ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Υλοποίηση διαδικτυακής εφαρμογής ανίχνευσης γεγονότων από δεδομένα προερχόμενα από μέσα κοινωνικής δικτύωσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

των

ΒΑΣΙΛΕΙΟΥ ΔΕΠΑΣΤΑ
ΑΝΔΡΕΑ-ΔΩΡΟΘΕΟΥ ΣΥΡΜΑΚΕΣΗ

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2017

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ
ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Υλοποίηση διαδικτυακής εφαρμογής ανίχνευσης γεγονότων από δεδομένα προερχόμενα από μέσα κοινωνικής δικτύωσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

των

ΒΑΣΙΛΕΙΟΥ ΔΕΠΑΣΤΑ
ΑΝΔΡΕΑ-ΔΩΡΟΘΕΟΥ ΣΥΡΜΑΚΕΣΗ

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή τη 13η Οκτωβρίου 2017.

.....	.....	.....
Δημήτριος Ασκούνης	Ιωάννης Ψαρράς	Χρυσόστομος Δούκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επ. Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2017

.....
ΒΑΣΙΛΕΙΟΣ ΔΕΠΑΣΤΑΣ

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

.....
ΑΝΔΡΕΑΣ-ΔΩΡΟΘΕΟΣ ΣΥΡΜΑΚΕΣΗΣ

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

Περίληψη

Η συνεχής αύξηση του αριθμού των χρηστών του μέσου κοινωνικής δικτύωσης Twitter στην Ελλάδα (σύμφωνα με το monitor.netsteps.gr - έναν ιστότοπο που αναλύει σε καθημερινή βάση τα μέσα κοινωνικής δικτύωσης Ελλήνων - ο τρέχων αριθμός ανέρχεται σε 697,678 τον Οκτώβριο του 2017 ενώ το 2013 ο αριθμός αυτός ανερχόταν σε 263,100) και η συνακόλουθη αύξηση του όγκου της πληροφορίας που διαμοιράζεται στο μέσο καθιστά το Twitter μια εξαιρετικά χρήσιμη πηγή δεδομένων σε πραγματικό χρόνο.

Στα πλαίσια της παρούσας διπλωματικής εργασίας πραγματοποιήθηκε α) μελέτη και σύγκριση μεθόδων ανίχνευσης προκαθορισμένων έκτακτων γεγονότων όπως σεισμών και πλημμύρων σε μικρό χρόνο από την πρώτη παρατήρηση τους από αναφορές στο μέσο κοινωνικής δικτύωσης Twitter και β) ανάπτυξη κατάλληλης μεθοδολογίας με επιλογή των πιο αποτελεσματικών μεθόδων για την υλοποίησή τους στα πλαίσια μιας διαδικτυακής πλατφόρμας και εφαρμογής για κινητά που ενημερώνει τους χρήστες για τέτοια γεγονότα.

Η υλοποίηση αφορά γεγονότα που ανιχνεύονται σε πραγματικό χρόνο μέσω της συγκέντρωσης και επεξεργασίας tweets με χρήση του Streaming API του Twitter με προκαθορισμένες λέξεις-κλειδιά στα ελληνικά. Ως εκ τούτου, γεγονότα τα οποία συζητούνται στο μέσο στην ελληνική γλώσσα είναι υποψήφια προς ανίχνευση από το σύστημα.

Η μεθοδολογία που αναπτύχθηκε μπορεί να χρησιμοποιηθεί για την ανίχνευση προκαθορισμένων γεγονότων με την τροποποίηση των λέξεων-κλειδιών για την λήψη των tweets μέσω του Twitter API και τη δημιουργία κατάλληλου dataset για την εκπαίδευση του συστήματος. Η γενικότητα της μεθοδολογίας έγκειται στο γεγονός ότι έχει υλοποιηθεί ένα γενικό σύστημα άντλησης και προεπεξεργασίας των tweets σε πραγματικό χρόνο όπως και η πλατφόρμα και εφαρμογή για κινητά που αμφότερες μπορούν να χρησιμοποιηθούν για ενημέρωση των χρηστών σχετικά με γεγονότα που τους αφορούν με κατάλληλη τροποποίηση.

Λέξεις Κλειδιά: <<Twitter, ανίχνευση έκτακτων γεγονότων, πραγματικό χρόνο, ελληνικά, Streaming API>>

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

The continuous increase of the Twitter social media users in Greece (according to monitor.netsteps.gr – a website analyzing Greek social media in a daily basis – the number of users is up to 697,678 as of October 2017, whereas in 2013 it was only 263,100) and the increase on data volumes exchanged through the platform makes Twitter an ideal real time data source.

In the context of this thesis, we a) studied and compared event detection methods for predefined unplanned events such as earthquakes or floodings in short time after their first report on Twitter and b) we developed an appropriate methodology to create this system along with an online platform and mobile app which informs its users regarding such events.

The system created detects events in real time after concentrating and processing tweets with the usage of Twitter's Streaming API and predefined keywords in Greek language. Therefore, events that are being discussed in Twitter in Greek are subject to detection by the system.
The methodology developed can also be used for detection of other predefined events with appropriate adjustments of the used keywords for tweets retrieval through Twitter API and the creation of an appropriate training dataset. The generality of this methods lies in the fact that we have developed a generic system for retrieving tweets and preprocessing in real time as well as a web platform and a mobile app in order to keep users updated regarding events of their interest with appropriate adjustments.

Keywords: <<Twitter, unscheduled events detection, real time, Greek, Streaming API>>

Η σελίδα αυτή είναι σκόπιμα λευκή.

Ευχαριστίες

Θα θέλαμε να ευχαριστήσουμε τον κ. Δημήτριο Ασκούνη για την εμπιστοσύνη και την ευκαιρία που μας έδωσε να εκπονήσουμε την παρούσα διπλωματική, όπως επίσης και τους Ευμορφία Μπιλίρη, Αριάδνη Μιχαλίτση – Ψαρρού και Δημήτρη Πανόπουλο για την καθοδήγησή τους όλο το χρονικό διάστημα της υλοποίησης και συγγραφής.

Ιδιαίτερα θα θέλαμε επίσης να ευχαριστήσουμε αμφότεροι τις οικογένειές μας για την πίστη και στήριξή τους κατά τη διάρκεια της ως τώρα ακαδημαϊκής μας πορείας όπως και στην περάτωση της παρούσας διπλωματικής εργασίας.

Τέλος, ευχαριστούμε όλους τους φίλους με τους οποίους μοιραστήκαμε χαρές και λύπες κατά τη διάρκεια των ακαδημαϊκών μας χρόνων.

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

1

Εισαγωγή

1.1 Η σημασία της ανίχνευσης γεγονότων στα μέσα κοινωνικής δικτύωσης

Τα μέσα κοινωνικής δικτύωσης αποτελούν σήμερα μια σημαντική πηγή πληροφοριών καθώς καθημερινά διακινείται τεράστιος όγκος πληροφορίας, σημαντικό μέρος της οποίας είναι άμεσα προσβάσιμο μέσω κατάλληλων APIs. Σύμφωνα με το [1] το 2012 στο μέσο Twitter μόνο δημιουργούνταν καθημερινά 170 εκατομμύρια tweets και διανέμονταν από εκατομμύρια ενεργούς χρήστες. Τον Αύγουστο του 2013, σύμφωνα με το επίσημο blog του Twitter περισσότερα από 500 εκατομμύρια tweets στέλνονταν καθημερινά κατά μέσο όρο [2].

Η ανάλυση δεδομένων από τα μέσα κοινωνικής δικτύωσης βρίσκει πολλές εφαρμογές με χαρακτηριστικά παραδείγματα τη χρήση από τα τμήματα marketing εταιρειών που αναλύουν την άποψη των χρηστών για τα προϊόντα τους μέσω της τεχνικής που ονομάζεται sentiment analysis και την παρακολούθηση και πρόβλεψη ασθενειών καθώς και πολλές άλλες εφαρμογές [3].

Τα δεδομένα που ανταλλάσσονται μεταξύ χρηστών έχουν ιδιαίτερη αξία λόγω της άμεσης σχέσης τους με γεγονότα (events) που συμβαίνουν στην πραγματικότητα. Έτσι τα μέσα κοινωνική δικτύωσης όπως το Facebook ή το Twitter καθίστανται πολύ σημαντικά εργαλεία στην παρατήρηση ενός πραγματικού συμβάντος καθώς και της χρονικής και χωρικής εξέλιξής του. Ωστόσο, παρά το ενδιαφέρον που έχει παρατηρηθεί στο χώρο της ανίχνευσης γεγονότων

από τα μέσα κοινωνικής δικτύωσης, δεν υπάρχει ξεκάθαρος κοινός ορισμός ενός γεγονότος. Στο [4] ως γεγονός ορίζεται ένα φαινόμενο το οποίο παρακινεί τους χρήστες ενός μέσου κοινωνικής δικτύωσης να κοινοποιήσουν μηνύματα για ένα συγκεκριμένο χρονικό διάστημα. Στο [5] νοείται ως γεγονός κάτι το οποίο συμβαίνει σε συγκεκριμένο χώρο και χρόνο με συνέπειες και πυροδοτεί ένα θέμα (topic). Στο [6] ως γεγονός θεωρείται κάτι σημαντικό που συμβαίνει σε συγκεκριμένο χρόνο και χώρο όπου σημαντικό νοείται οτιδήποτε μπορεί να συζητηθεί στα μέσα ενημέρωσης. Με τον προηγούμενο ορισμό περί σημαντικότητος, οι συγγραφείς προσπαθούν να φιλτράρουν καθημερινά προσωπικά και ασήμαντα γεγονότα που συζητούνται στα μέσα κοινωνικής δικτύωσης. Οι συγγραφείς στο [7] ορίζουν ένα γεγονός στα πλαίσια του Twitter ως ένα συμβάν του πραγματικού κόσμου e που σχετίζεται με μια χρονική περίοδο T_e και μια χρονοσειρά από μηνύματα tweets M_e σημαντικού όγκου που αφορούν το συμβάν και δημοσιεύονται εντός της περιόδου T_e.

Λαμβάνοντας υπόψην τα παραπάνω, στα πλαίσια των μέσων κοινωνικής δικτύωσης, θα θεωρηθεί ως γεγονός ένα συμβάν το οποίο είναι άμεσα συνδεδεμένο με την πραγματικότητα και επηρεάζει τη ζωή των ανθρώπων, έχει συγκεκριμένη χρονική διάρκεια, εντοπίζεται σε δεδομένο γεωγραφικό χώρο και συζητείται στα μέσα αυτά.

Σύμφωνα με το [3] τα γεγονότα που έχουν μελετηθεί στη βιβλιογραφία μπορούν να ταξινομηθούν στους εξής τύπους με βάση το χώρο και το χρόνο στον οποίο συμβαίνουν:

- **Προγραμματισμένα (Planned)** που αφορούν γεγονότα με προκαθορισμένο χρόνο και τόπο διεξαγωγής όπως για παράδειγμα μια θεατρική παράσταση
- **Μη προγραμματισμένα (Unplanned)**, δηλαδή έκτακτα γεγονότα που μπορούν να συμβούν ξαφνικά όπως ένας σεισμός ή μια πλημμύρα
- **Έκτακτες ειδήσεις (Breaking news)** που αφορούν έκτακτα γεγονότα τα οποία συζητούνται στα συμβατικά μέσα ενημέρωσης, όπως για παράδειγμα τα αποτελέσματα κάποιων εκλογών
- **Τοπικά (Local)** τα οποία λαμβάνουν χώρα σε συγκεκριμένο γεωγραφικό χώρο και ο χώρος αυτός είναι ο μόνος που επηρεάζεται όπως λόγου χάρη ένα αυτοκινητιστικό ατύχημα
- **Σχετιζόμενα με κάποια οντότητα (entity related)** που αφορά γεγονότα για μια συγκεκριμένη οντότητα όπως για παράδειγμα ένα νέο τραγούδι κάποιου γνωστού τραγουδιστή

Ιδιαίτερο ενδιαφέρον παρουσιάζουν τα μη προγραμματισμένα γεγονότα, στα οποία η απουσία γνώσης για την πραγματοποίησή τους απαιτεί τη συνεχή παρακολούθηση της ροής μηνυμάτων

ενός μέσου και την εξαγωγή συμπερασμάτων ως προς την ύπαρξη και εξέλιξη κάποιου γεγονότος από την επεξεργασία της ροής αυτής.

1.1.1 Μη προγραμματισμένα γεγονότα

Μη προγραμματισμένα γεγονότα είναι η κατηγορία έκτακτων γεγονότων που συμβαίνουν ξαφνικά χωρίς να υπάρχει γνώση εκ των προτέρων σχετικά με τον τόπο και χρόνο στον οποίο λαμβάνουν χώρα. Παραδείγματα τέτοιων γεγονότων είναι ένας σεισμός, μια πλημμύρα, μια πυρκαγιά, ένας πυροβολισμός ή ένα αυτοκινητιστικό ατύχημα. Ειδικότερα, τα μη προγραμματισμένα γεγονότα όπως σεισμοί και άλλα έκτακτα φυσικά φαινόμενα, παρουσιάζουν σημαντικό ενδιαφέρον αν ληφθεί υπόψην η κρίσιμη φύση τους όπως και η δυσκολία ανίχνευσής τους σε ένα μέσο κοινωνικής δικτύωσης σε εύλογο χρόνο από την εμφάνισή τους. Για παράδειγμα, ένας χρήστης του Twitter μπορεί να αναφέρει μια σεισμική δόνηση που συμβαίνει σε πραγματικό χρόνο μέσω ενός πολύ σύντομου tweet. Λόγω του μικρού μήκους του tweet, που από το ίδιο το μέσο περιορίζεται στους 140 χαρακτήρες, και της δυνατότητας retweet, δηλαδή της κοινοποίησης του tweet ενός άλλου χρήστη, καθίσταται πολύ εύκολη η διάδοση της πληροφορίας στο μέσο σε σχέση με άλλα μέσα διευκολύνοντας αυτό που στον προφορικό λόγο ονομάζεται word of mouth διάδοση ειδήσεων. Ακόμη, το Twitter επιστρέφει μια συνεχόμενη ροή tweets καθιστώντας έτσι εφικτή την παρακολούθηση της εξέλιξης ενός θέματος που συζητείται στο μέσο [8]. Μάλιστα, κατά τη διάρκεια έκτακτων γεγονότων, όπως στην περίπτωση ακραίων καιρικών φαινομένων που έχουν οδηγήσει σε κατάσταση έκτακτης ανάγκης, η δραστηριότητα στο Twitter είναι υψηλή και οι χρήστες είναι πιο πιθανό να μοιραστούν πληροφορίες για το γεγονός σύμφωνα με τους συγγραφείς του [9]. Επίσης, οι αναφορές σε γεγονός που στην πραγματικότητα συμβαίνει σε μια περιορισμένη περιοχή μπορούν να προέρχονται από οποιοδήποτε μέρος εντός της χώρας στην οποία συμβαίνει το γεγονός και όχι απαραίτητα μόνο από τη συγκεκριμένη περιοχή η οποία επηρεάζεται από αυτό.

1.1.2 Ανίχνευση γεγονότων

Από τα δεδομένα που λαμβάνονται από τα μέσα κοινωνικής δικτύωσης, προκύπτει το ερώτημα του πώς επιτυγχάνεται **αυτοματοποιημένη ανίχνευση γεγονότων (event detection)** και παρακολούθηση της εξέλιξής τους μέσα από αυτά. Έχοντας διαθέσιμη μια ροή μηνυμάτων χρηστών σε ένα μέσο, η ανίχνευση ενός γεγονότος έγκειται στην εύρεση ικανού όγκου μηνυμάτων χρηστών που αφορούν το γεγονός και μπορούν να το περιγράψουν εντός

συγκεκριμένου χρονικού διαστήματος και με καθορισμένη τοποθεσία που σχετίζεται με το γεγονός.

Με βάση το χρόνο στον οποίο πραγματοποιείται η ανίχνευση σε σχέση με το χρόνο που συμβαίνει ένα γεγονός, μπορεί να είναι είτε **«απευθείας ανίχνευση γεγονότος» (online detection)** είτε **«αναδρομική ανίχνευση γεγονότος» (retrospective detection)** [10]. Η απευθείας ανίχνευση γεγονότος αφορά άγνωστα ως την στιγμή της ανίχνευσης γεγονότα τα οποία βρίσκονται μέσα από την ανάλυση ζωντανής ροής δεδομένων από τα μέσα κοινωνικής δικτύωσης. Επομένως, η απευθείας ανίχνευση εξετάζει δεδομένα που μόλις λαμβάνονται από το σύστημα. Αντιθέτως, η αναδρομική ανίχνευση γεγονότος σχετίζεται με την εύρεση άγνωστων γεγονότων μέσα από την ανάλυση ιστορικών δεδομένων προερχόμενων από τα μέσα κοινωνικής δικτύωσης. Είναι εμφανές, ότι η απευθείας ανίχνευση γεγονότος απαιτεί την ικανότητα του συστήματος να λαμβάνει κρίσιμες αποφάσεις σχετικά με τα δεδομένα εισόδου κατά το χρόνο στον οποίο εισέρχονται σε αυτό. Οι περισσότερες προσεγγίσεις απευθείας ανίχνευσης μπορούν να εφαρμοστούν και σε ιστορικά δεδομένα για να βρουν και να αναλύσουν παρελθοντικά γεγονότα [11].

Έχουν προταθεί ποικίλες μέθοδοι στη βιβλιογραφία στις οποίες αξιοποιούνται τεχνικές εξόρυξης δεδομένων για τη μοντελοποίηση της ανίχνευσης γεγονότων και παρουσιάζονται στο κεφάλαιο 2 αναλυτικά. Ο μεγαλύτερος όγκος αυτών έχει επικεντρωθεί σε ανίχνευση γεγονότων από δεδομένα στο Twitter [3]. Σύμφωνα με τους συγγραφείς στο [12], ακόμα και στην περίπτωση που ένα tweet αναφέρεται σε ένα γεγονός-στόχο για ανίχνευση μπορεί να μην αποτελεί κατάλληλο αναφορικό στοιχείο για το γεγονός αν δεν γίνεται σε πραγματικό χρόνο. Για παράδειγμα, κατά τη διάρκεια ενός σεισμού, tweets όπως «Σεισμός τώρα» ή «Σείεται το έδαφος! Σεισμός!» είναι αντιπροσωπευτικά του γεγονότος και γίνονται σε πραγματικό χρόνο, ωστόσο ένα tweet το οποίο δημοσιεύεται μια βδομάδα μετά όπως για παράδειγμα «Ο σεισμός της προηγούμενης βδομάδας ήταν 5 ρίχτερ» δεν αποτελεί ζωντανή παρατήρηση του γεγονότος αλλά εκ των υστέρων διατύπωσή του. Αποτελεί, επομένως, ζητούμενο η σωστή διαλογή των μηνυμάτων από τη ροή που λαμβάνεται στο μέσο κοινωνικής δικτύωσης, τα οποία σχετίζονται ενδεχομένως με ένα γεγονός-στόχο της ανίχνευσης. Ο τρόπος με τον οποίο γίνεται αυτή η διαλογή σχετίζεται άμεσα από τον τύπο και τα χαρακτηριστικά του γεγονότος προς ανίχνευση όπως και από τα χαρακτηριστικά του μέσου δικτύωσης.

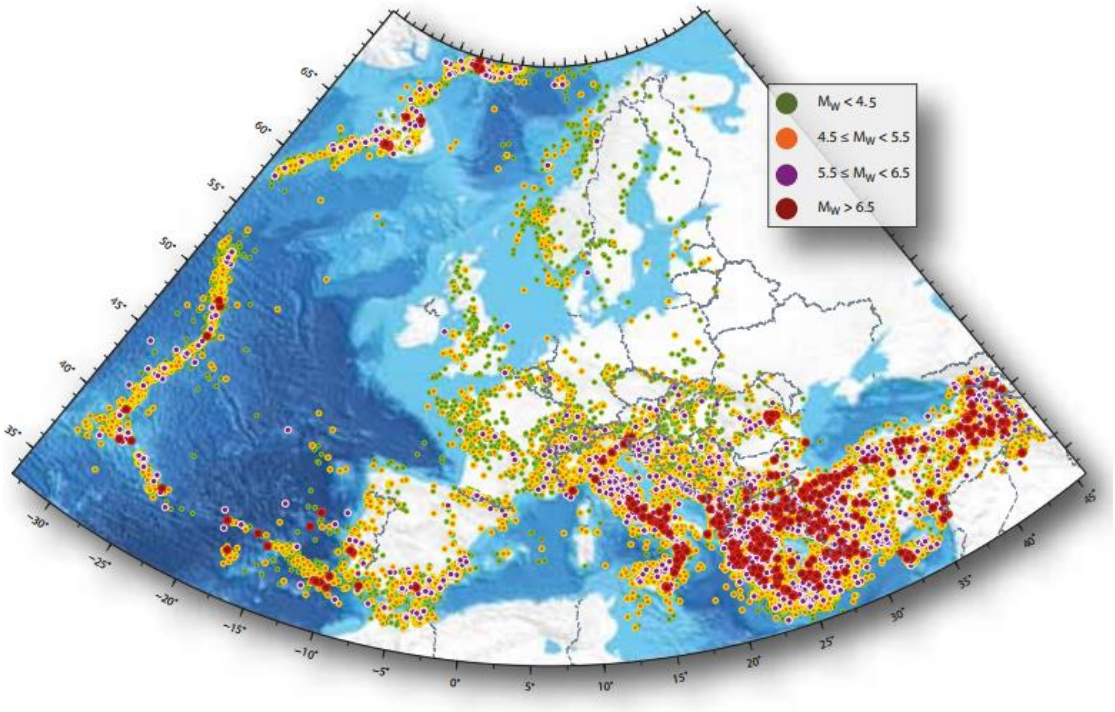
1.2 Αντικείμενο διπλωματικής

Αντικείμενο της παρούσας διπλωματικής αποτελεί η μελέτη των μεθόδων ανίχνευσης γεγονότων στα μέσα κοινωνικής δικτύωσης και η υλοποίηση ενός συστήματος ανίχνευσης συγκεκριμένων μη προγραμματισμένων γεγονότων που αφορούν φυσικές καταστροφές στην πλατφόρμα Twitter. Εν προκειμένω, λόγω του γεγονότος ότι η Ελλάδα είναι κατεξοχήν σεισμογενής χώρα [13] εξετάζονται φαινόμενα σεισμών ως παράδειγμα υλοποίησης – η Εικόνα 1.1 – Χάρτης σεισμών, έτη 1000 - 2007 του [13] επιβεβαιώνει τον ισχυρισμό αυτό παρουσιάζοντας έναν χάρτη σεισμών στην Ευρώπη για το χρονικό διάστημα 1000 – 2007 όπου φαίνεται ότι η Ελλάδα και η Ιταλία είναι οι χώρες με τους περισσότερους σεισμούς μεγάλης έντασης. Η υλοποίηση γίνεται στα πλαίσια μιας διαδικτυακής πλατφόρμας όπως επίσης και εφαρμογής για κινητές συσκευές η οποία έχει τη δυνατότητα ενημέρωσης των χρηστών για ανάλογα γεγονότα.

Πρόκληση αποτελεί το γεγονός ότι στη ροή tweets, ο μεγαλύτερος όγκος μηνυμάτων δεν αναφέρεται σε μη προγραμματισμένα γεγονότα και επομένως καλείται το σύστημα να είναι ιδιαίτερα επιλεκτικό όπως επίσης και να διακρίνει τις περιπτώσεις αναφοράς σε ένα γεγονός σε πραγματικό χρόνο και a posteriori, όπου στην τελευταία περίπτωση θα πρέπει η αναφορά να αγνοείται.

Σκοπός της εργασίας είναι μέσα από τη ζωντανή ροή των tweets που παρέχει το Twitter μέσω του Streaming API, το οποίο επιστρέφει σε πραγματικό χρόνο tweets με κατάλληλες λέξεις – κλειδιά, το σύστημα να διακρίνει εκείνα τα tweets που αφορούν πραγματικά ένα μη προγραμματισμένο γεγονός προς ανίχνευση καθώς επίσης να εξάγει πληροφορίες για αυτό από τα αντίστοιχα tweets, όπως την τοποθεσία και το χρόνο ενός γεγονότος.

Ακόμη, η εργασία αυτή αποσκοπεί στη δημιουργία ενός πρωτότυπου ολοκληρωμένου συστήματος ανίχνευσης έκτακτων γεγονότων που αφορά κατά κύριο λόγο ελληνικά tweets και κατ' επέκταση γεγονότα που λαμβάνουν χώρα στον ελλαδικό χώρο, αφού επικεντρώνεται σε αναφορές στο μέσο Twitter που χρησιμοποιούν ελληνικές λέξεις-κλειδιά.



Εικόνα 1.1 – Χάρτης σεισμών, έτη 1000 - 2007

1.2.1 Συνεισφορά

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Μελετήσαμε δύο είδη συστημάτων βασισμένα σε μηχανική μάθηση – συστήματα που αξιοποιούν αλγορίθμους ομαδοποίησης (clustering) και συστήματα που αξιοποιούν μεθόδους ταξινόμησης (classification) – για την ανίχνευση γεγονότων στο Twitter και επιλέξαμε προς υλοποίηση εκείνο που στηρίζεται στη μέθοδο της ταξινόμησης ως το καταλληλότερο για την ανίχνευση μη προγραμματισμένων γεγονότων συγκεκριμένου θέματος στηριζόμενοι στη βιβλιογραφία και τις παρατηρήσεις μας.
2. Υλοποιήσαμε αλγορίθμους καταγραφής και αποθήκευσης tweets από τη ροή του Streaming API, προεπεξεργασίας δεδομένων για tweets γραμμένα στην ελληνική γλώσσα και ειδοποίησης του συστήματος με πιθανοτικό τρόπο όσον αφορά την ύπαρξη ή όχι γεγονότος από τα δεδομένα αυτά. Ακόμη, δημιουργήσαμε μια πρωτότυπη μέθοδο εντοπισμού της τοποθεσίας ενός συμβάντος.
3. Δημιουργήσαμε ένα **δείγμα δεδομένων εκπαίδευσης (training dataset)** με ελληνικά tweets και εκπαιδεύσαμε έναν ταξινομητή (classifier) για την αντιστοίχιση των tweets σε κατάλληλες κλάσεις.
4. Δημιουργήσαμε χαρακτηριστικά για την εκπαίδευση του ταξινομητή προσαρμοσμένα στον τύπο των γεγονότων προς ανίχνευση και των tweets που τα περιγράφουν.

5. Υλοποιήσαμε διαδικτυακή πλατφόρμα και εφαρμογή για κινητές συσκευές για την ενημέρωση των χρηστών σχετικά με τα γεγονότα που ανιχνεύονται.

1.3 Οργάνωση κειμένου

Εργασίες σχετικές με το αντικείμενο της διπλωματικής παρουσιάζονται στο κεφάλαιο 2 όπου αναλύεται περισσότερο το θέμα της ανίχνευσης γεγονότων και σχολιάζονται μέθοδοι ανίχνευσης που έχουν προταθεί στη βιβλιογραφία.

Η περιγραφή των υποσυστημάτων ταξινόμησης, ειδοποίησης και εντοπισμού τοποθεσίας συμβάντος δίνονται στο κεφάλαιο 3.

Στο κεφάλαιο 4 αναλύεται η μορφή των δεδομένων που παρέχει το Twitter και επεξηγείται ο τρόπος δημιουργίας του δείγματος δεδομένων εκπαίδευσης από αυτά. Ακόμη, περιέχεται σε αυτή την ενότητα η επιλογή των κατάλληλων χαρακτηριστικών που αξιοποιήθηκαν για την εκπαίδευση του ταξινομητή του συστήματος και η προεπεξεργασία στην οποία υπόκεινται τα δεδομένα που τροφοδοτούνται στο υποσύστημα ταξινόμησης.

Ακολούθως, στο κεφάλαιο 5 δίνονται λεπτομέρειες για τη λειτουργία της διαδικτυακής πλατφόρμας και της εφαρμογής για κινητά καθώς και πληροφορίες για τα προγραμματιστικά εργαλεία που χρησιμοποιήθηκαν.

Στο κεφάλαιο 6 παρουσιάζονται τα πειράματα που πραγματοποιήθηκαν και η αξιολόγηση της επίδοσης του συστήματος ταξινόμησης και ειδοποίησης στην προσομοίωση.

Τέλος, στο κεφάλαιο 7 συνοψίζονται τα αποτελέσματα και συμπεράσματα της εργασίας και σχολιάζονται μελλοντικές βελτιώσεις και επεκτάσεις του χρησιμοποιούμενου συστήματος.

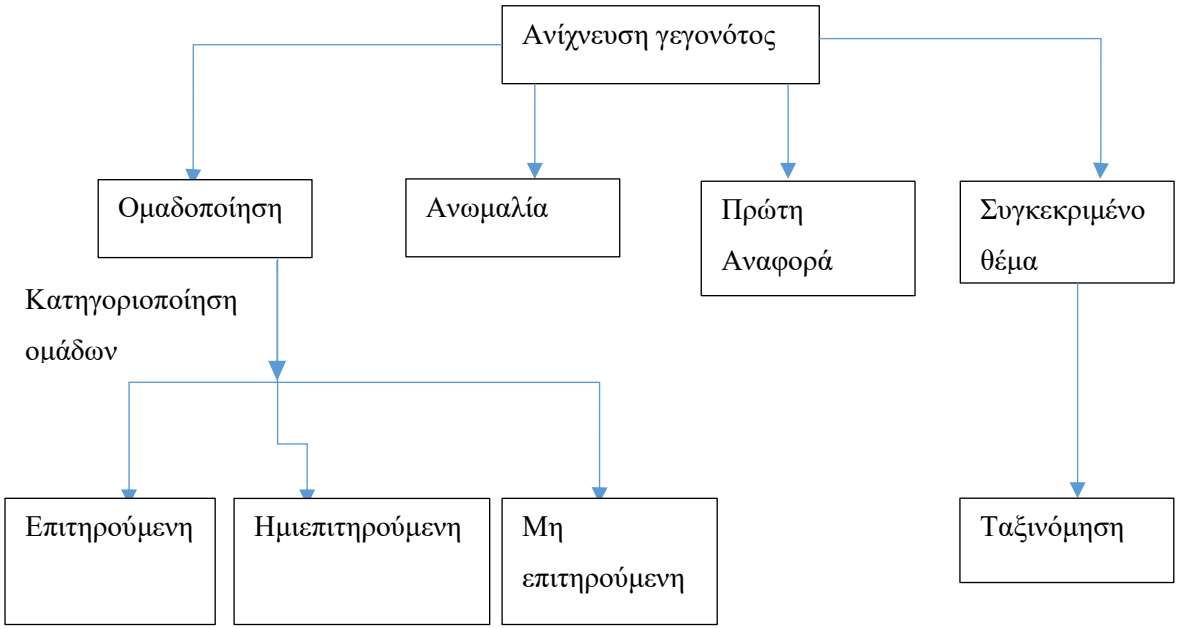
2

Σχετικές εργασίες

Στο παρόν κεφάλαιο αναλύονται οι διαφορετικές μέθοδοι ανίχνευσης γεγονότων από δεδομένα προερχόμενα από μέσα κοινωνικής δικτύωσης που έχουν προταθεί στη βιβλιογραφία και τις οποίες κατηγοριοποιήσαμε κατάλληλα με βάση τις θεμελιώδεις τεχνικές εξόρυξης δεδομένων που αξιοποιούνται στον αλγόριθμο ανίχνευσης. Οι κατηγορίες αυτές είναι οι εξής: ομαδοποίηση και ανίχνευση (clustering), ανίχνευση ανωμαλίας (anomaly), ανίχνευση πρώτης αναφοράς σε γεγονός (first story) και τέλος ανίχνευση γεγονότος συγκεκριμένου θέματος (topic specific).

2.1 Μέθοδοι ανίχνευσης γεγονότων από δεδομένα σε κοινωνικά δίκτυα

Στα επόμενα παρουσιάζονται λεπτομέρειες των κύριων τεχνικών ανίχνευσης γεγονότων που έχουν χρησιμοποιηθεί στη βιβλιογραφία με συγκεκριμένα παραδείγματα που παρέχουν πληροφορίες αναφορικά με την υλοποίηση των τεχνικών αυτών στα πλαίσια συστημάτων ανίχνευσης. Παρακάτω φαίνονται στο Σχήμα 2.1 οι διάφορες κατηγορίες ανίχνευσης σχηματικά σε ένα διάγραμμα πριν αυτές αναλυθούν. Η κατηγοριοποίηση έχει γίνει με βάση τη θεμελιώδεις αρχές εξόρυξης δεδομένων που χρησιμοποιούνται στον αλγόριθμο ανίχνευσης.



Σχήμα 2.1 – Ταξινόμηση μεθόδων εξόρυξης δεδομένων για ανίχνευση γεγονότων

2.1.1 Ομαδοποίηση και ανίχνευση

Οι περισσότερες προσεγγίσεις ανίχνευσης γεγονότος στο Twitter που είναι διαθέσιμες στη βιβλιογραφία βασίζονται σε τεχνικές **ομαδοποίησης (clustering)**, οι οποίες έχουν κατά κόρον αξιοποιηθεί σε συστήματα απευθείας ανίχνευσης γεγονότος όπου δεν είναι ορισμένα τα γεγονότα στα οποία στοχεύει η ανίχνευση [11]. Η τεχνική αυτή αποσκοπεί στην ταξινόμηση των tweets σε πραγματικό χρόνο σε ομάδες (εναλλακτικά συστάδες) με βάση κάποιο κριτήριο ομοιότητας και στην κατάταξή τους κατόπιν με βάση το αν αφορούν ή όχι κάποιο γεγονός. Η τελευταία κατάταξη προκύπτει μέσω μηχανισμών **χωρίς επιτήρηση ή με επιτήρηση**. Επομένως, η ομαδοποίηση γεγονότων είναι μια πολύ χρήσιμη μέθοδος για τον εντοπισμό θεμάτων που συζητούνται στο Twitter σε ζωντανό χρόνο όπου εν συνεχεία γίνεται διαλογή αυτών που αναφέρονται σε γεγονότα και δεν είναι τάσεις όπως memes που συζητούνται και διαμοιράζονται στο μέσο. Τα τελευταία αφορούν αμφιλεγόμενα θέματα συζήτησης που παρακινούν τους χρήστες να δημοσιοποιήσουν την άποψή τους στο μέσο. Σύμφωνα με το [14], οι περισσότερες καταγεγραμμένες προσεγγίσεις στη βιβλιογραφία αποτυγχάνουν να διαχωρίσουν τα Twitter memes από έκτακτα γεγονότα. Έτσι, μια σημαντική πρόκληση που αφορά την προσέγγιση της ομαδοποίησης είναι ότι οδηγεί σε ορισμένες περιπτώσεις σε θορυβώδεις περιγραφές των γεγονότων που ανιχνεύονται [15].

Στην περίπτωση που το σύστημα αποφασίζει **χωρίς επιτήρηση** για το αν μια ομάδα αναφέρεται σε κάποιο γεγονός, αποδίδεται σε κάθε ομάδα μια βαθμολόγηση (scoring) με βάση τα χαρακτηριστικά τους. Ακόμη, από τα tweets της ομάδας μπορούν να προκύψουν χρήσιμες πληροφορίες όπως ο χρόνος, ο τόπος και η περιγραφή του γεγονότος.

Στη δεύτερη περίπτωση, στην οποία το σύστημα αποφασίζει **επιτηρούμενο** σχετικά με την αναφορά μιας ομάδας σε γεγονός, αξιοποιούνται ταξινομητές, οι οποίοι αποφασίζουν με βάση τα χαρακτηριστικά των ομάδων και οι αποφάσεις αυτές είναι άμεσα συνδεδεμένες με τα ιστορικά δεδομένα τα οποία έχουν χρησιμοποιηθεί κατά την εκπαίδευσή του. Η εκπαίδευση επιτυγχάνεται μέσω ενός δείγματος δεδομένων εκπαίδευσης, δηλαδή ενός συνόλου δεδομένων το οποίο έχει χειροκίνητα ταξινομηθεί σε δύο κλάσεις, μία θετική όταν ο ταξινομητής αποφαίνεται ότι μια ομάδα αφορά ένα γεγονός προς ανίχνευση και μια αρνητική για την αντίθετη περίπτωση. Η περιοδική εκπαίδευση ενός συστήματος επιτηρούμενης μηχανικής μάθησης είναι επίσης μια σημαντική διαδικασία για την εξασφάλιση της εύρυθμη λειτουργίας ενός συστήματος σε βάθος χρόνου.

Υπάρχουν διάφορες τεχνικές ομαδοποίησης που χρησιμοποιούνται για την ομαδοποίηση δεδομένων όπως και ποικίλοι αλγόριθμοι που τις υλοποιούν. Η τεχνική της **ιεραρχικής** ομαδοποίησης **(hierarchical** clustering**)** μπορεί να είναι είτε **συσσωρευτική** ομαδοποίηση **(agglomerative** clustering**)**, όπου όλες οι οντότητες θεωρούνται εξ αρχής ως ανεξάρτητες συστάδες ενός μόνο στοιχείου και στη συνέχεια δημιουργούνται ευρύτερες συστάδες με συνένωσή τους μέσω ενός κριτηρίου, είτε **διαχωριστική** ομαδοποίηση (**divisive** clustering) όπου τα δεδομένα είναι αρχικά κατανεμημένα σε μια μεγάλη συστάδα και με συνεχή βήματα η συστάδα αυτή διαιρείται σε επιμέρους μικρότερες ομάδες [16]. Μια άλλη τεχνική είναι η **ομαδοποίηση βασισμένη στη χωρική πυκνότητα εφαρμογών με θόρυβο (density-based spatial clustering of applications with noise – DBSCAN)**. Ο αλγόριθμος DBSCAN κατατάσσει στην ίδια συστάδα δεδομένα τα οποία είναι κοντινά μεταξύ τους με βάσει μια μετρική, συνήθως την ευκλείδια απόσταση, και έχοντας ως δεδομένα έναν ελάχιστο αριθμό δεδομένων που ορίζουν μια συστάδα [17]. Ακόμη, η **διαμεριστική** ομαδοποίηση **(partitioning** clustering**)** προϋποθέτει τον ορισμό του αριθμού των ομάδων πριν την έναρξη της διαδικασίας ομαδοποίησης και αυτή τερματίζεται όταν ένα κριτήριο τερματισμού ικανοποιείται. Χαρακτηριστικός αλγόριθμος διαμεριστικής ομαδοποίησης είναι ο K-means, ο οποίος ομαδοποιεί δεδομένα σε k ομάδες [18]. Υπάρχουν και άλλοι μέθοδοι ομαδοποίησης όπως για παράδειγμα η ομαδοποίηση **βασισμένη σε πλέγμα (grid-based** clustering**)**.

