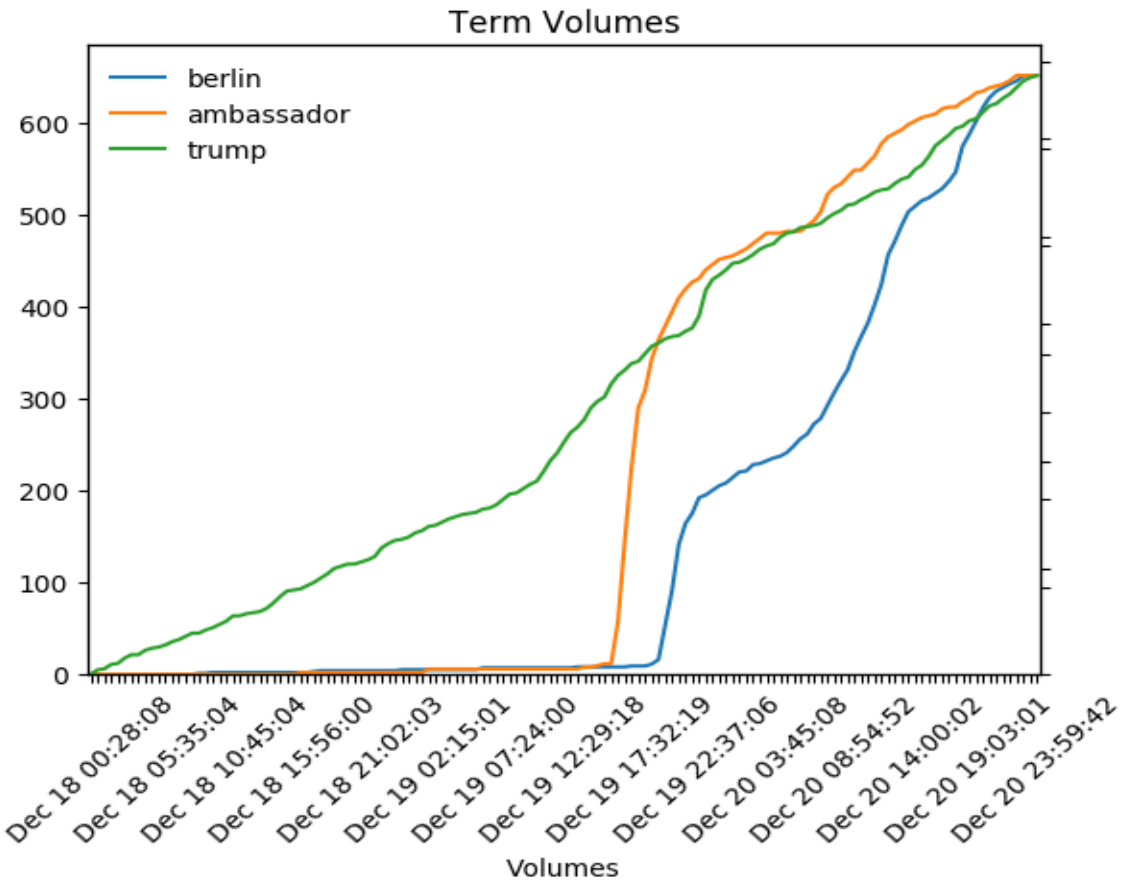


Γράφημα 5: Χρονικό παράθυρο 60 λεπτών

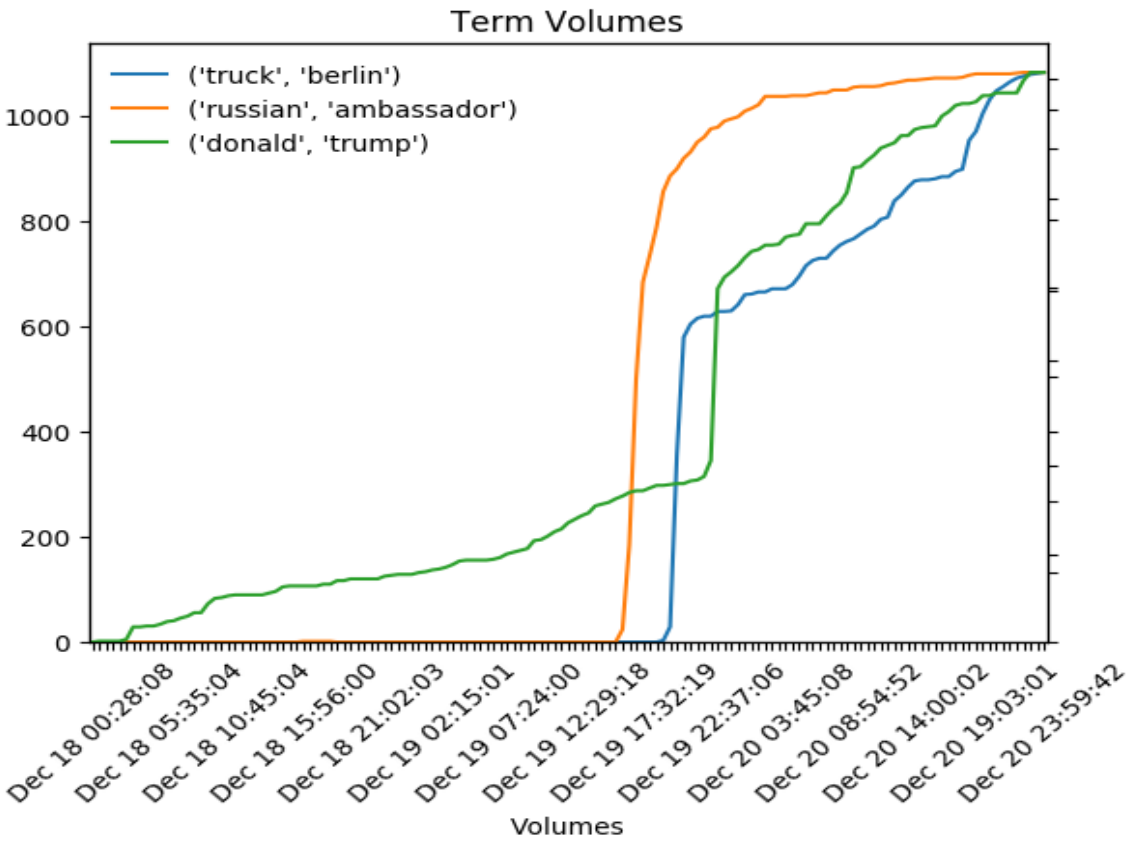
Οι παραπάνω γραφικές παραστάσεις απεικονίζουν την ταχύτητα διάδοσης συγκεκριμένων όρων μέσα σε κοινό χρονικό διάστημα για διαφορετικές τιμές του χρονικού παραθύρου. Μπορεί να γίνει εύκολα αντιληπτό πως ένα ιδιαίτερα μικρό παράθυρο παράγει θόρυβο, καθώς είναι υπερευαίσθητο σε μεταβολές στη ταχύτητα διάδοσης των όρων. Ένα μεγάλο παράθυρο από την άλλη φαίνεται να μην είναι ικανό να καταγράψει επακριβώς όλη την πληροφορία αναφορικά με την ταχύτητα διάδοσης. Τέλος, τα δεδομένα που λαμβάνουμε προέρχονται από λογαριασμούς χρηστών και ένα διάστημα μισής ώρας αποτελεί μία ρεαλιστική επιλογή για εντοπισμό νέων γεγονότων. Για όλους τους παραπάνω λόγους το σύστημα που υλοποιήσαμε κάνει χρήση παραθύρου ίσου με 30 λεπτά.

