Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Ανάλυση συναισθημάτων σύντομων μηνυμάτων σε εφαρμογή Android με χρήση τεχνολογιών Google Cloud Messaging και Data Mining (Μέρος Α)** **Sentiment analysis of instant messages on android application using Google Cloud Messaging technologies and Data Mining (Part A)**
Ονοματεπώνυμο Φοιτητή	**ΙΩΑΝΝΗΣ ΠΑΡΑΣΚΑΚΗΣ**
Πατρώνυμο	**ΧΡΗΣΤΟΣ**
Αριθμός Μητρώου	**ΜΠΣΠ13086**
Επιβλέπων	**Αλέπης Ευθύμιος, Επίκουρος Καθηγητής**

Ημερομηνία Παράδοσης **Σεπτέμβριος 2016**

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Αλέπης Ευθύμιος Επίκουρος Καθηγητής	Πατσάκης Κωνσταντίνος Επίκουρος Καθηγητής	Τσιχριντζής Γεώργιος Καθηγητής

Περίληψη

Εκατοντάδες εκατομμύρια ανθρώπων παγκοσμίως χρησιμοποιούν ιστοσελίδες microblogging και εφαρμογές αποστολής άμεσων μηνυμάτων προκειμένου να ανταλλάξουν ιδέες και απόψεις. Το φαινόμενο αυτό προσεγγίζει πολύ μεγάλο ενδιαφέρον στο τομέα της ανάλυσης των συναισθημάτων που κρύβονται πίσω από αυτές τις αλληλεπιδράσεις. Η ραγδαία εξέλιξη των φορητών συσκευών αλλά και των δυνατοτήτων τους (συνεχής τροφοδοσία από διάφορους αισθητήρες) δίνει στην ερευνητική κοινότητα ένα μεγάλο όγκο δεδομένων που συνοδεύουν τα μηνύματα των χρηστών και μπορούν να χρησιμοποιηθούν για τη καλύτερη κατανόηση των συνθηκών (περιβαλλοντικών και συναισθηματικών) υπό τις οποίες οι χρήστες χρησιμοποιούν τις προαναφερθείσες εφαρμογές και υπηρεσίες.

Η ανάλυση των συναισθημάτων (sentiment analysis), γνωστή και ως εξόρυξη γνώμης (opinion mining) είναι ο τομέας της επιστήμης που μελετά γνώμες, συναισθήματα, αξιολογήσεις, εκτιμήσεις, τις στάσεις και τα συναισθήματα των ανθρώπων προς οντότητες, όπως τα προϊόντα, τις υπηρεσίες, οργανισμούς, άτομα, τις εκδηλώσεις, τα διάφορα κοινωνικά θέματα και τα χαρακτηριστικά τους. Ο τομέας της ανάλυσης των συναισθημάτων ως τομέας της έρευνας είναι μεγάλος και σχετικά νέος και ανεξερεύνητος. Η πρώτη ίσως αναφορά του όρου εντοπίζεται στο όχι και τόσο μακρινό, για τα ερευνητικά δεδομένα 2003 (Jeonghee Yi, 2003).

Το έργο της εξαγωγής συμπερασμάτων από αλληλεπιδράσεις χρηστών σε διαδικτυακά κοινωνικά δίκτυα OSN και η ανάλυση της φυσικής γλώσσα NLP γίνεται ακόμα πιο δύσκολο αν αναλογιστούμε το εύρος των διαθέσιμων γλωσσών και διαλέκτων και την έλλειψη οργανωμένων και ολοκληρωμένων συνόλων δεδομένων, σε συνδυασμό με τη πληθώρα των συναισθημάτων που μπορεί να κρύβονται σε ένα μικρό κείμενο 140 χαρακτήρων (Twitter).

Στόχος της παρούσας διπλωματικής εργασίας είναι η δημιουργία μίας ολοκληρωμένης πλατφόρμας ανταλλαγής σύντομων μηνυμάτων με χρήση τεχνολογιών Cloud (Google Cloud Messaging). Επιπλέον η εν λόγω πλατφόρμα θα περιέχει ένα σύστημα αξιολόγησης και κατηγοριοποίησης των μηνυμάτων σε κλάσεις συναισθημάτων χρησιμοποιώντας τεχνικές μοντελοποίησης και κατηγοριοποίησης με τη χρήση του εργαλείου RapidMiner. Πρόκειται λοιπόν για ένα ολοκληρωμένο σύστημα μετάδοσης και αξιολόγησης σύντομων μηνυμάτων κειμένου που περιλαμβάνει μία εφαρμογή χρήστη για κινητά Android, μία εφαρμογή εξυπηρετητή που διαχειρίζεται την αποθήκευση και τη μετάδοση των μηνυμάτων μέσω του Google Cloud Messaging και τέλος ενός μοντέλου αξιολόγησης των μηνυμάτων.