Διάφορα συστήματα ομαδοποίησης χωρίς επιτήρηση έχουν προταθεί με βάση τις παραπάνω τεχνικές ομαδοποίησης στα πλαίσια της ανίχνευσης γεγονότων στα μέσα κοινωνικής δικτύωσης. Οι τεχνικές μη επιτηρούμενης ομαδοποίησης μπορούν να χρησιμοποιηθούν για την απευθείας ανίχνευση άγνωστων γεγονότων από το Twitter επειδή δεν απαιτούν εκπαίδευση και

χειροκίνητη ταξινόμηση και προσθήκη ετικετών (manual labelling) των δεδομένων [11]. Παρακάτω δίνονται ορισμένα παραδείγματα χρήσης **μεθόδων ομαδοποίησης χωρίς επιτήρηση** οι οποίες στηρίζονται σε μία συνάρτηση βαθμολόγησης (scoring function) ώστε να αποφανθούν για τις ομάδες που αφορούν γεγονότα.

Στο [19] παρουσιάζεται ένας τέτοιος αλγόριθμος διαδικτυακής ομαδοποίησης (online clustering) στο Twitter. Η διαδικτυακή ομαδοποίηση αφορά ομαδοποίηση στην οποία σε κάθε βήμα λαμβάνονται νέα δείγματα δεδομένων τα οποία σχηματίζουν μια νέα συστάδα ή κατατάσσονται σε μια ήδη υπάρχουσα από δείγματα που έχουν παρατηρηθεί προηγουμένως [20]. Ο αλγόριθμος που περιγράφεται στο [19] προτείνει ένα σύστημα ανίχνευσης σημαντικών ή γνωστών εκ των προτέρων γεγονότων, η υλοποίηση του οποίου έχει ως πρώτο βήμα τη διαδικτυακή διαμεριστική ομαδοποίηση. Ως εκ τούτου, δίνεται ως είσοδος ο αριθμός των ομάδων k στις οποίες κατανέμονται τα δείγματα της εισερχόμενης ροής από tweets. Ο αλγόριθμος, τον οποίο οι συγγραφείς ονομάζουν Cluster Summary τρέχει επαναληπτικά και σε κάθε επανάληψη υπολογίζεται η ομοιότητα των εισερχόμενων δειγμάτων με μια σύνοψη (summary) που αφορά κάθε ομάδα. Η ομοιότητα υπολογίζεται ως γραμμικός συνδυασμός της δομικής ομοιότητας με την ομοιότητα βάσει περιεχομένου, όπου η τελευταία δεν είναι παρά η tf-idf (term frequency – inverse document frequency) ομοιότητα. Το σχήμα tf-idf είναι ένα από τα πιο δημοφιλή σχήματα στον τομέα της εξόρυξης δεδομένων όσον αφορά την αντιστοίχιση βαρών σε όρους κειμένου [21]. Με βάση το σχήμα αυτό ένα έγγραφο κειμένου αναπαρίσταται με ένα διάνυσμα, η κάθε θέση του οποίου αντιστοιχεί σε έναν συγκεκριμένο όρο – λέξη και εντός της οποίας κείται μια αριθμητική τιμή, η οποία όσο υψηλότερη τιμή έχει τόσο περισσότερο πιθανό είναι να αντιπροσωπεύει καλύτερα το θέμα του κειμένου. Συγκεκριμένα, η αριθμητική αυτή τιμή είναι το γινόμενο $tf * idf$, όπου tf είναι ο αριθμός εμφανίσεων του όρου στο κείμενο και $idf = \log_2 \frac{N}{D}$, όπου N ο συνολικός αριθμός κειμένων, άρα και διανυσμάτων, προς επεξεργασία και D ο αριθμός των κειμένων που περιέχουν τον όρο αυτό.

Στο [4] περιγράφεται το σύστημα EvenTweet, το οποίο εντοπίζει τοπικά γεγονότα μέσα από τη συνεχή ανάλυση των πιο πρόσφατων tweets στο μέσο Twitter σε κυλιόμενα χρονικά παράθυρα. Τα γεγονότα που ανιχνεύονται περιγράφονται μέσω λέξεων – κλειδιών, ενώ ακόμη εκτιμούνται η γεωγραφική τοποθεσία στην οποία λαμβάνει χώρα το γεγονός καθώς και ο χρόνος έναρξης του γεγονότος. Το σύστημα αρχικά εξαγάγει εντός ενός χρονικού παραθύρου λέξεις – κλειδιά από tweets που επιδεικνύουν μεγάλη συχνότητα αναφοράς, ύστερα επιλέγει εκείνες που επιδεικνύουν γεωγραφική κατανομή τοπικά, τις ομαδοποιεί με βάση το χωρικό προφίλ τους και τέλος βαθμολογεί τις ομάδες αυτές – οι οποίες με το πέρασμα των χρονοπαραθύρων ανανεώνονται και αν η σημαντικότητά τους μειωθεί σημαντικά αφαιρούνται. Ως εκ τούτου, λέξεις – κλειδιά που αφορούν μια συγκεκριμένη γεωγραφική τοποθεσία τοποθετούνται στην ίδια συστάδα. Για κάθε τέτοια λέξη – κλειδί βρίσκεται η χωρική υπογραφή (spatial signature)

της μέσω των tweets, τα οποία εμπεριέχουν τοποθεσία και περικλείουν τις λέξεις – κλειδιά. Λέξεις – κλειδιά με υψηλή χωρική εντροπία και χαμηλή συχνότητα εμφάνισης δεν συμπεριλαμβάνονται αλλά απορρίπτονται ως θορυβώδεις. Η εντροπία της χωρικής κατανομής έχει οριστεί έτσι, ώστε υψηλή εντροπία ισοδυναμεί με λέξεις – κλειδιά διάσπαρτες στο χώρο, ενώ χαμηλή εντροπία σημαίνει εμφάνισή τους σε εντοπισμένες τοποθεσίες. Η ομαδοποίηση γίνεται διαδικτυακά σε πραγματικό χρόνο εντός των κυλιόμενων χρονοπαραθύρων σε ένα πέρασμα και ο αλγόριθμος υλοποίησης είναι παρόμοιος με τον αλγόριθμο Birch, όπως περιγράφεται στο [22]. Η ομαδοποίηση στηρίζεται στην ομοιότητα συνημιτόνου (cosine similarity) των χωρικών υπογραφών των λέξεων – κλειδιών, οπότε και με βάση ένα όριο απόστασης από τα κεντροειδή των ομάδων αντιστοιχίζεται κάθε μία στην κοντινότερη συστάδα. Εναλλακτικά, μια νέα ομάδα δημιουργείται από μια λέξη – κλειδί. Εν συνεχεία, οι ομάδες λαμβάνουν βαθμολογία με βάση ένα μοντέλο το οποίο αξιολογεί το πόσο πιθανό είναι μια ομάδα να αφορά κάποιο τοπικό γεγονός και στηρίζεται στα εξής (α) την υψηλή συχνότητα εμφάνισης των λέξεων – κλειδιών της (β) την σχετικά μακρόχρονη παραμονή των λέξεων – κλειδιών αυτών στην ομάδα ως μέλη της, και (γ) την πρόσφατη αντιστοίχισή τους σε αυτήν την ομάδα. Στο τέλος, οι k υψηλότερα βαθμολογημένες ομάδες θεωρούνται ως αντιπροσωπευτικές τοπικών γεγονότων και οπτικοποιούνται στη συνέχεια τα αποτελέσματα. Στο [23], προτείνεται ένα σύστημα το οποίο αξιοποιεί tweets που περιέχουν hashtags. Με βάση την μέθοδο της συσσωρευτικής ομαδοποίησης που βασίζεται στην ομοιότητα συνημιτόνου όπως αναφέρεται στο [24], προτείνεται ένα σύστημα που αντιστοιχεί ένα διάνυσμα σε κάθε tweet χρησιμοποιώντας τιμές $tf - idf$ και ύστερα υλοποιείται η ομαδοποίηση των tweets. Με βάση το σύστημα αυτό, συγκρίνονται τα αποτελέσματα χρησιμοποιώντας τις λέξεις των tweets στη μία περίπτωση για το μοντέλο και μόνο τα hashtags στην άλλη. Σύμφωνα με τα ευρήματα των συγγραφέων, η μέθοδος που αξιοποιεί τα hashtags στην $tf - idf$ οδηγεί σε βελτίωση των μετρικών ακρίβειας (precision), ανάκλησης (recall) και βαθμολογίας – F (F-score) και δείχνει ότι η μέθοδος αυτή που αξιοποιεί αποκλειστικά hashtags μπορεί να είναι ικανή από μόνη της στο να ανιχνεύει συγκεκριμένους τύπους γεγονότων.

Τα συστήματα ανίχνευσης που βασίζονται σε **ομαδοποίηση με επιτήρηση**, κατηγοριοποιούν τις ομάδες σε δύο κλάσεις αναφορικά με το αν σχετίζονται ή όχι με γεγονότα. Ένα παράδειγμα συστήματος ομαδοποίησης με επιτήρηση είναι το [7], που βασίζεται σε ομαδοποίηση tweets και επιτηρούμενη ταξινόμηση (supervised classification) μετέπειτα. Η ομαδοποίηση, γίνεται διαδικτυακά και αυξανόμενα, χωρίς να είναι προκαθορισμένος ο αριθμός των συστάδων. Η αναπαράσταση των tweets γίνεται μέσω στάθμισης των όρων με $tf - idf$ και όταν η απόσταση ενός tweet από τα κεντροειδή των ομάδων είναι μεγαλύτερη από ένα προκαθορισμένο όριο, δημιουργείται μια νέα ομάδα με μοναδικό στοιχείο αυτό το tweet. Μια μηχανή διανυσμάτων υποστήριξης (SVM – Support Vector Machine) αξιοποιείται ακόμη για την ταξινόμηση των

ομάδων στις κλάσεις. Η μηχανή διανυσμάτων υποστήριξης εκπαιδεύεται με βάση τα επιλεγμένα χαρακτηριστικά και τοποθετούνται χειροκίνητα κατάλληλες ετικέτες για να υποδείξουν την κλάση στην οποία ανήκει κάθε tweet του δείγματος δεδομένων εκπαίδευσης. Τα χαρακτηριστικά αυτά προκύπτουν από τις ομάδες και εμπεριέχουν τοπικά, χρονικά, σχετικά με το Twitter και κοινωνικά χαρακτηριστικά. Ως αποτέλεσμα, οι συγγραφείς αναφέρουν ότι η μέθοδος που βασίζεται σε ταξινομητή με χαρακτηριστικά που εξάγονται από τις ίδιες τις ομάδες είναι συγκριτικά καλύτερη από τη μέθοδο που αξιοποιεί αποκλειστικά λεκτικά χαρακτηριστικά.

2.1.2 Ανίχνευση ανωμαλίας

Στην παρούσα ενότητα σχολιάζονται μέθοδοι ανίχνευσης γεγονότων στα μέσα κοινωνικής δικτύωσης που αφορούν προσεγγίσεις βασισμένες στον εντοπισμό ανωμαλιών στα μέσα αυτά όπως λόγου χάρη η απότομη αύξηση των αναφορών σε συγκεκριμένη λέξη ή ομάδα λέξεων. Στις προσεγγίσεις αυτές ουσιαστικά αξιοποιούνται τυχόν εμφανίσεις ανωμαλιών στο περιεχόμενο που διακινείται στο μέσο, αφού πρώτα μοντελοποιηθεί το περιεχόμενο αυτό ώστε να είναι γνωστή η ομαλή ή αλλιώς τυπική κατάσταση. Η μοντελοποίηση αυτή προέρχεται από αξιοποίηση ιστορικών δεδομένων πάνω στα οποία δημιουργούνται γλωσσολογικά μοντέλα που καταγράφουν ένα χρονολόγιο της χρήσης λεξιλογικών όρων στο μέσο. Η ανωμαλία μπορεί για παράδειγμα να αφορά αύξηση του αριθμού των tweets στη μονάδα του χρόνου που κάνουν χρήση συγκεκριμένων όρων ή εντοπισμό αυξημένου αριθμού tweets που εκφράζουν συναισθήματα στα μέσα δικτύωσης.

Το πρώτο σύστημα που εξετάζουμε είναι το [25], στο οποίο παρουσιάζεται μια μέθοδος ανίχνευσης γεγονότων στο Twitter μέσω κυματιδιακής ανάλυσης σήματος (wavelet signal analysis) για τις εμφανίσεις των hashtags στη δημόσια ροή. Από τα APIs του Twitter παρέχεται για κάθε tweet μια λίστα με hashtags τα οποία έχουν χρησιμοποιηθεί και για τα οποία δημιουργούνται σήματα εμφάνισης, εν αντιθέσει με τεχνικές οι οποίες αξιοποιούν το κείμενο ενός tweet και είναι αρκετά πιο πολύπλοκες λόγω των σταδίων προεπεξεργασίας και του χωρισμού του κειμένου σε λεκτικά (tokens). Η προσέγγιση αυτή βασίζεται στον έλεγχο των αναφορών σημάτων hashtags αξιοποιώντας την τεχνική της κυματιδιακής ανάλυσης σήματος. Σε διαστήματα διάρκειας 5 λεπτών το σύστημα συλλέγει hashtags τα οποία αφαιρούνται από το κείμενο των tweets, καθώς αυτό αξιοποιείται για την περιγραφή των ανιχνευμένων γεγονότων σε επόμενο στάδιο, και κατόπιν προσμετράται ο αριθμός των εμφανίσεων των hashtags σε συνεχόμενα τέτοια διαστήματα. Μέσω του προγραμματιστικού μοντέλου MapReduce δημιουργούνται οι χρονοσειρές των σημάτων hashtags. Το μοντέλο MapReduce περιλαμβάνει μια συνάρτηση map, η οποία ορίζεται από το χρήστη και επεξεργάζεται ένα

ζεύγος κλειδιού – τιμής που δημιουργεί ένα σύνολο από ενδιάμεσα τέτοια ζεύγη και μια συνάρτηση reduce η οποία ενοποιεί όλες τις ενδιάμεσες τιμές που σχετίζονται με το ίδιο ενδιάμεσο κλειδί [26]. Στη συνέχεια μέσω του μετασχηματισμού κυματιδίων διακριτού χρόνου (discrete wavelet transformation) ανιχνεύονται ξεσπάσματα στη χρήση των hashtags που αντιστοιχούν σε πιθανά γεγονότα. Τέλος, τα ανιχνευόμενα γεγονότα περιγράφονται με μία σύνοψη αξιοποιώντας την τεχνική LDA (Latent Dirichlet Allocation), η οποία είναι ένας τρόπος εντοπισμού θεμάτων σε έναν αριθμό κειμένων.

Στο [8] οι συγγραφείς εισάγουν ένα καινοτόμο σύστημα ανίχνευσης γεγονότων βασισμένο κυρίως στη λειτουργία mentions του Twitter. Η λογική είναι πως tweets τα οποία περιέχουν τουλάχιστον ένα mention είναι πιο πιθανό να περιγράφουν ένα γεγονός επομένως το ενδιαφέρον επικεντρώνεται σε αυτά. Σε πρώτο σταδιο δημιουργείται ένα στατιστικό μοντέλο για να περιγράψει την συμπεριφορά του λεξιλογίου που χρησιμοποιείται στα tweets που περιέχουν mentions. Αν αυτή η συμπεριφορά εμφανίζει έντονες διακυμάνσεις από την αναμενόμενη τότε ένα σύνολο λέξεων θεωρείται πως σχετίζεται με κάποιο γεγονός. Για να εξακριβωθεί η διάρκεια του γεγονότος επιλύεται ένα πρόβλημα βελτιστοποίησης γνωστό και ως Maximum Contiguous Subsequence Sum (MCSS). Τέλος, ταξινομούνται τα γεγονότα με βάση το μέγεθος του αντίκτυπου που είχαν και ανιχνεύονται οι λέξεις που σχετίζονται περισσότερο με κάθε γεγονός μέσω υπολογισμού συντελεστών χρονοσειρών.

Στο [27], προτείνεται ένα σύστημα ανίχνευσης τοπικών γεγονότων που αφορούν την πραγματικότητα χρησιμοποιώντας πληροφορίες γεωγραφικής θέσης από tweets. Το σύστημα αυτό ονομάστηκε από τους δημιουργούς του Jasmine και θεωρεί ότι τα τοπικά γεγονότα προϋποθέτουν την ύπαρξη και συγκέντρωση ανθρώπων την ίδια ώρα και στο ίδιο μέρος. Για την υλοποίηση του συστήματος, εντοπίζονται ομάδες tweets που ασχολούνται με το ίδιο θέμα και τα οποία έχουν δημιουργηθεί εντός μικρού χρονικού διαστήματος και μικρής γεωγραφικής περιοχής. Βρίσκοντας κοινούς όρους που εμφανίζονται στα tweets επιτυγχάνεται η εύρεση κοινών θεμάτων προς συζήτηση στο μέσο Twitter. Το σύστημα έχει δομηθεί κατάλληλα ώστε να αλληλεπιδρά με το χρήστη, από τον οποίο απαιτεί την είσοδο ημερομηνίας, μεγέθους γεωγραφικής περιοχής καθώς και τον ελάχιστο αριθμό χρηστών του Twitter που σχετίζονται με την τοποθεσία αυτή. Όσον αφορά την υλοποίηση του συστήματος, ιαπωνικά tweets συλλέγονται μέσω του Streaming API και αποθηκεύονται χωριστά τα tweets με γεωγραφική θέση με εκείνα που δεν εμπεριέχουν γεωγραφική υπογραφή. Για τα τελευταία, μια μέθοδος αντιστοίχισης τοποθεσίας αναθέτει σε κάθε tweet χωρίς γεωγραφική υπογραφή μια τοποθεσία. Η μέθοδος αυτή βασίζεται στα υπόλοιπα tweets που εξαρχής διέθεταν γεωγραφική υπογραφή. Το σύστημα εξάγει σημαντικούς όρους που πιθανόν σχετίζονται με κάποιο γεγονός για κάθε δημοφιλή γεωγραφικό τόπο, όπου δημοφιλής είναι μια τοποθεσία η οποία παρατηρείται πολλάκις στα tweets που έχουν γεωγραφική υπογραφή εντός του χρονικού διαστήματος που

έχει ορίσει προηγουμένως ο χρήστης. Για την αποδοτική συλλογή γεωγραφικών υπογραφών διεσπαρμένων σε μια μικρή περιοχή αξιοποιείται ο αλγόριθμος Geohash [28]. Οι σημαντικοί όροι που αφορούν ένα δημοφιλές μέρος θα πρέπει να εμφανίζονται σε τουλάχιστον τρία tweets που σχετίζονται με την τοποθεσία. Κάθε τοποθεσία με έναν ή περισσότερους τέτοιους όρους ανιχνεύεται ως τοπικό γεγονός. Δημοφιλείς τοποθεσίες χωρίς σημαντικούς όρους απορρίπτονται. Ακόμη, το σύστημα δεν συμπεριλαμβάνει στην ανάλυση αυτή retweets καθώς θεωρείται ότι ένας χρήστης που κάνει retweet είναι πιθανό να μη βρίσκεται στην ίδια τοποθεσία με το δημιουργό του.

2.1.3 Ανίχνευση πρώτης αναφοράς σε γεγονός

Η εξέταση της ανίχνευσης **πρώτης αναφοράς** σε γεγονός (**first story** detection) έγινε για πρώτη φορά στο [5] το 2002, όπου στόχος της ανίχνευσης αυτής ήταν η εύρεση της πρώτης περίπτωσης δημοσίευσης ενός άρθρου σχετικού με κάποιο αναδυόμενο γεγονός ή θέμα. Αυτή η αρχική προσέγγιση είναι φανερό ότι επικεντρώθηκε σε περιεχόμενο άρθρων σχετιζόμενων με την επικαιρότητα. Σε μια πιο σύγχρονη προσέγγιση του θέματος, όπως στο [29], ως ανίχνευση πρώτης αναφοράς αναφέρεται η διαδικασία ανίχνευσης της πρώτης αναφοράς ή ιστορίας σε κάποιο νέο γεγονός από μια ροή κειμένων. Ένα γεγονός μπορεί επί παραδείγματι να θεωρηθεί ένας σεισμός στην Αθήνα ενώ δύο διαφορετικές αναφορές ή ιστορίες επί του γεγονότος είναι ο αριθμός των θυμάτων και το μέγεθος των υλικών ζημιών που προκάλεσε η εν λόγω φυσική καταστροφή.

Η αναγνώριση της πρώτης αναφοράς σχετικά με ειδησεογραφικά γεγονότα έχει αντιμετωπιστεί με διάφορες μεθόδους. Μια τυπική λύση στο πρόβλημα της ανίχνευσης πρώτης αναφοράς παρουσιάζεται στο [30] όπου υπολογίζεται η απόσταση ενός κειμένου που εξάγεται από το tweet από τους κοντινότερους γείτονές του. Αν η απόσταση είναι μικρότερη από ένα κατώφλι τότε το tweet θεωρείται πρωτότυπο και άρα θεωρείται ότι κάνει την πρώτη αναφορά σε ένα γεγονός.

Η προαναφερθείσα μέθοδος αποτολεί μια υλοποίηση μη-επιβλεπόμενης μηχανικής μάθησης. Η επιβλεπόμενης μάθησης εκδοχή αυτής της μεθόδου αναπτύσσεται στο [31] όπου ως χαρακτηριστικά του SVM ταξινομητή χρησιμοποιούνται η απόσταση από τον κοντινότερο γείτονα, η επικάλυψη μεταξύ διάφορων οντοτήτων και η επικάλυψη μεταξύ διάφορων όρων.

Μια πιο προηγμένη μέθοδος προτείνεται στο [32]. Οι συγγραφείς υποστηρίζουν πως τα προηγούμενα συστήματα αδυνατούν να κλιμακώσουν τις δυνατότητες τους σε συνθήκες συνεχούς ροής δεδομένων αφού ο υπολογισμός του κοντινότερου γείτονα έχει μεγάλο υπολογιστικό κόστος. Για να αντιμετωπιστεί αυτό το πρόβλημα καταστρώνεται μια τεχνική κατακερματισμού ώστε να υπολογιστεί σε σταθερό χρόνο O(1) ο πλησιέστερος γείτονας. Αυτή

η τεχνική κατακερματισμού αναφέρεται στον αλγόριθμο Locality Sensitive Hashing (LSH) που κατακερματίζει τα δεδομένα εισόδου έτσι ώστε παρόμοια αντικείμενα να αντιστοιχίζονται στους ίδιους «κουβάδες» (buckes) με μεγάλη πιθανότητα. Ένα πρόβλημα που προκύπτει από αυτή την προσέγγιση είναι ότι μπορούν να προκύψουν αρκετά λάθη εξαιτίας της τυχαιότητας του συστήματος. Για να μειωθεί η διακύμανση της αντιστοιχίας στο σωστό πλησιέστερο γείτονα χρησιμοποιούνται πολλαπλές LSH δομές δεδομένων και αναζητείται η καταλληλότερη για κάθε δεδομένο.

Μία ακόμη μέθοδος περιγράφεται στο [33], όπου οι συγγραφείς αναπτύσσουν μια βαθμολογία πρωτοτυπίας βασισμένη σε χρήση συγκεκριμένων όρων. Για να αποφευχθεί το υπολογιστικό κόστος της εύρεσης του πλησιέστερου γείτονα χρησιμοποιούν τον παράγοντα της αντίστροφης συχνότητας κειμένου (idf) ανα λέξη προκειμένου να σχεδιάσουν ένα βαθμολογικό σχήμα πρωτοτυπίας. Έτσι, κάθε κείμενο αποκτά μια βαθμολογία για το πόσο πρωτότυπο είναι σχετικά με τα υπόλοιπα δεδομένα η οποία αποτελείται από το άθροισμα των idf βαρών κάθε όρου (κάθε λέξης του κειμένου ουσιαστικά). Κατά συνέπεια, ένα κείμενο είναι πρωτότυπο εάν οι όροι του είναι πρωτότυποι. Θέτοντας ένα κατώφλι μπορούμε να αποφανθούμε για το αν ένα κείμενο είναι πρωτότυπο και να το ορίσουμε ως πρώτη αναφορά σε ένα γεγονός.

2.1.4 Ανίχνευση γεγονότος συγκεκριμένου θέματος

Οι προαναφερθείσες τεχνικές ανίχνευσης στοχεύουν σε γεγονότα οποιουδήποτε τύπου. Στην υποενότητα αυτή, ωστόσο, θα παρουσιαστούν μέθοδοι που έχουν αξιοποιηθεί στη βιβλιογραφία για την ανίχνευση γεγονότων τα οποία έχουν προκαθορισμένο θέμα και τύπο. Τέτοια γεγονότα είναι λόγου χάρη φυσικές καταστροφές, όπως καταιγίδες, τυφώνες και σεισμοί, αλλά και κοινωνικά γεγονότα όπως για παράδειγμα αθλητικά γεγονότα, εκθέσεις, ατυχήματα, πολιτικές καμπάνιες και μεγάλες γιορτές-εκδηλώσεις [12].

Ένα έργο το οποίο ασχολείται με την ανίχνευση γεγονότων συγκεκριμένου θέματος από δεδομένα του Twitter είναι το [12]. Οι συγγραφείς μελέτησαν φαινόμενα σεισμών και τυφώνων σε σχέση με τις αναφορές σχετικών γεγονότων στο μέσο και ανέπτυξαν ένα σύστημα ειδοποίησης που αφορά σεισμούς στην Ιαπωνία το οποίο ειδοποιεί τους εγγεγραμμένους χρήστες του μέσω ηλεκτρονικού μηνύματος. Αντικείμενο της εργασίας αυτής είναι η ανίχνευση ύπαρξης γεγονότων σε πραγματικό χρόνο εποπτεύοντας τη ροή tweets. Τα tweets συλλέγονται μέσω του API με χρήση λέξεων-κλειδιών. Με τη χρήση μιας μηχανής διανυσμάτων υποστήριξης (SVM) κατηγοριοποιείται κάθε tweet στη θετική ή αρνητική κλάση, όπου η ταξινόμηση στη θετική κλάση σημαίνει ότι ένα tweet αφορά γεγονός το θέμα του οποίου εμπίπτει στη θεματολογία ενδιαφέροντος. Για την εκπαίδευση του SVM τρία είδη

χαρακτηριστικών δοκιμάστηκαν ώστε να βρεθεί εκείνη η ομάδα με τις βέλτιστες μετρικές αξιολόγησης. Τέλος, η γεωγραφική υπογραφή των tweets αξιοποιείται για την εύρεση του επίκεντρου ενός γεγονότος με χρήση φίλτρων Kalman και Particle.

Στο [34], οι συγγραφείς υλοποίησαν ένα σύστημα ανίχνευσης αθλητικών γεγονότων κατά την περίοδο 2010 – 2011 που αφορούν το πρωτάθλημα NFL στις Ηνωμένες Πολιτείες Αμερικής. Κατά τη διάρκεια διεξαγωγής των αγώνων, συλλέγονται και αναλύονται tweets ώστε να εντοπιστούν σημαντικά γεγονότα που αφορούν τους αγώνες. Το δημιουργηθέν σύστημα υλοποιήθηκε σε δύο βήματα: πρώτον, ανιχνεύονται γεγονότα εντός μεταβλητών κυλιόμενων χρονικών παραθύρων με βάση τις μεταβολές στο ρυθμό δημοσίευσης tweets και δεύτερον, μέσω λεξικολογικής ανάλυσης των tweets που προκύπτουν μετά την ανίχνευση ενός γεγονότος από το πρώτο βήμα αναγνωρίζεται ο τύπος των γεγονότων. Παράδειγμα γεγονότων που ανιχνεύονται είναι η επίτευξη σκοραρίσματος. Οι συγγραφείς καταλήγουν στο συμπέρασμα ότι μπορούν να εντοπίσουν γεγονότα αρκετά αξιόπιστα έπειτα από 40 δευτερόλεπτα ζωντανής συλλογής tweets.

Ένα ακόμη σύστημα που εμπίπτει στην κατηγορία των γεγονότων συγκεκριμένου θέματος είναι το TEDAS [1], το οποίο στοχεύει σε γεγονότα που σχετίζονται με φυσικές καταστροφές και εγκλήματα. Τέτοια γεγονότα είναι αυτοκινητιστικά ατυχήματα, γεγονότα ανταλλαγής πυρών και τυφώνες. Το σύστημα προωθεί τα tweets που λαμβάνει σε έναν ταξινομητή που αποφαίνεται σχετικά με το αν πράγματι αυτά αφορούν κάποιο γεγονός ενδιαφέροντος. Έπειτα, ένα υποσύστημα εξάγει χρήσιμη πληροφορία από τα μεταδεδομένα των tweets της θετικής κλάσης σχετικά με το χρόνο και την τοποθεσία με τα οποία αυτά σχετίζονται. Εν συνεχεία, το κείμενο του tweet μαζί με αυτά τα μεταδεδομένα αποθηκεύονται σε μια βάση δεδομένων, η οποία αξιοποιείται για την ειδοποίηση σχετικά με γεγονότα που συμβαίνουν σε πραγματικό χρόνο αλλά και την ανάκληση δεδομένων με βάση κάποια είσοδο από το χρήστη του συστήματος. Το υλοποιηθέν μοντέλο αντιστοιχεί, ακόμη, τοποθεσίες στα tweets με βάση το δίκτυο του γράφοντος και οπτικοποιεί τα αποτελέσματα σε κατάλληλη πλατφόρμα. Τέλος,οι συγγραφείς αναφέρουν σημαντική βελτίωση της ακρίβειας του ταξινομητή του συστήματος με την προσθήκη επιπλέον χαρακτηριστικών πέραν των χαρακτηριστικών κειμένου των tweets.

3

Περιγραφή συστήματος ανίχνευση και επιμέρους υποσυστημάτων

Με βάση το προηγούμενο κεφάλαιο και στα πλαίσια της παρούσας διπλωματικής εργασίας, όπου είναι γνωστός ο τύπος γεγονότων προς ανίχνευση εκ των προτέρων, θα αντιμετωπίσουμε το πρόβλημα της ανίχνευσης ως ένα πρόβλημα εύρεσης γεγονότων συγκεκριμένου θέματος.

Στα επόμενα σχολιάζεται αναλυτικά το συνολικό σύστημα ανίχνευσης που υλοποιήθηκε καθώς επίσης και τα επιμέρους στοιχεία του, παραθέτοντας λεπτομέρειες που αφορούν το σύστημα ταξινόμησης των tweets, το σύστημα ειδοποίησης και τον τρόπο εντοπισμού τοποθεσίας ενός γεγονότος.

3.1 Υποσύστημα ταξινόμησης

Η επιλογή του κατάλληλου συστήματος ανίχνευσης προήλθε τόσο από τη μελέτη της βιβλιογραφίας όσο και από τις απαιτήσεις του συνολικού συστήματος προς υλοποίηση. Για ανίχνευση γεγονότος συγκεκριμένης θεματολογίας που αφορά φυσικές καταστροφές όπως σεισμοί επιλέχθηκε ένα σύστημα ανίχνευσης που βασίζεται σε ταξινομητές με εκπαίδευση για την διαδικτυακή και ζωντανή ταξινόμηση των tweets κατά την λήψη τους από το API ως σχετιζόμενα ή μη με τέτοια γεγονότα. Το βήμα αυτό είναι ιδιαίτερα σημαντικό για την εύρυθμη λειτουργία και του συστήματος ειδοποίησης, το οποίο όπως θα εξηγηθεί στα επόμενα βασίζεται στην στατιστική συμπεριφορά των εντοπιζόμενων tweet που έχουν ταξινομηθεί ως σχετικά με γεγονότα που αφορούν φυσικά φαινόμενα.

3.1.1 Κλάσεις ταξινομητή

Καθώς καταφθάνουν στο υποσύστημα ταξινόμησης tweets σε συνεχόμενη βάση, αυτά ταξινομούνται σε δύο κατηγορίες. Για παράδειγμα, ας υποθέσουμε ότι ένας σεισμός λαμβάνει χώρα στην πόλη της Αθήνας την 1η Αυγούστου 2017 και ώρα 12:00, τότε τα tweets «Σεισμός τώρα στην Αθήνα!» και «5.3 ρίχτερ, Αθήνα πριν λίγα λεπτά» τα οποία έχουν δημοσιευτεί σε διάστημα μερικών λεπτών μετά την πραγματοποίηση του σεισμού αποτελούν πραγματικές αναφορές σε ζωντανό χρόνο του γεγονότος. Η επιθυμητή λειτουργία του ταξινομητή είναι να αντιστοιχεί τέτοιες αναφορές στην κλάση που αφορά ζωντανές αναφορές σεισμών, δηλαδή τη θετική κλάση (True). Στον Πίνακας 3.1 παρατίθενται περισσότερα παραδείγματα tweets με τις κλάσεις στις οποίες είναι επιθυμητό να αντιστοιχίζονται από τον ταξινομητή.

Κείμενο tweet	Θετική κλάση (True, αναφορά σε γεγονός ενδιαφέροντος)	Αρνητική κλάση (False)	Επεξήγηση
Σεισμός σεισμός σοσιαλισμός		✓	Μεταφορική χρήση της λέξεως «σεισμός». Δεν αποτελεί αναφορά σε γεγονός
πολιτικός σεισμός μεταξύ των δύο χωρών		✓	Μεταφορική χρήση της λέξεως «σεισμός». Δεν αποτελεί αναφορά σε γεγονός
Στα 10 εκατομμύρια ευρώ οι ζημιές του περσινού σεισμού		✓	Αναφορά σε σεισμό αλλά όχι σε πραγματικό χρόνο (παρελθοντική αναφορά)
Σεισμός 4,3 Ρίχτερ νότια του Ηρακλείου - ΤΩΡΑ	✓		Πραγματική αναφορά σεισμού εντός εύλογου χρονικού διαστήματος

Σεισμός τώρα στην Ηλεία. Το επίκεντρο στη Ανδραβίδα	✓		Πραγματική αναφορά σεισμού εντός εύλογου χρονικού διαστήματος
Συμβαίνει τώρα: Εγκέλαδος 7,1 Ρίχτερ έπληξε την Ιαπωνία - Προειδοποίηση για τσουνάμι	✓		Πραγματική αναφορά σεισμού εντός εύλογου χρονικού διαστήματος
Γνωστός τρομοκράτης αυτό ο Εγκέλαδος. Καταστρέφει νοικοκυριά αιώνες τώρα.		✓	Δεν αποτελεί αναφορά σε γεγονός
Σεισμός εκπτώσεων 2017: Ποιο εμπορικό κέντρο «ξεπουλάει» με 70% έκπτωσεις. Πότε αρχίζει		✓	Μεταφορική χρήση της λέξεως «σεισμός». Δεν αποτελεί αναφορά σε γεγονός

Πίνακας 3.1

Για να φανεί ακόμη καλύτερα η διάκριση μεταξύ των tweets που ταξινομούνται στη θετική κλάση εν συγκρίσει με αυτά που ταξινομούνται στην αρνητική, δίνουμε ένα πραγματικό παράδειγμα σεισμού που έλαβε χώρα στις 15:28 ώρα Ελλάδος (12:28 UTC) την 12/06/2017 μεταξύ Χίου και Λέσβου και των τουρκικών ακτών και είχε ένταση 6.3 ρίχτερ [35], [36]. Για το σεισμό αυτό λάβαμε τα παρακάτω tweets του Πίνακας 3.2 τα οποία θεωρούμε ότι ανήκουν στη θετική κλάση. Από το κείμενο των tweets αυτών έχουν αφαιρεθεί τυχόν ηλεκτρονικές διευθύνσεις οι οποίες περιέχονταν.

Σεισμός 6.3 ρίχτερ μεταξύ Λέσβου, Χίου και τουρκικών ακτών, ώρα 12:28 UTC, 12/06/2017			
Κείμενο tweet	Ημερομηνία δημοσίευσης tweet (ώρα σε UTC)	Θετική κλάση (True)	Αρνητική κλάση (False)
Σεισμός 6,1 Ρίχτερ ανοιχτά της Μυτιλήνης #seismos #σεισμος via @giatigr	Mon Jun 12 12:56:55 +0000 2017	✓	

ΙΣΧΥΡΟΣ ΣΕΙΣΜΟΣ 6,3 ΡΙΧΤΕΡ ΝΟΤΙΑ ΤΗΣ ΜΥΤΙΛΗΝΗΣ	Mon Jun 12 12:53:25 +0000 2017	✓	
ΣΕΙΣΜΟΣ ΤΩΡΑ	Mon Jun 12 12:31:35 +0000 2017	✓	
σεισμος μεγαλος πριν λιγα λεπτα	Mon Jun 12 12:44:34 +0000 2017	✓	

Πίνακας 3.2

Όλα τα παραπάνω tweets αναφέρονται με εξαιρετικά μεγάλη πιθανότητα στον εν λόγω σεισμό είτε λόγω της περιγραφής τους όπου και γίνεται ξεκάθαρο σε ποιον τόπο αναφέρονται (δύο πρώτα tweets) είτε λόγω της χρονικής τους δημοσίευσης που βρίσκεται κοντά στο χρόνο που έλαβε χώρα ο σεισμός. Είναι αναμενόμενο οι χρήστες του Twitter να αναφέρουν ένα γεγονός όπως ένας μεγάλος σεισμός με κάποια χρονική καθυστέρηση, καθώς δεν έχουν άμεση πρόσβαση στο μέσο ώστε να δημοσιεύσουν την αναφορά τους. Ακόμη, είναι πιθανό να αναφέρονται σε κάποιο μετασεισμό που ακολουθεί τον κυρίως σεισμό.