4.2.4 Ζευγάρια Όρων

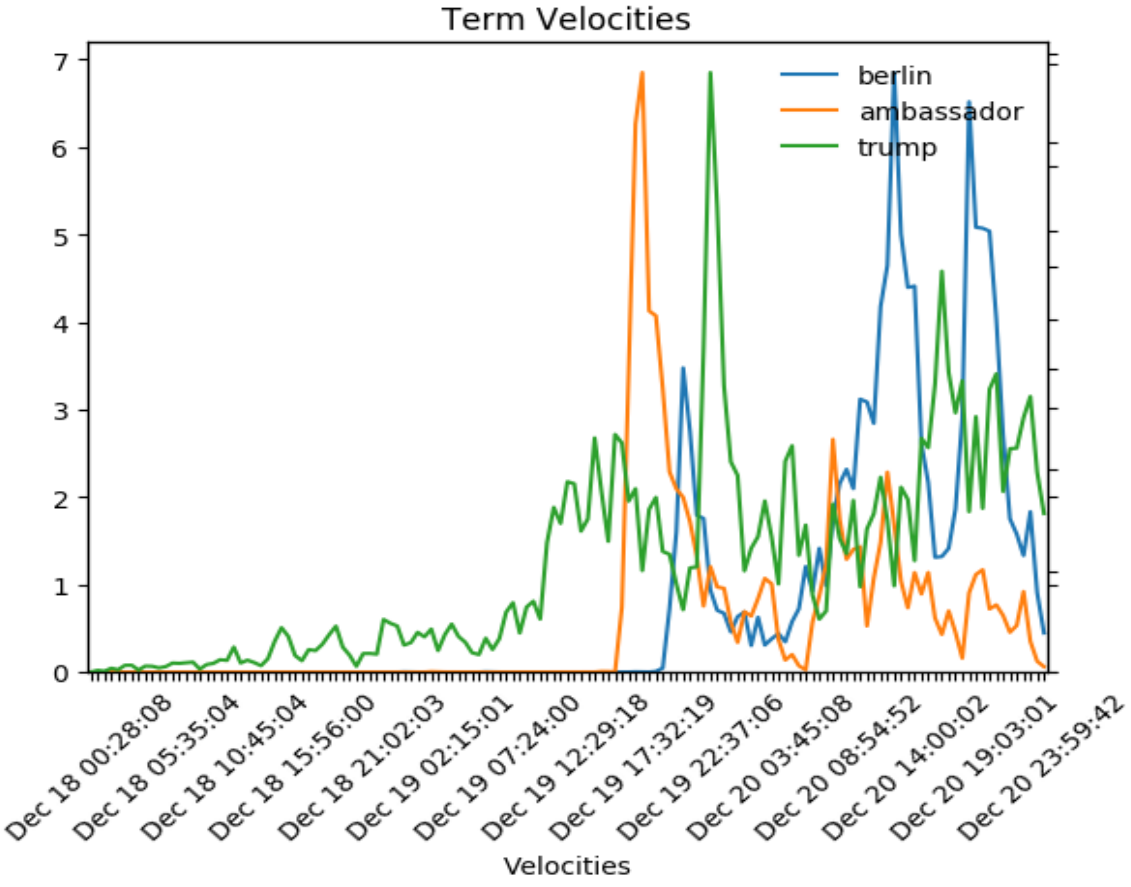
Μία ιδιαίτερα σημαντική παράμετρος για το σύστημα που υλοποιήθηκε αποτελεί η καταγραφή ζευγαριών όρων έναντι των μονών όρων. Πιο συγκεκριμένα, θέλοντας να παράγουμε πιο ρεαλιστικά και ποιοτικά δεδομένα, εστιάσαμε στον εντοπισμό όρων που συνεμφανίζονται ανάμεσα στις δημοσιεύσεις. Διατηρώντας τους ίδιους ορισμούς ως προς τον Όγκο Όρου και Ταχύτητα Όρου, θεωρήσαμε ως όρους τα ζευγάρια μονών όρων. Πλέον, θεωρούμε πως έχουμε εμφάνιση ενός όρου εντός μίας δημοσίευσης εφόσον και οι δύο υποόροι του εμφανίζονται εντός της. Ακολουθεί σύγκριση μεταξύ των αποτελεσμάτων που λάβαμε ως προς τον όγκο και την ταχύτητα διάδοσης για συγκεκριμένους όρους είτε ως μονούς όρους είτε ως ζευγάρια.



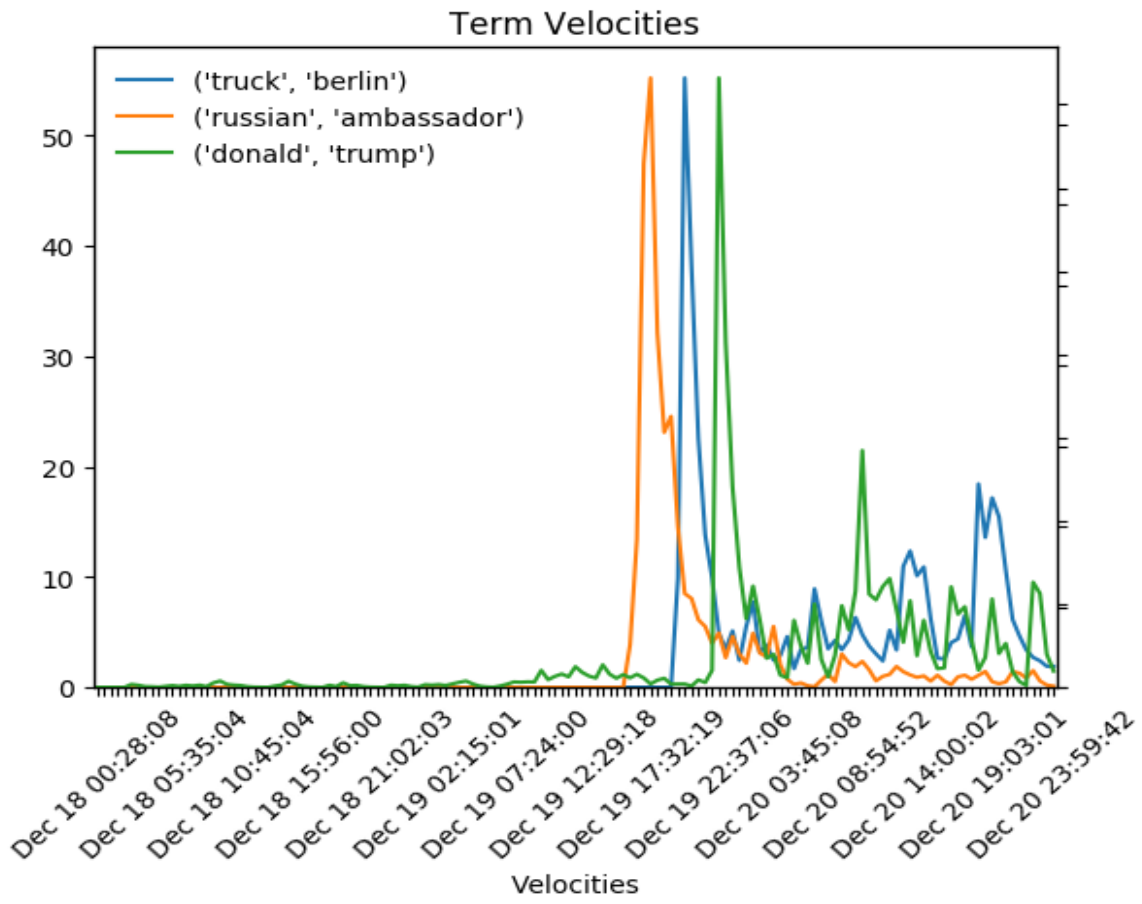
Γράφημα 6: Όγκος μονών όρων



Γράφημα 7: Όγκος ζευγαριών όρων



Γράφημα 8: Ταχύτητα μονών όρων



Γράφημα 9: Ταχύτητα ζευγαριών όρων

Μελετώντας τα αποτελέσματα εντοπίζουμε αισθητές διαφορές μεταξύ των δύο. Ως προς τον Όγκο όρων παρατηρούμε πως στα ζευγάρια όρων έχει προστεθεί συνιστώσα εξαιτίας και του όγκου του δεύτερου όρου, κάτι το οποίο ήταν αναμενόμενο. Παρόλα αυτά παρατηρούμε επίσης μεγαλύτερη καμπή σε συγκεκριμένα χρονικά σημεία για όλα τα ζευγάρια, η οποία υποδεικνύει την ύπαρξη κάποιων γεγονότων και με μονούς όρους δεν ήταν τόσο εμφανής. Ειδικά με τον όρο “Donald Trump” η καμπή αυτή δεν υπήρχε καθόλου κατά τη μελέτη του όρου “Trump”.