Abstract

Millions of people around the world use, on a daily basis, microblogging sites and instant messaging applications in order to communicate, exchange ideas and opinions. This phenomenon has attracted the interest of the research community in the realm of sentiment analysis of these interactions. The evolution of the handheld devices and their increasing capabilities (in terms of sensors and hardware) produces a vast amount of instant message meta-data which are available and can be used to better understand the context (environmental and sentimental) around the use of the aforementioned applications and services.

Sentiment analysis, also known as opinion mining is the area of science that studies the opinions, feelings, evaluations, assessments, attitudes and feelings of the people to entities such as products, services, organizations, people, events, social issues and their characteristics. The analysis of emotions as an area of research is large and relatively new and unexplored. The first mention perhaps the term found in the not too distant, for research data 2003 (Jeonghee Yi, 2003).

The task of inference from user interactions in online social networks (OSN) and the analysis of natural language (NLP) becomes even more difficult if we consider the range of available languages and dialects and the lack of organized and comprehensive data sets in combination with the plethora of emotions that can be hidden in a small text of 140 characters (Twitter).

The aim of this thesis is to create an integrated platform for exchanging short messages using Cloud technologies (Google Cloud Messaging). Furthermore the aforementioned platform will incorporate an evaluation and categorization system of messages in emotion classes using modeling and classification techniques build with the use of RapidMiner tool. It is thus an integrated transmission and evaluation of small text message that includes a user application for mobile Android devices, an application server that manages the storage and transmission of messages through the Google Cloud Messaging and a message evaluation model.

Περιεχόμενα

Εισαγωγή

Τα διαδικτυακά κοινωνικά δίκτυα OSN έχουν αποδειχθεί ιδιαίτερα δημοφιλή "σημεία" ανταλλαγής απόψεων, σκέψεων και συναισθημάτων που αφορούν διάφορα θέματα της καθημερινότητας. Ο αριθμός των ενεργών χρηστών και ο όγκος των δεδομένων που δημιουργούνται στα εν λόγω κοινωνικά δίκτυα είναι ιδιαίτερα μεγάλος. Ενδεικτικά, το Twitter διαθέτει 200 εκατομμύρια ενεργούς χρήστες οι οποίοι δημιουργούν καθημερινά περίπου 400 εκατομμύρια σύντομα μηνύματα (έως 140 χαρακτήρες) *(Wickre, 2013)*. Με το μεγαλύτερο μέρος αυτής της πληροφορίας να είναι διαθέσιμη στο ερευνητικό κοινό (π.χ. το 90% των προαναφερθέντων μηνυμάτων είναι προσβάσιμο από όλους) *(M. Cha, 2010)* δημιουργείται η ευκαιρία να δημιουργηθούν τα σύνολα δεδομένων που θα μας βοηθήσουν στην εξέλιξη της επιστήμης της επεξεργασίας φυσικής γλώσσας NLP αλλά και της εξόρυξης γνώμης μέσω της ανάλυσης αυτών των αλληλεπιδράσεων.