3.1.2 Μετρικές αξιολόγησης επιδόσεων ταξινομητή

Στο σημείο αυτό τονίζεται ότι ο ταξινομητής στην πραγματικότητα μπορεί να μην αντιστοιχίσει έτσι τα tweets στις δύο κλάσεις που έχουν επιλεγεί. Αυτό εξαρτάται από το είδος της εκπαίδευσης που έχει πραγματοποιηθεί και τα χαρακτηριστικά της ταξινόμησης όπως θα αναφερθεί παρακάτω. Σκοπός της μελέτης όσον αφορά τον κατάλληλο ταξινομητή είναι η ικανοποιητική απόδοσή του, η οποία και ποσοτικοποιείται με τις ακόλουθες μετρικές: **ακρίβειας (precision), ανάκλησης (recall) και βαθμολογίας – F (F-score)** .

Η μετρική ακρίβειας εκφράζει το πόσα εκ των tweets που ο ταξινομητής έχει αντιστοιχίσει στη θετική κλάση (κλάση True) ανήκουν πράγματι στην κλάση αυτή, δηλαδή περιγράφουν γεγονότα στα οποία στοχεύει η ανίχνευση. Μεγαλύτερη ακρίβεια σημαίνει και μικρότερο αριθμό ψευδώς θετικών αποτελεσμάτων (false positive – F_P).

Η μετρική ανάκλησης εκφράζει το πόσα εκ των θετικών παραδειγμάτων εκτίμησε ορθά ο ταξινομητής. Όσο μεγαλύτερο είναι το μέτρο της ανάκλησης, τόσο λιγότερα θετικά

παραδείγματα έχουν ταξινομηθεί λανθασμένα (ψευδώς αρνητικά αποτελέσματα, false negative – F_n). Ο Πίνακας 3.3 επεξηγεί τις προαναφερθείσες έννοιες των ψευδών αποτελεσμάτων.

	Πραγματική κλάση		
		TRUE	FALSE
Κλάση πρόβλεψης (από ταξινομητή)	TRUE	Αληθώς θετικό αποτέλεσμα T_p	Ψευδώς θετικό αποτέλεσμα F_p
	FALSE	Ψευδώς αρνητικό αποτέλσμα F_n	Αληθώς αρνητικό αποτέλεσμα T_n

Πίνακας 3.3

Η βαθμολογία – F, είναι μια μετρική που εκφράζει την ακρίβεια μιας δοκιμής. Λαμβάνει υπόψην τόσο την ακρίβεια όσο και την ανάκληση για τον υπολογισμό της τιμής της, ώστε να συνδυάσει αυτές τις δύο μετρικές. Μπορεί να ερμηνευθεί ακόμη ως ο αρμονικός μέσος των δύο προηγούμενων μετρικών και η βέλτιστη τιμή της βαθμολογίας – F είναι 1. Η μετρική αυτή τείνει να βρίσκεται κοντύτερα σε τιμή στη μετρική με την μικρότερη τιμή εκ των δύο.

Οι μαθηματικές σχέσεις που εκφράζουν τις τρεις αυτές μετρικές είναι οι εξής:

$$\text{Ακρίβεια (precision)} = \frac{T_p}{T_p + F_p}$$

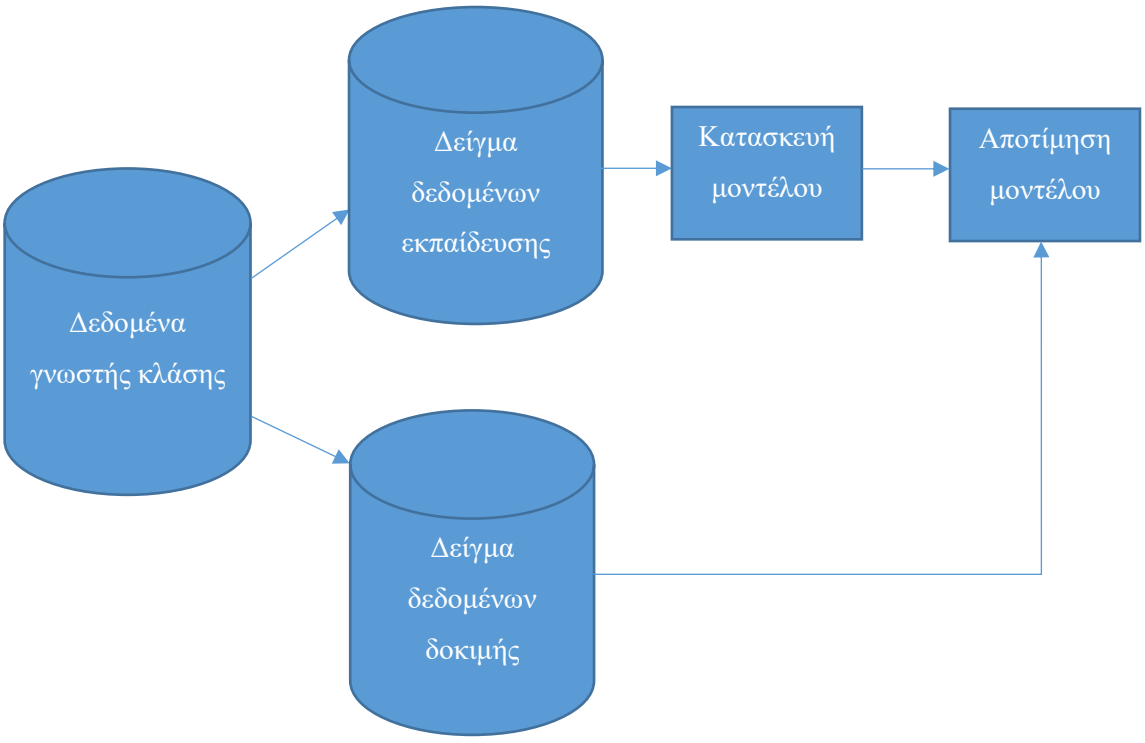
$$\text{Ανάκληση } (recall) = \frac{T_p}{T_p + F_n}$$

$$Βαθμολογία - F\ (F - score) = 2 \cdot \frac{1}{\frac{1}{ακρίβεια} + \frac{1}{ανάκληση}} =$$

$$= 2 \cdot \frac{\alpha\kappa\rho\acute{\iota}\beta\epsilon\iota\alpha \cdot \alpha\nu\acute{\alpha}\kappa\lambda\eta\sigma\eta}{\alpha\kappa\rho\acute{\iota}\beta\epsilon\iota\alpha + \alpha\nu\acute{\alpha}\kappa\lambda\eta\sigma\eta} = 2 \cdot \frac{T_p}{2 \cdot T_p + F_n + F_p}$$

3.1.3 Επιλογή κατάλληλου ταξινομητή

Το Σχήμα 3.1 παρουσιάζει το συνολικό υποσύστημα του ταξινομητή υπό τη σκοπιά των δεδομένων που χρησιμοποιούνται για την εκπαίδευσή του και την αξιολόγηση της απόδοσης του.



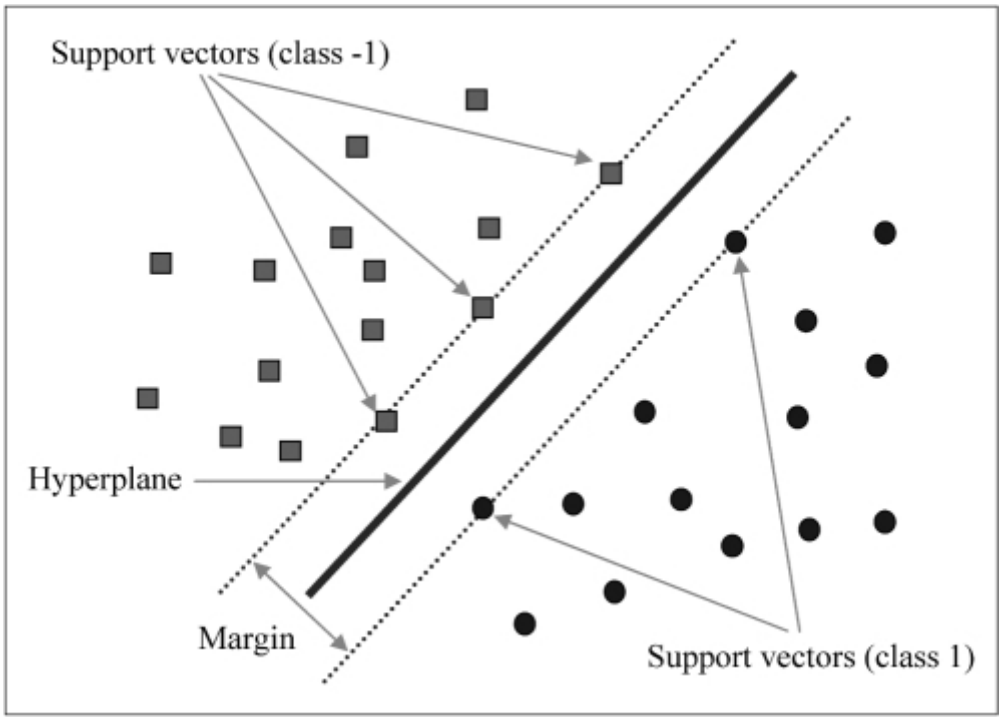
Σχήμα 3.1

Στο [12] οι συγγραφείς αξιοποιούν έναν ταξινομητή SVM με γραμμικό πυρήνα για τις ανάγκες της ανίχνευσης σεισμών και τυφώνων στην Ιαπωνία. Στα πλαίσια της παρούσας υλοποίησης εξετάσαμε τρεις διαφορετικούς ταξινομητές συμπεριλαμβανομένου του ταξινομητή SVM με γραμμικό πυρήνα όπως επίσης SVM με πυρήνα ακτινωτής συνάρτησης βάσης (RBF) και Multinomial Naive Bayes.

Στο τελικό υλοποιηθέν σύστημα αξιοποιήθηκε ένας ταξινομητής βασισμένος σε **Μηχανή Διανυσμάτων Υποστήριξης με πυρήνα ακτινωτής συνάρτησης βάσης (RBF kernel SVM)** καθώς έπειτα από σύγκριση με τους δύο άλλους ταξινομητές βρέθηκε με βάση τις μετρικές αξιολόγησης ότι συμπεριφέρεται καλύτερα ως προς την ταξινόμηση δεδομένων στο διαθέσιμο δείγμα δεδομένων.

Οι ταξινομητές SVM βασίζονται στην εύρεση ενός υπερεπιπέδου απόφασης, το οποίο διακρίνει το χώρο στον οποίο βρίσκεται το σύνολο των δεδομένων που δίνονται ως παραδείγματα (δείγμα δεδομένων εκπαίδευσης) με τέτοιο τρόπο ώστε τα δεδομένα ίδιας κλάσης να βρίσκονται στην ίδια πλευρά του υποχώρου. Τα δυνατά σύνορα – διαχωριστικά υπερεπίπεδα απόφασης είναι περισσότερα του ενός και επομένως μόνο ένα εξ αυτών επιλέγεται ως το βέλτιστο [37]. Αυτό είναι το υπερεπίπεδο απόφασης μέγιστου περιθωρίου (maximum margin separating hyperplane), ήτοι το υπερεπίπεδο εκείνο του οποίου η απόσταση από το κοντινότερο παράδειγμα είναι η μέγιστη δυνατή.

Στο Σχήμα 3.2 έχουν τοποθετηθεί τα δεδομένα εκπαίδευσης ενός γραμμικού ταξινομητή SVM στο δισδιάστατο χώρο με βάση δύο χαρακτηριστικά εκπαίδευσης x_1 και x_2 και έχει σχεδιαστεί το υπερεπίπεδο απόφασης με διακεκομμένες γραμές. Τα ζεύγη δεδομένων (x_1, x_2) που βρίσκονται κοντύτερα στο υπερεπίπεδο απόφασης ονομάζονται διανύσματα υποστήριξης (support vectors) και είναι τα δυσκολότερα σημεία προς ταξινόμηση. Ως γενικός κανόνας, όσο μεγαλύτερο είναι το περιθώριο (margin) τόσο μικρότερο είναι το σφάλμα γενίκευσης (generalization error) του ταξινομητή, δηλαδή το σφάλμα πρόβλεψης για νέα δεδομένα τα οποία δεν έχει συναντήσει προηγουμένως κατά την εκπαίδευσή του. Λόγω του πεπερασμένου του δείγματος δεδομένων εκπαίδευσης, είναι πρόδηλο ότι η επιλογή του δείγματος επηρεάζει την απόδοση του ταξινομητή και την ικανότητα πρόβλεψής του σε νέα δεδομένα.



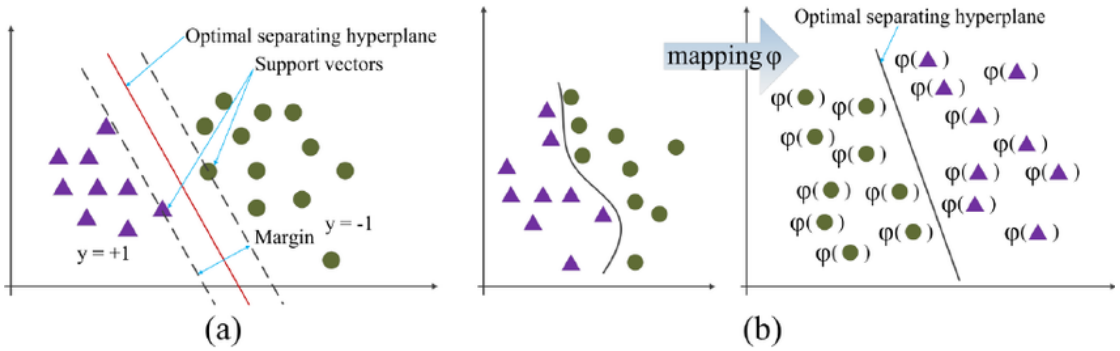
Σχήμα 3.2

Στο Σχήμα 3.2 παραπάνω, είναι φανερό ότι οι δύο κλάσεις είναι γραμμικώς διαχωρίσιμες. Παρακάτω ωστόσο (περίπτωση b στο Σχήμα 3.3) φαίνεται ένα παράδειγμα δύο κλάσεων που δεν μπορούν να διαχωριστούν με γραμμικό υπερεπίπεδο απόφασης. Για το λόγο αυτό χρησιμοποιούνται μέθοδοι πυρήνα (kernel functions) που στηρίζονται στο μετασχηματισμό ενός χώρου προτύπων σε γραμμικό χώρο μεγαλύτερης διάστασης.

3.1.4 Μέθοδοι πυρήνα SVM

Οι μέθοδοι πυρήνα στηρίζονται στον προσδιορισμό ενός θετικά ορισμένου πυρήνα στο χώρο προτύπων, η χρήση του οποίου οδηγεί σε μετασχηματισμό από το χώρο αυτό σε έναν μεγαλύτερης διάστασης χώρο συνήθως που είναι γραμμικός. Στον τελευταίο χώρο η υλοποίηση των πυρήνων ζευγών προτύπων επιτυγχάνεται μέσω του εσωτερικού γινομένου. Οι μέθοδοι πυρήνα είναι υπολογιστικά πολύ αποτελεσματικές. Τα βήματα της μεθόδου είναι τα εξής: αρχικά τα δεδομένα αντιστοιχίζονται σε νέο διανυσματικό χώρο που ονομάζεται χώρος χαρακτηριστικών (feature space), έπειτα αναζητούνται γραμμικές σχέσεις των αντιστοιχισμένων δεδομένων στο νέο χώρο όπου οι υπολογισμοί γίνονται χωρίς τη χρήση συντεταγμένων των δεδομένων, αλλά με χρήση εσωτερικών γινομένων σε ζεύγη χρησιμοποιώντας τα αρχικά δεδομένα και τη συνάρτηση πυρήνα. Μια συνάρτηση πυρήνα (kernel function) μαζί με μια μηχανή πυρήνα (kernel machine) αποτελούν μια μέθοδο πυρήνα.

Στην περίπτωση (b) στο Σχήμα 3.3 φαίνεται μια περίπτωση για την οποία δεν μπορεί να βρεθεί βέλτιστο γραμμικό υπερεπίπεδο απόφασης και ως εκ τούτου πρώτα πραγματοποιείται η αντιστοίχιση μέσω του μετασχηματισμού φ και ύστερα βρίσκεται γραμμικό υπερεπίπεδο στο νέο χώρο όπως έχει εξηγηθεί.



Σχήμα 3.3

Για κάθε πυρήνα υπάρχει και μια συσχετιζόμενη αντιστοίχιση χαρακτηριστικών φ. Η αντιστοίχιση

$$\varphi: \boldsymbol{x} \in \mathbb{R}^n \rightarrow \varphi(\boldsymbol{x}) \in S \subseteq \mathbb{R}^n$$

μεταφέρει τα δεδομένα από τον αρχικό χώρο στον χώρο των χαρακτηριστικών, όπου το εσωτερικό γινόμενο αποτελεί ένα μέτρο ομοιότητας και είναι πιο εύκολο να υπολογιστεί από την ίδια την τιμή της συνάρτησης φ. Η συνάρτηση πυρήνα k ορίζεται ως εξής:

$$k(\boldsymbol{x}, \boldsymbol{x}') = < \varphi(\boldsymbol{x}), \varphi(\boldsymbol{x}') > = \varphi^T(\boldsymbol{x}) \cdot \varphi(\boldsymbol{x}')$$

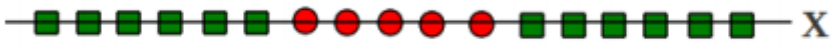
Οι πιο δημοφιλείς συναρτήσεις πυρήνα είναι οι ακόλουθες:

- Γραμμική, $k(\boldsymbol{x}, \boldsymbol{x}') = < \boldsymbol{x}, \boldsymbol{x}' >$. Στην περίπτωση αυτή η συνάρτηση φ είναι η ταυτοτική.
- Πολυωνυμική βαθμού d, $k(\boldsymbol{x}, \boldsymbol{x}') = (< \boldsymbol{x}, \boldsymbol{x}' > +1)^d$ ή $k(\boldsymbol{x}, \boldsymbol{x}') = (< \boldsymbol{x}, \boldsymbol{x}' >)^d$
- Ακτινωτής συνάρτησης βάσης (RBF – Radial Basis Function),

$$k(\boldsymbol{x}, \boldsymbol{x}') = e^{-\gamma \|x - x'\|^2},$$

όπου $\|\boldsymbol{x} - \boldsymbol{x}'\|$ είναι η ευκλείδεια απόσταση μεταξύ των δύο διανυσμάτων (feature vectors) $\boldsymbol{x}$ και $\boldsymbol{x}'$ στον $\mathbb{R}^n$ *και* $<\cdot,\cdot>$ η πράξη του εσωτερικού γινομένου στον ίδιο χώρο.
Για να φανεί ο τρόπος λειτουργίας των παραπάνω παρουσιάζουμε το εξής παράδειγμα.

Έστω το πρόβλημα δυαδικής ταξινόμησης στο Σχήμα 3.4 (το παράδειγμα προέρχεται από το [38]):

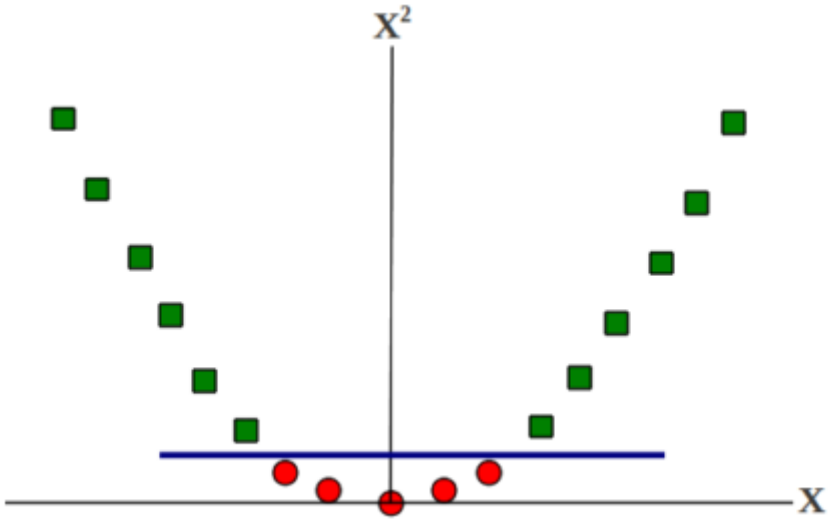


Σχήμα 3.4

στο οποίο τα δεδομένα αναπαριστώνται από ένα μόνο χαρακτηριστικό $\boldsymbol{x}$. Είναι εμφανές ότι δεν υπάρχει δυνατότητα κατασκευής γραμμικού υπερεπιπέδου απόφασης για τα δεδομένα αυτά. Ας θεωρήσουμε τώρα την εξής αντιστοίχιση

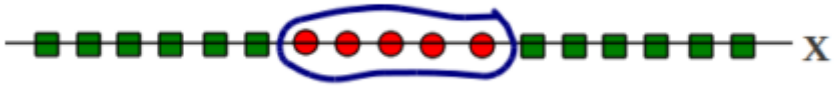
$$\varphi : \boldsymbol{x} \rightarrow \{\boldsymbol{x}, \boldsymbol{x^2}\}$$

όπου πλέον τα δεδομένα αναπαριστώνται από δύο χαρακτηριστικά που προκύπτουν απευθείας από την αρχική αναπαράσταση.



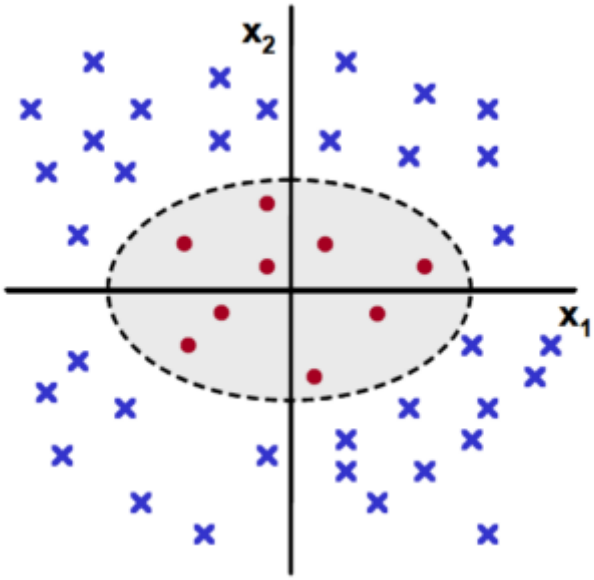
Σχήμα 3.5

Στην αναπαράσταση αυτή που παρουσιάζεται στο Σχήμα 3.5 φαίνεται ότι μπορεί να βρεθεί γραμμικό υπερεπίπεδο απόφασης που να διαχωρίζει τις δύο κλάσεις. Η γραμμικά διαχωρίσιμη αναπαράσταση στο Σχήμα 3.5 μετά την αντιστοίχιση είναι ισοδύναμη με την μη γραμμικά διαχωρίσιμη αρχική αναπαράσταση των δεδομένων στο Σχήμα 3.6.



Σχήμα 3.6

Ένα δεύτερο παράδειγμα ακολουθεί στο Σχήμα 3.7:

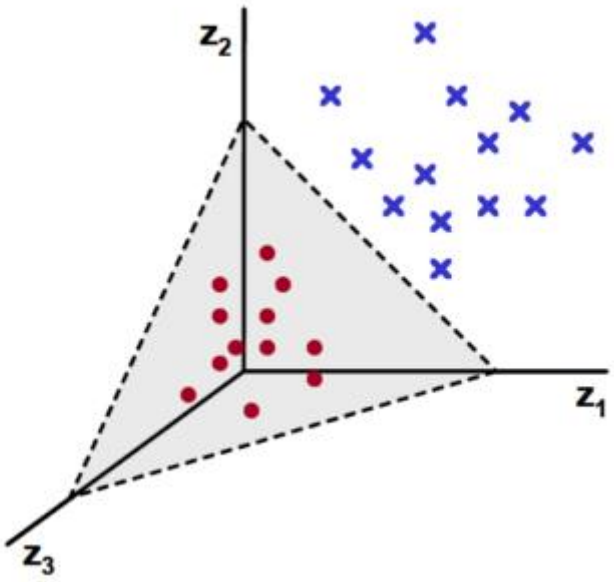


Σχήμα 3.7

Στην περίπτωση αυτή, κάθε δεδομένο έχει δύο χαρακτηριστικά, δηλαδή $\boldsymbol{x} = \{x_1, x_2\}$ και όπως φαίνεται δεν υπάρχει δυνατότητα γραμμικού διαχωρισμού των δύο κλάσεων. Μετά την αντιστοίχιση

$$\varphi : \boldsymbol{x} = \{x_1, x_2\} \rightarrow \boldsymbol{z} = \{x_1^2, \sqrt{2}x_1x_2 \;\; x_2^2\}$$

τα δεδομένα έχουν πλέον τρία χαρακτηριστικά και υπάρχει γραμμικό υπερεπίπεδο απόφασης (Σχήμα 3.8).



Σχήμα 3.8

3.2 Χρήση υποσυστήματος ειδοποίησης

Αντικείμενο του υποσυστήματος ειδοποίησης είναι η παρακολούθηση των tweets που κατατάσσονται στη θετική κλάση από τον ταξινομητή σε κατάλληλα χρονικά παράθυρα και η έγκαιρη ειδοποίηση του συστήματος για την πραγματοποίηση ενός σεισμού όταν ικανοποιούνται κάποιες προκαθορισμένες στατιστικές συνθήκες.

Παρόμοια με το [12], αναπτύχθηκε ένα στατιστικό μοντέλο το οποίο προσομοιώνει τη συμπεριφορά των χρηστών του Twitter. Το μοντέλο που επιλέχθηκε είναι αυτό της διαδικασίας Poisson, λόγω της ικανότητας μοντελοποίησης ανεξάρτητων συμβάντων που λαμβάνουν χώρα διαδοχικά και τα οποία με τη σειρά τους ακολουθούν μια κατανομή. Έτσι, κατασκευάζεται ένα πρότυπο που εξυπηρετεί την ανάγκη μοντελοποίησης της χρονολογικής σειράς εμφάνισης tweets που αφορούν σεισμούς.

3.2.1 Διαδικασία Poisson

Στην παρούσα υποενότητα παρουσιάζεται το θεωρητικό υπόβαθρο της διαδικασίας Poisson που χρησιμοποιείται στο μοντέλο. Αρχικά, ορίζουμε μια στοχαστική διαδικασία Poisson συνεχούς χρόνου $\{N(t), t \geq 0\}$. Αυτή, είναι μια συλλογή από τυχαίες μεταβλητές, δηλαδή για κάθε χρονική στιγμή $t \geq 0$, $N(t)$ είναι μια τυχαία μεταβλητή που αποτελεί την κατάσταση της διαδικασίας τη στιγμή t. Η Poisson είναι μια διαδικασία στην οποία γεγονότα συμβαίνουν συνεχώς και ανεξάρτητα και μοντελοποιεί τον αριθμό των συμβάντων που έχουν πραγματοποιηθεί ως τη χρονική στιγμή t.

Επομένως, σε οποιοδήποτε χρονικό διάστημα $(\Delta t, \Delta t + t], \Delta t > 0 \; και \; t \geq 0$ ο αριθμός των συμβάντων θα είναι $N(\Delta t + t) - N(\Delta t)$ και αποδεικνύεται ότι ο αριθμός αυτός ακολουθεί κατανομή Poisson παραμέτρου $\lambda \cdot t$, άρα η πιθανότητα ο αριθμός των συμβάντων που έχουν λάβει χώρα εντός του $(\Delta t, \Delta t + t]$ να είναι ίσος με n είναι:

$$P[N(\Delta t + t) - N(\Delta t) = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \quad , n \in \mathbb{N}$$

όπου $\lambda > 0$ είναι ο ρυθμός της διαδικασίας Poisson και $\lambda \cdot t$ είναι η έντασή της, η οποία εκφράζει το μέσο αριθμό εμφανίσεων στο χρόνο, δηλαδή $E[N(t)] = \lambda \cdot t$.

Έστω ότι για μια τέτοια διαδικασία Poisson ο χρόνος πραγματοποίησης ή άφιξης του πρώτου συμβάντος είναι T_1. Τότε για το n-οστό συμβάν με $n > 1$ ισχύει ότι ο χρόνος που μεσολαβεί από την πραγματοποίηση του συμβάντος υπ' αριθμόν $n - 1$ ως την πραγματοποίηση του $n - οστού$ συμβάντος είναι T_n. Για το λόγο αυτό η ακολουθία

$$\{T_n, n = 1,2,3, \dots\}$$

ονομάζεται ακολουθία των ενδιάμεσων χρόνων.

Επί παραδείγματι έστω $T_1 = 10s$ και $T_2 = 35s$, τότε το πρώτο συμβάν συνέβη 5 δευτερόλεπτα από την αρχή μέτρησης ενώ το δεύτερο 45 δευτερόλεπτα από την αρχή μέτρησης του χρόνου, δηλαδή 35 δευτερόλεπτα αργότερα από το πρώτο συμβάν.

Για να προσδιορίσουμε την πιθανότητα κανένα γεγονός να μην έχει συμβεί ως το χρόνο t ή αλλιώς στο διάστημα $(0, t]$ αρκεί να υπολογίσουμε την πιθανότητα:

$$P\{T_1 > t\} = P\{N(t) = 0\} = P\{N(0 + t) - N(0) = 0\} = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = \; e^{-\lambda t}$$

όπου έγινε η παραδοχή ότι κανένα συμβάν δε λαμβάνει χώρα τη στιγμή $t = 0$, και άρα $N(0) = 0$. Συνεπώς, προκύπτει:

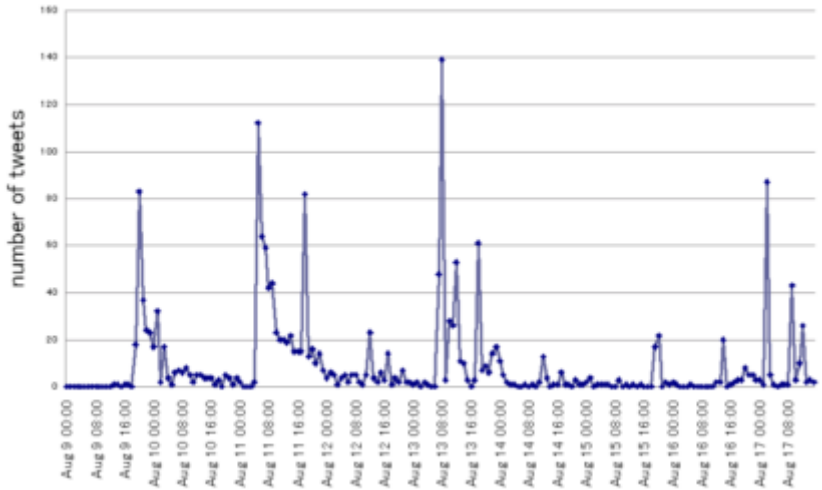
$$P\{T_1 \leq t\} = 1 - P\{T_1 > t\} = 1 - e^{-\lambda t}$$

Ως εκ τούτου η τυχαία μεταβλητή T_1 ακολουθεί εκθετική κατανομή μέσης τιμής $E[T_1] = \frac{1}{\lambda}$. Αποδεικνύεται ότι όλες οι τυχαίες μεταβλητές $T_n, n = 1,2,3, \ldots$ είναι ανεξάρτητες τυχαίες μεταβλητές μέσης τιμής $E[T_n] = \frac{1}{\lambda}$ που ακολουθούν εκθετική κατανομή.

3.2.2 Μοντελοποίηση αναφορών tweets

Αν υποτεθεί ότι ένας χρήστης του μέσου Twitter αντιλαμβάνεται την πραγματοποίηση ενός σεισμού τη χρονική στιγμή 0 και η πιθανότητα να δημοσιεύσει αναφορά για το συμβάν αυτό εντός του χρονικού διαστήματος $[t, \Delta t]$ είναι σταθερή και ίση με λ, τότε με βάση τα προηγούμενα ο χρόνος δημοσίευσης είναι τυχαία μεταβλητή που ακολουθεί εκθετική κατανομή. Αυτό πρακτικά ποσοτικοποιεί την πεποίθηση ότι οι περισσότεροι χρήστες είναι πιο πιθανό να δημοσιεύσουν την αναφορά ενός σημαντικού γεγονότος όπως έναν σεισμό όσο το δυνατόν πιο σύντομα από την πραγματοποίησή του, παρολαυτά διαφορετικοί χρήστες απαιτούν διαφορετικό χρόνο για τη δημοσίευση της αναφοράς τους εξαιτίας ποικίλων παραγόντων.

Ακόμη, με βάση τη μορφή του αριθμού των tweets που δημοσιεύονται ως προς το χρόνο αναφορικά με γεγονότα σεισμών που μελέτησαν οι συγγραφείς του [12], η οποία φαίνεται στο Σχήμα 3.9, παρατήρησαν οι ίδιοι ότι ταιριάζουν τα δεδομένα αρκετά καλά με εκθετική κατανομή που έχει $\lambda = 0.34$ κατά μέσο όρο.



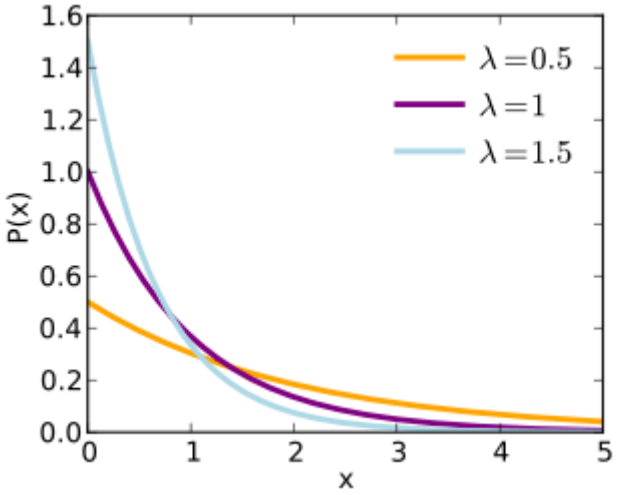
Σχήμα 3.9

Στο παραπάνω οι κορυφές αντιστοιχούν σε απότομη αύξηση των tweets και άρα αναφορά σεισμών με βάση τα tweets αυτά.

Η εκθετική κατανομή έχει συνάρτηση πυκνότητας πιθανότητας

$$f(t;\lambda) = \lambda \cdot e^{-\lambda t} \ , \lambda, t > 0$$

και η μορφή της φαίνεται στο Σχήμα 3.10 για διάφορες τιμές του λ.



Σχήμα 3.10

Με βάση και την προηγούμενη ενότητα, μοντελοποιούμε την εμφάνιση tweets σχετικών με τα γεγονότα τα οποία στοχεύει το σύστημα ως διαδικασία Poisson ανεξάρτητων χρονικών ακολουθιών T_n, όπου οι κατανομές των tweets αυτών εντός των χρονικών πλαισίων T_n είναι εκθετικές και ανεξάρτητες μεταξύ τους.

Ακόμη, θεωρούμε ότι κάθε αναφορά ενός χρήστη έχει μια συγκεκριμένη αξιοπιστία, καθώς είτε μπορεί να πρόκειται για αναφορά σε γεγονός το οποίο δεν υφίσταται πραγματικά είτε μπορεί να ταξινομηθεί στη λάθος κλάση από τον ταξινομητή του συστήματος. Για το λόγο αυτό δημιουργήθηκε ένα μέτρο αξιοπιστίας με τη μορφή πιθανότητας, όπου θεωρήθηκε ότι η πιθανότητα αναφοράς ενός χρήστη σε γεγονός σεισμού ενώ δε συμβαίνει σεισμός πραγματικά είναι ίση με P_f (false positive), τιμή που επιλέγεται με βάση την τιμή της ακρίβειας που επιτυγχάνει ο χρησιμοποιούμενος ταξινομητής και οι χρήστες που κάνουν αναφορές στο Twitter είναι ανεξάρτητοι μεταξύ τους και ομοιόμορφα κατανεμημένοι. Τότε, υποθέτοντας ακόμη ότι υπάρχουν n χρήστες οι οποίοι στα tweets τους αναφέρουν γεγονότα σεισμών, η πιθανότητα όλοι οι χρήστες να αναφέρονται σε σεισμό που δεν υφίσταται στην πράξη είναι P_f^n. Η πιθανότητα αυτή είναι και η πιθανότητα να έχουμε ψευδή ειδοποίηση (false alarm), ήτοι ειδοποίηση από το σύστημα ανίχνευσης σεισμού που όμως δεν υφίσταται πραγματικά εφόσον όλες οι αναφορές ήταν λανθασμένες. Από τα παραπάνω συμπεραίνεται

ότι η πιθανότητα να συμβαίνει ένας σεισμός υπολογίζεται ως $1 - P_f^n$, δηλαδή στην περίπτωση αυτή τουλάχιστον μία αναφορά χρήστη σε σεισμό ανταποκρίνεται σε πραγματικό συμβάν.