Ως προς την ταχύτητα διάδοσης, τα αποτελέσματα είναι ακόμα πιο εντυπωσιακά. Όσο αφορά τον όρο “Berlin” κάνοντας χρήση μονών όρων παρατηρούσαμε διάσπαρτα κάποιες μεγάλες κορυφές, υποδεικνύοντας κάποιο ή κάποια περιστατικά, χωρίς όμως να υπάρχει βεβαιότητα. Επιπλέον. Ως προς τους άλλους δύο όρους “Ambassador” και “Trump” παρατηρούμε πως περνώντας σε ζευγάρια όρων μηδενίζουμε τον θόρυβο που εμφανιζόταν. Το πιο σημαντικό όμως είναι η ξεκάθαρη υπόδειξη περιστατικών καθώς με τα ζευγάρια όρων διακρίνουμε πολύ εύκολα κορυφές στο πέρας του χρόνου. Κορυφές οι οποίες αντιστοιχίζονται σε πραγματικά γεγονότα και μάλιστα πολύ κοντά στη χρονική στιγμή που αυτά συνέβησαν. Ίσως το πιο σημαντικό είναι η κατά πολύ ταχύτερη αναγνώριση του γεγονότος που αφορά την τρομοκρατική επίθεση με φορτηγό στην αγορά του Βερολίνου.

Διακρίνοντας τα πολύ σημαντικά αυτά αποτελέσματα κρίναμε σκόπιμο να επικεντρωθούμε στη μελέτη ζευγαριών όρων καθώς μπορούν να οδηγήσουν στην αναγνώριση κάποιου γεγονότος πολύ ταχύτερα και με μεγαλύτερη σαφήνεια.

4.3 Αυτόματη ταχεία αναγνώριση διαδιδόμενων γεγονότων

Επόμενο βήμα στην αναγνώριση γεγονότων υπήρξε η αυτόματη αναγνώριση των γεγονότων. Αρχικά η επιλογή χρονικής περιόδου πραγματοποιήθηκε γνωρίζοντας εκ των υστέρων τα πραγματικά γεγονότα που συνέβησαν εντός της. Επιλέγοντας συγκεκριμένους όρους επιβεβαιώσαμε τον έγκαιρο και έγκυρο των γεγονότων από το σύστημα. Ο στόχος όμως παρέμενε ο αυτόματος εντοπισμός γεγονότων και όχι ο επιβλεπόμενος.

Για τις ανάγκες της αυτόματης αναγνώρισης υλοποιήθηκε ειδικός μηχανισμός. Ο μηχανισμός αυτός λαμβάνει υπόψιν του για κάθε χρονικά παράθυρο τους 20 όρους με τη μεγαλύτερη ταχύτητα διάδοσης. Επιπλέον, ορίζει ζώνες κρισιμότητας ως προς τους ταχέως διαδιδόμενους όρους. Θεωρεί κόκκινη ζώνη τις τρεις πρώτες θέσεις ως προς την ταχύτητα διάδοσης, πορτοκαλί τις τρεις επόμενες και κίτρινες τις τρεις που ακολουθούν. Είναι στην ευχέρεια του χρήστη του συστήματος να επιλέξει τον βαθμό ευαισθησίας ως προς τις εμφανιζόμενες ζώνες.

Ο μηχανισμός μετά το πέρας της επεξεργασίας ενός χρονικού παραθύρου λαμβάνει τους 20 ταχύτερα διαδιδόμενους όρους σαν είσοδο. Από αυτούς επιλέγει μόνο όσους υπερβαίνουν το όριο ταχύτητας 1 όρος ανά δευτερόλεπτο. Στη συνέχεια τους συγκρίνει με τους αντίστοιχους όρους του προηγούμενου παραθύρου, για τις ζώνες που έχει επιλέξει ο χρήστης να ενημερώνεται. Ολοκληρώνοντας τον παραπάνω έλεγχο καταλήγει σε γεγονότα που είτε έχει ενημερώσει στο παρελθόν για αυτά είτε βρίσκονται για πρώτη φορά εντός των κρίσιμων ζωνών. Προκειμένου να μην ενημερώσει εκ νέου για κάποιο γεγονός διατηρεί ιστορικό με τις ενημερώσεις που έχουν ήδη γίνει στον χρήστη. Αν τα γεγονότα αυτά δεν βρίσκονται στο ιστορικό, αποστέλλονται στον χρήστη ως νέα γεγονότα που προέκυψαν εντός του συγκεκριμένου παραθύρου. Εφόσον τα γεγονότα βρίσκονται στο ιστορικό, σβήνει από το λεξικό όρων τους όρους που τα αντιπροσωπεύουν, ώστε να μην διατηρούν υψηλές θέσεις δύο όροι που αναφέρονται στο ίδιο γεγονός.

Με τον παραπάνω μηχανισμό το σύστημα μπορεί να εντοπίζει αυτόματα νέα γεγονότα βασιζόμενο στην ταχύτητα διάδοσής του και το ιστορικό γεγονότων που διατηρεί. Επόμενο ζητούμενο της μελέτης ήταν η διερεύνηση της δυνατότητας εξαγωγής θεμάτων μέσω των θεμάτων αυτών.

4.4 Εξαγωγή θεμάτων μέσω συνένωσης

Μελετώντας τα αποτελέσματα που προέκυψαν από τον εντοπισμό γεγονότων μέσω ταχύτητας διάδοσης των όρων παρατηρήθηκε πως υπήρχαν ζευγάρια τα οποία εννοιολογικά ήταν ταυτόσημα. Επιπλέον, σε περίπτωση καταγραφής κάποιου σημαντικού γεγονότος καταλάμβαναν την πλειοψηφία των πρώτων θέσεων ως προς τη ταχύτητα διάδοσης των όρων. Απέκτησε έτσι ενδιαφέρον, η δυνατότητα συνένωσης ζευγαριών όρων που εννοιολογικά ήταν ταυτόσημα και η από κοινού μελέτη τους ως ένας όρος. Εισάγοντας τη λειτουργικότητα αυτή στο σύστημα, επιτεύχθηκε όχι μόνο η παραγωγή ποιοτικότερων αποτελεσμάτων αλλά και η εξαγωγή θεμάτων μέσα από τα αναδυόμενα γεγονότα.

Προκειμένου να γίνει εφικτή η συνένωση ζευγαριών όρων αναπτύχθηκε σχετική λειτουργικότητα. Με το πέρας της επεξεργασίας ενός χρονικού παραθύρου, το σύστημα συγκεντρώνει τους 20 ταχύτερους όρους ως προς τη διάδοση. Στη συνέχεια εφαρμόζεται ο κανόνας συνένωσης και υπολογίζεται εκ νέου η ταχύτητα διάδοσης όλων των όρων του παραθύρου. Ο κανόνας που περιγράφεται παρακάτω. Αξίζει να σημειωθεί πως πλέον δεν έχουμε ζευγάρια όρων αλλά σύνολα από όρους.

4.4.1 Κανόνας συνένωσης

Έστω τα παρακάτω σύνολα όρων:

$$A = \{a_k\},\ B = \{b_k\} \text{ και } C = \{c_m\}$$

Εφόσον υπάρχει $A \cap B$, $A \cap C$ και $B \cap C$ και το σύνολο $A \cup B \cup C$ δεν ξεπερνάει τους 5 μονούς όρους τότε οι όροι A, B, C αντικαθίστανται από τον όρο:

$$D = A \cup B \cup C = \{a_i, b_k, c_m\}$$

Ο κανόνας συνένωσης εφαρμόζεται εντός των 20 ταχύτερων όρων και στη συνέχεια υπολογίζεται εκ νέου η ταχύτητα διάδοσης. Επίσης, προκειμένου να αποδώσει καλύτερα εφαρμόζεται μετά το πέρας των πρώτων 10 χρονικών παραθύρων, αποφεύγοντας το φαινόμενο κρύας εκκίνησης, γνωστό ως “cold start”.

Η εφαρμογή του παραπάνω κανόνα οδήγησε στην εξαγωγή θεμάτων μέσω των γεγονότων που εντοπίστηκαν. Το σύστημα αντί να εμφανίζει στον χρήστη ταχέως διαδιδόμενα γεγονότα, επιστρέφει θέματα των 5 το πολύ όρων, τα οποία μπορούν να γίνουν ευκολότερα αντιληπτά από τον άνθρωπο. Επιπλέον, παρατηρήθηκε βελτίωση και στον εντοπισμό των θεμάτων καθώς παρατηρήθηκε πως κάποια από αυτά εντοπίστηκαν σε προηγούμενο παράθυρο από το αντίστοιχο του ζευγαριού όρων. Τέλος, αισθητή ήταν και η βελτίωση στον χρόνο που χρειάστηκε το σύστημα για την παραγωγή αποτελέσματος.