Ταυτόχρονα η ραγδαία εξέλιξη, αναβάθμιση και διείσδυση των έξυπνων τηλεφώνων και κινητών συσκευών στη καθημερινότητα των χρηστών των διαδικτυακών κοινωνικών δικτύων OSN δημιουργεί άλλη μία ιδιαίτερα σημαντική ροή δεδομένων που μπορεί να μας βοηθήσει στη καλύτερη κατανόηση των συναισθημάτων και των απόψεων των χρηστών. Αυτή η ροή δεν είναι άλλη από την ροή δεδομένων που προσφέρουν οι εγκατεστημένοι αισθητήρες των σύγχρονων έξυπνων κινητών συσκευών. Τα σύγχρονα κινητά τηλέφωνα και κινητές συσκευές είναι εφοδιασμένες με αισθητήρες που μπορούν να μας παρέχουν πολύτιμες πληροφορίες σχετικά με τη τοποθεσία του χρήστη, τη ταχύτητα και την επιτάχυνση της συσκευής, τη φωτεινότητα, τη κλίση της συσκευής και άλλα στοιχεία που μας βοηθούν να καταλάβουμε καλύτερα το γενικό πλαίσιο της χρήσης της εκάστοτε συσκευής. Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν στην εξαγωγή χρήσιμων συμπερασμάτων ως προς τα εν λόγω μηνύματα. Η σημασία αυτής της περιφερειακής πληροφορίας (πληροφορίας που δε συμπεριλαμβάνεται στο μήνυμα αυτό καθαυτό αλλά στη πληροφορία που υπάρχει και συλλέγεται στη συσκευή ή στην αλληλεπίδραση του χρήστη με τη συσκευή) μπορεί να γίνει κατανοητή αν αναλογιστούμε το εξής παράδειγμα. Δύο χρήστες γράφουν κριτικές για ένα προϊόν, ας θεωρήσουμε ότι αφορά σε ένα κρεβάτι που και οι δύο χρήστες αγόρασαν πρόσφατα και κλήθηκαν να αξιολογήσουν. Η κριτική του πρώτου χρήστη είναι θετική ενώ αυτή του δεύτερου χρήστη είναι αρνητική ως προς το γενικό αίσθημα του κειμένου. Με τα δεδομένα που έχουμε αυτή τη στιγμή, και αφορούν μόνο το περιεχόμενο της κριτικής, η βαρύτητα των δύο κριτικών είναι ίδια. Ας προσθέσουμε τώρα στα δεδομένα που έχουμε την εξής πληροφορία.

Τοποθεσία Κριτή #1 (GPS & WiFi)	Ο χρήστης βρίσκεται στο σπίτι του και έχει συνδεθεί σε οικιακό δίκτυο
Ώρα συγγραφής από Κριτή #1	23:00 (Ισχυροποιείται ο ισχυρισμός ότι ο χρήστης βρίσκεται στο σπίτι του)
Keylogging Κριτή \#1	Ορθογραφικά και συντακτικά ορθό κείμενο, λίγες διαγραφές χαρακτήρων
Τοποθεσία Κριτή \#2 (GPS & WiFi)	Ο χρήστης βρίσκεται στο χώρο εργασίας του και έχει συνδεθεί σε επαγγελματικό δίκτυο
Ώρα συγγραφής από Κριτή #2	12:00 (Ισχυροποιείται ο ισχυρισμός ότι ο χρήστης βρίσκεται στο χώρο εργασίας του)
Keylogging Κριτή #2	Ορθογραφικά και συντακτικά λάθη, πολλές διαγραφές χαρακτήρων

Πίνακας 1: Πληροφορία που παράγεται από τους αισθητήρες της συσκευής

Από τη πληροφορία που παρουσιάζεται στον πίνακα 1 μπορούμε να υποθέσουμε πως ο κριτής #2 βρίσκεται σε μία κατάσταση σύγχυσης (ίσως πίεσης από την εργασία του) με αποτέλεσμα να μη δίνει την απαραίτητη προσοχή στη συγγραφή της κριτικής ενώ ο κριτής #1 φαίνεται να είναι σε πιο καλή περιβάλλουσα κατάσταση. Αναλύοντας τις επιπρόσθετες πληροφορίες μπορούμε εύκολα να συμπεράνουμε πως η ισορροπία μεταβάλλεται και πως η δεύτερη κριτική έχει μικρότερη αξία από τη πρώτη.

Τόσο σε αυτό το μέρος της διπλωματικής εργασίας (Α Μέρος) όσο και στο επόμενο μέρος του Μποζιονέλου Στέφανου (Β Μέρος) παρουσιάζουμε τόσο σε θεωρητικό όσο και σε τεχνολογικό επίπεδο τη πλατφόρμα Thesis, τις συνιστώσες της αλλά και τις σχεδιαστικές αποφάσεις που πήραμε κατά την έρευνα και την υλοποίηση προκειμένου να δημιουργήσουμε ένα ασφαλές σύνολο δεδομένων αναγνώρισης συναισθημάτων από σύντομα μηνύματα στην Ελληνική γλώσσα.

1.1 Σκοπός και στόχος της διπλωματικής

Σκοπός της διπλωματικής εργασίας αυτής είναι να μπορέσουν να αξιοποιηθούν οι δυνατότητες των έξυπνων συσκευών στον τομέα της αναγνώρισης συναισθημάτων και εξόρυξης γνώσης από κείμενα. Πιο συγκεκριμένα σκοπός μας είναι δημιουργήσουμε ένα ασφαλές σύνολο δεδομένων που αποτελείται τόσο από μηνύματα όσο και από metadata που μπορούν να επαυξήσουν τη πληροφορία που είναι διαθέσιμη σε κάθε χρονική στιγμή.