Με την υπόθεση ότι υπάρχουν N_0 tweets – άρα και χρήστες – που αναφέρονται σε σεισμό τη στιγμή $t = 0$ και $N_0 \cdot \mathrm{e}^{-\lambda t}$ τη στιγμή t, τότε προκύπτει άμεσα από το γεγονός ότι ο αριθμός των tweets ακολουθεί την εκθετική κατανομή, ότι τη στιγμή t το συνολικό πλήθος των tweets που έχουν δημοσιευτεί στο διαστημα $[0, t]$ είναι:

$$N_{total} = \sum_{\tau=0}^{t} N_0 \cdot \mathrm{e}^{-\lambda \tau} = N_0 \cdot \frac{1 - \mathrm{e}^{-\lambda(t+1)}}{1 - \mathrm{e}^{-\lambda}}$$

Έτσι, έπεται ότι η πιθανότητα ένας σεισμός να συμβαίνει τη χρονική στιγμή t είναι ίση με

$$P_{\sigma\varepsilon\iota\sigma\mu o\acute{\upsilon}}(t) = 1 - P_f^{N_{total}}, \text{ όπου } N_{total} = \sum_{\tau=0}^{t} N_0 \cdot \mathrm{e}^{-\lambda \tau} = N_0 \cdot \frac{1 - \mathrm{e}^{-\lambda(t+1)}}{1 - \mathrm{e}^{-\lambda}}, \text{ άρα}$$

$$P_{\sigma\varepsilon\iota\sigma\mu o\acute{\upsilon}}(t) = 1 - P_f^{N_0 \cdot \frac{1 - \mathrm{e}^{-\lambda(t+1)}}{1 - \mathrm{e}^{-\lambda}}}$$

Για να θεωρήσουμε ότι ένας σεισμός λαμβάνει χώρα με δεδομένες N_{total} αναφορές σε σεισμό στο Twitter που έχουν τοποθετηθεί στη θετική κλάση από τον ταξινομητή σε χρονικό διάστημα μήκους t, απαιτήσαμε η πιθανότητα

$$P_{\sigma\varepsilon\iota\sigma\mu o\acute{\upsilon}}(t) \geq 0.99 \Leftrightarrow 1 - P_f^{N_{total}} \geq 0.99 \Leftrightarrow P_f^{N_{total}} \leq 0.01 \underset{log\uparrow}{\Longleftrightarrow} N_{total} \cdot \log(P_f)$$

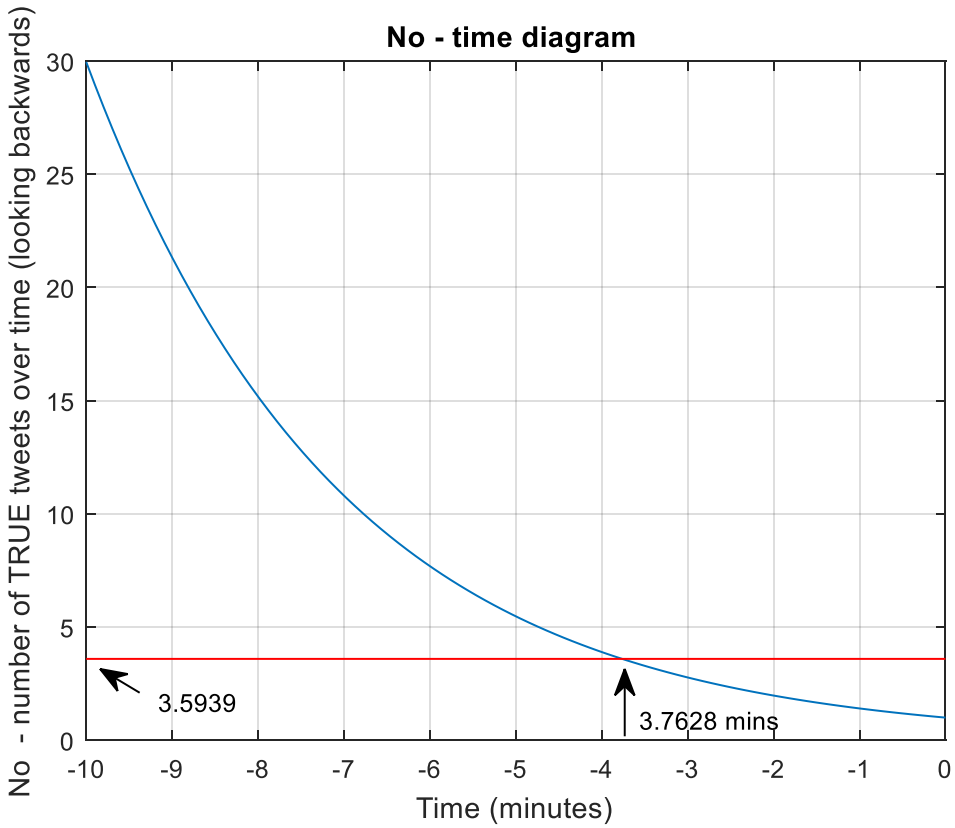
$$\leq \log(0.01) \underset{\log(P_f)<0}{\Longleftrightarrow} \Rightarrow N_{total} \geq \frac{\log(0.01)}{\log(P_f)} = \frac{-2}{\log(P_f)}$$

Από την προηγούμενη σχέση, με γνωστή την πιθανότητα P_f και με δεδομένο ότι η ποσότητα N_{total} είναι φυσικός αριθμός φαίνεται ότι το υποσύστημα ειδοποίησης χρειάζεται να εντοπίσει συνολικά, σε ένα χρονικό διάστημα διάρκειας t, τουλάχιστον $\frac{-2}{\log(P_f)}$ tweets θετικής κλάσης ώστε να είναι αξιόπιστη οποιαδήποτε ειδοποίηση για σεισμό.

3.2.3 Αλγόριθμος ειδοποίησης

Με βάση την παραπάνω θεωρητική θεμελίωση και με δεδομένη την εκθετική κατανομή που ακολουθούν τα tweets τα οποία καταφθάνουν μέσω του Streaming API και ταξινομούνται στη θετική κλάση θα περιγραφεί η τελική υλοποίηση του συστήματος ειδοποίησης.

Αρχικά, έστω ότι σε μια χρονική στιγμή παραλαμβάνει το υποσύστημα ειδοποίησης ένα tweet θετικής κλάσης. Τότε, έχοντας κρατήσει τα προηγούμενα θετικά ταξινομημένα tweets που είχαν καταφθάσει προηγουμένως, αναζητά αναδρομικά ένα χρονικό διάστημα στο οποίο τα δημοσιευμένα tweets παρουσίασαν τη ζητούμενη εκθετική συμπεριφορά όπως επίσης και τη συνθήκη $N_{total} \geq \frac{-2}{\log(P_f)}$ σε όλο το διάστημα αυτό. Στο Σχήμα 3.11 φαίνεται η κατανομή των tweets που αναμένεται να έχουν καταγραφεί έως και 10 λεπτά πριν την παραλαβή του τρέχοντος tweet για $\lambda = 0.34$.



Σχήμα 3.11

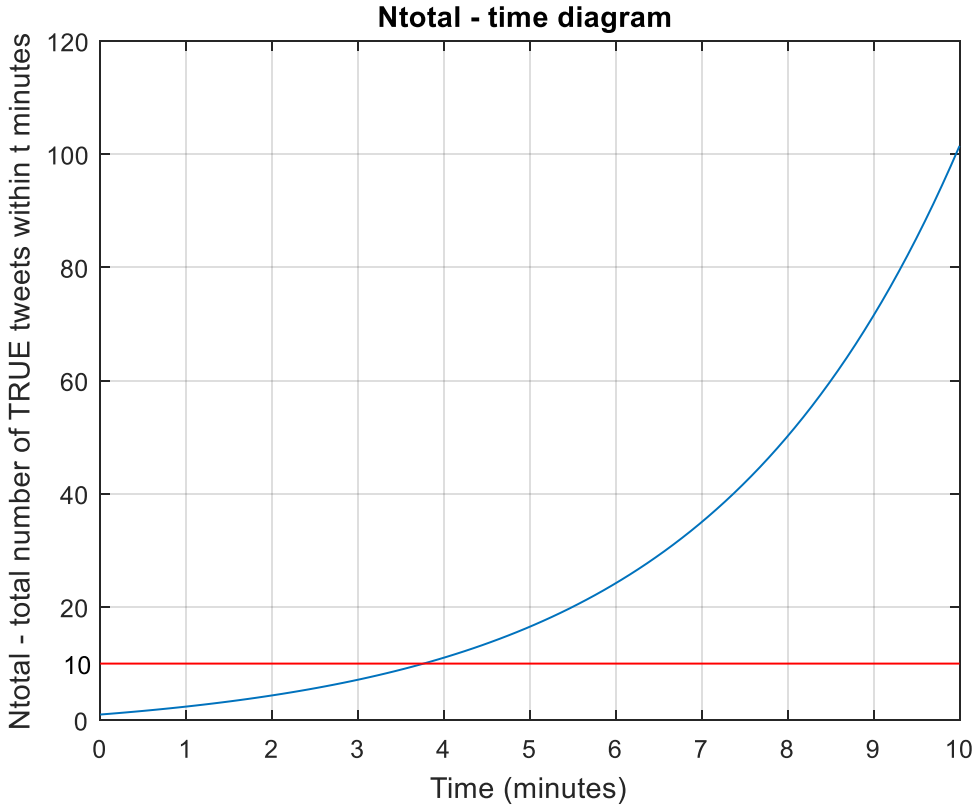
Από τη σχέση $N_{total} \geq \frac{-2}{\log(P_f)}$ προκύπτει:

$$N_{total} \geq \frac{-2}{\log(P_f)} \Leftrightarrow N_0 \cdot \frac{1 - \mathrm{e}^{-\lambda(\mathrm{t}+1)}}{1 - \mathrm{e}^{-\lambda}} \geq \frac{-2}{\log(P_f)}$$

Όμως, εφόσον έχουμε 1 ακριβώς tweet τη στιγμή που εξετάζουμε, τότε t λεπτά πριν την άφιξή του αναμένονται να έχουν καταγραφεί $N_0 = 1 \cdot e^{\lambda t} = e^{\lambda t}$ tweets. Συνεπώς τα συνολικά tweets ως και t λεπτά πριν θα είναι:

$$N_{total}(t) = N_0 \cdot \frac{1 - \mathrm{e}^{-\lambda(\mathrm{t}+1)}}{1 - \mathrm{e}^{-\lambda}} = e^{\lambda t} \cdot \frac{1 - \mathrm{e}^{\lambda(\mathrm{t}+1)}}{1 - \mathrm{e}^{-\lambda}} = \frac{e^{\lambda t} - e^{-\lambda}}{1 - e^{-\lambda}} = \frac{1}{1 - e^{-\lambda}} \cdot \left(e^{\lambda t} - e^{-\lambda}\right)$$

Για παράδειγμα, στο Σχήμα 3.12 φαίνεται η γραφική παράσταση των συνολικών θετικών αναφορών N_{total} στη μονάδα του χρόνου για την περίπτωση όπου $P_f = 0.6$, οπότε ισχύει για το συνολικό αριθμό θετικών αναφορών: $N_{total} \geq \frac{-2}{\log(P_f)} \geq 9.015$, δηλαδή πρακτικά στην περίπτωση αυτή χρειάζονται $N_{total} \geq 10$ για την ειδοποίηση αναφορικά με σεισμό έχοντας υποθέσει $\lambda = 0.34$. Λύνοντας την $N_{total} \geq 10$ προκύπτει ότι $t \geq 3.7628$ λεπτά.



Σχήμα 3.12

Δηλαδή για να είναι αξιόπιστη η αξιολόγηση των tweets που αναφέρονται σε σεισμό πριν οποιαδήποτε ειδοποίηση προκύπτει η απαίτηση της εξέτασης των εισερχόμενων αναφορών σε χρονικό διάστημα τουλάχιστον 4 λεπτών και η εξέταση του αριθμού τους που πρέπει να είναι τουλάχιστον 10 με βάση το παραπάνω παράδειγμα.

Συμπερασματικά, το βήματα του αλγορίθμου ειδοποίησης με βάση τη μοντελοποίηση που προηγήθηκε είναι τα ακόλουθα:

1. Παραλαβή ενός νέου tweet θετικής κλάσης
2. Εξέταση του αριθμού των tweets θετικής κλάσης του προηγούμενου λεπτού ($t = 1\,min$) και αποθήκευση του αριθμού τους $N(t=1) = N(1)$. Ακόμη αρχικοποίηση $N_{total}(t=1) = N(1) + N(0) = N(1) + 1$, ώστε να συμπεριληφθεί το tweet που μόλις καταφθάνει από τον ταξινομητή $N(0) = 1$.
3. Αύξηση του χρόνου t κατά 1 λεπτό και υπολογισμός του συνολικού αριθμού tweets θετικής κλάσης ως και τα t λεπτά

$$N_{total}(t) = \sum_{\tau=0}^{\tau=t} N(\tau)$$

Αν $t \leq 10$ λεπτά και $P_f^{N_{total}(t)} \leq 0.01$, τότε ειδοποίησε το σύστημα για ύπαρξη σεισμού και πήγαινε στο βήμα 1, αλλιώς επανάλαβε το βήμα 3.

Στην πράξη, με βάση τις εμπειρικές παρατηρήσεις μας στα δείγματα δεδομένων που λάβαμε μέσω του Twitter και του κανόνα 10 tweets εντός παραθύρων 10 λεπτών προκύπτει ότι ικανοποείται η συνθήκη $P_{σεισμού} = 1 - P_f^n \geq 0.99$.

3.2.4 Χρησιμότητα συστήματος ειδοποίησης

Το υλοποιηθέν σύστημα ειδοποίησης εκμεταλλεύεται τη στατιστική συμπεριφορά των αναφορών σεισμών στο Twitter για την αναγνώριση ενός συμβάντος. Ο τρόπος με τον οποίο κατασκευάστηκε εγγυάται τόσο την εξασφάλιση υψηλής πιθανότητας ύπαρξης σεισμού στην πραγματικότητα όσο και την ικανότητα αναζήτησης σεισμών σε χρονικά παράθυρα που δεν έχουν συγκεκριμένο μήκος, αλλά μπορεί να είναι οποιουδήποτε μήκους έως και 10 λεπτών με βάση τον αριθμό των αναφορών που λαμβάνονται. Τα 10 λεπτά τίθενται ως άνω όριο για την αναζήτηση καθώς σκοπός του συστήματος είναι η έγκαιρη αναγνώριση ενός σεισμού.
Η μέθοδος καθιστά σαφές ότι απαιτείται ικανός αριθμός αναφορών για την ανίχνευση ενός σεισμού με βεβαιότητα καθώς ένα ή δύο μόλις tweets δεν μπορούν να εγγυηθούν την ύπαρξη

ενός συμβάντος. Η αλληλουχία συμβάντων σεισμών που αντιλαμβάνεται το σύστημα ειδοποίησης με βάση τις κατανομές των tweets σε συνεχόμενα χρονικά διαστήματα είναι διαδικασίες Poisson με βάση τα προηγούμενα.

3.2.5 Γνωστά προβλήματα υποσυστήματος ειδοποίησης

Ένα από τα σημαντικότερα προβλήματα από τα οποία πάσχει η υλοποίηση του ειδοποιητήριου υποσυστήματος είναι αυτό της πολλαπλής αναφοράς (multiple reporting), όπου μετά την ανίχνευση ενός σεισμού και την αντίστοιχη ειδοποίηση του ευρύτερου συστήματος ανίχνευσης και αναφοράς το σύστημα αντιλαμβάνεται και πάλι τον ίδιο σεισμό από αναφορές στο μέσο δικτύωσης που είναι χρονικά κοντά με την προηγούμενη ειδοποίηση με αποτέλεσμα το σύστημα να αναφέρει έναν σεισμό πάνω από μία φορές εντός ενός χρονικού διαστήματος.

3.3 Χρήση υποσυστήματος εντοπισμού τοποθεσίας συμβάντος

Αφότου ένας σεισμός ανιχνευθεί από το σύστημα ειδοποίησης, σημαντικό βήμα αποτελεί ο προσδιορισμός της τοποθεσίας του συμβάντος, έτσι ώστε ο χρήστης της εφαρμογής να μπορεί να ενημερωθεί κατάλληλα.

3.3.1 Τοποθεσία συμβάντος στα tweets

Εξετάζοντας το δείγμα δεδομένων των 10,996 tweets τα οποία συγκεντρώθηκαν την περίοδο 6 Ιουνίου 2017 έως 17 Ιουλίου 2017, παρατηρήθηκε πως το 66% εξ αυτών περιέχουν τοποθεσία, όπως αυτή επιστρέφεται στο πεδίο user => location από το API, ενώ μόλις 1.18% εξ αυτών περιέχουν μη κενό πεδίο place => name. Τα πεδία αυτά φαίνονται στην Εικόνα 4.1. Η τοποθεσία του πεδίου place είναι συγκεκριμένη τοποθεσία που αντιστοιχεί σε γεωγραφικές συντεταγμένες. Μπορεί να αντιστοιχισθεί από το χρήστη κατά τη δημοσίευση ενός tweet και δε εσχετίζεται κατ' ανάγκη με το μέρος από το οποίο δημοσιεύεται αλλά μπορεί να αφορά μια τοποθεσία ενδιαφέροντος. Το πεδίο place => name, είναι μια σύντομη αναπαράσταση της τοποθεσίας σε μορφή κειμένου που είναι κατανοητή από τον άνθρωπο. Τέτοια παραδείγματα είναι τα ακόλουθα που έχουν προκύψει από το προαναφερθέν δείγμα δεδομένων: Kos, Ελλάς, Attica, Αγρίνιο, Greece, Μυτιλήνη, Αττική, Θεσσαλονίκη. Παρόλαυτα, κρίνεται ακατάλληλο το πεδίο αυτό για την περιγραφή ενός γεγονότος αφού μόλις 1.18% αυτών

περιέχουν μη κενό αυτό το πεδίο. Ανάλογη παρατήρηση έγινε από τους συγγραφείς του [27], οι οποίοι αναφέρουν πως μόλις το 0.7% των tweets περιέχουν γεωγραφικό στίγμα.

Επιπλέον, η τοποθεσία στο πεδίο user => location ορίζεται ως μια συμβολοσειρά από το χρήστη και αφορά το προφίλ του, ενώ δεν είναι απαραίτητα κάποια πραγματική τοποθεσία. Για παράδειγμα παραθέτουμε 10 παραδείγματα τιμών από το προαναφερθέν δείγμα στον Πίνακας 3.4:

Τοποθεσία Χρήστη (user => location)	Πραγματική τοποθεσία
ΒΟΡΕΙΟΣ ΚΟΡΕΑ	
Parga-Greece	✓
δίπλα από το σπίτι του Γιώργου	
Αθήνα, Ελλάδα	✓
Στο Δικο Μου Κοσμο	
Schwäbisch Gmünd, Deutschland	✓
θα δειξει...	
GLORIOUS GREECE	
Ιωάννινα	✓
Winterfell	

Πίνακας 3.4

Είναι φανερό από τα παραπάνω παραδείγματα ότι μικρό μέρος των χρηστών καταχωρούν πραγματική τοποθεσία στο προφίλ τους. Σε ποσοστό 33.6% , με βάση το συνολικό δείγμα, οι χρήστες προτιμούν να μην δηλώνουν τοποθεσία, ενώ από το υπόλοιπο 66,4% το μεγαλύτερο μέρος δηλώνει μη έγκυρη τοποθεσία. Έτσι, έπειτα από παρατήρηση της μορφής των tweets αναφορικά με σεισμούς, επιλέχθηκε η υλοποίηση υποσυστήματος αναγνώρισης τοποθεσίας να βασίζεται στο κυρίως κείμενο των tweets. Ο Πίνακας 3.2 περιέχει το ακόλουθο tweet, το οποίο όπως φαίνεται εμπεριέχει την τοποθεσία του σεισμού αναφοράς: «Σεισμός 6,1 Ρίχτερ ανοιχτά της Μυτιλήνης #seismos #σεισμος via @giatigr». Σκοπός, λοιπόν, του παρόντος

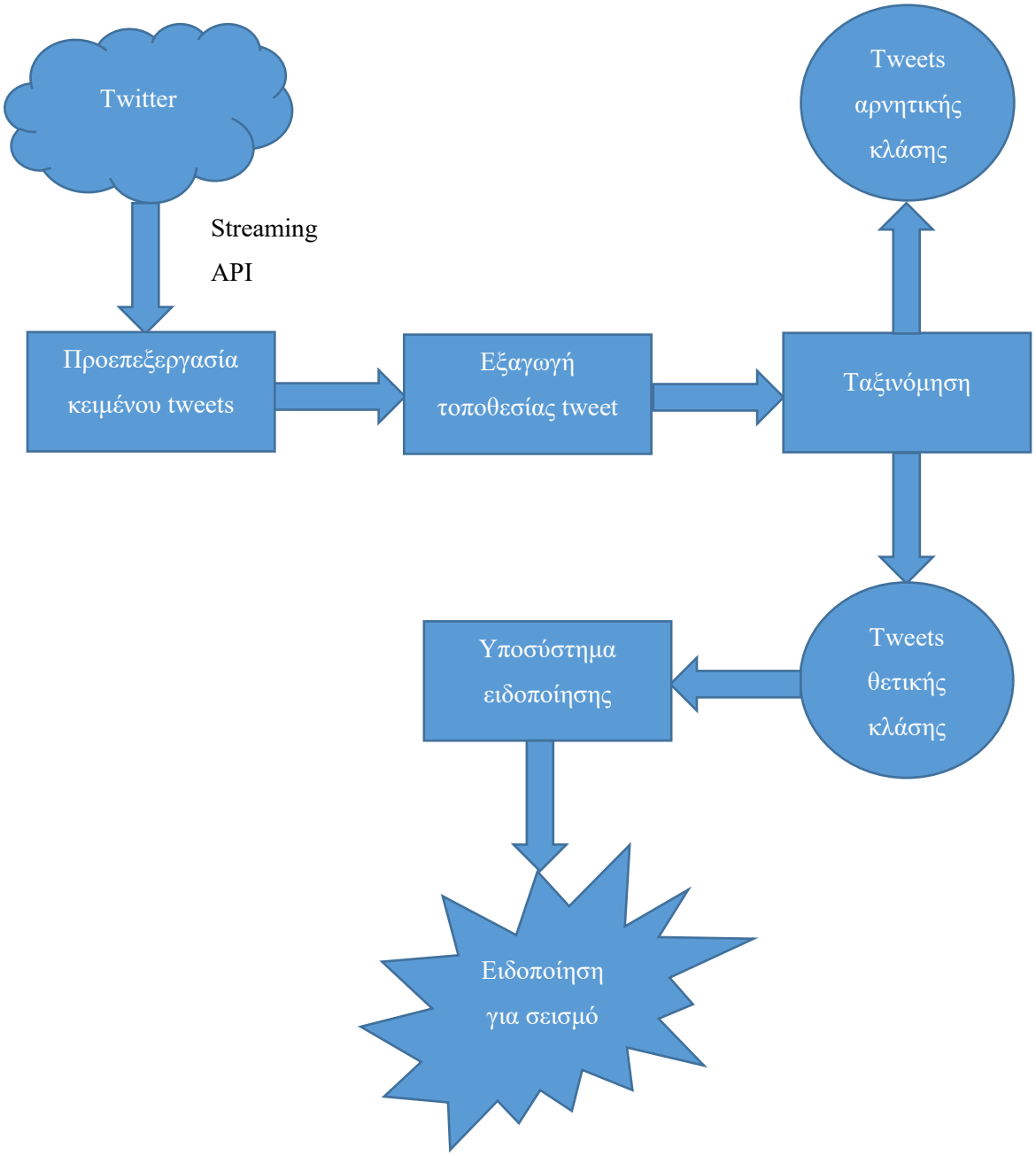
υποσυστήματος είναι η εξαγωγή χρήσιμης πληροφορίας όσον αφορά την τοποθεσία στην οποία λαμβάνει χώρα ένας σεισμός από το κείμενο των tweets που λαμβάνονται.

3.3.2 Υλοποίηση υποσυστήματος εντοπισμού τοποθεσίας

Η βιβλιοθήκη «Geotext» [39] για τη γλώσσα Python έχει ως αντικείμενο την εξαγωγή χωρών και πόλεων από ελεύθερο κείμενο. Στην πράξη αντλεί δεδομένα από τη γεωγραφική βάση δεδομένων GeoNames [40],η οποία περιέχει περισσότερα από 10 εκατομμύρια καταχωρήσεις και ελέγχει πιθανές τοποθεσίες που υπάρχουν στο κείμενο. Η υλοποιούμενη ιδέα είναι ανάλογη. Χρησιμοποιήσαμε τη βάση τοποθεσιών [41], η οποία περιέχει τους νομούς της Ελλάδος και τις περιοχές κάθε νομού και δημιουργήσαμε ένα λεξικό (dictionary) στο οποίο αναζητείται – αν υπάρχει – η τοποθεσία αναφοράς κάθε tweet θετικής κλάσης.

Κατά την ειδοποίηση του συστήματος, υπολογίζονται οι κορυφαίες τοποθεσίες αναφοράς από το πλήθος των tweets που τις περιέχουν και εμφανίζονται χωρισμένες με κόμμα και σε σειρά από την πιο δημοφιλή στη λιγότερο. Επί παραδείγματι, ακολουθεί μια αναφορά του συστήματος που περιέχει τοποθεσία: «Earthquake Now in Μυτιληνη, Λεσβος, Χιος».

Στο Σχήμα 3.13 ακολουθεί ένα γενικό σχήμα της συνολικής υλοποίησης του συστήματος όπου φαίνεται και η ροή πληροφορίας μεταξύ των διαφορετικών υποσυστημάτων.


Twitter
Streaming
API
Προεπεξεργασία
κειμένου tweets
Εξαγωγή
τοποθεσίας tweet
Ταξινόμηση
Tweets
αρνητικής
κλάσης
Tweets
θετικής
κλάσης
Υποσύστημα
ειδοποίησης
Ειδοποίηση
για σεισμό

Σχήμα 3.13

4

Δείγμα δεδομένων και χαρακτηριστικά εκπαίδευσης

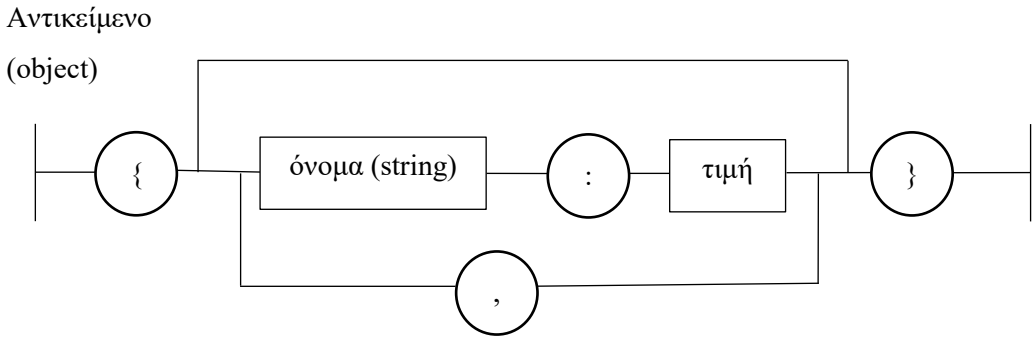
Στο κεφάλαιο αυτό αρχικά αναλύουμε τη μορφή των δεδομένων που λαμβάνονται από το Twitter, τον τρόπο απόκτησής τους μέσω του κατάλληλου API και την επιλογή των χρήσιμων για την παρούσα υλοποίηση δεδομένων και μεταδεδομένων. Στη συνέχεια, σχολιάζεται η δημιουργία κατάλληλου δείγματος δεδομένων εκπαίδευσης (training dataset) για τον ταξινομητή, η τοποθέτηση ετικετών (labelling) χειροκίνητα στις χρησιμοποιούμενες κλάσεις ώστε να πραγματοποιηθεί η επιτηρούμενη εκπαίδευσή του, καθώς επίσης περιγράφονται οι διαφορετικές ομάδες χαρακτηριστικών του ταξινομητή που δημιουργήθηκαν και δοκιμάστηκαν κατά τον πειραματισμό. Τέλος, περιγράφεται το υλοποιηθέν υποσύστημα προεπεξεργασίας των δεδομένων, το οποίο επεξεργάζεται το κείμενο των tweets πριν αυτά τροφοδοτηθούν στο υποσύστημα ταξινόμηση.

4.1 Μορφή και επιλογή δεδομένων και μεταδεδομένων

Τα δεδομένα τα οποία παρέχονται μέσω των APIs του Twitter και πιο συγκεκριμένα του Streaming API που αξιοποιήθηκε εν προκειμένω δίνονται σε μορφή json. Στα επόμενα αναλύεται η μορφή json και δίνεται μια περιγραφή του τρόπου άντλησης tweets από το API σε αυτή τη μορφή. Τέλος, πραγματοποιείται κατάλληλη επιλογή αυτών ώστε να εξυπηρετούν τους σκοπούς της παρούσας εργασίας.

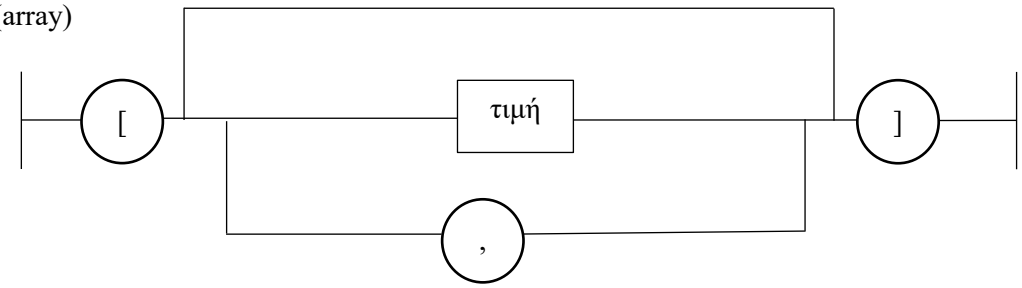
4.1.1 Μορφή json

Με βάση τις προκαθορισμένες λέξεις – κλειδιά της προηγούμενης υποενότητας το Twitter μέσω του API επιστρέφει ένα δείγμα δημόσιων tweets που τις εμπεριέχουν σε μορφή json (Javascript Object Notation), η δομή της οποίας διευκολύνει την ανάγνωση. Json είναι κατά βάση ένα πρότυπο ανταλλαγής δεδομένων που λόγω της απλότητας του όσον αφορά την ανάγνωση και γραφή από άνθρωπο και της ευκολίας ανάλυσης (parsing) και παραγωγής (generating) του από μηχανές χρησιμοποιείται ευρέως. Παρότι είναι βασισμένο πάνω σε ένα υποσύνολο της γλώσας προγραμματισμού JavaScript, το json είναι ένα πρότυπο κειμένου ανεξάρτητο από γλώσσες προγραμματισμού και χτίζεται πάνω σε δύο δομές. Η πρώτη αφορά μια συλλογή ονομάτων – τιμών, η οποία νοείται ως ένα αντικείμενο (object), εγγραφή (record), δομή δεδομένων (struct), λεξικό (dictionary), πίνακας κατατεμαχισμού (hash table), ή πίνακας συσχετισμών (associative array) και η δεύτερη αφορά μια σειρά τιμών που στις περισσότερες προγραμματιστικές γλώσσες γίνεται αντιληπτή ως ένας πίνακας (array), λίστα (list) ή διάνυσμα (vector). Στο πρότυπο json, οι προαναφερθείσες δομές παίρνουν τη μορφή ενός αντικειμένου που αποτελείται από ένα μη ταξινομημένο σύνολο ζευγών ονομάτων – τιμών. Ένα αντικείμενο εμπεριέχεται εντός των χαρακτήρων «{» (left brace) και «}» (right brace) και κάθε όνομα ακολουθείται από άνω – κάτω τελεία (χαρακτήρας «:») ενώ τα ζεύγη ονομάτων – τιμών χωρίζονται μεταξύ τους με κόμμα (χαρακτήρας «,»). Ακολουθούν ορισμένα σχηματικά που επεξηγούν τα παραπάνω.



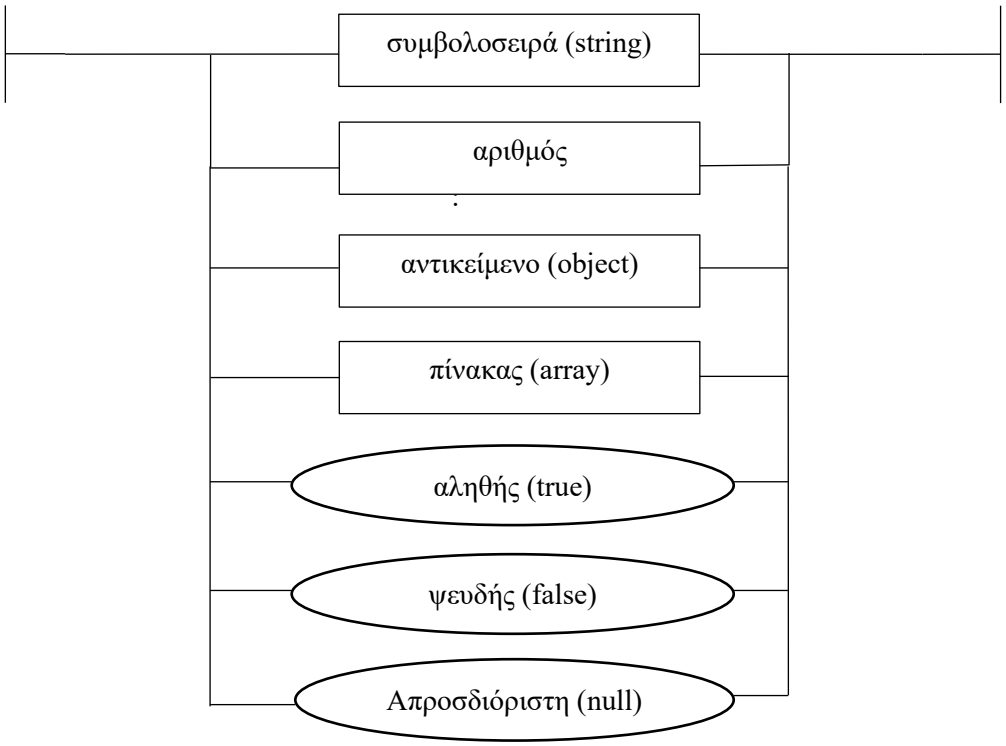
Σχήμα 4.1

Πίνακας
(array)

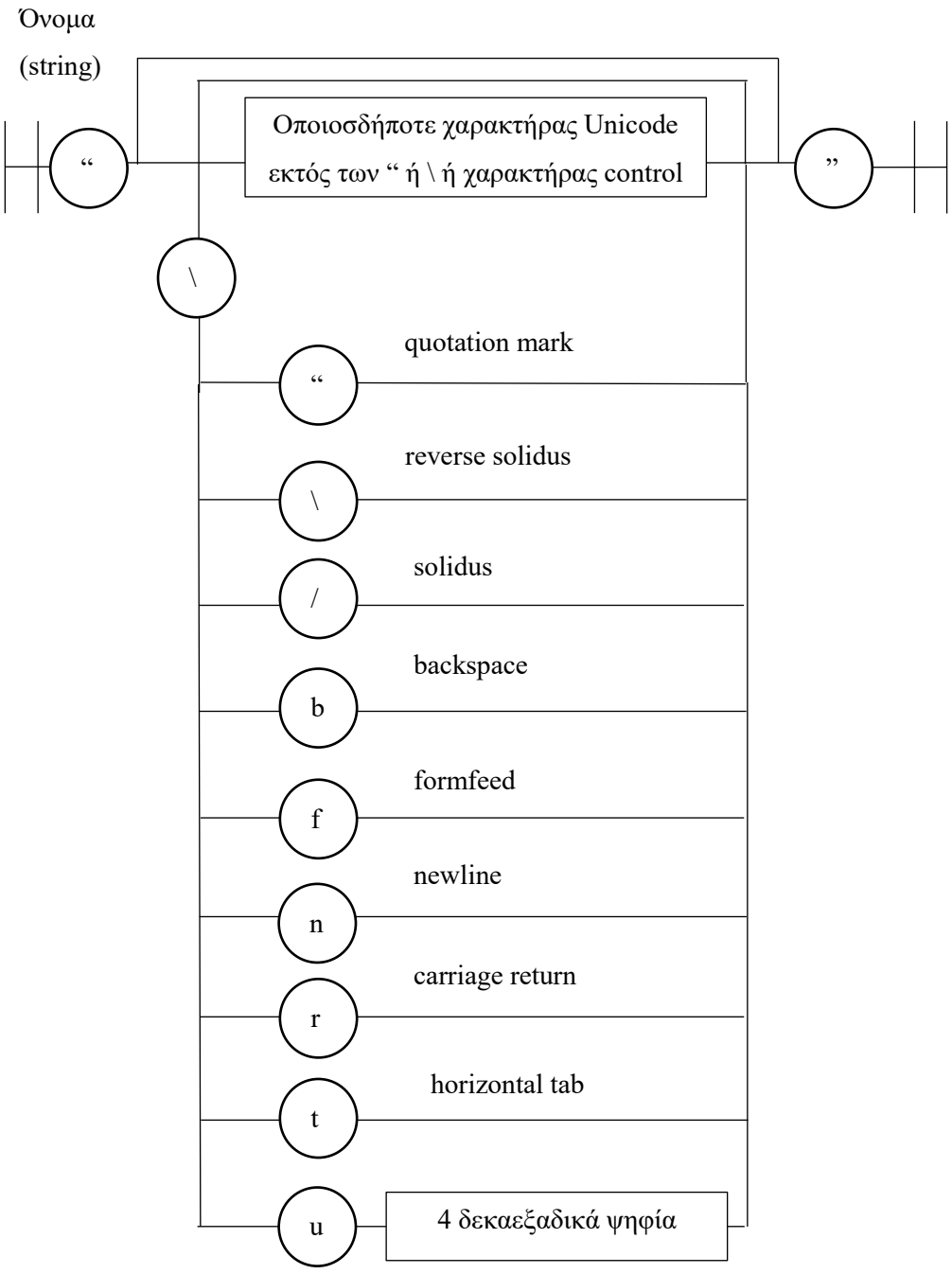


Σχήμα 4.2

Τιμή
(value)



Σχήμα 4.3

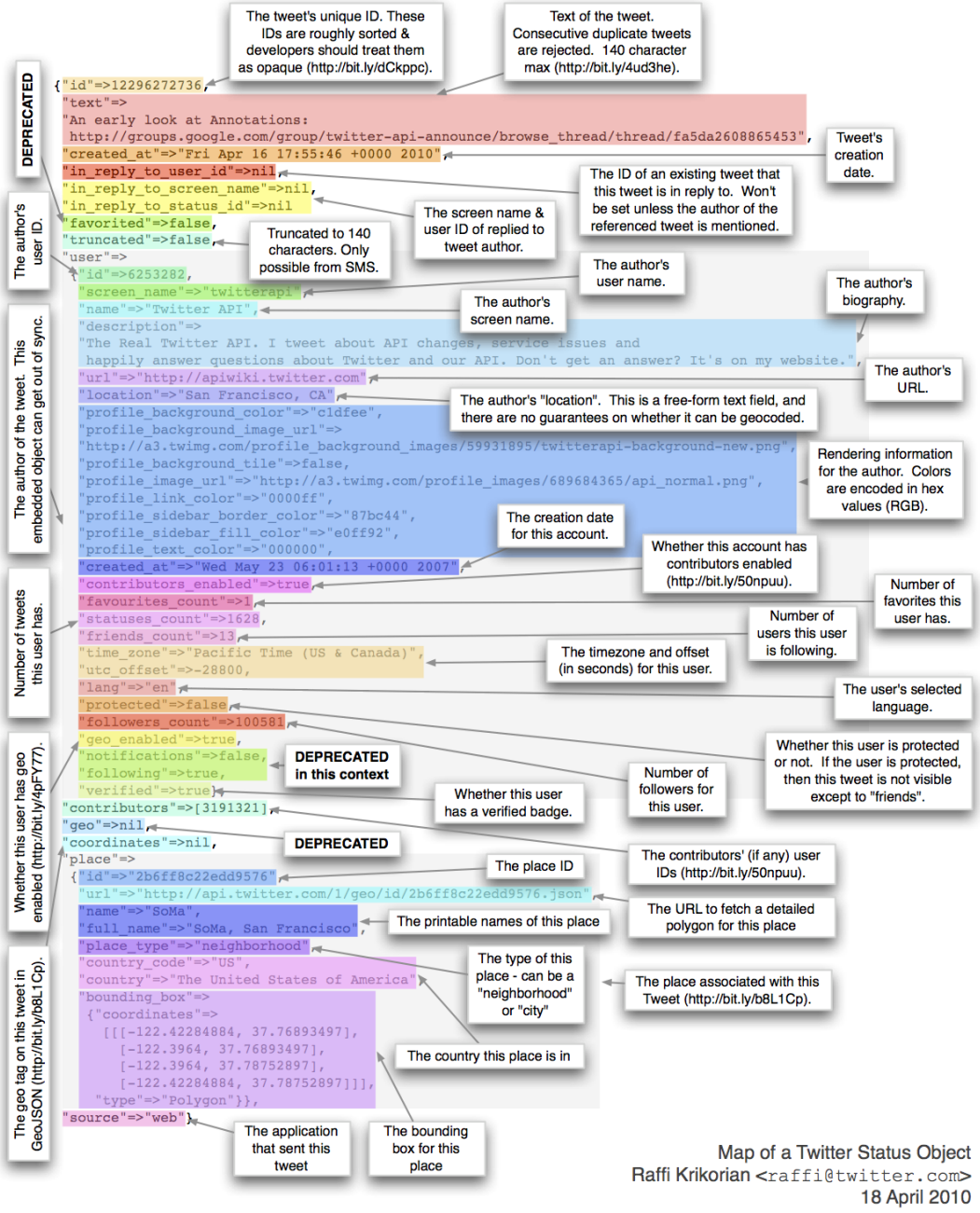

Όνομα
(string)
“
Οποιοσδήποτε χαρακτήρας Unicode
εκτός των “ ή \ ή χαρακτήρας control
”
\
quotation mark
“
reverse solidus
\
solidus
/
backspace
b
formfeed
f
newline
n
carriage return
r
horizontal tab
t
u
4 δεκαεξαδικά ψηφία

Σχήμα 4.4

Παρά την απλότητά της, η μορφή αυτή έχει την ιδιαιτερότητα ότι ως τιμή των συλλογών ονομάτων – τιμών μπορεί να χρησιμοποιηθεί ένα άλλο ζεύγος ονομάτων – τιμών οδηγώντας σε εμφωλευμένες δομές που είναι περίπλοκες. Το json που επιστρέφει το Twitter είναι εμπλουτισμένο με ένα μεγάλο πλήθος δεδομένων όπως η τοποθεσία του γράφοντος για κάποιο tweet ή ο αριθμός των retweets, δηλαδή των αναδημοσιεύσεων σε αυτούσια μορφή ενός tweet από άλλους χρήστες. Όπως διαπιστώθηκε, σε πολλές περιπτώσεις οι τιμές στα ζεύγη ονομάτων – τιμών στα json που σχετίζεται με ένα tweet απουσιάζουν είτε επειδή η συγκεκριμένη τιμή δεν έχει καθοριστεί από το χρήστη του Twitter είτε επειδή δεν έχουν λάβει τιμή ως εκείνη την ώρα. Χαρακτηριστικό παράδειγμα αποτελεί η απουσία αριθμού στο πεδίο του πλήθους των retweets για κάποιο tweet, καθώς τα λαμβάνουμε σε πραγματικό χρόνο μέσω της ροής του Twitter οπότε είναι επόμενο να μην έχει πραγματοποιηθεί κάποιο retweet ακόμα.