4.5 Ευριστικές Μέθοδοι - Βελτιστοποιήσεις

Έχοντας ένα σύστημα το οποίο εντοπίζει έγκαιρα και έγκυρα θέματα και γεγονότα με αυτόματο τρόπο προέκυψε η ανάγκη βελτίωσης ως προς τον χρόνο που χρειάζεται για την παραγωγή των αποτελεσμάτων. Σε αυτή την κατεύθυνση υλοποιήθηκαν βελτιστοποιήσεις οι οποίες περιγράφονται στην ενότητα αυτή.

4.5.1 Αφαίρεση Εξαιρούμενων Λέξεων

Κάθε μία δημοσίευση αποτελείται από ένα ελεύθερο κείμενο της αγγλικής, το οποίο αποτυπώνει προτάσεις που έχουν καταχωριστεί από κάποιον χρήστη. Άμεσο επακόλουθο του παραπάνω κατά τη μελέτη όρων είναι η ύπαρξη όρων, οι οποίοι δεν περιέχουν καμία ουσιαστική πληροφορία και προσθέτουν θόρυβο στα αποτελέσματα. Άρθρα, κοινότυπα ρήματα και επιρρήματα είναι χαρακτηριστικό παράδειγμα τέτοιων λέξεων.

Προκειμένου να μην επηρεάσουν τα αποτελέσματα, αυτές οι λέξεις έχουν αφαιρεθεί από τις δημοσιεύσεις. Μαζί τους έχει αφαιρεθεί και ο όρος “RT” που αποτελεί κοινότυπη συντομογραφία μεταξύ των χρηστών του twitter και σχετίζεται με αναδημοσίευση (ή αλλιώς “retweet”). Τέλος αφαιρέθηκαν οι εμφανίσεις του έτους 2016 και 2017 καθώς λόγω της περιόδου μελέτης εμφανίζονταν πολύ συχνά δημιουργώντας μη αποδεκτό θόρυβο. Η αφαίρεση των εξαιρούμενων λέξεων έγινε με τη βοήθεια των stopwords του πακέτου NLTK.

4.5.2 Αφαίρεση στοιχείων Twitter

Οι δημοσιεύσεις που προέρχονται από το Twitter περιέχουν εκτός του κειμένου και ειδικά στοιχεία του twitter, τα οποία δημιουργούν θόρυβο στα αποτελέσματα του συστήματος. Τα στοιχεία αυτά αφορούν υπερσυνδέσμους (urls), αναφορές χρηστών (user mentions), “hashtags”, ειδικά σύμβολα και αρχεία εικόνων και βίντεο (media). Μέσω των μεταδεδομένων του tweet, όπου εμφανίζονται οι θέσεις τους εντός της συμβολοσειράς, εντοπίστηκαν και αφαιρέθηκαν.

4.5.3 Λήμμα όρου - Μέρους του λόγου

Προκειμένου να υπάρχει ομοιογένεια μεταξύ των όρων, εκτός από τη μετατροπή τους σε πεζά γράμματα χρειάστηκε να αποφευχθούν διπλοεγγραφές όρων. Πιο συγκεκριμένα, αν δεν χειριστεί κατάλληλα, το σύστημα θεωρεί ένα ρήμα διαφορετικό από το ομόριζό του ουσιαστικό. Κατά την επεξεργασία των δημοσιεύσεων, δόθηκε ιδιαίτερη προσοχή ώστε να μην παρατηρηθούν τέτοιου είδους φαινόμενα. Με τη βοήθεια της μεθόδου pos_tag του πακέτου NLTK εντοπίστηκε το μέρος του λόγου για κάθε έναν από τους όρους. Στη συνέχεια με τη βοήθεια της μεθόδου WordNetLemmatizer του ίδιου πακέτου, γνωρίζοντας το μέρος του λόγου, μετατρέψαμε κάθε όρο στο λήμμα του αγγλικού λεξικού που αντιστοιχίζεται.

4.5.4 Διαγραφή σπάνιων όρων

Εκτός της προ-επεξεργασίας του ελεύθερου κειμένου των δημοσιεύσεων υπήρξε επιπλέον βελτιστοποίηση ως προς τους όρους υπό μελέτη. Πιο συγκεκριμένα, διαπιστώθηκε ότι όροι οι οποίοι επιτυγχάνουν πολύ χαμηλή ταχύτητα διάδοσης εντός κάποιου παραθύρου δεν έχουν ιδιαίτερη αξία για την αναγνώριση γεγονότων.

Όταν ένας όρος έχει πολύ χαμηλή ταχύτητα διάδοσης δύο ενδεχόμενα μπορούν να συμβούν. Το πρώτο είναι ο όρος αυτός να είναι ιδιαίτερα σπάνιος και να μην υπάρξει καμία νέα εμφάνισή του σε επόμενα χρονικά παράθυρα. Το άλλο ενδεχόμενο είναι αυτός ο όρος να αποτελέσει γεγονός σε κάποιο από τα παράθυρα που ακολουθούν. Στην περίπτωση αυτή όμως ο όρος δεν εντοπίζεται ως γεγονός λόγω της ταχύτητας που είχε στο παράθυρο αυτό. Χαρακτηρίζεται γεγονός επειδή υπάρχει πληθώρα εμφανίσεών του, η οποία υπερκαλύπτει την μικρή αυτή ταχύτητα σε βαθμό που να την εκμηδενίζει.

Όπως είναι αντιληπτό διατηρώντας όρους με μικρές ταχύτητες διάδοσης το σύστημα καταναλώνει εκτός από μνήμη και υπολογιστική ισχύ. Αυτό συμβαίνει καθώς για κάθε νέα δημοσίευση που επεξεργάζεται, πρέπει να γίνεται έλεγχος για το αν αυτή περιέχει τον συγκεκριμένο όρο. Για όλους τους παραπάνω λόγους η απόφαση που λήφθηκε ήταν η διαγραφή των σπάνιων όρων.

Προκειμένου να γίνει εφικτή η διαγραφή σπάνιων όρων καθορίστηκε ένα κατώφλι. Στο πέρας της επεξεργασίας ενός χρονικού παραθύρου όσοι όροι έχουν ταχύτητα κατώτερη του

κατωφλίου διαγράφονται από το λεξικό όρων. Το κατώφλι αυτό επιλέχθηκε βάσει δύο κριτηρίων. Το πρώτο ήταν κατά πόσο επηρεάζονται τα τελικά αποτελέσματα των πρώτων θέσεων στην ταχύτητα διάδοσης και το δεύτερο η βελτίωση της απόδοσης του συστήματος ως προς τον χρόνο επεξεργασίας. Η τιμή που επιλέχθηκε ήταν τέτοια ώστε να σέβεται και τα δύο κριτήρια. Μεγαλύτερο κατώφλι δεν βελτίωνε αισθητά τον χρόνο και επίσης φαινόταν να επηρεάζει τα αποτελέσματα του συστήματος. Η τιμή της καθορίστηκε στους 0.00001 όρους ανά δευτερόλεπτο.

4.6 Ποιοτική Αξιολόγηση

Προκειμένου να επιβεβαιώσουμε την αποτελεσματικότητα αλλά και την αξιοπιστία του προτεινόμενου συστήματος προχωρήσαμε στη δοκιμή του σε προσομοίωση πραγματικού χρόνου. Κάνοντας χρήση των δεδομένων που είχαν συλλεχθεί από το Twitter API και αφού τα ταξινομήσαμε ως προς τον χρόνο δημοσίευσης πραγματοποιήσαμε μετρήσεις εντός συγκεκριμένης χρονικής περιόδου. Στις παραγράφους που ακολουθούν παρουσιάζονται λεπτομέρειες σχετικά με την προσομοίωση που πραγματοποιήθηκε καθώς και τα αποτελέσματα που παρήχθησαν.

4.6.1 Περίοδος Μελέτης

Η περίοδος μελέτης που επιλέχθηκε ήταν το διάστημα 17 έως 30 Δεκεμβρίου του 2016. Πρόκειται για ένα διάστημα 14 ημερών μέσα στο οποίο αναμέναμε να εντοπίσουμε όσο το δυνατόν περισσότερα γνωστά εκ των προτέρων γεγονότα. Ζητούμενο υπήρξε επίσης ο εντοπισμός τους όσο το δυνατόν πιο κοντά στη χρονική στιγμή που αυτά αναφέρονται για πρώτη φορά σε κάποια δημοσίευση.

4.6.2 Κριτήριο Επιλογής

Η συγκεκριμένη περίοδος επιλέχθηκε ύστερα από μελέτη των γεγονότων που συνέβησαν εντός του έτους 2016. Τη συγκεκριμένη περίοδο υπήρξαν γεγονότα με παγκόσμιο αντίκτυπο, τα οποία απασχόλησαν για μεγάλο διάστημα την ανθρωπότητα. Όπως ήταν αναμενόμενο αυτό αντανακλάστηκε και στις δημοσιεύσεις των χρηστών του Twitter για την περίοδο αυτή. Η βάση δεδομένων που χρησιμοποιήθηκε για τον εντοπισμό των γεγονότων ήταν το Αρχείο του πρακτορείου Reuters (http://www.reuters.com/news/archive/topNews).

Τα γεγονότα που συνέβησαν είναι τα παρακάτω:

1. Ο Ρώσος πρέσβης Αντρέι Καρλόφ δολοφονείται στην Τουρκία στις 19/12/2016
2. Επίθεση του ISIS στην χριστουγεννιάτικη αγορά του Βερολίνου στις 19/12/2016
3. Ο Ντόναλντ Τραμπ κερδίζει τη ψήφο του Κολεγίου των Εκλεκτόρων στις 19/12/2016
4. Η επικεφαλής του IMF Κριστίν Λαγκάρντ κρίνεται ένοχη σε υπόθεση της γαλλικής δικαιοσύνης στις 19/12/2016
5. Μεγάλη έκρηξη στην αγορά πυροτεχνημάτων του Μεξικό στις 21/12/2016
6. Αεροπειρατεία πτήσης με προορισμό τη Μάλτα στις 23/12/2016
7. Θάνατος του τραγουδιστή Τζορτζ Μάικλ στις 25/12/2016

Το σύστημα που αναπτύχθηκε αναγνώρισε επιτυχώς όλα τα σημαντικά γεγονότα της περιόδου και μάλιστα λίγο αφότου αυτά αναφέρθηκαν πρώτη φορά σε κάποια δημοσίευση.

4.6.3 Ανεύρεση δευτερευόντων γεγονότων εντός περιόδου

Εντός της περιόδου που επιλέχθηκε για τη μελέτη συνέβησαν εξίσου σημαντικά γεγονότα με τα προαναφερθέντα έχοντας όμως δευτερεύουσα σημασία ως προς αυτά. Με μεγάλο ενδιαφέρον είδαμε το σύστημα να τα εντοπίζει και να μας ενημερώνει για την ύπαρξή τους. Παρότι δεν αποτέλεσαν κριτήριο για την επιλογή της περιόδου μελέτης ήταν εξίσου σημαντικά. Για τον λόγο αυτό τα αναφέρουμε ενδεικτικά:

1. Μεγάλος σεισμός στην Παπούα στις 17/12/2016
2. Θάνατος της ηθοποιού Ζα Ζα Γκαμπόρ στις 18/12/2016
3. Το Χαλέπι επιστρέφει στην κατοχή των κυβερνητικών δυνάμεων στις 22/12/2016
4. Η ηθοποιός Κάρι Φίσερ υπέστη καρδιακό επεισόδιο στις 24/12/2016
5. Θάνατος του κιθαρίστα των Status Quo, Ρικ Παρφίτ στις 24/12/2016
6. Εξαφάνιση ρωσικού στρατιωτικού αεροσκάφους στη Μαύρη Θάλασσα στις 25/12/2016
7. Μεγάλος σεισμός στη Χιλή στις 25/12/2016
8. Θάνατος της ηθοποιού Κάρι Φίσερ στις 27/12/2016
9. Ιστορική επίσκεψη στο Περλ Χάρμπορ του πρωθυπουργού της Ιαπωνίας στις 27/12/2016
10. Αποσύρεται από τη ποδηλασία ο Σερ Μπράντλεϊ Γουίγκινς στις 28/12/2016
11. Ομιλία Τζον Κέρι για το Ισραήλ στις 28/12/2016
12. Κατάπαυση πυρός στη Συρία στις 29/12/2016
13. Απέλαση 35 Ρώσων διπλωματών από τις ΗΠΑ στις 29/12/2016
14. Απάντηση Πούτιν στην απέλαση των Ρώσων διπλωματών στις 30/12/2016

4.6.4 Πρώτη εμφάνιση και Παράθυρο εντοπισμού γεγονότων

Οι πίνακες που παρουσιάζονται παρακάτω περιέχουν πληροφορίες σχετικά με τον έγκαιρο εντοπισμό των γεγονότων της μελετώμενης χρονικής περιόδου. Για κάθε ένα από τα γεγονότα παρουσιάζεται η ακριβής ώρα που πραγματοποιήθηκε η πρώτη δημοσίευση που το αναφέρει. Ο τίτλος της στήλης αναφέρεται ως “Πρώτη Εμφάνιση”. Η επόμενη στήλη “Παράθυρο Εντοπισμού” περιέχει το χρονικό παράθυρο στο οποίο το σύστημα εντόπισε το γεγονός και πραγματοποίησε σχετική ενημέρωση.

α/α	Γεγονός	Ημερομηνία	Πρώτη Εμφάνιση	Παράθυρο Εντοπισμού
1	Αποτέλεσμα δίκης Κ. Λαγκάρντ	19/12/2016	14:13:45	14:10 - 14:41
2	Δολοφονία Ρώσου Πρέσβη	19/12/2016	16:14:53	16:12 - 16:43
3	Επίθεση στο Βερολίνο	19/12/2016	19:29:42	19:16 - 19:46
4	Ψήφος Κολεγίου Εκλεκτόρων	19/12/2016	22:29:01	22:18 - 22:49
5	Έκρηξη στην αγορά του Μεξικό	20/12/2016	22:19:41	22:40 - 23:11
6	Αεροπειρατεία στη Μάλτα	23/12/2016	10:47:09	10:55 - 11:25
7	Θάνατος Τζορτζ Μάικλ	25/12/2016	22:57:49	23:08 - 23:38

Πίνακας 7: Αποτελέσματα κύριων γεγονότων

α/α	Γεγονός	Ημερομηνία	Πρώτη Εμφάνιση	Παράθυρο Εντοπισμού
1	Μεγάλος σεισμός στην Παπούα	17/12/2016	11:09:09	11:15 - 11:45
2	Θάνατος της ηθοποιού Ζα Ζα Γκαμπόρ	18/12/2016	22:47:45	22:54 - 23:28
3	Το Χαλέπι επιστρέφει στην κατοχή των κυβερνητικών δυνάμεων	22/12/2016	18:27:47	18:33 - 19:04
4	Η ηθοποιός Κάρι Φίσερ υπέστη καρδιακό επεισόδιο	24/12/2016	21:36:56	22:06 - 22:36
5	Θάνατος του κιθαρίστα των Status Quo, Ρικ Παρφίτ	24/12/2016	15:14:50	15:14 - 15:44
6	Εξαφάνιση ρωσικού στρατιωτικού αεροσκάφους στη Μαύρη Θάλασσα	25/12/2016	04:36:43	05:35 - 06:07
7	Μεγάλος σεισμός στη Χιλή	25/12/2016	14:36:38	14:25 - 14:55
8	Θάνατος της ηθοποιού Κάρι Φίσερ	27/12/2016	17:55:18	17:45 - 18:15
9	Ιστορική επίσκεψη στο Περλ Χάρμπορ του πρωθυπουργού της Ιαπωνίας	27/12/2016	21:00:06	21:18 - 21:48
10	Αποσύρεται από τη ποδηλασία ο Σερ Μπράντλεϊ Γουίγκινς	28/12/2016	14:54:13	15:06 - 15:36
11	Ομιλία Τζον Κέρι για το Ισραήλ	28/12/2016	16:21:11	16:06 - 16:36
12	Κατάπαυση πυρός στη Συρία	29/12/2016	11:22:01	11:33 - 12:03
13	Απέλαση 35 Ρώσων διπλωματών από τις ΗΠΑ	29/12/2016	19:03:45	19:07 - 19:39
14	Απάντηση Πούτιν στην απέλαση των Ρώσων διπλωματών	30/12/2016	12:30:58	12:30 - 13:01

Πίνακας 8: Αποτελέσματα δευτερευόντων γεγονότων

4.6.4.1 Σχόλια

Παρατηρούμε πως το σύστημα ήταν ιδιαίτερα αποτελεσματικό ως προς τον εντοπισμό των σημαντικών γεγονότων της περιόδου. Για τα 9 από τα 21 γεγονότα, το σύστημα πετυχαίνει τον εντοπισμό τους εντός του χρονικού παραθύρου που ανήκει η αρχική δημοσίευσή τους. Για τα 11 από τα 21 γεγονότα παρατηρούμε πως τα εντοπίζει στο ακριβώς επόμενο χρονικό παράθυρο, το οποίο όμως ξεκινάει ελάχιστα λεπτά αργότερα από την αρχική δημοσίευση. Τέλος το μοναδικό γεγονός που παρατηρούμε να εντοπίζεται με καθυστέρηση μίας ώρας είναι η εξαφάνιση του ρωσικού αεροσκάφους πάνω από τη Μαύρη Θάλασσα. Ακόμα και σε αυτή την περίπτωση όμως το γεγονός εντοπίζεται με επιτυχία, καθιστώντας το σύστημα αξιόπιστο ως προς τα αποτελέσματά του.