Στα πλαίσια της παρούσας εργασίας, θεωρήθηκε σκόπιμο να κρατούνται συγκεκριμένα πεδία ενός json που επιστρέφει το Streaming API, όπως το όνομα χρήστη, η ημερομηνία δημιουργίας του tweet, το κείμενο του tweet, τα hashtags και οι διευθύνσεις URL που αναφέρονται εντός του καθώς και η τοποθεσία του χρήστη αν αυτή είναι διαθέσιμη. Η επιλογή αυτή έγινε ούτως ώστε να λαμβάνουμε μόνο εκείνο το μέρος της πληροφορίας που παρέχει το API το οποίο θα αξιοποιηθεί για τη μελέτη και υλοποίηση του συστήματος ανίχνευσης.

Στην Εικόνα 4.1 παρουσιάζεται ένα διάγραμμα το οποίο δημοσιεύτηκε το 2010 από το μηχανικό του Twitter εκείνη την περίοδο, Raffi Krikorian, στο οποίο φαίνεται ένας μεγάλος αριθμός μεταδεδομένων ενός tweet όπως τα κατέγραψε ο ίδιος στη μορφή json. Η βασική μορφή του json ενός tweet παραμένει η ίδια τη στιγμή που γράφεται η παρούσα εργασία.



Εικόνα 4.1

4.1.2 Χρήση λέξεων – κλειδιών στο Streaming API

Το Streaming API χρησιμοποιήθηκε για τις ανάγκες του συστήματος προς υλοποίηση όσον αφορά την απόκτηση δεδομένων. Συγκεκριμένα, προκαθορισμένες λέξεις – κλειδιά σχετικές με τον τύπο των γεγονότων προς ανίχνευση χρησιμοποιήθηκαν για την άντληση tweets από τη

δημόσια ροή του Twitter σε πραγματικό χρόνο. Τα tweets που επιστρέφονται εμπεριέχουν μία ή περισσότερες εκ των προκαθορισμένων αυτών λέξεων – κλειδιών.

Συγκεκριμένα, οι λέξεις – κλειδιά που χρησιμοποιήθηκαν για την άντληση των tweets είναι οι εξής: σεισμός, σεισμικη δονηση, ριχτερ, εγκελαδος, πυρκαγια, εμπρησμος, μεγαλη φωτια, φωτια μαινεται, τροχαιο, αυτοκινητιστικο, συγκρουση οχηματων, συγκνοινωνιες, ΟΑΣΑ, ΟΑΣΘ, Αστεροσκοπειο, Ιδρυμα Σταυρος Νιαρχος, συναυλια, εκθεση τεχνης, επενδυση, καιρος. Οι λέξεις αυτές επιλέχθηκαν έτσι ώστε να δημιουργηθεί ένα δείγμα που θα περιέχει tweets σχετικά με το γεγονός ενδιαφέροντος αλλά και άλλα γεγονότα σχετιζόμενα με φυσικές καταστροφές όπως και ειδήσεις ή τάσεις που αφορούν την επικαιρότητα.

4.1.3 Χρησιμοποιούμενα δεδομένα

Η επιλογή των χρησιμοποιούμενων από τα δεδομένα που παρέχει το Twitter έγινε με βάση τον τύπο των επιθυμητών προς ανίχνευση γεγονότων για το υλοποιούμενο σύστημα, τη γενικότερη μορφή των μηνυμάτων στο μέσο αλλά και την ειδικότερη όσον αφορά αντίστοιχα γεγονότα που έχουν συμβεί. Με βάση την εξέταση που προηγήθηκε πριν την επιλογή των δεδομένων αυτών, διαπιστώθηκε ότι η σημαντικότερη συνεισφορά σε πληροφορία για ένα γεγονός προέρχεται από το κείμενο των tweets. Επιπλέον, πληροφορία για τη γεωγραφική τοποθεσία του γράφοντος περιέχεται στο κείμενο. Στο [27], οι συγγραφείς αναφέρουν ότι μόνον το 0.7% των tweets περιέχουν γεωγραφική υπογραφή. Στην περίπτωση γεγονότων που συνδέονται με φυσικές καταστροφές, αυτά τείνουν να συζητούνται και εκτός των γεωγραφικών ορίων μέσα στα οποία λαμβάνουν χώρα λόγω της ιδιαίτερης προσοχής που λαμβάνουν. Ως εκ τούτου είναι αναμενόμενο αναφορές που αφορούν συγκεκριμένο γεγονός που αφορά εντοπισμένη γεωγραφική περιοχή να προέρχονται από διάσπαρτα σημεία εντός της χώρας. Έτσι, πολύ συχνά η τοποθεσία στην οποία εξελίσσεται το αναφερόμενο γεγονός δίνεται εντός του κειμένου του tweet. Χαρακτηριστικά είναι τα παρακάτω δύο παραδείγματα tweets που αφορούν σεισμούς και στο κείμενο των οποίων δίνεται η τοποθεσία: «Μεγάλος σεισμός 6,3 ρίχτερ ανάμεσα σε Χίο και Λέσβο!», «Σεισμός 3,1 Ρίχτερ με επίκεντρο τη Ραφήνα».

4.1.4 Χρησιμοποιούμενα μεταδεδομένα

Ως μεταδεδομένα θεωρούνται συμφραζόμενα δεδομένα ή πληροφορία που περιγράφει τα δεδομένα (κείμενο tweets) ή τα χαρακτηριστικά τους. Ορισμένα παραδείγματα είναι η τοποθεσία του χρήστη που κοινοποιεί ένα κείμενο – tweet, ο αριθμός των retweets του, η

ημερομηνία και ώρα δημοσίευσης του, η επιλεγμένη γλώσσα του χρήστη και ο αριθμός των συνολικών tweets που έχει δημοσιεύσει ο χρήστης.

Από τα μεταδεδομένα αυτά που περιέχονται στο json αρχείο που επιστρέφει το Streaming API, επιλέχθηκαν ως δυνητικά χρήσιμα για περαιτέρω ανάλυση τα εξής μεταδεδομένα: διαδικτυακές διευθύνσεις (urls) που περιέχονται σε ένα tweet, λίστα των hashtags,τοποθεσία χρήστη καθώς και ημερομηνία και ώρα δημοσίευσης.

4.2 Δείγμα δεδομένων

Ένα δείγμα δεδομένων ή σύνολο δεδομένων (dataset ή data set) είναι μια συλλογή από δεδομένα που αντιστοιχεί συνήθως στα περιεχόμενα ενός πίνακα μιας βάσης δεδομένων, όπου κάθε στήλη του πίνακα αντιστοιχεί σε συσγκεκριμένη μεταβλητή και κάθε γραμμή σε κάθε στοιχείο του εν λόγω δείγματος δεδομένων. Για παράδειγμα, έστω μια βάση δεδομένων των πελατών μιας εταιρείας κινητής τηλεφωνίας στην οποία υπάρχουν πληροφορίες επικοινωνίας με τους πελάτες, πληροφορίες για λογαριασμούς και πληρωμές κ.α. Αν ληφθεί ένα δείγμα δεδομένων με τα στοιχεία επικοινωνίας των πελατών τότε αυτό θα είναι ένας δισδιάστατος πίνακας που θα μπορούσε να μοιάζει με τον Πίνακας 4.1, όπου φαίνεται ένα δείγμα με 2 στοιχεία και 7 μεταβλητές. Ακόμη, κάθε μια μεταβλητή του δείγματος δεδομένων αντιστοιχεί σε μια λίστα τιμών που είναι τόσες σε αριθμό όσα και τα στοιχεία του.

Όνομα	Επώνυμο	Ημερομηνία γέννησης	Οδός	Αριθμός	ΤΚ	Τηλέφωνο
Νίκος	Δέτας	16/03/1980	Σαλαμίνος	2	16238	22500 53100
Ιωάννης	Ιωάννου	25/09/1950	Μακεδονομάχων	121	13123	28314 00851

Πίνακας 4.1

4.2.1 Δημιουργία δείγματος δεδομένων εκπαίδευσης

Για τη λειτουργία της ταξινόμησης είναι απαραίτητη η δημιουργία ενός δείγματος δεδομένων εκπαίδευσης. Αυτό είναι ένα σύνολο δεδομένων, το οποίο αξιοποιείται στην εξεύρεση χαρακτηριστικών ανάμεσα στα δεδομένα που μπορούν να βοηθήσουν στην εργασία της πρόβλεψης της κλάσης μελλοντικών δεδομένων [42]. Έτσι, τα δεδομένα αυτά χρησιμοποιούνται στην εκμάθηση του ταξινομητή μιας μεθόδου αντιστοίχισης κάθε νέου στοιχείου που λαμβάνει προς ταξινόμηση σε προεπιλεγμένες κλάσεις με βάση τις τιμές των χαρακτηριστικών κάθε τέτοιου στοιχείου και τις κλάσεις στις οποίες τα στοιχεία του δείγματος εκπαίδευσης έχουν αντιστοιχισθεί. Επιπροσθέτως, το δείγμα δεδομένων θα αξιοποιηθεί στην αξιολόγηση του συστήματος ταξινόμησης όπως και του συνολικού συστήματος ανίχνευσης και ειδοποίησης.

Εν προκειμένω, κατά το χρονικό διάστημα 6 Ιουνίου ως 17 Ιουλίου 2017 λάβαμε 10,996 tweets συνολικά με βάση τις προκαθορισμένες λέξεις – κλειδιά που έχουν οριστεί. Στον Πίνακας 4.2 παρουσιάζεται η μορφή του δείγματος δεδομένων δίνοντας 5 tweets ως παραδείγματα.

Created_at	User	Tweet	Hastags_list	Expanded_url_list	User_Location	Country	Place_Name
Tue Jun 06 14:53:59 +0000 2017	GigiV46	ενα respect στο γιωργο μαργαριτη που εδωσε αφιλοκερδως συναυλια στις φυλακες κορυδαλου αυτος ειναι μαγκας .	[]	[]	Trelleborg, Sverige		
Tue Jun 06 17:32:35 +0000 2017	ban_mg	RT @tsipouridhs: Θα έρθει ο καιρος τα θελω να νικησουν ολα εκείνα τα μυαλά τις λογικής που βάζουνε τα πρέπει στη ζωή μας και δεν αφήνουν τ…	[]	[]			

Tue Jun 06 17:43:34 +0000 2017	athenstransport	Μηνιαίες κάρτες για 30 μέρες και όχι για ένα ημερολογιακό μήνα υπόσχεται ο ΟΑΣΑ https://t.co/roPy8WdU8i	[]	['http://www.athenstransport.com/2017/06/miniaies-kartes/']	Αθήνα, Ελλάδα		
Tue Jun 06 17:52:54 +0000 2017	eZCj3qkIVUbo5pD	@roussosap @chrystallia1 @marios_chr @Marios_skorpios @toxotakiisback @AndreasMouyis1 @Antantonis @ms_msms Καλός ο καιρος στο Λονδινο?	[]	[]	Εκεί που αγγίζει τη ψυχή		
Tue Jun 06 18:03:10 +0000 2017	paidisi	ΑΓΩΝΙΑ ΑΠΟ ΝΕΑ ΠΥΡΚΑΓΙΑ ΣΤΗ ΒΙΟΚΕΡΑΛ · https://t.co/MtgKhLJ2d7 https://t.co/Iu9TwwIiXd	[]	['http://Paidis.com', 'http://fb.me/8AEmOPDif']	ΛΑΡΙΣΑ		

Πίνακας 4.2

4.2.2 Μη ισορροπημένο δείγμα δεδομένων εκπαίδευσης

Στο [43], οι συγγραφείς εξετάζουν ένα σύστημα εξόρυξης πληροφορίας από κείμενο αναφορικά με την αλληλεπίδραση μεταξύ φαρμάκων (Drug-Drug Interaction) έχοντας ένα μη ισορροπημένο δείγμα δεδομένων εκπαίδευσης (imbalanced dataset). Σε αυτό εξετάζεται ένα δείγμα 23,827 ζευγών φαρμάκων εκ των οποίων μόλις το 9.89% ανήκουν στη θετική κλάση με βάση την χειροκίνητη ταξινόμηση που πραγματοποιήθηκε ενώ τα υπόλοιπα κατατάσσονται στην αρνητική. Όταν υπάρχει τέτοια αναντιστοιχία στην εκπροσώπηση των δύο κλάσεων από τα δεδομένα, το σύνολο δεδομένων καλείται μη ισορροπημένο [44]. Το πρόβλημα της ύπαρξης μη ισορροπημένου δείγματος είναι σύνηθες και σχετίζεται άμεσα με τον τύπο του προβλήματος

το οποίο αφορά. Για την περιορισμό της επίδρασης του μη ισορροπημένου δείγματος, που οδηγεί σε «μεροληπτική» πρόβλεψη νέων δεδομένων υπέρ της δεσπόζουσας κλάσης από τον ταξινομητή υπολογίζονται οι μετρικές ακρίβειας, ανάκλησης και βαθμολογίας-F. Προς το σκοπό αυτό, δοκιμάζονται επίσης διαφορετικοί αλγόριθμοι ταξινόμησης.

4.2.3 Αντιστοίχιση ετικετών στο δείγμα δεδομένων

Για το δείγμα δεδομένων που λήφθηκε και με βάση όσα έχουν προηγηθεί για την ταξινόμηση tweets στις δύο κλάσεις, θετική και αρνητική, προσθέσαμε χειροκίνητα τις ετικέτες True και False που αντιστοιχούν στις δύο αυτές κλάσεις εντός του συνόλου δεδομένων κάτω από την κατάλληλη στήλη – μεταβλητή για κάθε στοιχείο.

Με δεδομένες τις ετικέτες αυτές εκπαιδεύεται ο ταξινομητής στην πρόβλεψη της κλάσης νέων tweets με βάση τα χαρακτηριστικά εκπαίδευσης.

4.3 Χαρακτηριστικά εκπαίδευσης

Στην ενότητα αυτή παρουσιάζουμε τις ομάδες χαρακτηριστικών των δεδομένων οι οποίες εξετάστηκαν ως προς την αποτελεσματικότητά τους στη μάθηση του ταξινομητή και την ικανότητα πρόβλεψης των κλάσεων νέων δεδομένων. Έγινε αξιολόγηση στην ενότητα του πειραματισμού της αποτελεσματικότητας του επιλεγμένου ταξινομητή τόσο για ομάδες χαρακτηριστικών που έχουν καταγραφεί στη βιβλιογραφία όσο και για ομάδες οι οποίες καταρτίστηκαν στα πλαίσια της εργασίας αυτής.

- **Ομάδα Α** (2 χαρακτηριστικά): το πλήθος λέξεων του κειμένου ενός tweet και η θέση μιας εκ των λέξεων-κλειδιών εντός του κειμένου όπως αυτές έχουν οριστεί.

- **Ομάδα Β** (3 χαρακτηριστικά): το πλήθος λέξεων του κειμένου ενός tweet και η θέση μιας εκ των λέξεων-κλειδιών εντός του κειμένου όπως αυτές έχουν οριστεί καθώς και η θέση εντός του κειμένου τοποθεσίας από τη λίστα τοποθεσιών αν υπάρχει. Αν δεν υπάρχει τοποθεσία το αντίστοιχο χαρακτηριστικό λαμβάνει τιμή 0, ενώ το ίδιο συμβαίνει και σε περίπτωση απουσίας λέξεως-κλειδιού από το κείμενο για το αντίστοιχο χαρακτηριστικό.

- **Ομάδα Γ** (tf-idf, τόσα χαρακτηριστικά όσες οι μοναδικές λέξεις μεταξύ του πλήθους των κειμένων που επεξεργάζονται): αναπαράσταση του κειμένου ως bag of words που χρησιμοποιείται ως εργαλείο για τη δημιουργία χαρακτηριστικών [45]. Κάθε μοναδική λέξη είναι ένα χαρακτηριστικό και η tf-idf τιμή της είναι το μέτρο της σημαντικότητας αυτής της λέξης.

- **Ομάδα Δ** (3 χαρακτηριστικά): Η ομάδα Δ' διαφοροποιείται από την ομάδα Β' στο ότι χρησιμοποιούνται ως χαρακτηριστικά η ύπαρξη ή όχι λέξης-κλειδιού και τοποθεσίας στο κείμενο αντί της θέσης αυτών. Κρίνεται σκόπιμο εδώ να σχολιαστεί ότι η ομάδα χαρακτηριστικών Β' εμπεριέχει την έννοια της ύπαρξης λέξης-κλειδιού και τοποθεσίας αφού στην περίπτωση που δεν υπάρχει εντός του κειμένου μια εκ των ζητούμενων λέξεων ή τοποθεσία, η τιμή του χαρακτηριστικού τίθεται μηδέν, ενώ σε αντίθετη περίπτωση λαμβάνει την τιμή ενός φυσικού αριθμού μεγαλύτερου ή ίσου με 1.

Στον Πίνακας 4.3 φαίνεται ένα παράδειγμα κειμένου tweet και οι τιμές που λαμβάνουν τα χαρακτηριστικά για τις διάφορες ομάδες που αναφέρθηκαν.

Παράδειγμα κειμένου tweet	ΟΜΑΔΑ Α'		ΟΜΑΔΑ Β'			ΟΜΑΔΑ Γ'	ΟΜΑΔΑ Δ'		
	Πλήθος λέξεων	Θέση λέξης-κλειδιού	Πλήθος λέξεων	Θέση λέξης-κλειδιού	Θέση τοποθεσίας	tf-idf	Πλήθος λέξεων	Ύπαρξη λέξης-κλειδιού	Ύπαρξη τοποθεσίας
Ισχυρός σεισμός τώρα στη Χίο	5	2	5	2	5	Κάθε λέξη χρησιμοποιείται ως χαρακτηριστικό	5	1	1

Πίνακας 4.3

4.4 Προεπεξεργασία δεδομένων

Εδώ περιγράφουμε την προεπεξεργασία στην οποία υπόκειται κάθε κείμενο tweet πριν αυτό οδηγηθεί προς ταξινόμηση από το σύστημα ανίχνευσης. Κατ' αναλογία όλα τα δεδομένα στο δείγμα προς εκπαίδευση προεπεξεργάζονται πριν πραγματοποιηθούν τα πειράματα που διενεργήθηκαν.

4.4.1 Αφαίρεση τονισμού και μετατροπή χαρακτήρων σε πεζούς

Αρχικά, κάθε χαρακτήρας εντός του κειμένου ενός tweet που αποτελεί μια συμβολοσειρά μετατρέπεται σε πεζό ούτως ώστε λέξεις όπως «Σεισμός», «ΣΕΙΣΜΟΣ» και «σεισμός» να θεωρούνται ταυτόσημες από τον ταξινομητή. Προκειμένου ακόμη να θεωρούνται και οι λέξεις «Σεισμος» και «σεισμος» ισοδύναμες με τις προαναφερθείσες, αφαιρούνται και οι τονισμοί από το κείμενο. Ακολουθεί στον Πίνακας 4.4 ένα παράδειγμα χρήσης της εν λόγω επεξεργασίας.

Ανεπεξέργαστο κείμενο tweet	Επεξεργασμένο κείμενο tweet
RT @pkarkatsoulis: Δείτε τις οδηγίες της πολιτικής προστασίας για την περίπτωση σεισμού: https://t.co/WjPZG9irIj Η γνώση σώζει. #earthquake…	rt @pkarkatsoulis: δειτε τις οδηγιες της πολιτικης προστασιας για την περιπτωση σεισμου: https://t.co/wjpzg9irij η γνωση σωζει. #earthquake…

Πίνακας 4.4

4.4.2 Αφαίρεση συνδέσμων, συμβόλων hashtags και ονομάτων χρηστών

Ακόμη, το κείμενο απαλάσσεται από συνδέσμους (urls), σύμβολα hashtags και ονόματα άλλων χρηστών (όπως για παράδειγμα στην περίπτωση ενός retweet, όπου διαγράφεται το κείμενο που ακολουθεί το χαρακτήρα «@»). Στον Πίνακας 4.5 παρουσιάζεται ένα παράδειγμα tweet πριν και μετά την συγκεκριμένη επεξεργασία. Η αφαίρεση του συμβόλου hashtag αλλά η διατήρηση του ίδιου του αναφερόμενου hashtag γίνεται σκοπίμως, καθώς παρατηρήσαμε ότι τα hashtags εμπεριέχουν πλούσιες περιεχομένου λέξεις όπως λέξεις-κλειδιά ή τοποθεσία. Στη βιβλιογραφία έχει μελετηθεί η επίδραση των hashtags στην ανίχνευση γεγονότων στο Twitter μέσω ομαδοποίησης [23].

Ανεπεξέργαστο κείμενο tweet	Επεξεργασμένο κείμενο tweet
RT @pkarkatsoulis: Δείτε τις οδηγίες της πολιτικής προστασίας για την περίπτωση σεισμού: https://t.co/WjPZG9irIj Η γνώση σώζει. #earthquake…	RT Δείτε τις οδηγίες της πολιτικής προστασίας για την περίπτωση σεισμού: Η γνώση σώζει. earthquake…

Πίνακας 4.5

4.4.3 Αφαίρεση σημείων στίξης

Εν συνεχεία, αφαιρούνται τα σημεία στίξης, όπως για παράδειγμα χαρακτήρες τελείας, κόμματος, θαυμαστικών και ερωτηματικών μιας και θεωρούμε ότι δεν προσθέτουν αξία στο εννοιολογικό περιεχόμενο του κειμένου. Παρόλαυτα, τα στοιχεία στίξης όπως θαυμαστικό και ερωτηματικό έχουν αξιοποιηθεί στη βιβλιογραφία κατά την ανάλυση συναισθημάτων (sentiment analysis) σε κοινωνικά δίκτυα [46], [47]. Στον Πίνακας 4.6 φαίνονται τα σημεία στίξης που λήφθηκαν υπόψην. Ακόμη, στον Πίνακα 4.7 δίνεται ένα παράδειγμα αφαίρεσης σημείων στίξης από κείμενο tweet.

!	"	#	$
%	&	'	(
)	*	+	,
-	.	/	:
;	<	=	>
?	@	[	\
]	^	_	`
{	¦	}	~

Πίνακας 4.6

Ανεπεξέργαστο κείμενο tweet	Επεξεργασμένο κείμενο tweet
RT @pkarkatsoulis: Δείτε τις οδηγίες της πολιτικής προστασίας για την περίπτωση σεισμού: https://t.co/WjPZG9irIj Η γνώση σώζει. #earthquake…	RT pkarkatsoulis Δείτε τις οδηγίες της πολιτικής προστασίας για την περίπτωση σεισμού httpstcoWjPZG9irIj Η γνώση σώζει earthquake

Πίνακα 4.7

4.4.4 Αφαίρεση περιττών λέξεων

Επιπροσθέτως, δημιουργήθηκε μια λίστα περιττών λέξεων (stop words) στα ελληνικά, οι οποίες είναι κοινότοπες και ως εκ τούτου δεν προσφέρουν εννοιολογικό περιεχόμενο σε ένα tweet. Αφαιρούνται, επομένως, από το κείμενο αφού η πιθανότητα εύρεσης τους σε ένα tweet σχετικό με το γεγονός που ενδιαφέρει προς ανίχνευση είναι ίδια σε σχέση με ένα οποιοδήποτε άλλο tweet [48]. Καταρτήσαμε μια λίστα με τέτοιες λέξεις στην ελληνική και αγγλική γλώσσα και προσθέσαμε ακόμη λέξεις όπως «rt», «via» και «amp». Αυτές οι λέξεις αναζητούνται κατά την παραλαβή ενός tweet από το API και αφαιρούνται άμεσα. Παράδειγμα περιττών λέξεων προς αφαίρεση είναι αντωνυμίες και σύνδεσμοι.

Η λίστα των 268 ελληνικών περιττών λέξεων που χρησιμοποιήθηκαν δίνεται στην Εικόνα 4.2.

```
["μου", "σου", "του", "αλλα", "αν", "αντι", "απο", "αυτα", "αυτεσ", "αυτη", "αυτο", "αυτοι", "αυτοσ",
 "αυτουσ", "αυτων", "αἱ", "αἳ", "αἵ", "αὐτόσ",
"αὐτὸς", "αὖ", "γάρ", "γα", "γα^", "γε", "για", "γοῦν", "γὰρ", "δ'", "δέ", "δή", "δαί", "δαίσ", "δαὶ",
"δαὶς", "δε", "δεν", "δι'", "διά", "διὰ", "δὲ",
"δὴ", "δ’", "εαν", "ειμαι", "ειμαστε", "ειναι", "εισαι", "ειστε", "εκεινα", "εκεινεσ", "εκεινη", "εκεινο",
"εκεινοι", "εκεινοσ", "εκεινουσ",
"εκεινων", "ενω", "επ", "επι", "εἰ", "εἰμί", "εἰμὶ", "εἰς", "εἰσ", "εἴ", "εἶμι", "εἶτε", "η", "θα", "ισωσ",
"κ", "καί", "καίτοι", "καθ", "και", "κατ",
"κατά", "κατα", "κατὰ", "καὶ", "κι", "κἀν", "κᾶν", "μέν", "μή", "μήτε", "μα", "με", "μεθ", "μετ", "μετά",
"μετα", "μετὰ", "μη", "μην", "μέν", "μὲν",
"μὴ", "μὴν", "να", "ο", "οι", "ομωσ", "οπωσ", "οσο", "οτι", "οἱ", "οἳ", "οἷς", "οὐ", "οὐδ", "οὐδέ",
"οὐδείσ", "οὐδεὶς", "οὐδὲ", "οὐδὲν", "οὐκ", "οὐχ",
"οὐχὶ", "οὓς", "οὔτε", "οὕτω", "οὕτως", "οὕτωσ", "οὖν", "οὗ", "οὗτος", "οὗτοσ", "παρ", "παρά", "παρα",
"παρὰ", "περί", "περὶ", "ποια", "ποιεσ",
"ποιο", "ποιοι", "ποιοσ", "ποιουσ", "ποιων", "ποτε", "που", "ποῦ", "προ", "προσ", "πρόσ", "πρὸ", "πρὸς",
"πωσ", "πωσ", "σε", "στη", "στην", "στο",
"στον", "σόσ", "σύ", "σύν", "σὸς", "σὺ", "σὺν", "τά", "τήν", "τί", "τίς", "τίσ", "τα", "ταῖς", "τε", "την",
"τησ", "τι", "τινα", "τις", "τισ", "το",
"τοί", "τοι", "τοιοῦτος", "τοιοῦτοσ", "τον", "τοτε", "του", "τούσ", "τοὺς", "τοῖς", "τοῦ", "των", "τό",
"τόν", "τότε", "τὰ", "τὰς", "τὴν", "τὸ",
"τὸν", "τῆς", "τῆσ", "τῇ", "τῶν", "τῷ", "ωσ", "ἀλλ'", "ἀλλά", "ἀλλὰ", "ἀλλ’", "ἀπ", "ἀπό", "ἀπὸ", "ἀφ",
"ἂν", "ἃ", "ἄλλος", "ἄλλοσ", "ἄν", "ἄρα",
"ἅμα", "ἐάν", "ἐγώ", "ἐγὼ", "ἐκ", "ἐμόσ", "ἐμὸς", "ἐν", "ἐξ", "ἐπί", "ἐπεὶ", "ἐπὶ", "ἐστι", "ἐφ", "ἐὰν",
"ἑαυτοῦ", "ἔτι", "ἡ", "ἦ", "ἦ", "ἦ", "ἦ", "ἧς",
"ἵνα", "ὁ", "ὃ", "ὃν", "ὅς", "ὅ", "ὅδε", "ὅθεν", "ὅπερ", "ὅς", "ὅσ", "ὅστις", "ὅστισ", "ὅτε", "ὅτι",
"ὑμόσ", "ὑπ", "ὑπέρ", "ὑπό", "ὑπὲρ", "ὑπὸ",
"ὡς", "ὡσ", "ὥς", "ὥστε", "ὦ", "ὦ"]
```

Εικόνα 4.2

Ακολουθεί ένα παράδειγμα στον Πίνακας 4.8 που δείχνει την επίδραση της συγκεκριμένης επεξεργασίας.

Ανεπεξέργαστο κείμενο tweet	Επεξεργασμένο κείμενο tweet
RT @pkarkatsoulis: Δείτε τις οδηγίες της πολιτικής προστασίας για την περίπτωση σεισμού: https://t.co/WjPZG9irIj Η γνώση σώζει. #earthquake…	@pkarkatsoulis: Δείτε οδηγίες πολιτικής προστασίας περίπτωση σεισμού: https://t.co/WjPZG9irIj γνώση σώζει. #earthquake…

Πίνακας 4.8

4.4.5 Αφαίρεση emojis

Η χρήση emojis, δηλαδή ενός μεγάλου σετ χαρακτήρων που έχουν καθοριστεί από το πρότυπο Unicode έχει ενσωματωθεί επιτυχώς τα τελευταία χρόνια στον ψηφιακό γραπτό λόγο. Τα emojis είναι πρακτικά εικονίδια που μπορούν να αναπαριστούν πρόσωπα και διαθέσεις, αντικείμενα, οχήματα κ.α. και χρησιμοποιούνται ευρέως στα μέσα κοινωνικής δικτύωσης όπως το Twitter. Σύμφωνα με το [49], το 19.6% των tweets στο Twitter περιέχουν ένα ή περισσότερα emoji. Τεχνικές ανάλυσης συναισθημάτων σε κοινωνικά δίκτυα αξιοποιούν τη χρήση emojis, τα οποία είναι πολυάριθμα, για να βγάλουν χρήσιμα συμπεράσματα όσον αφορά τα συναισθήματα των μηνυμάτων και συνεπώς των χρηστών των μέσων αυτών απέναντι σε προϊόντα ή γεγονότα.

Σύμφωνα με τις διαπιστώσεις των συγγραφέων στο [50], τα περισσότερα emojis εκφράζουν θετικά συναισθήματα. Ως εκ τούτου, στα πλαίσια της παρούσας διπλωματικής για την ανίχνευση γεγονότων φυσικών καταστροφών στο Twitter – γεγονότα που από τη φύση τους δημιουργούν αρνητικά συναισθήματα στο κοινωνικό σύνολο – θεωρήθηκε σκόπιμο να αφαιρεθούν οι χαρακτήρες αυτοί από το κυρίως κείμενο ενός tweet, και να επικεντρωθεί η ανάλυση στο περιεχόμενό τους. Δημιουργήθηκε για το σκοπό αυτό ένα φίλτρο συγκεκριμένων χαρακτήρων Unicode, το οποίο αφαιρεί μεγάλο μέρος των emojis από το κείμενο των tweets που λαμβάνονται. Παραδείγματα emojis είναι τα ακόλουθα: 🔌, 👆.

Τελικά, εφαρμόζοντας όλα τα παραπάνω βήματα προεπεξεργασίας κειμένου στη σειρά που παρουσιάστηκαν, θα προκύψει ένα επεξεργασμένο tweet όπως φαίνεται στον Πίνακας 4.9.

Ανεπεξέργαστο κείμενο tweet	RT @pkarkatsoulis: Δείτε τις οδηγίες της πολιτικής προστασίας για την περίπτωση σεισμού: https://t.co/WjPZG9irIj Η γνώση σώζει. #earthquake…
Αφαίρεση τονισμού και μετατροπή χαρακτήρων σε πεζούς	rt @pkarkatsoulis: δειτε τις οδηγιες της πολιτικης προστασιας για την περιπτωση σεισμου: https://t.co/wjpzg9irij η γνωση σωζει. #earthquake…
Αφαίρεση συνδέσμων, συμβόλων hashtags και ονομάτων χρηστών	rt: δειτε τις οδηγιες της πολιτικης προστασιας για την περιπτωση σεισμου: η γνωση σωζει. earthquake…

Αφαίρεση σημείων στίξης	rt δειτε τις οδηγιες της πολιτικης προστασιας για την περιπτωση σεισμου η γνωση σωζει earthquake
Αφαίρεση περιττών λέξεων	δειτε οδηγιες πολιτικης προστασιας περιπτωση σεισμου γνωση σωζει earthquake
Αφαϊρεση emojis	δειτε οδηγιες πολιτικης προστασιας περιπτωση σεισμου γνωση σωζει earthquake

Πίνακας 4.9

5

Διαδικτυακή πλατφόρμα και προγραμματιστικά εργαλεία

Στην ενότητα αυτή γίνεται αναφορά στην υλοποίηση της διαδικτυακής πλατφόρμας αναφοράς των γεγονότων που ανιχνέυονται καθώς και στα σημαντικότερα προγραμματιστικά εργαλεία που χρησιμοποιήθηκαν.