5 Επίλογος

Ολοκληρώνοντας την παρουσίαση της μελέτης αξίζει να αναφερθούμε στη σημασία των αποτελεσμάτων που επιτεύχθηκαν με τη χρήση του προτεινόμενου συστήματος. Η παράγραφος αυτή περιλαμβάνει μία σύνοψη των όσων υλοποιήθηκαν και αναδεικνύει τη χρησιμότητά τους σε ένα περιβάλλον γεμάτο προκλήσεις όπως τα κοινωνικά δίκτυα. Τέλος,

περιγράφονται ιδέες για πιθανές επεκτάσεις στο σύστημα που αποσκοπούν στην όσο το δυνατόν μεγαλύτερη εκμετάλλευση των δυνατοτήτων του.

5.1 Σύνοψη και συμπεράσματα

Μέσω της μελέτης αυτής υλοποιήθηκε ένα ολοκληρωτικό σύστημα επεξεργασίας ειδήσεων που προέρχονται από δημοσιεύσεις του Twitter. Το ένα από τα συστατικά του μέρη αξιολογεί τις νέες δημοσιεύσεις, εκτιμώντας την αξιοπιστία των δημιουργών τους. Αποτελείται από ένα μοντέλο ταξινομητή, το οποίο δοθέντος ενός διανύσματος χαρακτηριστικών, προερχόμενο από κάποια δημοσίευση, αποφαίνεται για την κλάση στην οποία ανήκει. Προϋπόθεση για τη λειτουργία του μοντέλου είναι ένας μηχανισμός ικανός να μετασχηματίζει δημοσιεύσεις σε διανύσματα τιμών, ανάλογα τα εκάστοτε χαρακτηριστικά τους. Οι δημοσιεύσεις μαζί με τα μεταδεδομένα που τις πλαισιώνουν συλλέγονται απευθείας από την πηγή (Twitter) μέσω της ειδικής διαπροσωπείας που το μέσο αυτό προσφέρει. Λόγω του μεγάλου όγκου δεδομένων είναι επιτακτική και η χρήση βάσης δεδομένων. Κριτήριο επιλογής για τη βάση δεδομένων αποτελεί η ικανότητά της να αποθηκεύει και να ανακτά αποδοτικά αρχεία σε μορφή εγγράφου κειμένου JSON.

Το δεύτερο συστατικό μέρος του συστήματος μελετά τη ροή δεδομένων στο Twitter και αποσκοπεί στην τάχιστη εύρεση νέων ειδήσεων. Ανά τακτά χρονικά παράθυρα, το υποσύστημα αυτό καταγράφει τη συχνότητα των διαδιδόμενων όρων. Κατόπιν, υπολογίζει την ταχύτητα διάδοσής τους, ενώ κάνει χρήση τακτικών που το οδηγούν σε αποδοτικότερα αποτελέσματα. Διατηρώντας μηχανισμό εντοπισμού ειδήσεων, το σύστημα ειδοποιεί τον χρήστη όταν αυτά εμφανίζονται, ανάλογα τον βαθμό ευαισθησίας που έχει τεθεί από αυτόν. Προκειμένου να καταλήξει ταχύτερα στην εύρεση νέων, το σύστημα ενώ εκκινεί από μοναδικούς όρους, καταλήγει να συγκεντρώνει θέματα όρων. Όσο το σύστημα βρίσκεται σε λειτουργία, ο χρήστης όχι μόνο ενημερώνεται για νέες ειδήσεις, αλλά λαμβάνει ένα σύνολο όρων που χαρακτηρίζουν την είδηση αυτή.

Τα συμπεράσματα που προκύπτουν από την εργασία αυτή είναι ποικίλα. Αρχικά, επιβεβαιώνεται η ικανότητα των μοντέλων ταξινομητών να ταξινομούν με επιτυχία δημοσιεύσεις, εφόσον αυτές μετασχηματιστούν κατάλληλα. Επιπλέον, η θεώρηση των διαδιδόμενων όρων ως φυσικά σωματίδια με όγκο και ταχύτητα διάδοσης, μπορεί να εντοπίσει με επιτυχία ειδήσεις ανάμεσα στη ροή πληροφοριών ενός κοινωνικού δικτύου. Εφόσον εφαρμοστούν κατάλληλες ευριστικές μέθοδοι στους όρους, κάνοντας χρήση συνόλων, αποδεικνύεται πως τα αποτελέσματα βελτιώνονται αισθητά. Τέλος, η ικανοποιητική απόδοση του συστήματος και στις δύο επιμέρους διαδικασίες, καταδεικνύει την επιτακτική ανάγκη χρήσης ευφυών συστημάτων για τη μελέτη των κοινωνικών δικτύων. Γνωρίζοντας τα αποτελέσματα που μόνο ένα ευφυές σύστημα μπορεί να παράξει θέτουμε ισχυρές βάσεις για την προστασία του θεσμού ανταλλαγής νέων, προστατεύοντάς τον από τους κινδύνους που ελλοχεύουν.

5.2 Μελλοντικές επεκτάσεις και βελτιώσεις

Η ενότητα αυτή περιλαμβάνει ιδέες και προτάσεις σχετικά με τη βελτίωση του προτεινόμενου συστήματος. Μέρος των προτάσεων έχει ήδη υλοποιηθεί ως ένα βαθμό και είναι προσβάσιμες μέσω του ηλεκτρονικού αποθετηρίου κώδικα github, όπου φιλοξενείται ολόκληρος ο κώδικας του συστήματος.

5.2.1 Συνδυαστικό Σύστημα

Έχοντας δημιουργήσει ένα σύστημα ταχείας εύρεσης γεγονότων και ένα αναγνώρισης αξιοπιστίας των δημοσιεύσεων το επόμενο βήμα είναι η συνεργασία των δύο. Στα πλαίσια αυτά, προτείνεται η δημιουργία ενός ενιαίου συστήματος, αποτελούμενο από τα δύο αυτά συστατικά μέρη.

Όσο αφορά το πρώτο μέρος, αυτό της εύρεσης γεγονότων, το σύστημα πρέπει ανά πάσα στιγμή να μπορεί να εντοπίσει ένα πραγματικό γεγονός, τη στιγμή που αυτό συμβαίνει. Το παραπάνω προϋποθέτει πως το σύστημα πρέπει, εκτός του να είναι διαρκώς ενεργό, να διατηρεί ένα μεγάλο λεξικό όρων, το οποίο ενημερώνεται σε παράθυρα της μισής ώρας. Όπως μπορεί να γίνει αντιληπτό, για της ανάγκες της μελέτης, τα δεδομένα είχαν ήδη συλλεχθεί και επεξεργαστεί προτού εισαχθούν στο σύστημα. Φυσικά, προηγήθηκε χρονική ταξινόμηση ώστε να αποτελέσουν είσοδο του συστήματος. Σε κάθε περίπτωση όμως ήταν δεδομένα εκ των προτέρων γνωστά και είχαν διαχειρίσιμο μέγεθος.

Για τη μετάβαση σε σύστημα ζωντανού χρόνου απαιτούνται δύο προσθήκες. Η πρώτη έχει να κάνει με την επεξεργασία των δεδομένων και μετατροπή τους σε κατάλληλη μορφή, προκειμένου να είναι άμεσα διαθέσιμα στο σύστημα αναγνώρισης γεγονότων. Έχοντας υλοποιήσει ήδη τον μηχανισμό επεξεργασίας, αρκεί αυτός να τοποθετηθεί σε έναν πράκτορα (agent), διαρκώς σε δράση. Ο πράκτορας θα έχει ως σκοπό την άντληση δεδομένων από τη MongoDB, την επεξεργασία αυτών και τέλος την τροφοδότηση του συστήματος αναγνώρισης με τα επεξεργασμένα δεδομένα.