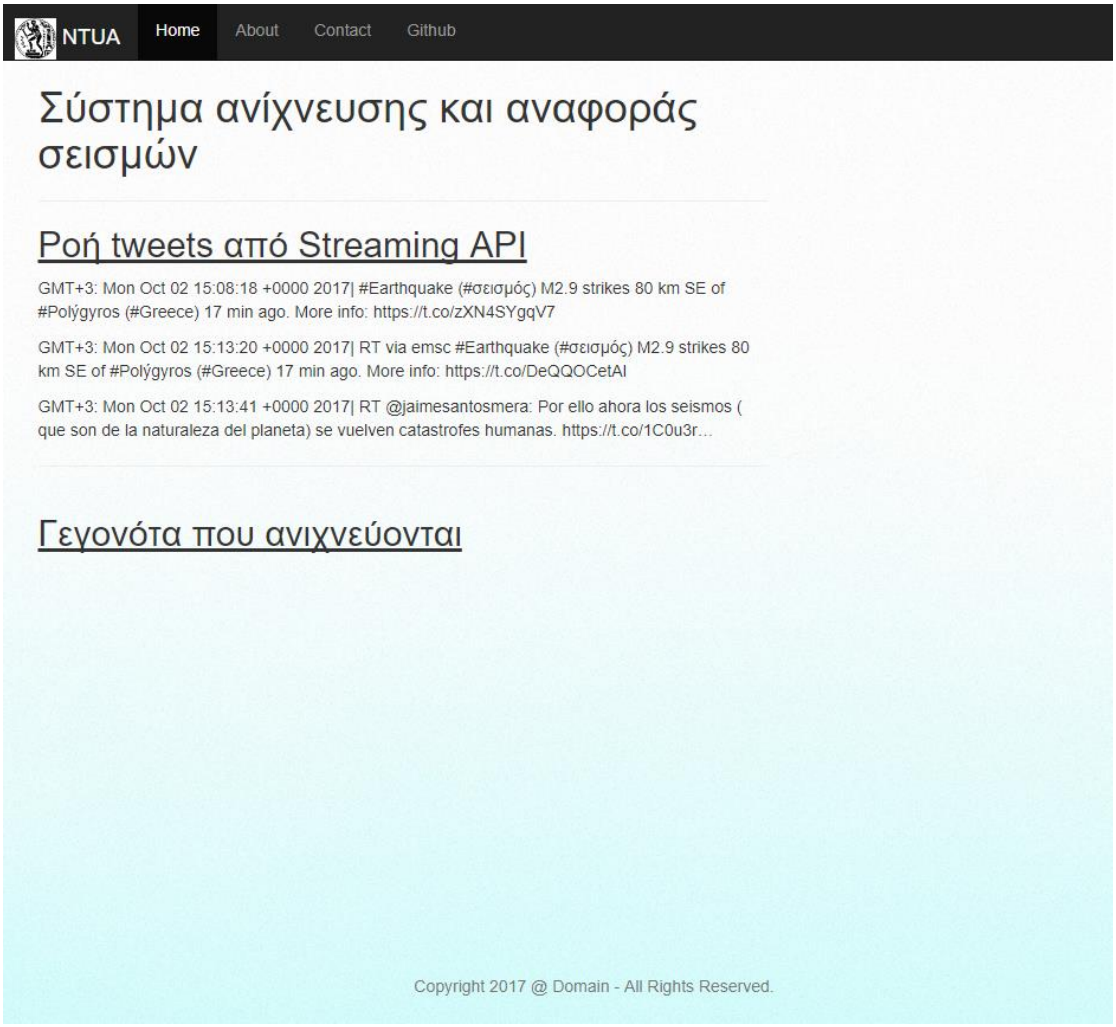
5.1 Διαδικτυακή πλατφόρμα

Τα αποτελέσματα της ανίχνευσης γεγονότων μαζί με τη ροή tweets που λαμβάνει το υλοποιηθέν σύστημα ανίχνευσης παρουσιάζονται σε μία ιστοσελίδα.

5.1.1 Διαδικτυακή σελίδα

Η ανάπτυξη της στατικής ιστοσελίδας αναφοράς γεγονότων έγινε αξιοποιώντας τη γλώσσα html, ενώ το δυναμικό κομμάτι της υλοποιήθηκε σε Javascript. Για την υλοποίηση της διαδικτυακής εφαρμογής χρησιμοποιήσαμε το μικρο-πλαίσιο για διαδικτυακές εφαρμογές (micro web framework) Flask. Το Flask είναι γραμμένο στη γλώσσα υψηλού επιπέδου Python και στηρίζεται σε δύο εξωτερικές βιβλιοθήκες, τη μηχανή προτύπων Jinja2 και την WSGI εργαλειοθήκη Werkzeug. Με τον όρο WSGI (Web Server Gateway Interface) νοείται ο μεσολαβητής μεταξύ του web server που φιλοξενεί την ιστοσελίδα και της εφαρμογής Python.

Στην Εικόνα 5.1 ακολουθεί ένα στιγμιότυπο της ιστοσελίδας όπου σε ζωντανό χρόνο παρουσιάζεται η ροή των tweets που παραλαμβάνονται από το σύστημα μέσω του Streaming API. Στο συγκεκριμένο παράδειγμα δεν έχει αναφερθεί ακόμη κάποιος σεισμός.



Εικόνα 5.1

5.1.2 Εφαρμογή για κινητές συσκευές

Για την εφαρμογή για κινητές συσκευές χρησιμοποιήσαμε την πλατφόρμα GoNative.io για τη μετατροπή ιστοσελίδας σε εφαρμογή για συσκευές Android και iOS. Έτσι, δημιουργήθηκε μια εφαρμογή για φορητές συσκευές μέσω της οποίας οι χρήστες έχουν άμεση πρόσβαση στον ιστότοπο παρουσίασης των αποτελεσμάτων του συστήματος ανίχνευσης. Παρακάτω φαίνεται στην Εικόνα 5.2 η αρχική οθόνη της εφαρμογής.

Εικόνα 5.2

5.2 Προγραμματιστικά εργαλεία

Η υλοποίηση του συστήματος που περιγράφεται στην παρούσα διπλωματική έγινε στη γλώσσα Python. Συγκεκριμένα, χρησιμοποιήθηκε η έκδοση της Python Anaconda 3.6. Ιδιαίτερη μνεία γίνεται στη βιβλιοθήκη scikit-learn που αξιοποιήθηκε για την υλοποίηση του τμήματος μηχανικής μάθησης του συνολικού συστήματος.

5.2.1 Twitter Streaming API

Για τη λήψη tweets στο μέσο Twitter, απαιτείται η χρήση ενός από τα διαθέσιμα APIs (Application Programming Interface), δηλαδή μιας διεπαφής προγραμματισμού εφαρμογών που επιτρέπει την άντληση δεδομένων μέσω ενός συνόλου προκαθορισμένων διαδικασιών. Τα διαθέσιμα APIs είναι τα ακόλουθα:

- REST APIs, που δίνουν τη δυνατότητα μέσω προγραμματιστικών εργαλείων της εγγραφής ή ανάγνωσης δεδομένων του μέσου. Τα APIs αυτά επιτρέπουν τη λήψη δεδομένων και μεταδεδομένων για συγκεκριμένους χρήστες τους οποίους καθορίζει ο προγραμματιστής.
- Webhook APIs. Το μοναδικό API αυτής της κατηγορίας που προσφέρεται είναι το Account Activity API, το οποίο βρίσκεται σε δοκιμαστική έκδοση και χρησιμοποιείται σε εφαρμογές εξυπηρέτησης πελατών όπως αποστολή απευθείας μηνυμάτων.
- Ads API, το οποίο επιτρέπει τη διαχείριση και εκτέλεση διαφημιστικών ενεργειών στην πλατφόρμα του Twitter.
- Streaming APIs, τα οποία δίνουν στους προγραμματιστές πρόσβαση στη παγκόσμια ροή δεδομένων tweets με μικρό χρόνο απόκρισης.

Για τις ανάγκες της παρούσας εργασίας, έγινε χρήση των Streaming APIs και πιο συγκεκριμένα των Public streams, δηλαδή της ροής δεδομένων που έχουν δημόσια διαμοιραστεί στην πλατφόρμα. Το συγκεκριμένο API είναι κατάλληλο για ανάκτηση σημαντικού όγκου δεδομένων σε πραγματικό χρόνο από διαφορετικούς χρήστες με βάση κάποια κριτήρια αναζήτησης. Συγκεκριμένα, στην παρούσα υλοποίηση συστήματος χρησιμοποιήθηκαν συγκεκριμένες λέξεις – κλειδιά για την κατάλληλη επιστροφή σχετικών δεδομένων.

5.2.2 Tweepy

Για την απόκτηση δεδομένων από το Streaming API του Twitter έγινε χρήση της βιβλιοθήκης tweepy της Python, η οποία παρέχει εύκολη πρόσβαση στα APIs μέσω κλειδιών ταυτοποίησης του χρήστη-προγραμματιστή. Για τη χρήση των APIs και την παροχή των κλειδιών αυτών, απαιτείται πρώτα η δημιουργία ενός λογαριασμού χρήστη του Twitter και η έναρξη μιας εφαρμογής προγραμματιστή.

5.2.3 Python scikit-learn

Η scikit – learn είναι μια ελεύθερα προσβάσιμη βιβλιοθήκη της προγραμματιστικής γλώσσας Python που περιλαμβάνει μια ευρεία γκάμα αλγορίθμων μηχανικής μάθησης για μεσαίου μεγέθους προβλήματα με ή χωρίς επιτήρηση [51]. Μεταξύ άλλων περιλαμβάνει αλγορίθμους

ταξινόμησης και ομαδοποίησης οι οποίοι είναι εύκολο να χρησιμοποιηθούν άμεσα μέσω μιας υψηλού επιπέδου γλώσσας όπως η Python λόγω της ευκολίας χρήσης, της ύπαρξης τεκμηρίωσης και της συνεχούς συντήρησης αλλά και εμπλούτισης.

Επίσης, διανέμεται υπό την άδεια BSD, που ενθαρρύνει την ακαδημαϊκή χρήση παρέχοντας ελεύθερα όλα τα απαραίτητα αρχεία κώδικα και τεκμηρίωσης.

6

Πειράματα και αξιολόγηση

Με βάση το σύνολο δεδομένων που δημιουργήθηκε στα πλαίσια της παρούσας εργασίας και στο οποίο τα δεδομένα ταξινομήθηκαν χειροκίνητα στις επιθυμητές κλάσεις, στο κεφάλαιο αυτό αξιολογούμε τους ταξινομητές που χρησιμοποιήθηκαν ως προς την επίδοση τους στην ταξινόμηση νέων δεδομένων. Παρουσιάζουμε, ακόμη, τα πειραματικά αποτελέσματα του συστήματος ειδοποίησης και ως εκ τούτου την αξιολόγηση του συστήματος στην ανίχνευση σεισμών και σε σχέση με τον πραγματικό χρόνο πραγματοποίησής τους.

6.1 Επιλογή και οργάνωση πειραμάτων

6.1.1 Επιλογή συνόλου δεδομένων για πειραματισμό

Για το πείραμα της αξιολόγησης του συστήματος δημιουργήθηκε δείγμα δεδομένων μεγέθους 10,996 tweets το οποίο προέκυψε από συγκέντρωση tweets την περίοδο 6 Ιουνίου ως 17 Ιουλίου του έτους 2017 με βάση τις λέξεις – κλειδιά που αναφέρθηκαν στην ενότητα 4.1.2. Το δείγμα αυτό μπορεί να χωριστεί σε δύο υποσύνολα, οποιοδήποτε εκ των οποίων μπορεί να χρησιμοποιηθεί είτε ως σύνολο εκπαίδευσης (training set), είτε ως σύνολο δοκιμής (test set).

6.1.2 Οργάνωση πειράματος για αξιολόγηση ταξινόμησης

Αρχικά, εκπαιδεύουμε τρεις διαφορετικούς ταξινομητές – RBF kernel SVM, linear kernel SVM και Multinomial Naive Bayes – με το συνολικό δείγμα δεδομένων των 10,996 tweets και επιλέγεται ο καταλληλότερος ταξινομητής με βάση τις μετρικές αξιολόγησης του καθενός για μια συγκεκριμένη ομάδα χαρακτηριστικών.

Στη συνέχεια, ο επιλεγμένος ταξινομητής εκπαιδεύεται με τα δεδομένα του δείγματος εκπαίδευσης για διάφορες ομάδες χαρακτηριστικών, ώστε να βρεθεί η ομάδα εξ αυτών που αποδίδει καλύτερα στην ταξινόμηση νέων δεδομένων.Για τις ομάδες χαρακτηριστικών που αναφέρθηκαν στην ενότητα 4.3 υπολογίστηκαν τα αποτελέσματα της ταξινόμησης και αξιολογήθηκαν με βάση τις μετρικές εκτίμησης.

Τα αποτελέσματα που παρουσιάζονται έπονται όλων των βημάτων προεπεξεργασίας που έχουν παρουσιαστεί.

6.1.3 Οργάνωση πειράματος για αξιολόγηση ειδοποιητήριου συστήματος

Το συνολικό δείγμα δεδομένων του πειράματος, με βάση την εξέτασή του, περιέχει αναφορές σε τουλάχιστον 5 σεισμούς μικρής έντασης ως προς το επίπεδο συζήτησής τους στο μέσο του Twitter και τουλάχιστον 3 σεισμούς μεγάλης έντασης. Παρουσιάζουμε στην ενότητα αυτή διαγράμματα που παρουσιάζουν το πλήθος των δημοσιευμένων αναφορών θετικής κλάσης στη μονάδα του χρόνου απ' όπου εποπτικά γίνεται αντιληπτός ο σχολιασμός γεγονότων σεισμών στο μέσο και συζητείται η επίδοση του υποσυστήματος ειδοποίησης όσον αφορά τον εντοπισμό αυτών.

6.2 Παρουσίαση αποτελεσμάτων

6.2.1 Επιλογή ταξινομητή

Κατόπιν της εκπαίδευσης των τριών ταξινομητών προς σύγκριση, RBF kernel SVM, linear kernel SVM και Multinomial Naive Bayes, παρουσιάζουμε παρακάτω ορισμένα στιγμιότυπα από την εκτέλεση της διαδικασίας εκπαίδευσης και τον υπολογισμό των μετρικών αξιολόγησής τους ως προς την ικανότητα ταξινόμησης των δεδομένων.

Η εκτίμηση των μετρικών των ταξινομητών προκύπτει με τη μέθοδο της διασταυρούμενης επικύρωσης k-στρώσεων (k-fold validation). Η εκμάθηση και δοκιμή ενός ταξινομητή στα ίδια δεδομένα θεωρείται μεθοδολογικό λάθος μιας και θα αδυνατούσε να προβλέψει την κλάση νέων δεδομένων που δεν έχουν εξεταστεί προηγουμένως [52]. Η περίπτωση αυτή αντιστοιχεί στο πρόβλημα της υπερπροσαρμογής (overfitting), δηλαδή στον κίνδυνο να προστεθεί στα δεδομένα θόρυβος μέσω της απομνημόνευσης ιδιαίτερων περιπτώσεων στα δεδομένα εκπαίδευσης αντί ενός γενικού κανόνα πρόβλεψης της κλάσης δεδομένων [53]. Για να αποφύγουμε την υπερπροσαρμογή, αφαιρούμε από το σύνολο των δεδομένων που προορίζονται για εκπαίδευση (10,996 tweets) ένα μέρος το οποίο το θεωρούμε ως σύνολο

δεδομένων δοκιμής. Έτσι, υλοποιείται ένας διαχωρισμός του συνόλου δεδομένων εκπαίδευσης σε ένα υποσύνολο εκπαίδευσης και ένα υποσύνολο δοκιμής. Η τελευταία μέθοδος ονομάζεται holdout και η λειτουργία της βασίζεται στην εκπαίδευση του ταξινομητή στο υποσύνολο εκπαίδευσης και στη δοκιμή του στο αντίστοιχο υποσύνολο. Τα αποτελέσματα της μεθόδου εξαρτώνται από το διαχωρισμό που γίνεται εφόσον διαφορετικοί διαχωρισμοί οδηγούν σε διαφορετικά υποσύνολα εκπαίδευσης.

Κατ' αναλογία, η μέθοδος διασταυρούμενης επικύρωσης k-στρώσεων αποτελεί μια πιο σύνθετη υλοποίηση της holdout μεθόδου, η οποία πρακτικά είναι μέθοδος διασταυρούμενης επικύρωσης 2-στρώσεων. Ένα σύνολο δεδομένων εκπαίδευσης χωρίζεται σε k υποσύνολα και η μέθοδος holdout επαναλαμβάνεται k φορές όπου σε κάθε επανάληψη μόνο ένα εκ των υποσυνόλων θεωρείται ως δείγμα ελέγχου-δοκιμής και η ένωση των υπόλοιπων k-1συνιστούν το δείγμα εκπαίδευσης. Τελικά, η αξιολόγηση του ταξινομητή προκύπτει ως μέσος όρος των επιμέρους αποτελεσμάτων.

Σε όλα τα παρακάτω πειράματα που αφορούν την απόφαση επιλογής κατάλληλου αλγορίθμου ταξινόμησης, η ομάδα χαρακτηριστικων που χρησιμοποιείται είναι η B που αναφέρεται στην ενότητα 4.3. Ο αρχικός αλγόριθμος ταξινόμησης που εφαρμόστηκε είναι ο SVM με γραμμικό πυρήνα όπως περιγράφεται στο [12]. Στις παραμέτρους του αλγορίθμου ορίστηκε να δίνεται περισσότερο έμφαση στη θετική κλάση έτσι ώστε να αντιμετωπιστούν τα γνωστά προβλήματα του μη-ισορροπημένου συνόλου δεδομένων. Στην Εικόνα 6.1 εμφανίζονται τα αποτελέσματα των μετρικών ακρίβειας, ανάκλησης και βαθμολογίας–F για διασταυρούμενη επικύρωση 10-στρώσεων καθώς και ο μέσος όρος των προαναφερθέντων μετρικών.

```
C:\Users\user\Anaconda3\python.exe "C:/Users/user/PycharmProjects/Thesis/Classification Testing (final).py"
Precision: [ 0.24434389  0.37714286  0.26356589  0.60377358  0.40588235  0.43884892
  0.11392405  0.3006993   0.4952381   0.29059829]
Recall: [ 0.675       0.825       0.85        0.8         0.8625      0.7625
  0.34177215  0.5443038   0.65822785  0.43037975]
F-Measure: [ 0.35880399  0.51764706  0.40236686  0.68817204  0.552       0.55707763
  0.17088608  0.38738739  0.56521739  0.34693878]
Average Precision: 0.353402
Average Recall: 0.674968
Average F-Measure: 0.454650

Process finished with exit code 0
```

Εικόνα 6.1

Όπως παρατηρούμε η ανάκληση κατά μέσο όρο είναι στο 67.5%, κάτι που είναι σχετικά ικανοποιητικό, ωστόσο η ακρίβεια του αλγορίθμου είναι στη μέση περίπτωση στο 35%. Αυτά μεταφράζονται στο ότι ο αλγόριθμος ταξινόμησης βρίσκει κυρίως tweets που σχετίζονται με

σεισμό αλλά δεν μπορεί να αναγνωρίσει όλα τα διαφορετικά είδη tweet που αναφέρονται σε σεισμό.

Στην προσπάθεια μας να βελτιώσουμε την απόδοση του συστήματος ταξινόμησης, ο επόμενος αλγόριθμος που δοκιμάσαμε ήταν πάλι ένας SVM αλλα με πυρήνα RBF με την προοπτική να εκμεταλευτούμε την ευελιξία ενός τέτοιου πυρήνα. Έμφαση δόθηκε ξανά στη θετική κλάση, όπως και στην προηγούμενη περίπτωση και η μέθοδος αξιολόγησης παρέμεινε ίδια με τον SVM γραμμικού πυρήνα. Τα αποτελέσματα παρατίθενται στην Εικόνα 6.2.

```
C:\Users\user\Anaconda3\python.exe "C:/Users/user/PycharmProjects/Thesis/Classification Testing (final).py"
Precision: [ 0.62068966  0.44311377  0.45341615  0.83529412  0.75        0.74285714
  0.97183099  0.98507463  0.87837838  0.72941176]
Recall: [ 0.9         0.925       0.9125      0.8875      0.8625      0.975
  0.87341772  0.83544304  0.82278481  0.78481013]
F-Measure: [ 0.73469388  0.59919028  0.60580913  0.86060606  0.80232558  0.84324324
  0.92        0.90410959  0.8496732   0.75609756]
Average Precision: 0.741007
Average Recall: 0.877896
Average F-Measure: 0.787575

Process finished with exit code 0
```

Εικόνα 6.2

Τα αποτελέσματα του πειράματος εμφάνισαν αισθητά σημαντική βελτίωση συγκριτικά με τον προηγούμενο αλγόριθμο και όπως είναι λογικό στην περίπτωση της μηχανής διανυσμάτων υποστήριξης αγλορίθμου επιλέγεται ο RBF πυρήνας έναντι του γραμμικού.

Ένας ακόμα αλγόριθμος ταξινόμησης που χρησιμοποιήθηκε είναι ο πολυωνυμικός ταξινομητής του Naive Bayes, ο οποίος είναι ευρέως χρησιμοποιούμενος για κατηγοριοποίηση διακριτών χαρακτηριστικών. Για μία ακόμα φορά, χρησιμοποιήσαμε την ίδια μέθοδο αξιολόγησης με αυτή των προαναφερθέντων αλγορίθμων και τα αποτελέσματα φαίνονται στην Εικόνα 6.3.

```
C:\Users\user\Anaconda3\python.exe "C:/Users/user/PycharmProjects/Thesis/Classification Testing (final).py"
Precision: [ 0.24886878  0.32512315  0.26436782  0.55462185  0.40828402  0.33888889
  0.24875622  0.31690141  0.41935484  0.29059829]
Recall: [ 0.6875      0.825       0.8625      0.825       0.8625      0.7625
  0.63291139  0.56962025  0.65822785  0.43037975]
F-Measure: [ 0.3654485   0.4664311   0.40469208  0.66331658  0.55421687  0.46923077
  0.35714286  0.40723982  0.51231527  0.34693878]
Average Precision: 0.341577
Average Recall: 0.711614
Average F-Measure: 0.454697

Process finished with exit code 0
```

Εικόνα 6.3

Τα αποτελέσματα αυτά δεν είναι ιδιαίτερα ενθαρρυντικά και για το λόγο αυτό απορρίφθηκε ως επιλογή ο εν λόγω ταξινομητής. Τα προβλήματα που εμφανίζονται είναι ανάλογα με του SVM γραμμικού πυρήνα. Στον Πίνακας 6.1 παρουσιάζονται τα αποτελέσματα των μετρικών για τους εξεταζόμενους αλγορίθμους ταξινόμησης σε ποσοστά επί τοις εκατό, όπου είναι φανερή η υπεροχή του ταξινομητή SVM με πυρήνα RBF έναντι των υπολοίπων.

Ομάδα χαρακτηριστικών B	Ακρίβεια	Ανάκληση	Βαθμολογία - F
Linear kernel SVM	35.3	67.5	45.5
RBF kernel SVM	74.1	87.8	78.6
Multinomial Naive Bayes	34.2	71.1	45.5

Πίνακας 6.1

6.2.2 Αξιολόγηση ταξινομητή SVM πυρήνα RBF

Η επόμενη παράμετρος του συστήματος ταξινόμησης με την οποία πειραματιστήκαμε είναι τα χαρακτηριστικά του ταξινομητή. Ο ταξινομητής που θα χρησιμοποιηθεί στα επόμενα είναι ο SVM με πυρήνα RBF με βάση τα προηγούμενα. Οι διάφορες ομάδες χαρακτηριστικών που ελέγχθηκαν περιγράφονται στην ενότητα 4.3. Στα επόμενα πειράματα η μέθοδος αξιολόγησης που επιλέγχθηκε είναι αυτή της διασταυρούμενης επικύρωσης 10-στρώσεων με τις μετρικές ακρίβειας, ανάκλησης και βαθμολογίας–F.

Η ομάδα Α των χαρακτηριστικών προς πειραματισμό είναι αυτή που αναλύεται στο [12]. Ως χαρακτηριστικά επιλέγονται το πλήθος των λέξεων του tweet και η θέση μιας λέξης-κλειδί όπως «σεισμός», «ρίχτερ», «εγκέλαδος» και «σεισμική δόνηση» εντός του tweet. Η λογική πίσω από αυτή την ομάδα στατιστικών χαρακτηριστικών είναι πως μικρά σε μηκος tweet που περιλαμβάνουν λέξεις σχετικές με το γεγονός του σεισμού σε συγκεκριμένες θέσεις εντός του tweet (π.χ. στην αρχή του tweet) είναι πιο πιθανό να αναφέρουν έναν αληθινό σεισμό. Τα αποτελέσματα αυτής της ομάδας χαρακτηριστικών φαίνεται στην Εικόνα 6.4.

```
C:\Users\user\Anaconda3\python.exe "C:/Users/user/PycharmProjects/Thesis/Classification Testing (final).py"
Precision: [ 0.56034483  0.38983051  0.40935673  0.69565217  0.62886598  0.56923077
  0.91566265  0.97222222  0.8875      0.66666667]
Recall: [ 0.8125      0.8625      0.875       0.8         0.7625      0.925
  0.96202532  0.88607595  0.89873418  0.83544304]
F-Measure: [ 0.66326531  0.53696498  0.55776892  0.74418605  0.68926554  0.7047619
  0.9382716   0.92715232  0.89308176  0.74157303]
Average Precision: 0.669533
Average Recall: 0.861978
Average F-Measure: 0.739629

Process finished with exit code 0
```

Εικόνα 6.4

Η απόδοση των χαρακτηριστικών φαίνεται ικανοποιητική με βάση τις υπολογισμένες μετρικές, ωστόσο εξαιτίας της υψηλής ανάκλησης, tweets όπως «ΠΟΛΙΤΙΚΟΣ ΣΕΙΣΜΟΣ στο Κόσσοβο από ενδεχόμενη επικράτηση της συμμαχίας εξτρεμιστών…» τείνουν να αντιστοιχίζονται στη θετική κλάση, γεγονός το οποίο μας οδηγεί σε αναζήτηση νέων ομάδων χαρακτηριστικών που θα κατηγοριοποιούν τα tweets πιο αποτελεσματικά.

Αξιοποιώντας και τα υπόλοιπα χαρακτηριστικά που έλεγξαν οι συγγραφείς του [12], δοκιμάσαμε την λειτουργικότητα της ομάδας Γ των χαρακτηριστικών της ενότητας 4.3, η οποία πρόκειται ουσιαστικά για τη χρήση των ίδιων των λέξεων του κειμένου ενός tweet ως χαρακτηριστικά μέσω της tf-idf μετατροπής. Η Εικόνα 6.5 περιλαμβάνει τα αποτελέσματα της εν λόγω δοκιμής.

```
C:\Users\user\Anaconda3\python.exe "C:/Users/user/PycharmProjects/Thesis/Classification Testing (BoW).py"
Precision: [ 0.83673469  0.64285714  0.62745098  0.90909091  0.85507246  0.81818182
  0.88372093  0.88235294  0.88636364  0.82926829]
Recall: [ 0.5125      0.9         0.8         0.75        0.7375      0.7875
  0.48101266  0.75949367  0.49367089  0.43037975]
F-Measure: [ 0.63565891  0.75        0.7032967   0.82191781  0.79194631  0.80254777
  0.62295082  0.81632653  0.63414634  0.56666667]
Average Precision: 0.817109
Average Recall: 0.665206
Average F-Measure: 0.714546

Process finished with exit code 0
```

Εικόνα 6.5

Να σημειώσουμε εδώ πως η συγκεκριμένη ομάδα χαρακτηριστικών δοκιμάστηκε με τον πολυωνυμικό ταξινομητή του Naive Bayes ώστε να επιτευχθεί καλύτερη επίδοση. Παρόλο που η απόδοση των χαρακτηριστικών αυτών μοιάζει ικανοποιητική κρίνοντας από τις τιμές των τριών μετρικών, εντούτοις μεγαλύτερη ανάκληση έναντι ακρίβειας είναι προτιμητέα αφού τα μειονεκτήματα της μειωμένης ακρίβειας αντισταθμίζονται από το σύστημα ειδοποίησης ενώ μειωμένη τιμή ανάκλησης συνεπάγεται μικρότερο ποσοστό εύρεσης των πραγματικών αναφορών σεισμών εντός του δείγματός μας.

Για να αντιμετωπίσουμε τις αδυναμίες των προηγούμενων ομάδων χαρακτηριστικών εξετάσαμε τις επιδόσεις της ομάδας Β των χαρακτηριστικών ταξινόμησης. Τα στατιστικά αυτά χαρακτηριστικά είναι ίδια με της ομάδας Α με την προσθήκη ενός νέου στοιχείου το οποίο είναι η θέση μιας λέξης που αναφέρεται σε περιοχή στην Ελλάδα εντός του κειμένου του tweet, όπως αναλύθηκε στην ενότητα 3.3. Η φιλοσοφία αυτής της επιλογής είναι πως ένα tweet που περιλαμβανει λέξεις-κλειδια όπως «σεισμός», «ρίχτερ», «εγκέλαδος» και «σεισμική δόνηση». και αναφέρεται σε μια ελληνική περιοχή είναι πιο πιθανό να περιγράφει έναν πραγματικό σεισμό στην Ελλάδα συγκριτικά με ένα tweet που περιλαμβάνει απλά τις προαναφερθείσες λέξεις-κλειδιά. Στην Εικόνα 6.6 παραθέτουμε τις επιδόσεις του ταξινομητή με την νέα ομάδα χαρακτηριστικών.

```
C:\Users\user\Anaconda3\python.exe "C:/Users/user/PycharmProjects/Thesis/Classification Testing (final).py"
Precision: [ 0.62608696  0.52112676  0.45341615  0.83529412  0.75        0.75
  0.97222222  0.98529412  0.88        0.73255814]
Recall: [ 0.9         0.925       0.9125      0.8875      0.8625      0.975
  0.88607595  0.84810127  0.83544304  0.79746835]
F-Measure: [ 0.73846154  0.66666667  0.60580913  0.86060606  0.80232558  0.84782609
  0.92715232  0.91156463  0.85714286  0.76363636]
Average Precision: 0.750600
Average Recall: 0.882959
Average F-Measure: 0.798119

Process finished with exit code 0
```

Εικόνα 6.6

Η απόδοση του ταξινομητή παρουσιάζει αισθητή βελτίωση σε σχέση με τις προηγούμενες ομάδες χαρακτηριστικών μετά την εισαγωγή του νέου στατιστικού στοιχείου της περιοχής. Έτσι, συμπεραίνουμε με ασφάλεια πως η προσθήκη του συγκεκριμένου χαρακτηριστικού εξασφαλίζει έναν πιο αξιόπιστο ταξινομητή για το σύστημά μας.

Το τελευταίο πείραμα των χαρακτηριστικών κατηγοριοποίησης που πραγματοποιήθηκε αφορούσε την ομάδα Δ της ενότητας 4.3 η οποία συγκροτήθηκε για να ελέγξουμε το κατά πόσο επηρεάζουν τον ταξινομητή μας οι θέσεις εντός του κειμένου του tweet της λέξης-κλειδί και της λέξης αναφερόμενης σε ελληνική περιοχή. Πρακτικά, η διαφορά αυτής της ομάδας με την ομάδα Β είναι ότι ως στατιστικά στοιχεία πλέον δεν χρησιμοποιούνται οι θέσεις των προαναφερθέντων λέξεων εντός του tweet αλλά το αν αυτές υπάρχουν ή όχι μέσα στο tweet. Η δοκιμή του ταξινομητή με αυτή την ομάδα χαρακτηριστικών παρουσιάζεται στην Εικόνα 6.7.

```
C:\Users\user\Anaconda3\python.exe "C:/Users/user/PycharmProjects/Thesis/Classification Testing (final).py"
Precision: [ 0.92771084  0.61904762  0.55555556  0.87209302  0.74757282  0.76        1.
  0.98360656  0.95588235  0.73684211]
Recall: [ 0.9625      0.975       0.9375      0.9375      0.9625      0.95
  0.82278481  0.75949367  0.82278481  0.70886076]
F-Measure: [ 0.94478528  0.75728155  0.69767442  0.90361446  0.84153005  0.84444444
  0.90277778  0.85714286  0.88435374  0.72258065]
Average Precision: 0.815831
Average Recall: 0.883892
Average F-Measure: 0.835619

Process finished with exit code 0
```

Εικόνα 6.7

Η εικόνα μας δείχνει ότι η θέση της λέξης-κλειδί και της τοποθεσίας όχι μόνο δεν παίζει σημαντικό ρόλο για τον ταξινομητή αλλά μείωνει ελαφρώς την απόδοσή του. Τα συμπεράσματα αυτά φαίνονται να συμβαδίζουν με την πραγματικότητα αφού εξετάζοντας το σύνολο δεδομένων μας διαπιστώνουμε ότι tweets τα οποία αναφέρουν πραγματικό σεισμό μπορούν να περιέχουν τη λέξη-κλειδί και την τοποθεσία σε διάφορες θέσεις μέσα στο κείμενο του tweet. Από τα τελευταία αποτελέσματα καταλήγουμε πως η καταλληλότερη ομάδα χαρακτηριστικών είναι η Δ την οποία και θα χρησιμοποιήσουμε για το τελικό μας σύστημα. Στον Πίνακας 6.2 έχουν συγκεντρωθεί οι τιμές των μετρικών αξιολόγησης για όλες τις εξεταζόμενες ομάδες χαρακτηριστικών.

Ομάδα χαρακτηριστικών	Ακρίβεια	Ανάκληση	Βαθμολογία - F
Α	67	86.2	74
Β	75	88.3	79.8
Γ	81.7	66.5	71.5
Δ	81.6	88.4	83.6

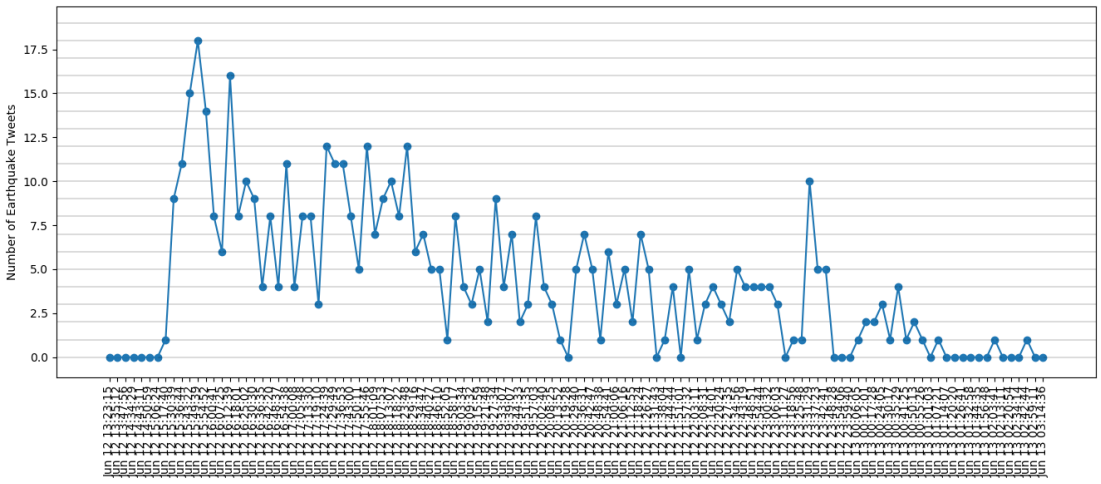
Πίνακας 6.2

6.2.3 Αξιολόγηση συστήματος ειδοποίησης

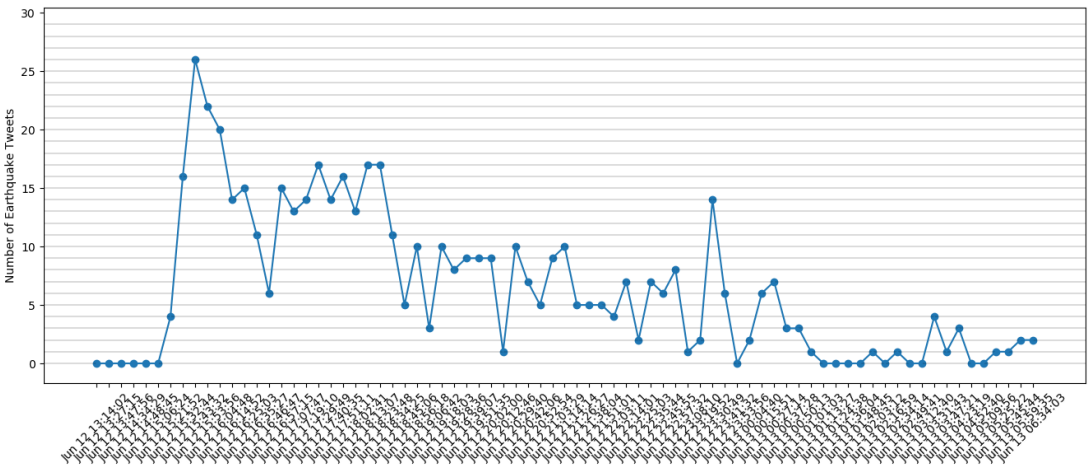
Για την καλύτερη κατανόηση της χρονικής εξέλιξης των tweets που περιγράφουν έναν πραγματικό σεισμό παραθέτουμε μια σειρά από στιγμιότυπα που παρουσιάζουν το πλήθος των tweets στη μονάδα του χρόνου που έχουμε χειροκίνητα ταξινομήσει στη θετική κλάση.

Επομένως, τα tweets αυτά αναφέρονται σε γεγονός σεισμού και συγκεντρώνονται σε παράθυρα (κατά προσέγγιση) πέντε και δέκα λεπτών.

Τα επόμενα στιγμιότυπα καλύπτουν το χρονικό διάστημα 12/06/2017 και ώρα 13:23:15 έως 13/06/2017 και ώρα 03:24:36 όπου κυρίως αναφέρεται ο ισχυρός σεισμός που έγινε στο νησί της Λέσβου στις 12 Ιουνίου 2017 για χρονικά παράθυρα διάρκειας πέντε (*Εικόνα* 6.8) και δέκα λεπτών (*Εικόνα* 6.9) αντίστοιχα.