Η δεύτερη απαραίτητη προσθήκη είναι η διαχείριση του λεξικού όρων από το σύστημα αναγνώρισης. Πιο συγκεκριμένα, όταν το σύστημα εκτεθεί σε τεράστιο όγκο πληροφορίας, όντας διαρκώς ενεργό, τίθεται σημαντικό ζήτημα με τη μνήμη που καταναλώνει. Επιπλέον, σε περίπτωση σφάλματος, όταν το σύστημα ανανήψει, οτιδήποτε βρισκόταν στη μνήμη έχει πλέον χαθεί, με αποτέλεσμα να πρέπει να δημιουργήσει νέο λεξικό. Για τους λόγους αυτούς, κρίνεται απαραίτητος ένας μηχανισμός αποθήκευσης του λεξικού όρων που παράγονται από το σύστημα αναγνώρισης. Εξαιτίας της ανάγκης άμεσης ανάκτησης των όρων λόγω των απαιτητικών πράξεων σε αυτούς, πρέπει να υιοθετηθεί μία cached προσέγγιση αποθήκευσης. Ιδανική τεχνολογία αποτελεί η Redis (https://redis.io/), ένα σύστημα αποθήκευσης λεξικών όρων με εξαιρετική ταχύτητα ανάγνωσης των δεδομένων. Η ύπαρξη βιβλιοθήκης σε γλώσσα Python για συνεργασία με τη Redis, την κάνει ακόμα πιο ελκυστική, καθώς προσφέρει ομοιογένεια στο υπάρχον σύστημα.

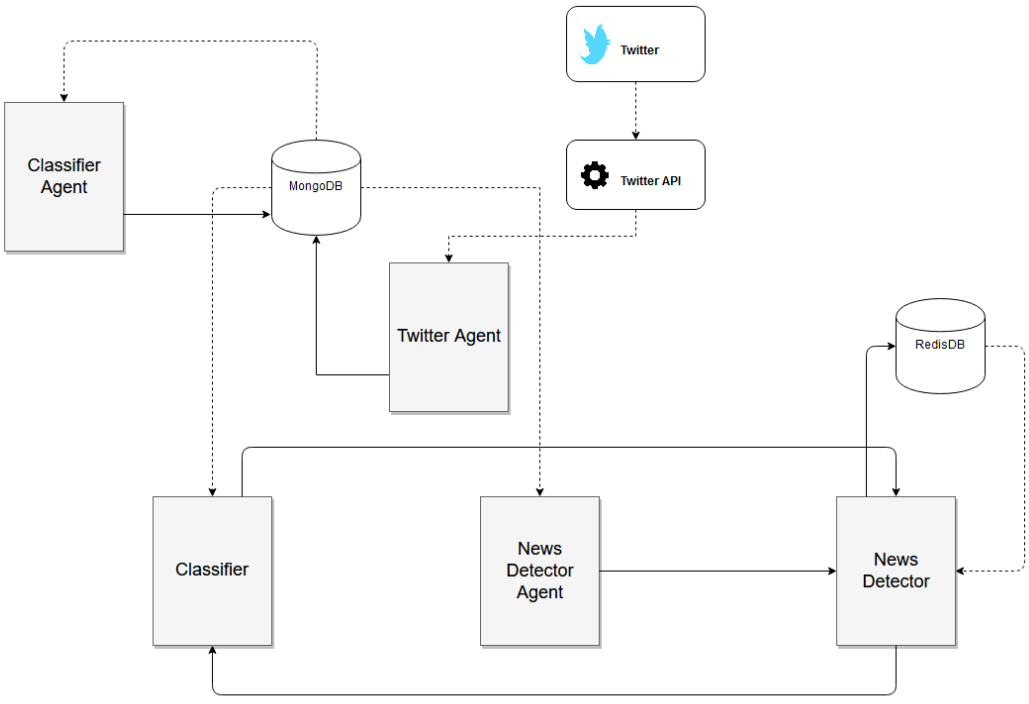
Το σύστημα αξιολόγησης αξιοπιστίας των δημοσιευμάτων χρειάζεται εξίσου κάποιες προσθήκες ώστε να μπορεί να συνεργαστεί με το σύστημα αναγνώρισης. Πιο συγκεκριμένα, το σύστημα πρέπει να είναι άμεσα διαθέσιμο ώστε οποιαδήποτε στιγμή του ζητηθεί, να μπορεί να αποφανθεί για την αξιοπιστία μίας ή περισσότερων δημοσιεύσεων. Όπως και στην περίπτωση του συστήματος αναγνώρισης, έτσι και με αυτό τα δεδομένα μας είχαν ήδη συλλεχθεί και επεξεργαστεί προτού αποτελέσουν είσοδο για το σύστημα. Συνεπώς και για αυτό το σύστημα χρειάζονται δύο επιπλέον βελτιώσεις.

Η πρώτη αφορά στην παραγωγή διανυσμάτων χαρακτηριστικών. Εξαιτίας της πολυπλοκότητας υπολογισμού ενός διανύσματος χαρακτηριστικών, αυτά είχαν ήδη παραχθεί πριν γίνει η εκμάθηση του μοντέλου ταξινόμησης. Σε ένα σύστημα πραγματικού

χρόνου όμως αυτά δεν μπορούν να υπολογιστούν εκ των προτέρων. Όπως και στην περίπτωση του συστήματος αναγνώρισης έτσι και σε αυτό, πρέπει να υλοποιηθεί ένας πράκτορας ικανός να παράγει διανύσματα χαρακτηριστικών και στη συνέχεια να τροφοδοτεί με αυτά το σύστημα ταξινόμησης. Ο μηχανισμός υπολογισμού ενός διανύσματος χαρακτηριστικών είναι ήδη υλοποιημένος. Το μόνο που χρειάζεται ο πράκτορας είναι να γνωρίζει ποια είναι η βάση δεδομένων που περιέχει τις ίδιες τις δημοσιεύσεις καθώς και που θα αποστείλει τα διανύσματα.

Η δεύτερη προσθήκη αφορά την εξασφάλιση της διαρκούς διαθεσιμότητας. Το μοντέλο ταξινόμησης πρέπει εκτός του να είναι σε διαρκή λειτουργία, να λαμβάνει αιτήματα για χαρακτηρισμό δημοσιεύσεων και να τα διεκπεραιώνει. Φυσικά, αποστολέας των αιτημάτων μπορεί να είναι ο οποιοσδήποτε καθώς μπορεί να υπάρχουν άνω του ενός ενδιαφερόμενοι για την αξιοπιστία κάποιας δημοσίευσης. Προτείνεται η τοποθέτηση του ταξινομητή σε κάποιον εξυπηρετητή και η συσχέτισή του με μία συγκεκριμένη πόρτα αυτού. Ο ταξινομητής στην περίπτωση αυτή είναι διαρκώς διαθέσιμος ενώ, μόλις αποφανθεί για την κατηγορίας της δημοσίευσης, ενημερώνει τον αποστολέα για αυτή. Το παραπάνω προϋποθέτει πως ο ταξινομητής έχει ολοκληρώσει τη διαδικασία εκμάθησης σε πρότερο στάδιο. Συνεπώς, η διαδικασία εκμάθησης πρέπει να ολοκληρωθεί όταν ο ταξινομητής είναι εκτός δικτύου. Μόλις αυτή ολοκληρωθεί, ο ταξινομητής μπορεί να περιμένει αιτήματα σε συγκεκριμένη διεύθυνση και αναλόγως να τα εξυπηρετεί.

Ακολουθεί ένα συνοπτικό σχήμα του συνολικού συστήματος, στο οποίο φαίνονται τα συστατικά του μέρη καθώς και η επικοινωνία μεταξύ τους.



Εικόνα 9: Συνολικό σύστημα

Έχοντας υλοποιήσει τους παραπάνω μηχανισμούς τα δύο συστήματα μπορούν να αλληλοεπιδρούν και να συνεργάζονται με επιτυχία. Προκειμένου να ανταλλάσσουν μηνύματα κοινώς αποδεκτά, χρειάζεται και μία σύμβαση ως προς τη μορφή τους. Προτείνεται η χρήση λεξικών JSON προκαθορισμένης μορφής, μέσω των οποίων τα συστήματα στέλνουν μηνύματα και λαμβάνουν απαντήσεις.