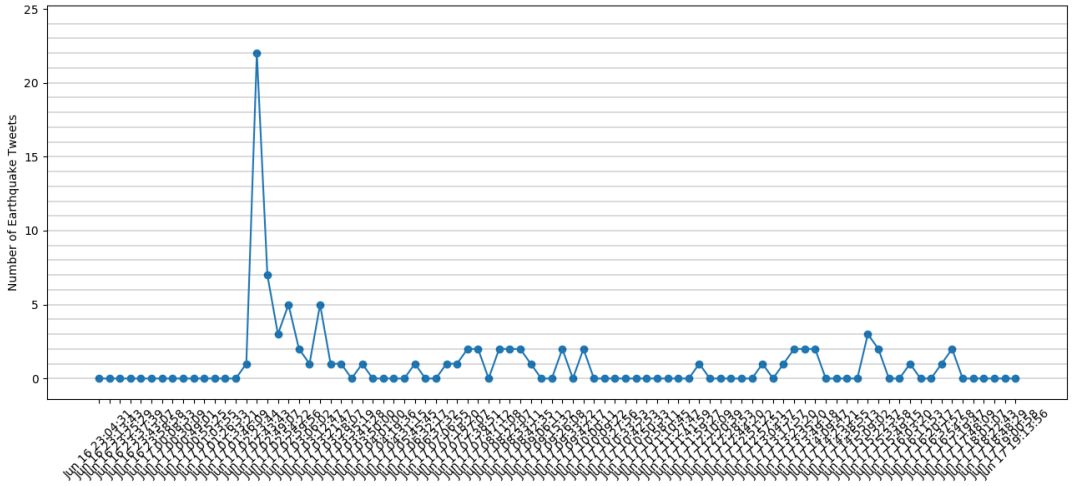


Εικόνα 6.8

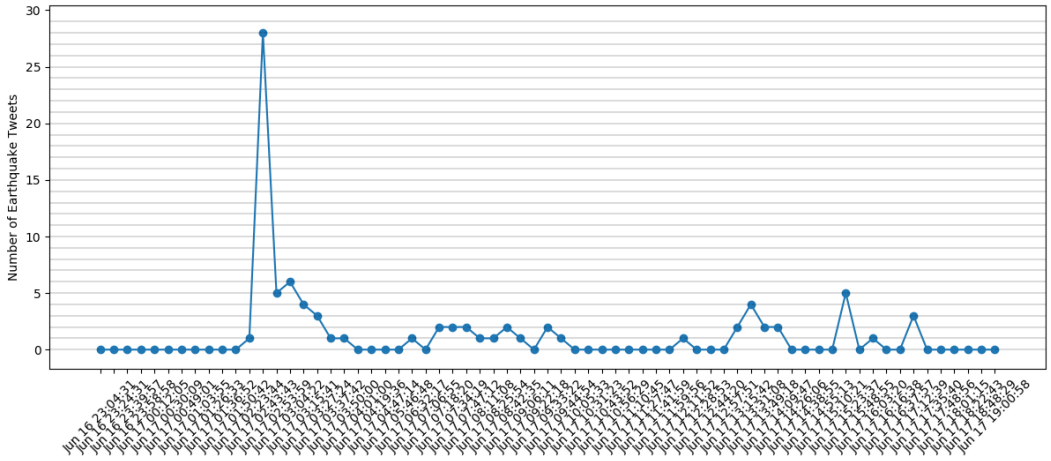


Εικόνα 6.9

Οι παρακάτω εικόνες δείχνουν την κινητικότητα του Twitter στο χρονικό διάστημα 16/06/2017 και ώρα 23:04:31 έως 17/06/2017 και ώρα 19:23:56 σχετικά με τον έντονο σεισμό που συνέβη στην Αθήνα για χρονικά παράθυρα διάρκειας πέντε (*Εικόνα* 6.10) και δέκα λεπτών (*Εικόνα* 6.11) αντίστοιχα.

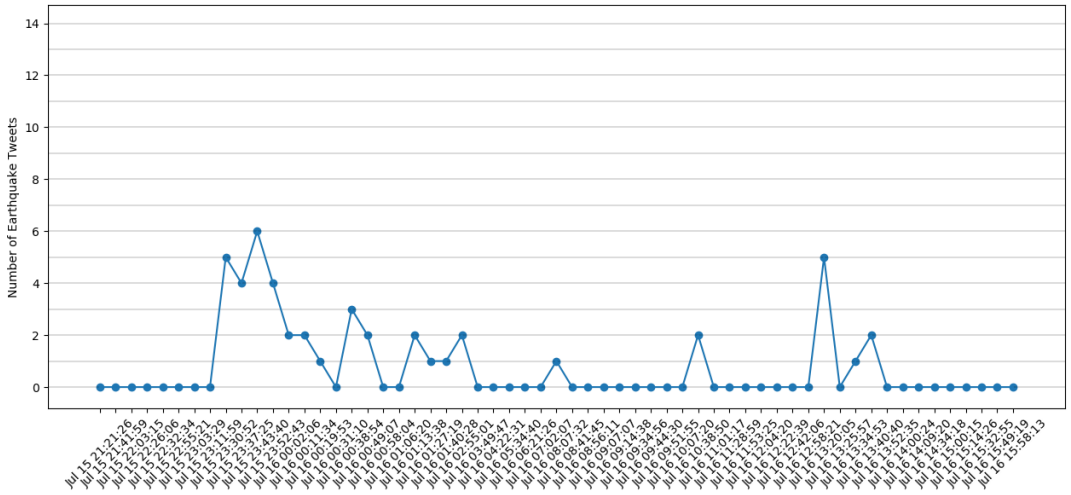


Εικόνα 6.10

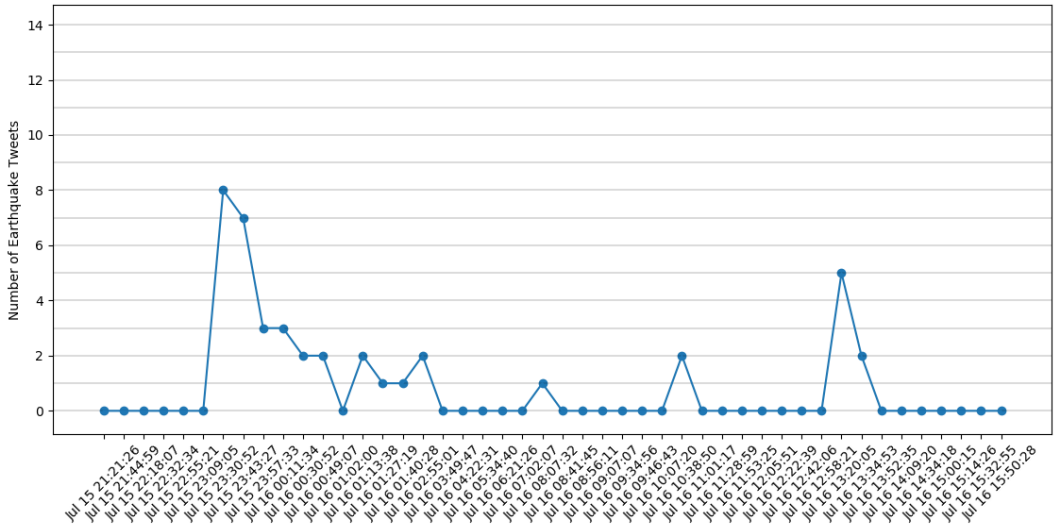


Εικόνα 6.11

Οι επόμενες γραφικές παράστασεις αναφέρονται στο διάστημα μεταξύ 15/07/2017 και ώρα 21:21:26 έως 16/07/2017 και ώρα 16:08:13 και δείχνει τις αναφορές των tweets για τον μεσαίας κλίμακας σεισμό που πραγματοποιήθηκε τα μεσάνυχτα της 15ης Ιουλίου 2017 στην Κρήτη για χρονικά παράθυρα διάρκειας πέντε (*Εικόνα* 6.12) και δέκα λεπτών (*Εικόνα* 6.13) αντίστοιχα.

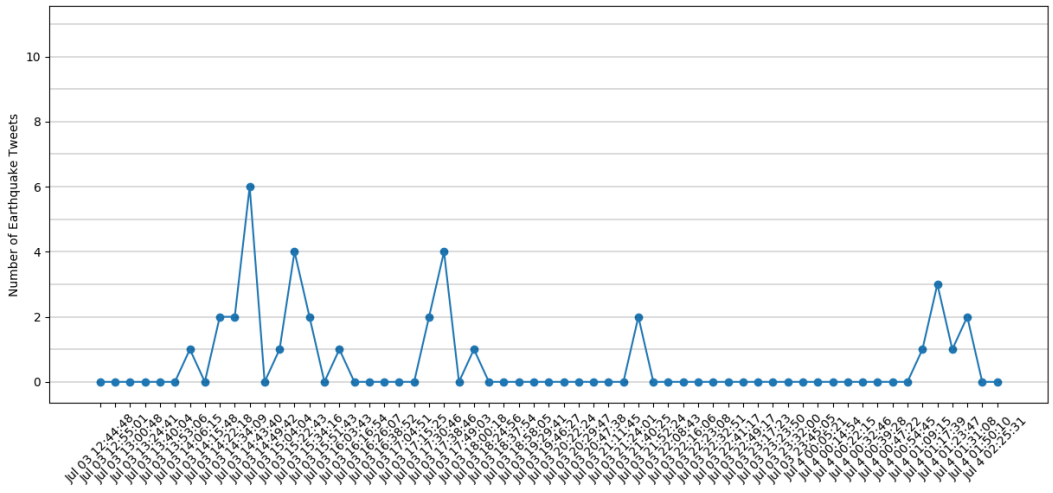


Εικόνα 6.12

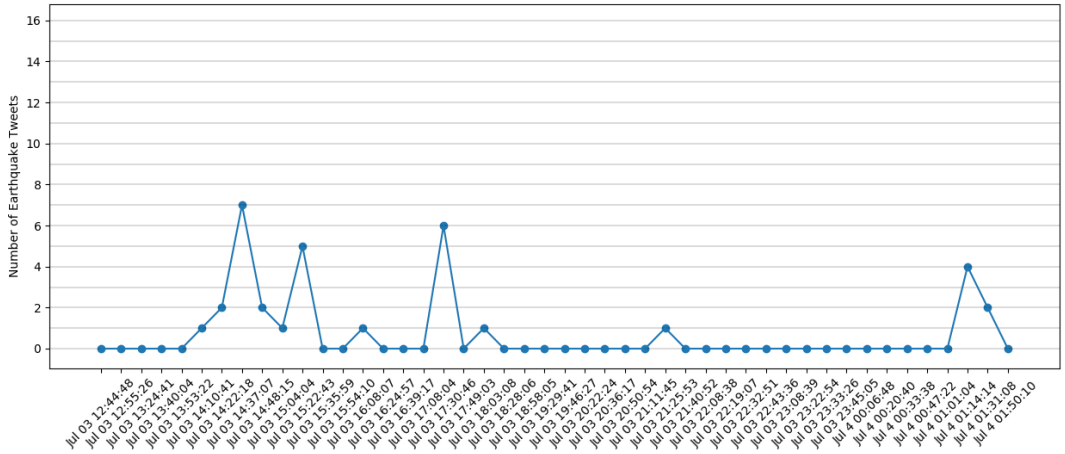


Εικόνα 6.13

Στην ίδια λογική με τα προηγούμενα, παραθέτουμε παρακάτω την ανταπόκριση των χρηστών του Twiter για το σεισμό που έγινε στη Φλώρινα στις 3 Ιουλίου 2017 μεταξύ των χρονικών διαστημάτων που φαίνονται στις εικόνες για χρονικά παράθυρα διάρκειας πέντε (*Εικόνα* 6.14) και δέκα (*Εικόνα* 6.15) λεπτών αντίστοιχα.

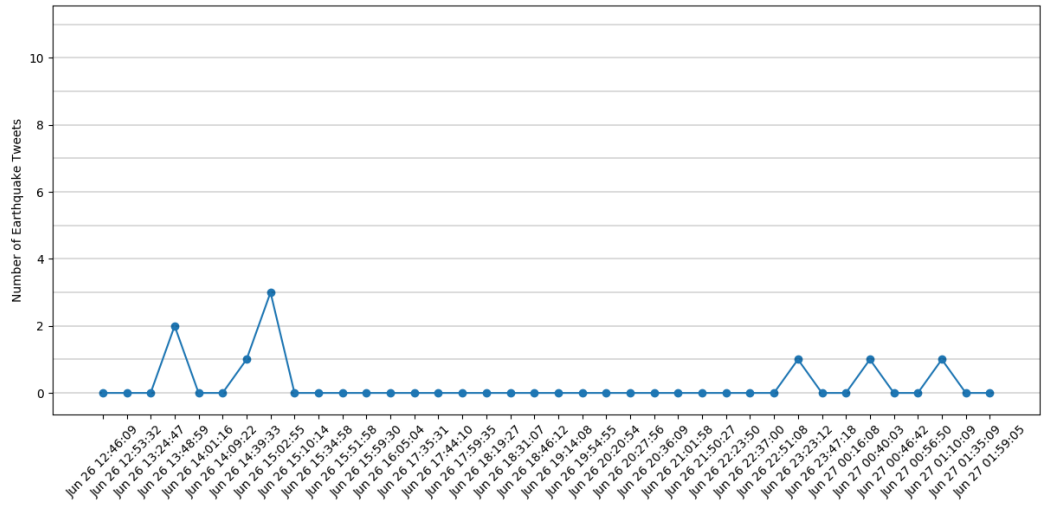


Εικόνα 6.14

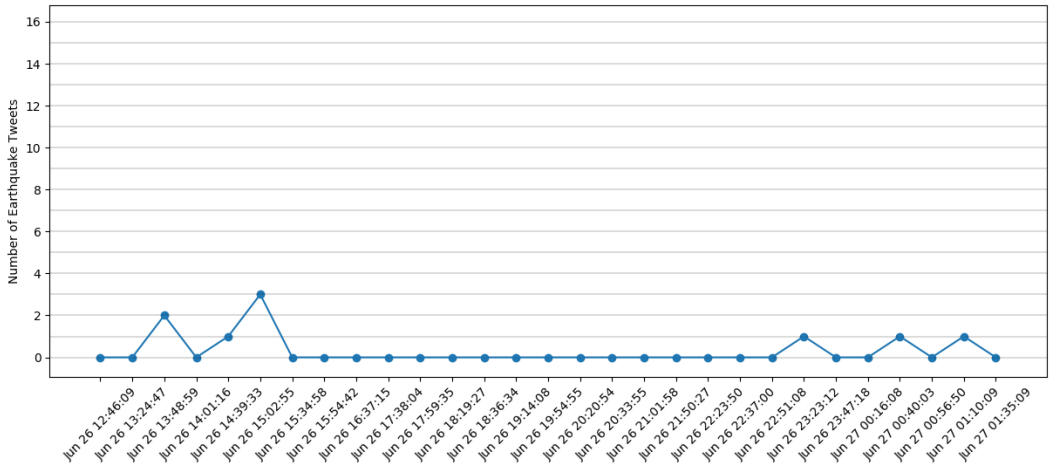


Εικόνα 6.15

Ο τελευταίος σεισμός του οποίου οπτικοποιήσαμε τις σχετικές αναφορές σε tweet είναι ο μικρής κλίμακας σεισμός της Κεφαλονιάς στις 26 Ιουνίου 2017 για χρονικά παράθυρα διάρκειας πέντε (*Εικόνα* 6.16) και δέκα (*Εικόνα* 6.17) λεπτών αντίστοιχα.



Εικόνα 6.16



Εικόνα 6.17

Όλα τα προηγούμενα στιγμιότυπα παρουσιάστηκαν για να δοθεί μια ιδέα περί της χρονικής συμπεριφοράς των tweets που αναφέρουν σεισμό. Το πλεονέκτημα αυτής της οπτικοποίησης είναι πως δίνει μια ξεκάθαρη εικόνα για πόσα tweets αναμένονται να εμφανιστούν σε περίπτωση σεισμού ανάλογα με το μέγεθος του σεισμού για διάφορες χρονικές μονάδες. Όπως αναφέραμε ήδη, οι χρονικές αυτές μονάδες στις οποίες μετρώνται τα True tweets είναι διάρκειας πέντε λεπτών και δέκα λεπτών. Οι χρόνοι αυτοί δεν είναι αυθαίρετοι και σχετίζονται άμεσα με την συνθήκη ενεργοποίησης του συστήματος ειδοποίησης. Στα πλαίσια του

πειραματισμού εξετάστηκαν χρονικά παράθυρα διάρκειας 5 έως 10 λεπτών αφού κρίναμε πως έχει νόημα να εντοπίζουμε τα γεγονότα εντός διαστημάτων σχετικά μικρής διάρκειας μετά την πραγματοποίηση τους.

Στην ενότητα 3.2 που περιγράψαμε την υλοποίηση του σύστηματος ειδοποίησης καταλήξαμε μέσω της θεωρητικής ανάλυσης ότι απαιτούνται τουλαχιστον δέκα tweets σε διάστημα το πολύ δέκα λεπτών για να εξασφαλίσουμε πως ο συναγερμός ενεργοποιείται σε περίπτωση σεισμού με μεγάλη πιθανότητα. Εξετάζοντας τη χρονική συμπεριφορά των tweets όταν γίνεται σεισμός μπορούμε ελέγξουμε την ορθότητα του παραπάνω ισχυρισμού καθώς και να δοκιμάσουμε άλλα κατώφλια για το πόσα True tweets έρχονται τα οποία ενδεχομένως να είναι περισσότερο λειτουργικά. Η συνθήκη ενεργοποίησης στη γενική της μορφή είναι **«N tweets σε t λεπτά»** και πειραματιζόμαστε μεταβάλλοντας κάθε φορά το N που δηλώνει πλήθος tweet και κατηγοριοποιήθηκαν ως True και t είναι η χρονική μονάδα που επιλέξαμε.

Η πρώτη περίπτωση που θα μελετήσουμε είναι αυτή της θεωρητικής προσέγγισης, δηλαδή με συνθήση ενεργοποίησης «δέκα tweets σε διάστημα δέκα λεπτών». Για το λόγο αυτό θα μοιράσουμε το σύνολο δεδομένων σε δύο ομάδες, την πρώτη που περιέχει δύο μεγάλους σε ένταση σεισμούς (38% του αρχικού συνόλου δεδομένων) και τη δεύτερη που περιέχει άλλους μικρότερους (62% του αρχικού συνόλου δεδομένων). Θα πραγμοτοποιήσουμε δύο δοκιμές όπου οι προαναφερθείσες ομάδες θα χρησιμοποιηθούν ως σύνολο εκπαίδευσης και ως σύνολο δοκιμής εναλλάξ. Στα επόμενα παρουσιάζονται τα αποτελέσματα των πειραμάτων δίνοντας κατάλληλα στιγμιότυπα.

<u>1η περίπτωση: «δέκα tweets σε διάστημα δέκα λεπτών»</u>

- Σύνολο δεδομένων εκπαίδευσης: μικροί σεισμοί, σύνολο δεδομένων δοκιμής: μεγάλοι σεισμοί (*Εικόνα* 6.18).

```
C:\Users\user\Anaconda3\python.exe "C:/Users/user/PycharmProjects/Thesis/Event Detection (Simulation).py"
Σεισμός τώρα! Περιοχές: Ελλαδα, Κομοτηνη, Σαμοθρακη. Ωρα: Jun 12 15:37:42
Σεισμός τώρα! Περιοχές: Μυτιληνη, Αθηνα, Χιος, Ελλαδα. Ωρα: Jun 12 15:44:14
Σεισμός τώρα! Περιοχές: Μυτιληνη. Ωρα: Jun 12 15:49:17
Σεισμός τώρα! Περιοχές: Μυτιληνη, Πλωμαρι, Θεσσαλονικη, Χιος. Ωρα: Jun 12 15:53:25
Σεισμός τώρα! Περιοχές: Μυτιληνη, Χιος, Λεσβος, Θεσσαλονικη. Ωρα: Jun 12 15:55:38
Σεισμός τώρα! Περιοχές: Μυτιληνη, Λεσβος, Αιγαίο. Ωρα: Jun 12 15:58:40
Σεισμός τώρα! Περιοχές: Μυτιληνη, Λεσβος, Χιος. Ωρα: Jun 12 16:05:35
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη, Χιος. Ωρα: Jun 12 16:13:03
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη, Πλωμαρι, Χιος. Ωρα: Jun 12 16:16:15
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη, Πλωμαρι. Ωρα: Jun 12 16:25:21
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη, Χιος. Ωρα: Jun 12 16:32:49
Σεισμός τώρα! Περιοχές: Λεσβος, Αιγαίο. Ωρα: Jun 12 16:38:59
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 16:59:34
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη, Ελλαδα. Ωρα: Jun 12 17:27:38
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 17:31:44
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 17:37:55
Σεισμός τώρα! Περιοχές: Λεσβος, Ελλαδα. Ωρα: Jun 12 17:46:02
Σεισμός τώρα! Περιοχές: Λεσβος, Κλημα, Ελλαδα. Ωρα: Jun 12 17:54:52
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη, Χιος. Ωρα: Jun 12 17:58:38
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 18:08:11
Σεισμός τώρα! Περιοχές: Λεσβος, Αθηνα, Μυτιληνη. Ωρα: Jun 12 18:16:39
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 18:24:55
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 18:31:39
Σεισμός τώρα! Περιοχές: Λεσβος, Ελλαδα, Μυτιληνη. Ωρα: Jun 12 18:41:58
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 20:05:20
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 20:39:32
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 22:36:50
Σεισμός τώρα! Περιοχές: Αθηνα. Ωρα: Jun 17 02:44:46
Σεισμός τώρα! Περιοχές: Αθηνα, Χανια. Ωρα: Jun 17 02:47:51
Σεισμός τώρα! Περιοχές: Αθηνα, Καλαματα. Ωρα: Jun 17 02:51:07
Σεισμός τώρα! Περιοχές: Αθηνα. Ωρα: Jun 17 02:57:02
Σεισμός τώρα! Περιοχές: Αθηνα, Μονεμβασια. Ωρα: Jun 17 03:00:55
Σεισμός τώρα! Περιοχές: Λακωνια, Αθηνα, Μονεμβασια, Καλαματα. Ωρα: Jun 17 03:07:17
Σεισμός τώρα! Περιοχές: Αθηνα, Λακωνια, Καλαματα, Μονεμβασια. Ωρα: Jun 17 03:12:46
```

Εικόνα 6.18

- Σύνολο δεδομένων εκπαίδευσης: μεγάλοι σεισμοί, σύνολο δεδομένων δοκιμής: μικροί σεισμοί

Οι σεισμοί μικρότερης έντασης δεν ανιχνεύονται αφού όπως βλέπουμε και στα στιγμιότυπα των δέκα λεπτών, που παρουσιάστηκαν προηγουμένως, εμφανίζονται λιγότερα από δέκα tweets.

Παρατηρούμε πως η συνθήκη ενεργοποίησης του συναγερμού που χρησιμοποιήσαμε λειτουργεί αρκετά ικανοποιητικά για την περίπτωση των μεγάλων σεισμών αφού ειδοποιεί πως συνέβη σεισμός σε λιγότερο από δέκα λεπτά από την πραγματοποίηση του. Ωστόσο, για την περίπτωση των σεισμών μικρής κλίμακας το σύστημα αστόχησε πλήρως και δεν ανίχνευσε κανέναν από αυτούς.

Η επόμενη περίπτωση που εξετάζεται αφορά συνθήκη στην οποία η χρονική μονάδα παραμένει ίδια (δέκα λεπτά) αλλά μειώνεται ο αριθμός των True tweets που χρειάζονται να εμφανιστούν εντός του χρονικού αυτού πλαισίου. Αυτό γίνεται για να μπορέσουμε να ανιχνεύσουμε και σεισμούς μικρότερης κλίμακας αφού είδαμε πως κάτι τέτοιο δεν είναι εφικτό με τη θεωρητική συνθήκη. Παρατηρώντας τα στιγμιότυπα που δείχνουν πόσα True tweets εμφανίστηκαν ανα δέκα λεπτά για μικρότερους σεισμούς ορίζουμε ως νέα συνθήκη ενεργοποίησης την εξής:

«πέντε tweets σε διάστημα δέκα λεπτών». Ακολουθώντας την ίδια μέθοδο πειραματισμού με πριν παρουσιάζουμε τα αποτελέσματα στα επόμενα.

2η περίπτωση: «πέντε tweets σε διάστημα δέκα λεπτών»

- Σύνολο δεδομένων εκπαίδευσης: μικροί σεισμοί, σύνολο δεδομένων δοκιμής: μεγάλοι σεισμοί (*Εικόνα* 6.19)

```
C:\Users\user\Anaconda3\python.exe "C:/Users/user/PycharmProjects/Thesis/Event Detection (Simulation).py"
Σεισμός τώρα! Περιοχές: . Ώρα: Jun 10 18:48:06
Σεισμός τώρα! Περιοχές: . Ώρα: Jun 12 15:32:56
Σεισμός τώρα! Περιοχές: Κομοτηνη, Σαμοθρακη. Ώρα: Jun 12 15:36:44
Σεισμός τώρα! Περιοχές: Αθηνα, Χιος, Ελλαδα, Μυτιληνη. Ώρα: Jun 12 15:39:11
Σεισμός τώρα! Περιοχές: Μυτιληνη, Αθηνα. Ώρα: Jun 12 15:43:44
Σεισμός τώρα! Περιοχές: Μυτιληνη. Ώρα: Jun 12 15:45:46
Σεισμός τώρα! Περιοχές: Μυτιληνη. Ώρα: Jun 12 15:47:10
Σεισμός τώρα! Περιοχές: Μυτιληνη. Ώρα: Jun 12 15:49:59
Σεισμός τώρα! Περιοχές: Μυτιληνη, Θεσσαλονικη, Χιος. Ώρα: Jun 12 15:53:01
Σεισμός τώρα! Περιοχές: Θεσσαλονικη, Μυτιληνη, Πλωμαρι. Ώρα: Jun 12 15:53:38
Σεισμός τώρα! Περιοχές: Μυτιληνη, Λεσβος, Χιος, Θεσσαλονικη. Ώρα: Jun 12 15:54:12
Σεισμός τώρα! Περιοχές: Μυτιληνη, Χιος, Λεσβος. Ώρα: Jun 12 15:55:38
Σεισμός τώρα! Περιοχές: Μυτιληνη, Λεσβος. Ώρα: Jun 12 15:56:51
Σεισμός τώρα! Περιοχές: Μυτιληνη, Λεσβος, Αιγαίο. Ώρα: Jun 12 15:58:40
Σεισμός τώρα! Περιοχές: Μυτιληνη, Χιος. Ώρα: Jun 12 16:04:01
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη, Χιος. Ώρα: Jun 12 16:05:16
Σεισμός τώρα! Περιοχές: Μυτιληνη, Λεσβος. Ώρα: Jun 12 16:07:34
Σεισμός τώρα! Περιοχές: Λεσβος, Χιος. Ώρα: Jun 12 16:12:49
Σεισμός τώρα! Περιοχές: Λεσβος, Χιος. Ώρα: Jun 12 16:13:32
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη, Πλωμαρι. Ώρα: Jun 12 16:15:38
Σεισμός τώρα! Περιοχές: Λεσβος, Πλωμαρι. Ώρα: Jun 12 16:18:01
Σεισμός τώρα! Περιοχές: Λεσβος, Πλωμαρι. Ώρα: Jun 12 16:22:26
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ώρα: Jun 12 16:25:21
Σεισμός τώρα! Περιοχές: Μυτιληνη, Λεσβος, Χιος. Ώρα: Jun 12 16:25:41
Σεισμός τώρα! Περιοχές: Λεσβος. Ώρα: Jun 12 16:32:38
Σεισμός τώρα! Περιοχές: Λεσβος, Αιγαίο, Μυτιληνη. Ώρα: Jun 12 16:35:11
Σεισμός τώρα! Περιοχές: Λεσβος. Ώρα: Jun 12 16:36:42
Σεισμός τώρα! Περιοχές: Λεσβος, Αιγαίο. Ώρα: Jun 12 16:43:55
Σεισμός τώρα! Περιοχές: Λεσβος, Χιος, Αιγαίο. Ώρα: Jun 12 16:47:19
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ώρα: Jun 12 16:57:53
Σεισμός τώρα! Περιοχές: Λεσβος. Ώρα: Jun 12 16:58:59
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ώρα: Jun 12 17:02:31
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ώρα: Jun 12 17:05:48
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ώρα: Jun 12 17:13:30
Σεισμός τώρα! Περιοχές: Λεσβος. Ώρα: Jun 12 17:19:10
```

Σεισμός τώρα! Περιοχές: Λεσβος, Ελλαδα. Ωρα: Jun 12 17:24:39
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 17:27:46
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 17:29:43
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 17:31:44
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 17:34:28
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 17:37:05
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 17:40:02
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 17:43:00
Σεισμός τώρα! Περιοχές: Λεσβος, Ελλαδα. Ωρα: Jun 12 17:46:02
Σεισμός τώρα! Περιοχές: Λεσβος, Ελλαδα. Ωρα: Jun 12 17:51:11
Σεισμός τώρα! Περιοχές: Κλημα. Ωρα: Jun 12 17:54:11
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 17:55:55
Σεισμός τώρα! Περιοχές: Χιος, Μυτιληνη, Λεσβος. Ωρα: Jun 12 17:57:48
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη, Χιος. Ωρα: Jun 12 18:00:28
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 18:04:52
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 18:09:11
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 18:13:00
Σεισμός τώρα! Περιοχές: Λεσβος, Αθηνα, Μυτιληνη. Ωρα: Jun 12 18:17:44
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 18:19:50
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 18:24:55
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 18:28:20
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 18:31:39
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 18:37:21
Σεισμός τώρα! Περιοχές: Λεσβος, Ελλαδα. Ωρα: Jun 12 18:40:33
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 18:48:49
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 18:52:06
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 19:02:57
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 19:10:33
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 19:20:48
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 19:30:23
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 19:39:01
Σεισμός τώρα! Περιοχές: Λεσβος, Πλωμαρι. Ωρα: Jun 12 19:43:39
Σεισμός τώρα! Περιοχές: Λεσβος, Πλωμαρι. Ωρα: Jun 12 19:53:27
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 19:59:23
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 20:02:40
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 20:09:51

```
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 20:37:15
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 20:39:32
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 20:43:32
Σεισμός τώρα! Περιοχές: Λεσβος, Μυτιληνη. Ωρα: Jun 12 21:03:29
Σεισμός τώρα! Περιοχές: Λεσβος, Πλωμαρι, Μυτιληνη. Ωρα: Jun 12 21:07:58
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 21:14:14
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 21:19:30
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 21:30:40
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 22:02:46
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 22:17:35
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 22:22:19
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 22:36:05
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 22:36:50
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 22:37:46
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 22:45:55
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 22:50:32
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 23:03:10
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 23:34:17
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 12 23:37:54
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 15 09:55:36
Σεισμός τώρα! Περιοχές: . Ωρα: Jun 15 16:09:43
Σεισμός τώρα! Περιοχές: Αθηνα. Ωρα: Jun 17 02:44:13
Σεισμός τώρα! Περιοχές: Αθηνα. Ωρα: Jun 17 02:44:39
Σεισμός τώρα! Περιοχές: Αθηνα. Ωρα: Jun 17 02:46:25
Σεισμός τώρα! Περιοχές: Αθηνα, Χανια. Ωρα: Jun 17 02:47:09
Σεισμός τώρα! Περιοχές: Αθηνα. Ωρα: Jun 17 02:48:32
Σεισμός τώρα! Περιοχές: Αθηνα, Καλαματα. Ωρα: Jun 17 02:49:29
Σεισμός τώρα! Περιοχές: Αθηνα, Καλαματα. Ωρα: Jun 17 02:53:17
Σεισμός τώρα! Περιοχές: Αθηνα. Ωρα: Jun 17 02:55:20
Σεισμός τώρα! Περιοχές: Αθηνα. Ωρα: Jun 17 02:57:02
Σεισμός τώρα! Περιοχές: Αθηνα, Μονεμβασια. Ωρα: Jun 17 02:57:28
Σεισμός τώρα! Περιοχές: Αθηνα. Ωρα: Jun 17 03:00:40
Σεισμός τώρα! Περιοχές: Λακωνια, Αθηνα. Ωρα: Jun 17 03:03:04
Σεισμός τώρα! Περιοχές: Λακωνια, Μονεμβασια, Καλαματα. Ωρα: Jun 17 03:06:02
Σεισμός τώρα! Περιοχές: Αθηνα, Καλαματα, Μονεμβασια. Ωρα: Jun 17 03:08:15
Σεισμός τώρα! Περιοχές: Λακωνια, Καλαματα, Μονεμβασια. Ωρα: Jun 17 03:10:32

Σεισμός τώρα! Περιοχές: Λακωνια. Ωρα: Jun 17 03:14:18
Σεισμός τώρα! Περιοχές: Μονεμβασια, Αθηνα. Ωρα: Jun 17 03:22:33
Σεισμός τώρα! Περιοχές: Αττικη, Μονεμβασια. Ωρα: Jun 17 03:33:16
Σεισμός τώρα! Περιοχές: Ελλαδα, Μονεμβασια. Ωρα: Jun 17 03:53:04
Σεισμός τώρα! Περιοχές: Ελλαδα. Ωρα: Jun 17 09:37:20
Σεισμός τώρα! Περιοχές: Λακωνια, Αθηνα. Ωρα: Jun 17 11:00:38
Σεισμός τώρα! Περιοχές: Μυτιληνη. Ωρα: Jun 17 23:05:59
Σεισμός τώρα! Περιοχές: Μυτιληνη, Λεσβος. Ωρα: Jun 17 23:12:59
```

Εικόνα 6.19

- Σύνολο δεδομένων εκπαίδευσης: μεγάλοι σεισμοί, σύνολο δεδομένων δοκιμής: μικροί σεισμοί (*Εικόνα* 6.20)

```
C:\Users\user\Anaconda3\python.exe "C:/Users/user/PycharmProjects/Thesis/Event Detection (Simulation).py"
Σεισμός τώρα! Περιοχές: Λεσβος. Ωρα: Jun 26 20:27:56
Σεισμός τώρα! Περιοχές: Ελλαδα, Φλωρινα. Ωρα: Jul 03 14:48:15
Σεισμός τώρα! Περιοχές: Φλωρινα. Ωρα: Jul 03 22:50:55
Σεισμός τώρα! Περιοχές: Ηρακλειο, Κρήτη. Ωρα: Jul 15 23:38:04
Σεισμός τώρα! Περιοχές: Κρήτη, Ηρακλειο. Ωρα: Jul 15 23:44:27
Σεισμός τώρα! Περιοχές: Κρήτη, Ηρακλειο. Ωρα: Jul 15 23:52:43
Σεισμός τώρα! Περιοχές: Κρήτη. Ωρα: Jul 16 00:38:54
Σεισμός τώρα! Περιοχές: Κρήτη, Ηρακλειο. Ωρα: Jul 16 00:58:04
```

Εικόνα 6.20

Τα προβλήματα που παρουσιάστηκαν στην πρώτη περίπτωση φαίνεται να εξαλείφονται με τη νέα συνθήκη. Οι περισσότεροι σεισμοί που συνέβησαν ανιχνεύονται πράγματι από το σύστημα

και δίνεται ειδοποίηση για το ποίες περιοχές επηρεάστηκαν και την ώρα εντοπισμού του γεγονότος. Παρόλα αυτά βλέπουμε ότι οι ειδοποιήσεις που στέλνει το σύστημά μας είναι πολύ περισσοτερες από εκείνες της συνθήκης στην πρώτη περίπτωση. Αυτό εγκυμονεί τον κίνδυνο να εμφανίζονται ορισμένες λανθασμένες ειδοποιήσεις και επίσης εισάγει το πρόβλημα των πολλαπλών συγαγερμών όπως περιγράψαμε στην ενότητα 3.2.5.

Στον *Πίνακας* 6.3 φαίνονται οι τοποθεσίες και οι χρόνοι στους οποίους συνέβησαν 5 από τους σεισμούς που περιείχε το σύνολο δεδομένων και το αποτέλεσμα της ανίχνευσης όταν το υποσύστημα ειδοποίησης χρησιμοποιεί χρονικά παράθυρα των 10 λεπτών και 5 ή 10 tweets ως όριο για την ανίχνευση γεγονότος.

Τόπος/τόποι που επηρεάζονται από σεισμό	Ώρα πραγματοποίησης σεισμού (ώρα Ελλάδος)	Ώρα 1ου εντοπισμού σεισμού από το σύστημα (10 tweets εντός 10 λεπτών)	Ώρα 1ου εντοπισμού σεισμού από το σύστημα (5 tweets εντός 10 λεπτών)	Ώρα δημοσίευσης 1ης σχετικής αναφοράς σε tweet (1ο True tweet)	Τόπος/τόποι πραγματοποίησης σεισμού που από το σύστημα
Λέσβος, Μυτιλήνη, Χίος, Κομοτηνή, Σαμοθράκη	12-06-2017 15:28	15:37	15:36	15:30:39	Λέσβος, Μυτιλήνη, Χίος, Κομοτηνή, Σαμοθράκη, Αθήνα, Θεσσαλονίκη
Νεάπολη, Λακωνία, Αθήνα	17-06-2017 02:42	02:44	02:44	02:43	Αθήνα, Λακωνία, Χανιά, Καλαμάτα, Μονεμβασιά
Κεφαλονιά	26-06-2017 13:17	-	-	-	-
Φλώρινα, Σκόπια	03-07-2017 14:18	-	14:48	14:22	Ελλάδα, Φλώρινα
Κρήτη, Ηράκλειο	15-07-2017 23:30	-	23:38	23:30	Κρήτη, Ηράκλειο

Πίνακας 6.3

Τα προηγούμενα πειράματα μας οδηγούν με ασφάλεια στο συμπέρασμα πως για τη συνθήκη ενεργοποίησης του συναγερμού «**N tweets σε t λεπτά**» δεν υπάρχει ιδανική τιμή για τη μεταβλητή N. Εξαρτάται κάθε φορά από το είδος της εφορμογής που θα χρησιμοποιηθεί η προτεινόμενη μέθοδος μας. Αν για παράδειγμα ο στόχος είναι η ανίχνευση όσο το δυνατόν περισσότερων σεισμών τότε χρειάζεται μια σχετικά μικρή τιμή για το N. Στις δοκιμές μας χρησιμοποιήσαμε την τιμή πέντε και ενδέχομένως για άλλες εφαρμογές να λειτουργούν καλύτερα μικρότερες τιμές από αυτή. Η ουσία είναι πως όσο μειώνουμε το N τόσο αυξάνονται τα γεγονότα που ανιχνεύουμε με αντάλλαγμα τον κίνδυνο να ενεργοποιηθεί ο συναγερμός σε περιπτώσεις άσχετες με το γεγονός ενδιαφέροντος. Από την άλλη, μια μεγαλύτερη τιμή για το N εξασφαλίζει πως τα γεγονότα που ανιχνεύει το σύστημα μας περιγράφουν έναν σεισμό με μεγάλη πιθανότητα ωστόσο οι μεγαλύτερες αυτές τιμές δυσχεραίνουν τον εντοπισμό σεισμών μικρότερης κλίμακας. Υπάρχει ουσιαστικά ένα trade-off μεταξύ της απόδοσης των μετρικών ανάκλησης και ακρίβειας για το πρόβλημα του εντοπισμού σεισμών σε ένα σύνολο δεδομένων καθώς μεταβάλλεται το N.

Σχετικά με την τιμή του t για τη συνθήκη αξίζει να σημειωθεί για μια ακόμα φορα πως εξαρτάται από την ταχύτητα εντοπισμού του γεγονότος που επιθυμούμε για το σύστημα. Όπως φαίνεται και από τα στιγμιότυπα της χρονικής εξέλιξης των tweets, μειώνοντας το χρόνο που παρακολουθούμε τα True tweets ελοχεύει ο κίνδυνος αστοχίας στον εντοπισμό γεγονότων μικρότερης έντασης παρόλο που το σύστημα ειδοποιεί για σεισμό ακόμα πιο γρήγορα. Γενικά, η τιμή του t πρέπει να είναι εντός λογικών πλαισίων και με βάση τη μελέτη που έγινε οι τιμές έως και δέκα λεπτά λειτουργούν ικανοποιητικά για το ελλήνικό Twitter.

6.2.4 Μελέτη περίπτωσης: γεγονότα πυρκαγιάς

Μέχρι τώρα εξετάσαμε τις επιδόσεις της μεθόδου μας σε ένα συγκεκριμένο γεγονός ενδιαφέροντος, το σεισμό. Τα γεγονότα που σχετίζονται με φυσικές καταστροφές όμως δεν περιορίζονται μόνο σε σεισμούς αλλά επεκτείνονται και σε πυρκαγιες, πλημμύρες κ.α. Κατά συνέπεια, έχει ιδιαίτερο ενδιαφέρον να μελετήσουμε κατά πόσο το σύστημα μας μπορεί να γενικευτεί για διάφορες φυσικές καταστροφές.

Σε συστήματα μηχανικής μάθησης που κατατάσσουν δεδομένα σε κατηγορίες, όπως το δικό μας, μια σημαντική παράμετρος αξιολογησής της λειτουργικότητάς τους είναι η ικανότητά να γενικεύουν. Για να κατανοήσουμε καλύτερα την έννοια αυτή θα χρησιμοποιήσουμε ένα παράδειγμα: ένας πολύ απλός ταξινομητής είναι αυτός που διακρίνει αν ένα χρώμα είναι μπλε ή όχι. Αν ο ταξινομητής δέχεται δεδομένα που απεικονίζουν π.χ. το κίτρινο ή το κόκκινο χρώμα και δεν τα κατατάσει ως μπλε μοιάζει να λειτουργεί ικανοποιητικά. Ωστόσο, η πραγματική αξία ενός τέτοιου ταξινομητή έγγειται στο να είναι ικανός να ξεχωρίζει όσο το δυνατόν

περισσότερες αποχρώσεις του μπλε. Την ικανότητα αυτή την ονομάζουμε ικανότητα γενίκευσης του ταξινομητή.

Κατ' αναλογία με τη δικιά μας περίπτωση, είναι κρίσιμο να δοκιμάσουμε το κατά πόσο η μέθοδος που αναπτύξαμε μπορεί να γενικευθεί στο θέμα των φυσικών καταστροφών, εφόσον χρησιμοποιεί έναν αλγόριθμο ταξινόμησης. Για το λόγο αυτό θα πειραματιστούμε με το πώς ανταποκρίνεται το υποσύστημα ταξινόμησης σε γεγονότα που αφορούν πυρκαγιές. Αίσθηση μας είναι πως τα γεγονότα του σεισμού και της πυρκαγιάς εμφανίζουν παρόμοια μοτίβα στις αναφορές τους στο Twitter. Αν με ελάχιστες ή καθόλου τροποποιήσεις στα χαρακτηριστικά του ταξινομητή καταφέρουμε να εξάγουμε ικανοποιητικά αποτελέσματα θα έχουμε μια ισχυρή ένδειξη πως το σύστημα μας μπορεί να γενικεύσει αποδοτικά και σε άλλα γεγονότα συγκεκριμένου θέματος σχετικά με φυσικές καταστροφές.

Σε αυτό το σημείο, είναι ουσιώδες να κάνουμε μια επισήμανση. Μιλήσαμε για δυνατότητα γενίκευσης του συνολικού συστήματος με μόνη προϋπόθεση να δύναται να γενικεύσει το υποσύστημα ταξινόμησης, ανεξάρτητα του υποσυστήματος ειδοποίησης. Ο πρώτος λόγος που έγινε αυτό είναι επειδή το σύνολο δεδομένων που συλλέξαμε δεν περιλαμβάνει μεγάλο πλήθος tweets σχετικών με πυρκαγία. Συγκεκριμένα επι συνόλου 10,996 tweets μόνο τα 181 μιλάνε για κάποια φωτιά που ξέσπασε και χωρίς να εμφανίζουν έντονη συγκέντρωση σε ορισμένο χρονικό διάστημα. Έτσι οποιοσδήποτε πειραματισμός με το σύστημα συναγερμού δεν μπορεί να οδηγήσει σε αξιόπιστα συμπεράσματα. Πέρα από τα πρακτικά ζητήματα, ο δεύτερος λόγος για τον οποίον δεν εξετάστηκε το σύστημα ειδοποίησης ως προς τη γενίκευση είναι η αντίληψη πως δεν χρείαζεται ιδιαίτερες αλλαγές το κομμάτι του συναγερμού εφόσον τα αποτελέσματα του ταξινομητή για μία συγκεκριμένη φυσική καταστροφή είναι ικανοποιητικά. Η φύση του στατιστικού μοντέλου παραμένει ίδια για όλες τις περιπτώσεις με μόνη διαφορά την συνθήκη που ορίζουμε εμείς πως ταιριάζει ανάλογα με τις συνθήκες, δηλαδή το πόσα σχετικά tweets χρειάζονται να δημοσιευθούν σε συγκεκριμένο χρονικό παράθυρο.

Με όλα τα παραπάνω υπόψη μας, μπορούμε πλεόν να πειραματιστούμε με την ανίχνευση γεγονότων σχετικών με πυρκαγιές. Αρχικά, τοποθετούμε ετικέτες σε κάθε tweet του συνόλου δεδομένων ανάλογα με το αν σχετίζονται με πυρκαγιά ή όχι: ετικέτα True αν αναφέρονται σε πυρκαγιά και ετικέτα False αν όχι, όπως και στην περίπτωση του σεισμού. Να σημειθεί πως τα tweets που μιλάνε για σεισμό χαρακτηρίζονται ως False σε αυτό το πείραμα αφού εξετάζουμε ξεχωριστά το κάθε γεγονός ενδιαφέροντος. Τα χαρακτηριστικά του ταξινομητή παραμένουν ακριβώς τα ίδια με τη μόνη διαφορά ότι τώρα έχουμε διαφορετικες λέξεις-κλειδιά. Το νέο σύνολο των λέξεων-κλειδιών περιλαμβάνει λέξεις όπως «πυρκαγιά», «φωτιά», «μαίνεται» κ.λ.π. Στα ορίσματα του αλγορίθμου SVM δεν δόθηκε επιπλέον βαρύτητα σε κάποια κλάση όπως έγινε στην περίπτωση του σεισμού. Ο τρόπος αξιολόγησης του ταξινομητή για το γεγονός της πυρκαγιάς γίνεται με τις μετρικές ακρίβειας, ανάκλησης και βαθμολογίας–F για

διασταυρούμενη επικύρωση 10-στρώσεων όπως και στα προηγούμενα πειράματα. Τα αποτελέσματα φαίνονται στην Εικόνα 6.21 που ακολουθεί.

```
C:\Users\user\Anaconda3\python.exe "C:/Users/user/PycharmProjects/Thesis/Classification Testing (fire).py"
Precision: [ 0.89473684  0.92307692  0.89473684  0.93333333  0.66666667  0.9
  0.61538462  0.94444444  0.78947368  1.        ]
Recall: [ 0.94444444  0.66666667  0.94444444  0.77777778  0.88888889  1.
  0.88888889  0.94444444  0.88235294  0.82352941]
F-Measure: [ 0.91891892  0.77419355  0.91891892  0.84848485  0.76190476  0.94736842
  0.72727273  0.94444444  0.83333333  0.90322581]
Average Precision: 0.856185
Average Recall: 0.876144
Average F-Measure: 0.857807

Process finished with exit code 0
```

Εικόνα 6.21

Τα αποτελέσματα του πειράματος φαίνονται ιδιαίτερα ικανοποιητικά ειδικά αν αναλογιστούμε πως οι αλλαγές στο σύστημα ταξινόμησης περιορίστηκαν στις απολύτως απαραίτητες, δηλαδή μόνο στις λέξεις-κλειδία που περιγράφουν το γεγονός ενδιαφέροντος. Παρατηρώντας το σύνολο δεδομένων που συλλέξαμε συμπεραίνουμε πως πλειοψηφία των tweets που αναφέρονται σε πυρκαγιές περιέχουν τις σχετικές λέξεις-κλειδιά και την περιοχή όπου ξέσπασε η πυρκαγιά. Επομενως, τα υψηλά ποσοστά των μετρικών του ταξινομητή είναι αναμενόμενα αφού το μοτίβο της ανταπόκρισης των χρηστών του Twitter σε μια πυρκαγιά μοιάζει με αυτό του σεισμού.

Ενδεικτικά παραθέτουμε στο ακόλουθο στιγμιότυπο (Εικόνα 6.22) όρισμένα tweets και την κλάση στην οποία κατατάσσονται σύμφωνα με τον ταξινομητή.

```
3731: 'ΦΩΤΟ &amp  ΒΙΝΤΕΟ ΜΕΓΑΛΗ ΦΩΤΙΑ Από τις πόρτες προς Αντιμάχεια – Καρδάμαινα https://t.co/YdgN0HZbrB via @aegeanews' => True
3736: 'ΣΥΜΒΑΙΝΕΙ ΤΩΡΑ: ΜΕΓΑΛΗ ΦΩΤΙΑ ΣΤΗΝ ΑΝΤΙΜΑΧΕΙΑ (φωτο- βίντεο) - Τελευταία νέα - Το Βήμα της Κω https://t.co/UKf3xo222Y' => True
3746: 'ΦΩΤΟ &amp  ΒΙΝΤΕΟ ΜΕΓΑΛΗ ΦΩΤΙΑ Από τις Πόρτες προς Αντιμάχεια – Καρδάμαινα https://t.co/YdgN0HZbrB via @aegeanews' => True
3747: 'ΤΩΡΑ...!!! ΜΕΓΑΛΗ ΦΩΤΙΑ ΣΕ ΕΡΓΟΣΤΑΣΙΟ ΣΤΑ ΟΙΝΟΦΥΤΑ - ΠΛΗΡΟΦΟΡΙΕΣ ΓΙΑ ΤΡΑΥΜΑΤΙΕΣ https://t.co/U4OEVk8QBQ https://t.co/mdnsoyvkk5' => True
3751: 'ΕΚΤΑΚΤΟ! ΠΥΡΚΑΓΙΑ ΣΕ ΕΡΓΟΣΤΑΣΙΟ ΣΤΗΝ ΕΘΝΙΚΟ ΟΔΟ-ΥΠΑΡΧΟΥΝ\xa0ΤΡΑΥΜΑΤΙΕΣ https://t.co/q62C2a6yNa https://t.co/TsLfxS6ucI' => True
3759: 'ΕΠΙΣΗΜΗ ΕΝΗΜΕΡΩΣΗ ΑΠΟ ΤΗΝ ΠΥΡΟΣΒΕΣΤΙΚΗ: 2 ΤΡΑΥΜΑΤΙΕΣ ΑΠΟ ΜΕΓΑΛΗ ΠΥΡΚΑΓΙΑ ΣΤΑ ΟΙΝΟΦΥΤΑ! https://t.co/XNRNC6etQ8 https://t.co/DxVJHpQBxI' => True
3760: 'ΤΩΡΑ ! ΜΕΓΑΛΗ ΦΩΤΙΑ ΣΤΟ ΜΠΟΛΑΤΙ https://t.co/2UZvDWtoUA' => True
3777: 'Μεγάλη πυρκαγιά σε εργοστάσιο στα Οινόφυτα – Τουλάχιστον δύο εργαζόμενοι έχουν τραυματιστε  #πυρκαγια https://t.co/WAlpFkNQJ3' => True
3796: 'ΦΩΤΟ &amp  ΒΙΝΤΕΟ ΜΕΓΑΛΗ ΦΩΤΙΑ Από τις Πόρτες προς Αντιμάχεια – Καρδάμαινα - Ρίψεις νερού από ελικόπτερο https://t.co/YdgN0HZbrB via @aegeanews' => True
3797: 'Πάτρα: Πυρκαγια τώρα στα Νιφορέϊκα https://t.co/z5ryvvxzcp9' => True
3802: 'Ανεξέλεγκτη η πυρκαγια που ξεκίνησε από τις Πόρτες με κατεύθυνση την Καρδάμαινα - Αναμένονται ... [ΔΗΛΩΣΕΙΣ-ΒΙΝΤΕΟ] https://t.co/rMcElRZx2O' => False
3809: 'ΤΩΡΑ! ΜΕΓΑΛΗ ΦΩΤΙΑ ΣΤΟ ΖΕΥΓΟΛΑΤΙΟ-ΤΡΑΥΜΑΤΙΣΤΗΚΑΝ ΠΥΡΟΣΒΕΣΤΕΣ (ΦΩΤΟ &amp \xa0ΒΙΝΤΕΟ) https://t.co/H77rNFiojt https://t.co/RQ2VFO6N5m' => True
3811: 'https://t.co/vHbvBCIGU7 | Πυρκαγιά με δύο τραυματίες σε εργοστάσιο https://t.co/mPksqxqXC3 #Ελλαδα #Πυρκαγια' => True
3812: 'RT @hellasforcecom: ΤΩΡΑ! ΜΕΓΑΛΗ ΦΩΤΙΑ ΣΤΟ ΖΕΥΓΟΛΑΤΙΟ-ΤΡΑΥΜΑΤΙΣΤΗΚΑΝ ΠΥΡΟΣΒΕΣΤΕΣ (ΦΩΤΟ &amp \xa0ΒΙΝΤΕΟ) https://t.co/H77rNFiojt https://t.co/RQ2VF…' => True
3814: 'RT @hellasforcecom: ΤΩΡΑ! ΜΕΓΑΛΗ ΦΩΤΙΑ ΣΤΟ ΖΕΥΓΟΛΑΤΙΟ-ΤΡΑΥΜΑΤΙΣΤΗΚΑΝ ΠΥΡΟΣΒΕΣΤΕΣ (ΦΩΤΟ &amp \xa0ΒΙΝΤΕΟ) https://t.co/H77rNFiojt https://t.co/RQ2VF…' => True
```

Εικόνα 6.22

Η παραπάνω ανάλυση δείχνει πως οι πρώτες ενδείξεις για το αν το σύστημα μας μπορεί να γενικεύσει αποδοτικά είναι κάτι περισσότερο από θετικές. Οι αλλαγές που κάναμε στις παραμέτρους του ταξινομητή ήταν αμελητέες ενώ οι επιδόσεις του συστήματος ταξινόμησης παρουσίασαν ικανοποιητικά αποτελέσματα. Εφόσον καταφέραμε με κατάλληλες προσαρμογές να πετύχουμε το επιθυμητό αποτέλεσμα μπορούμε να επεκτείνουμε το σύστημα ταξινόμησης

ώστε να περίεχει ένα επιπλέον χαρακτηριστικό που θα είναι η ύπαρξη λέξεων-κλειδιών σχετικές με πυρκαγιά. Σε αυτή την περίπτωση θα πρέπει να δημιουργηθούν τρεις κλάσεις, μία για το αν ένα tweet μιλάει για σεισμό, άλλη μία για πυρκαγιά και μία τρίτη που θα δηλώνει πως το tweet δεν έχει σχέση με κανένα γεγονός.

Ωστόσο, για να μιλήσουμε με πλήρη ασφάλεια για τη δυνατότητα γενίκευσης του συστήματος χρειάζεται ένα σύνολο δεδομένων με περισσότερα γεγονότα πυρκαγιών για να καλύπτονται όλες οι περιπτώσεις των αντιδράσεων των χρηστών απέναντι σε ένα τέτοιο γεγονός. Επίσης, ένα αντίστοιχο σύνολο δεδομένων με αυτό που προαναφέρθηκε είναι απαραίτητο ώστε να δοκιμαστούν οι διάφορες συνθήκες για το σύστημα ειδοποίησης.

7

Επίλογος

Στο παρόν κεφάλαιο συνοψίζουμε τα βασικά συμπεράσματα που προέκυψαν από την περάτωση της παρούσας διπλωματικής εργασίας και προτείνουμε πιθανές επεκτάσεις του υλοποιηθέντος συστήματος, ώστε να βελτιωθεί περαιτέρω η λειτουργία υπερκεράζοντας προβληματισμούς τους οποίους συναντήσαμε κατά την υλοποίηση.

7.1 Σύνοψη και συμπεράσματα

Στην παρούσα εργασία εξετάστηκε η πραγματικού-χρόνου φύση του κοινωνικού δικτύου Twitter ως μέσου για την αναγνώριση γεγονότων στα πλαίσια της ευρύτερης κατηγορίας των φυσικών καταστροφών.

Πιο συγκεκριμένα, μελετήθηκε η ανίχνευση και η αναφορά γεγονότων που σχετίζονται με πραγματικούς σεισμούς στην Ελλάδα από αναφορές στο Twitter στην ελληνική γλώσσα και αναπτύχθηκε ένα κατάλληλο πληροφοριακό σύστημα σε διαδικτυακή πλατφόρμα για την αυτοματοποίηση της διαδικασίας.

Αρχικά, ερευνήθηκαν οι διάφορες μέθοδοι που υπάρχουν στην διεθνή βιβλιογραφία για την ανίχνευση γεγονότων από μέσα κοινωνικής δικτύωσης και από αυτές η πλέον κατάλληλη για το δεδομένο πρόβλημα κρίθηκε η κατηγορία ανίχνευσης γεγονότος συγκεκριμένου θέματος αφού ήταν γνωστό εκ των προτέρων το είδος των γεγονότων προς αναζήτηση.

Από την προαναφερθήσα κατηγορία ανίχνευσης γεγονότων επιλέχθηκε μία που ειδικεύεται σε φυσικές καταστροφές. Η μέθοδος που περιγράφεται δέχεται μια ροή tweets, τα προεπεξεργάζεται κατάλληλα, κατηγοριοποιεί κάθε tweet σύμφωνα με το αν σχετίζεται με σεισμό ή όχι μέσω ενός ταξινομητή μηχανικής μάθησης και χρησιμοποιώντας ένα στατιστικό μοντέλο ειδοποίησης, ενημερώνει τους χρήστες για το αν συμβαίνει σεισμός στην Ελλάδα.

Στο κομμάτι του ταξινομητή διαπιστώθηκε πως τα χαρακτηριστικά που λειτουργούν περισσότερο αποδοτικά είναι ορισμένα στατιστικά του tweet όπως μήκος κειμένου και η θέση λέξεων-κλειδιών εντός του tweet. Η μελέτη του συνόλου δεδομένων οδήγησε στη χρήση επιπλέον χαρακτηριστικών του ταξινομητή όπως το αν το κείμενο του tweet περιλαμβάνει μια ελληνική περιοχή. Τα πειράματα εμφάνισαν σημαντικά καλύτερες επιδόσεις. Για την εξόρυξη των πρόσθετων χαρακτηριστικών δημιουργήθηκε μια διαδικάσία εξαγωγής περιοχής αποκλειστικά από το κείμενο ενός tweet όπως έχει περιγραφεί στην ενότητα 3.3.

Για το μοντέλο ειδοποίησης, ορίστηκαν σχετικά αυστηρές συνθήκες ενεργοποίησης με την λογική να εντοπίζονται σεισμοί ισχυρής έντασης για να ενημερώνονται οι χρήστες αν κάπου υπάρχει πρόβλημα λόγω σεισμού.

Τέλος, ελέχθηκε η δυνατότητα γενίκευσης του συστήματος μας και σε επιπλέον φυσικές καταστροφές, όπως η πυρκαγιά, και τα πειράματα παρουσίασαν ικανοποιήτική απόδοση με μείζονες προσαρμογές.

7.2 Μελλοντικές επεκτάσεις

7.2.1 Δυνατότητα γενίκευσης του συστήματος

Στην ενοτητα 6.2.4 εξετάσαμε την δυνατότητα του συστήματος να κλιμακώνει τη χρήση του και σε άλλες φυσικές καταστροφές πέραν του σεισμού. Τα πειράματα σχετικά με τον ταξινομητή ήταν ιδιαίτερα ενθαρρυντικά ωστόσο το σύνολο δεδομένων που συλλέχθηκε δεν περιλάμβανε αναφορές σε κάποια μεγάλης έκτασης πυρκαγιά ωστε να πυροδοτηθεί ο συναγερμός. Γενικότερα, διατηρούνται ορισμένες επιφυλάξεις για την ικανόνητα γενίκευσης του συστήματος εξαιτίας του περιορισμένου αριθμού αναφορών σε πυρκαγιές στο διαθέσιμο σύνολο δεδομένων.

Με βάση τα παραπάνω, προτεινόμενο αντικείμενο περαιτέρω μελέτης αποτελεί η δοκιμή της μεθόδου που αναπτύχθηκε στην παρούσα εργασία σε άλλου είδους γεγονότα όπως η πυρκαγιά, τα τροχαία ατυχήματα, κυκλοφοριακή συμφόρηση κ.λ.π. Ιδιαίτερη έμφαση θα μπορούσε να δοθεί στην άντληση συνόλου δεδομένων επαρκούς μεγέθους τις κατάλληλες περιόδους (για την περίπτωση π.χ. της πυρκαγιάς ιδανικό είναι το καλοκαίρι), στην αναζήτηση αντιπροσωπευτικών λέξεων-κλειδιών για το εκάστοτε γεγονός και στη μελέτη των tweets που περιγράφουν το γεγονός για να εξαχθούν χαρακτηριστικά που θα χρησιμοποιηθούν στον ταξινομητή. Επιπρόσθετα, επειδή μια πυρκαγιά δεν σχετίζεται αποκλειστικά με περιοχή αλλά μπορεί να ξεσπάσει σε κτίριο, εργαστάσιο, μεταφορικό μέσο κ.λ.π. κρίνεται σκόπιμο να

μελετηθεί και η χρήση μιας αντίστοιχης ομάδας χαρακτηριστικών που θα περιλαμβάνει και τέτοιες περιπτώσεις.

7.2.2 Στάθμιση των tweets με βάση την αξιοπιστία των χρηστών που τα δημοσιεύουν

Κατά την περιγραφή των υποσυστημάτων ταξινόμησης και ειδοποίησης κάναμε μια σιωπηρή υπόθεση: όλα τα δημοσιευμένα tweets έχουν την ίδια βαρύτητα στη συνολική διαδικασία, ανεξαρτήτως του χρήστη που τα δημοσίευσε. Ως εκ τούτου, η επιρροή χρηστών όπως @LastQuake που πρόκειται για ανεξάρτητο επιστημονικό οργανισμό ο οποίος παρέχει μέσω των εφαρμογών του και των λογαριασμών του στα κοινωνικά δίκτυα προειδοποιήσεις πραγματικού-χρόνου για σεισμούς ή όπως @amna news που αποτελεί το Αθηναϊκό - Μακεδονικό Πρακτορείο Ειδήσεων, εξισώνεται με αυτή που έχει ένας οποιοσδήποτε άλλος λογαριασμός του Twitter. Όπως είναι εύλογο, η εγκυρότητα και η αξιοπιστία μιας δημοσίευσης που περιγράφει έναν σεισμό είναι ισχυρότερες από λογαριασμούς οργανισμών, ειδησεογραφικών πρακτορείων κ.λ.π.

Για να ενισχυθεί η αξιοπιστία του συνολικού συστήματος προτείνουμε την υλοποίηση μιας μεθόδου ιεράρχησης των tweets με βάση το επίπεδο σημαντικότητας του χρήστη που το δημοσίευσε. Ανάλογες μέθοδοι έχουν μελετηθεί στη βιβλιογραφία όπως στο [54] , στο οποίο περιγράφεται ένας συστηματικός τρόπος υπολογισμού της αυθεντίας του χρήστη (user authority) μέσω του αλγορίθμου PageRank [55].

Η προαναφερθείσα ιεράρχηση των tweets θα μπορούσε να βρει εφαρμογή και στο κομμάτι του ταξινομητή και στο σύστημα ειδοποίησης. Από τη μία πλευρά, ο ταξινομητής θα μπορούσε να έχει την δυνατότητα να λαμβάνει υπόψη το επίπεδο αυθεντίας του χρήστη πριν αξιολογήσει ένα tweet ως προς τις κλάσεις του σεισμού. Από την άλλη, με έναν κατάλληλο εμπειρικό κανόνα το σύστημα ειδοποίησης θα μπορούσε να ενεργοποιείται όταν επαρκής αριθμός αξιόπιστων χρηστών δημοσιεύσει για έναν σεισμό.

Αξίζει να σημειωθεί πως την περίοδο που πραγματοποιήσαμε την συλλογή των tweets δεν αποθηκεύσαμε πληροφόριες σχετικές με τους χρήστες που τα δημοσίευαν παρά μόνο τα ψευδώνυμα των λογαριασμών τους. Χωρίς βασικές πληροφορίες όπως το πλήθος των «ακόλουθων» (ακόλουθος ενός χρήστη στην ορολογία του Twitter είναι ένας άλλος χρήστης παρακολουθεί τις δημοσιεύσεις του πρώτου) κ.α. καθίσταται αδύνατη η οποιαδήποτε στάθμιση χρηστών.

7.2.3 Προσθήκη νέων λειτουργιών

Τα tweets τα οποία συλλέγονται και ενεργόποιούν το σύστημα ειδοποίησης είναι πολύ πιθανό να περίεχουν επιπλέον πληροφορίες, ιδιαίτερα ενδιαφέρουσες για να αναφερθούν όπως π.χ. ζημιές που προκλήθηκαν, αριθμός τραυματιών και νεκρών, μέτρα που πρόκειται να ληφθούν κ.λ.π. Το σύστημα ανίχνευσης γεγονότων που αναπτύχθηκε παρακολουθεί τη ροή των tweets με μοναδικό σκοπό να αναφέρει το αν συνέβη σεισμός, χωρίς να αποθηκεύει τις οποιεσδήποτε επιπλέον πληροφορίες.

Επομένως, εξαιτίας της ειδησεογραφικής φύσης του Twitter, προτείνεται η έρευνα για προσθήκη νέων λειτουργιών στο σύστημα όπως η καταγραφή των συνεπειών που είχε το χτύπημα του εγκέλαδου σε μια περιοχή.

7.2.4 Προστασία από εσφαλμένους συναγερμούς

Η υψηλή απόδοση του ταξινομητή και η απαίτηση του συστήματος ειδοποίησης για απότομη αύξηση των tweets που αναφέρονται σε σεισμό εντός μικρού χρονικού διαστήματος διασφαλίζουν τα αξιόπιστα αποτελέσματα της παρούσας μεθόδου. Ουσιαστικά, τα δύο αυτά υποσυστήματα λειτουργούν ως δικλείδες ασφαλείας αυξάνοντας την πιθανότητα η ακολουθία των tweets που θα πυροδοτήσει το συναγερμό να αναφέρεται σε σεισμό που όντως συνέβη. Παρόλα αυτά, υπάρχουν περιπτώσεις όπου εξασφαλισμένα το σύστημα αστοχεί να δώσει έγκυρα αποτελέσματα.

Στην περίπτωση όπου ο ίδιος χρήστης δημοσιεύσει επανειλημμένα και σε μικρό χρονικό διάστημα ένα tweet που κρίνεται ως True από τον ταξινομητη ή αν ένα tweet όπως το «Σεισμός στον Πειραιά από την νέα μεταγραφή του Ολυμπιακού!» γίνει retweet πολλές φορές υπάρχει πιθανότητα ο συναγερμός να ενεργοποιηθεί εσφαλμένα (αν όλα αυτά συμβούν σε χρονικό διάστημα που ικανοποιεί τις συνθήκες πυροδότησης του συναγερμού).

Για να μην είναι το σύστημα επιρρεπές σε τέτοιες συμπεριφορές χρηστών του Twitter προτείνεται για εκτενέστερη μελέτη η αναζήτηση επιπρόσθετων επιπέδων ασφάλειας που θα μειώνουν την πιθανότητα ο συναγερμός να ενεργοποιείται λανθασμένα. Ορισμένες προτεινόμενες λύσεις για να επιτευχθουν τα προηγούμενα είναι να μειώνεται το αντίκτυπο που έχουν στο σύστημα tweets που έχουν γίνει retweet και το να μην λαμβάνονται υπόψη πολλαπλά tweets από τον ίδιο λογαριασμό εντός ενός χρονικού παραθύρου. Σε κάθε περίπτωση η προστασία του συστήματος από τέτοια φαινόμενα είναι μείζονος σημασίας.

7.2.5 Συναγερμός βασισμένος σε κανόνες

Το σύστημα του συναγερμού προκύπτει αποκλειστικά από ένα στατιστικό μοντέλο εφαρμοσμένο σε χρονικά παράθυρα. Για να διευρύνουμε τις δυνατότητές του μπορούμε να αναπτύξουμε πρόσθετους κανόνες διαμορφωμένους έτσι ώστε να ενεργοποιείται ανάλογα με την εφαρμογή που θέλουμε να χρησιμοποιήσουμε τη μέθοδο μας.

Παράδειγμα τέτοιων κανόνων είναι η ενεργοποίηση του συναγερμού στις περιπτώσεις που ο σεισμός ήταν μεγάλης έντασης (επομένως έγινε έντονος λόγος για αυτόν). Μία άλλη ενδιαφέρουσα περίπτωση είναι να συμβαίνουν ταυτόχρονα πολλαπλά γεγονότα ενδιαφέροντος. Στο σύστημα που αναπτύξαμε υποθέσαμε ότι κάθε φορά συμβαίνει ένας σεισμός ή μία πυρκαγιά αντίστοιχα, πράγμα που μοίαζει λογικό για την κατηγορία των φυσικών καταστροφών. Ωστόσο στις περιπτώσεις ατυχημάτων, κυκλοφοριακής συμφόρησης κ.α. αυτό μπορεί να μη συμβαίνει.

Γενικά, η επιπρόσθετη μελέτη των στατιστικών μοντέλων με σκοπό να χρησιμοποιηθούν προηγμένα συστήματα ειδοποίησης είναι απαραίτητη ούτως ώστε να επιτευχθεί αποτελεσματική αναγνώριση πολλαπλών γεγονότων που εξελίσσονται παράλληλα.

8

Βιβλιογραφία

[1] R. Li, K. H. Lei, R. Khadiwala, and K. C. C. Chang, "TEDAS: A twitter-based event detection and analysis system," in *Proceedings - International Conference on Data Engineering*, 2012, pp. 1273–1276.

[2] R. Krikorian, "New Tweets per second record, and how," *Official Twitter Blog*. p. 1, 2013.

[3] N. Panagiotou, I. Katakis, and D. Gunopulos, "Detecting events in online social networks: Definitions, trends and challenges," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9580, pp. 42–84.

[4] H. Abdelhaq, C. Sengstock, and M. Gertz, "EvenTweet: Online Localized Event Detection from Twitter," *Proc. VLDB Endow.*, vol. 6, no. 12, pp. 1326–1329, 2013.

[5] J. Allan, "Introduction to Topic Detection and Tracking," in *Topic Detection and Tracking: Event-based Information Organization*, vol. 12, 2002, pp. 1–16.

[6] A. J. McMinn, Y. Moshfeghi, and J. M. Jose, "Building a large-scale corpus for evaluating event detection on twitter," in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013)*, 2013, pp. 409–418.

[7] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on Twitter," *Icwsm*, pp. 1–17, 2011.

[8] A. Guille and C. Favre, "Event detection, tracking, and visualization in Twitter: a mention-anomaly-based approach," *Soc. Netw. Anal.*

Min., vol. 5, no. 1, pp. 1–18, 2015.

[9] C. Rogers-Pettie and J. Herrmann, “Information Diffusion : A Study of Twitter During Large Scale Events,” *IIE Annu. Conf. Proceedings. Inst. Ind. Syst. Eng. (IISE*, 2015.

[10] Y. Yang, T. Pierce, and J. Carbonell, “A Study of Retrospective and On-line Event Detection,” *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR ’98*, pp. 28–36, 1998.

[11] F. Atefeh and W. Khreich, “A survey of techniques for event detection in Twitter,” *Comput. Intell.*, vol. 31, no. 1, pp. 133–164, 2015.

[12] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors,” *Proc. 19th Int. Conf. World Wide Web*, pp. 851–860, 2010.

[13] SHARE European Earthquake Catalog (SHEEC), “Earthquakes in Europe 1000 - 2007,” 2013.

[14] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang, “Discover breaking events with popular hashtags in twitter,” in *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM ’12*, 2012, p. 1794.

[15] G. Valkanas and D. Gunopulos, “How the Live Web Feels About Events,” *ACM Int. Conf. Inf. Knowl. Manag.*, no. June, pp. 639–648, 2013.

[16] “Hierarchical Clustering,” *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Hierarchical_clustering.

[17] “DBSCAN,” *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/DBSCAN.

[18] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[19] C. C. Aggarwal and K. Subbian, “Event Detection in Social Streams,” *Proc. SIAM Int. Conf. Data Min.*, pp. 624–635, 2012.

[20] A. Khaleghi, D. Ryabko, J. Mary, and P. Preux, “Online Clustering of Processes,” *Aistats’12*, vol. XX, pp. 601–609, 2012.

[21] A. Aizawa, “An information-theoretic perspective of tf-idf measures,” *Inf. Process. Manag.*, vol. 39, no. 1, pp. 45–65, 2003.

[22] T. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: A New Data

Clustering Algorithm and Its Applications," *Data Min. Knowl. Discov.*, vol. 1, no. 2, pp. 141–182, 1997.

[23] O. Ozdikis, P. Senkul, and H. Oguztuzun, "Semantic Expansion of Hashtags for Enhanced Event Detection in Twitter," in *2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2012, pp. 20–24.

[24] F. Can, S. Kocberber, O. Baglioglu, S. Kardas, H. C. Ocalan, and E. Uyar, "New event detection and topic tracking in turkish," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 4, pp. 802–819, 2010.

[25] M. Cordeiro, "Twitter event detection: combining wavelet analysis and topic inference summarization," *Proc. Dr. Symp. Informatics Eng.*, 2012.

[26] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Proc. OSDI - Symp. Oper. Syst. Des. Implement.*, pp. 137–149, 2004.

[27] K. Watanabe, M. Ochi, M. Okabe, and R. Onai, "Jasmine: A Real-time Local-event Detection System Based on Geolocation Information Propagated to Microblogs," *Proc. 20th ACM Int. Conf. Inf. Knowl. Manag.*, pp. 2541–2544, 2011.

[28] G. Niemeyer, "Geohash," 2008. [Online]. Available: http://geohash.org.

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

[30] J. Allan, J. Allan, R. Papka, R. Papka, V. Lavrenko, and V. Lavrenko, "On-line New Event Detection and Tracking," *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 37–45, 1998.

[31] G. Kumaran and J. Allan, "Using names and topics for new event detection," *Proc. Conf. Hum. ...*, pp. 121–128, 2005.

[32] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," *NAACL HLT 2010 - Hum. Lang. Technol. 2010 Annu. Conf. North Am. Chapter Assoc. Comput. Linguist. Proc. Main Conf.*, no. June, pp. 181–189, 2010.

[33] M. Karkali, F. Rousseau, A. Ntoulas, and M. Vazirgiannis, “Efficient online novelty detection in news streams,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8180 LNCS, no. PART 1, pp. 57–71.

[34] S. Zhao, L. Zhong, J. Wickramasuriya, and V. Vasudevan, “Human as Real-Time Sensors of Social and Physical Events: A Case Study of Twitter and Sports Games,” *arXiv Prepr. arXiv1106.4300*, no. June, p. 9, 2011.

[35] E. Sokos and J. Zahradník, “Lesvos June 12, 2017, Mw 6.3 event, a quick study of the source,” 2017.

[36] K. I. Papadimitriou P., Tselentis G.A., Voulgaris N., Kouskouna V., Lagios E., K. Kaviris G., Pavlou K., Sakkas V., Moumoulidou A., Karakonstantis A., S. I. V., Sakkas G., Kazantzidou D., Aspiotis T., Fountoulakis I., Millas C., and A. E. Lekkas E., Antoniou V., Mavroulis S., Skourtsos E., “Preliminary report on the Lesvos 12 June 2017 M,” 2017.

[37] W. S. Noble, “What is a support vector machine?,” *Nat. Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, 2006.

[38] P. Rai, “Kernel Methods and Nonlinear Classification,” *CS5350/6350: Machine Learning, School of Computing, The University of Utah*, vol. 2011. pp. 1–16, 2011.

[39] Y. M. Palenzuela, “Geotext, https://pypi.python.org/pypi/geotext.” 2014.

[40] “GeoNames.” [Online]. Available: http://www.geonames.org/.

[41] “Vrisko.gr νομοί Ελλάδος.” [Online]. Available: http://www.vrisko.gr/localdir-nomoi.

[42] “Feature extraction,” *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Feature_extraction.

[43] J. Mata, R. Santano, and D. Blanco, “A Machine Learning Approach to Extract Drug-Drug Interactions in an Unbalanced Dataset,” *Proc. 1st Chall. task Drug-Drug Interact. Extr. (DDIExtraction 2011)*, pp. 59–65, 2011.

[44] J. Brownlee, “8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset,” 2015. [Online]. Available:

https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/.

[45] Wikipedia, "Bag-of-words model." [Online]. Available: https://en.wikipedia.org/wiki/Bag-of-words_model.

[46] L. Barbosa and J. Feng, "Robust Sentiment Detection on Twitter from Biased and Noisy Data," *Coling*, no. August, pp. 36–44, 2010.

[47] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 2011, pp. 30–38.

[48] W. J. Wilbur and K. Sirotkin, "The automatic identification of stop words," *J. Inf. Sci.*, vol. 18, no. 1, pp. 45–55, 1992.

[49] L. Zhao and C. Zeng, "Using Neural Networks to Predict Emoji Usage from Twitter Data," pp. 1–6.

[50] P. K. Novak, J. Smailović, B. Sluban, and I. Mozetič, "Sentiment of emojis," *PLoS One*, vol. 10, no. 12, 2015.

[51] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2012.

[52] "Cross-validation: evaluating estimator performance," 2015. [Online]. Available: http://scikit-learn.org/stable/modules/cross_validation.html.

[53] T. Dietterich, "Overfitting and undercomputing in machine learning," *ACM Comput. Surv.*, vol. 27, no. 3, pp. 326–327, 1995.

[54] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation," in *Proceedings of the Tenth International Workshop on Multimedia Data Mining - MDMKDD '10*, 2010, pp. 1–10.

[55] WikiPedia, "PageRank." [Online]. Available: https://en.wikipedia.org/wiki/PageRank.