Έχοντας συμφωνήσει και στον τρόπο επικοινωνίας των δύο συστημάτων ένα σενάριο συνεργασίας είναι το παρακάτω:

1. Το ενιαίο σύστημα λαμβάνει tweets από το Tweeter API και τα αποθηκεύει σε μορφή JSON στη MongoDB.
2. Ο πράκτορας του συστήματος ταξινόμησης δημιουργεί τα αντίστοιχα διανύσματα χαρακτηριστικών, τα οποία αποθηκεύει εξίσου στη MongoDB.
3. Ο πράκτορας του συστήματος αναγνώρισης εξάγει τους όρους των δημοσιεύσεων, βάσει της σχετικής διαδικασίας και τους αποστέλλει στο σύστημα αναγνώρισης.
4. Το σύστημα αναγνώρισης ανά μισή ώρα υπολογίζει τον όγκο και την ταχύτητα διάδοσης κάθε όρου και αποθηκεύει τα αποτελέσματα στη Redis
5. Το σύστημα αναγνώρισης εντοπίζει μία είδηση και αποστέλλει τα αναγνωριστικά όλων των σχετικών με αυτή δημοσιεύσεων στο σύστημα ταξινόμησης μέσω JSON.
6. Το σύστημα ταξινόμησης λαμβάνει το αίτημα για ταξινόμηση και αναζητεί στη MongoDB τα διανύσματα χαρακτηριστικών που αντιστοιχίζονται στις υπό μελέτη δημοσιεύσεις.
7. Το σύστημα ταξινόμησης χαρακτηρίζει κάθε μία από τις δημοσιεύσεις ως προς την αξιοπιστία και αποστέλλει τα αποτελέσματα στο σύστημα αναγνώρισης.
8. Το σύστημα αναγνώρισης πραγματοποιεί συνολική εκτίμηση αξιοπιστίας αναλόγως της πλειοψηφίας από τα αποτελέσματα που έλαβε από το σύστημα ταξινόμησης.
9. Το σύστημα αναγνώρισης ενημερώνει τον χρήστη για τη νέα είδηση καθώς και για τη συνολική εκτίμηση για την αξιοπιστία αυτής.

5.2.2 Λογαριασμός Ρομπότ στο Twitter

Η φιλοσοφία του Twitter ως μέσο κοινωνικής δικτύωσης είναι η ενημέρωση των χρηστών για δημοσιεύσεις χρηστών που ακολουθούν και αντιστρόφως. Ένα πολύ ενδιαφέρον χαρακτηριστικό ως προς τα προτεινόμενα συστήματα είναι η δυνατότητα αλληλεπίδρασής τους με χρήστες του Twitter. Μία πρόταση βελτίωσης αποτελεί η δημιουργία λογαριασμού ρομπότ στο Twitter, ο οποίος θα δημοσιεύει τα αποτελέσματα των δύο συστημάτων ανεξάρτητα μεταξύ τους.

Προκειμένου να γίνει εφικτό το παραπάνω απαιτείται ένας μηχανισμός αυτόματης δημιουργίας δημοσιεύσεων στο Twitter, ο οποίος θα είναι μη επιβλεπόμενος. Για την υλοποίησή του αρκεί να γίνει χρήση του Twitter API, το οποίο επιτρέπει στους χρήστες να μπορούν να δημοσιεύουν αναρτήσεις μέσω ειδικών αιτήσεων προς αυτό.

Στην περίπτωση του συστήματος ταξινόμησης προτείνεται η υλοποίηση ενός μηχανισμού ερωταπαντήσεων. Πιο συγκεκριμένα, ένας χρήστης του Twitter δημιουργεί μία δημοσίευση ειδικής μορφής, η οποία περιλαμβάνει αναφορά στον λογαριασμό του ρομπότ του συστήματος μαζί με ένα μοναδικό αναγνωριστικό κάποιας δημοσίευσης. Μόλις γίνει αυτό, το σύστημα ταξινόμησης ειδοποιείται, λαμβάνοντας το αναγνωριστικό της δημοσίευσης. Στη

συνέχεια, ανακτά τη δημοσίευση μέσω του Twitter API και υπολογίζει το διάνυσμα χαρακτηριστικών αυτής. Τέλος, αποφαίνεται για την αξιοπιστία της και δημιουργεί μία νέα δημοσίευση. Με αυτή ενημερώνει τον χρήστη της αρχικής δημοσίευσης σχετικά με την αξιοπιστία της δημοσίευσης που τον ενδιαφέρει.

Στην περίπτωση του συστήματος αναγνώρισης η διαδικασία είναι πιο απλή. Έχοντας το σύστημα αναγνώρισης να επεξεργάζεται διαρκώς νέα δεδομένα σε χρονικά παράθυρα μισής ώρας, αυτό εντοπίζει νέες ειδήσεις. Μόλις μία νέα είδηση εντοπιστεί, τότε το σύστημα αναγνώρισης αναλαμβάνει να την κοινοποιήσει στους χρήστες του Twitter. Πιο συγκεκριμένα, μέσω του Twitter API δημοσιεύει τα αποτελέσματα της έρευνάς του, τα οποία είναι άμεσα διαθέσιμα στους χρήστες του Twitter που είναι ακόλουθοι του λογαριασμού ρομπότ του συστήματος.

6 Βιβλιογραφία

Stephens, History of News (1988), pp. 14, 305

Alpaydin, E. (2010) *Introduction to Machine Learning*, *Machine Learning*.
Benhardus, J. and Kalita, J. (2013) 'Streaming trend detection in twitter', *International Journal of Web Based Communities*, 9(1), pp. 122–139.

Bird, S. and Loper, E. (2004) 'NLTK: The Natural Language Toolkit', in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 1–4.

Chih-Wei Hsu, Chih-Chung Chang, and C.-J. L. (2008) 'A Practical Guide to Support Vector Classification', *BJU international*, 101(1), pp. 1396–400.

Fellbaum, C. (1998) *WordNet: An Electronic Lexical Database*, *MIT Press, Cambridge, London, England*.

Gupta, A., Kumaraguru, P. and Castillo, C. (2014) 'TweetCred : Real-Time Credibility Assessment', *Springer International Publishing Switzerland*, pp. 228–243.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) 'The Elements of Statistical Learning', *Elements*, 1, pp. 337–387.

Hunter, J. D. (2007) 'Matplotlib: A 2D graphics environment', *Computing in Science and Engineering*, 9(3), pp. 99–104.

Jin, F., Dougherty, E., Saraf, P., Cao, Y. and Ramakrishnan, N. (2013) 'Epidemiological Modeling of News and Rumors on Twitter', *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, p. 8:1-8:9.

Liu, X. (2015) 'Real-time Rumor Debunking on Twitter', *Cikm*, (March 2016), pp. 1867–1870.

Mendoza, M., Poblete, B. and Castillo, C. (2010) 'Twitter Under Crisis: Can we trust what we RT?', *Workshop on Social Media Analytics*, p. 9.

MongoDB Inc. (2016) *The MongoDB 3.2 Manual*, *MongoDB Manual 3.2*. Available at: https://docs.mongodb.com/manual/.

MySQL AB (2002) 'MySQL Reference Manual', *Notes*, p. 239.

Osborne, M., Petrovic, S. and McCreadie, R. (2012) 'Bieber no more: First Story Detection using Twitter and Wikipedia', *Redirect.Subscribe.Ru*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2012) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Petrović, S., Osborne, M. and Lavrenko, V. (2010) 'Streaming first story detection with application to twitter', *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, (June), pp. 181–189.

Petrović, S., Osborne, M., Lavrenko, V. and Petrovic, S. (2012) 'Using paraphrases for improving first story detection in news and Twitter', *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, pp. 338–346.

Poblete, B., Castillo, C. and Mendoza, M. (2011) 'Information Credibility on Twitter', pp. 45–59.

Safavian, S. R. and Landgrebe, D. (1991) 'A Survey of Decision Tree Classifier Methodology', *IEEE Transactions on Systems, Man and Cybernetics*, 21(3), pp. 660–674.

Tolos, L., Tagarev, A. and Georgiev, G. (2016) 'An Analysis of Event-Agnostic Features for Rumour Classification in Twitter', pp. 151–158.

Twitter Developers (2013) *Using the Twitter Search API*, *Twitter, Inc*. Available at: https://dev.twitter.com/docs/using-search.

Xie, W., Zhu, F., Jiang, J., Lim, E. P. and Wang, K. (2016) 'TopicSketch: Real-time bursty topic detection from twitter', *IEEE Transactions on Knowledge and Data Engineering*, 28(8), pp. 2216–2229.

Yang, F., Liu, Y., Yu, X. and Yang, M. (2012) 'Automatic detection of rumor on Sina Weibo', *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, 2, p. 13:1--13:7.