Στόχος της παρούσας διπλωματικής είναι η σχεδίαση και η υλοποίηση μίας πλατφόρμας που θα δίνει τη δυνατότητα στους χρήστες να ανταλλάσσουν μηνύματα μέσω της προσωπικής τους κινητής έξυπνης συσκευής χρησιμοποιώντας τεχνολογίες Cloud. Επίσης, στόχος μας με την υλοποίηση της συγκεκριμένης πλατφόρμας, είναι η διευκόλυνση του ερευνητικού έργου της ανάλυσης συναισθημάτων και απόψεων των χρηστών δημιουργώντας ένα σύνολο δεδομένων στην Ελληνική γλώσσα για περεταίρω μελέτη.

1.2 Διάρθρωση της διπλωματικής

Η παρούσα διπλωματική αποτελείται από 4 βασικά κεφάλαια. Στο επόμενο και δεύτερο κεφάλαιο της παρούσας διπλωματικής περιγράφουμε αναλυτικά το σύστημα που υποστηρίζει την αναγνώριση των συναισθημάτων στην πλατφόρμα και αναλύουμε το πρωτόκολλο επικοινωνίας μεταξύ των συνιστωσών του συστήματος αλλά και τις διεργασίες που εκτελούνται σε κάθε βήμα του κάθε πρωτοκόλλου. Παρέχουμε κομμάτια κώδικα για κατανόηση των διαδικασιών και εξηγούμε την υλοποίηση και την προσαρμογή μοντέλων αναγνώρισης συναισθημάτων στην υλοποίηση μας.

Στο τρίτο κεφάλαιο με τίτλο related work συνοψίζουμε την πορεία της ερευνητικής κοινότητας στον τομέα του sentiment analysis και opinion mining με χρήση "έξυπνων κινητών συσκευών" και εξηγούμε τους λόγους που αποφασίσαμε να αναπτύξουμε αυτήν την εφαρμογή προκειμένου να συνεισφέρουμε στον τομέα αυτό.

Στο τέταρτο και τελευταίο μέρος της εργασίας μας παρουσιάζουμε συγκεντρωτικά τα αποτελέσματα της χρήσης της εφαρμογής σε πραγματικό χρόνο και αναλύουμε τα συμπεράσματα που προέκυψαν τόσο κατά την υλοποίηση όσο και κατά τη δοκιμή.

Περιγραφή του συστήματος αναγνώρισης συναισθημάτων

2.1 Απαραίτητο θεωρητικό υπόβαθρο

Προκειμένου να κατανοήσει ο αναγνώστης το περιεχόμενο αυτής της διπλωματικής εργασίας και δεδομένης της λειτουργίας του πληροφοριακού συστήματος που ονομάζεται Thesis Platform, σε διαφορετικές πλατφόρμες, λειτουργικά συστήματα και συσκευές καθώς και τη χρήση διαφόρων εργαλείων, θα πρέπει να έχει κάποιο επίπεδο γνώσεων στα παρακάτω θέματα.

2.1.1 Εξόρυξη γνώσης από δεδομένα

Η πλατφόρμα Thesis υποστηρίζει την αναγνώριση των συναισθημάτων του χρήστη μέσω της κατηγοριοποίησης των μηνυμάτων του. Για την υλοποίηση αυτού του συστήματος χρησιμοποιούμε την πλατφόρμα του RapidMiner για την επεξεργασία κειμένου αλλά και την κατηγοριοποίηση των νέων εισερχόμενων μηνυμάτων. Ο αναγνώστης πρέπει να έχει τις βασικές γνώσεις εξόρυξης γνώσεων από δεδομένα που συμπεριλαμβάνουν την επεξεργασία κειμένου για εξαγωγή διανύσματος των χαρακτηριστικών της και τους αλγορίθμους μάθησης και δημιουργίας μοντέλων.

2.2 Λογισμικό

Το λογισμικό που χρησιμοποιήσαμε για τη δημιουργία του συστήματος που υποστηρίζει την εφαρμογή Thesis αλλά και για τη δημιουργία της ίδιας της εφαρμογής και των λειτουργιών της είναι:

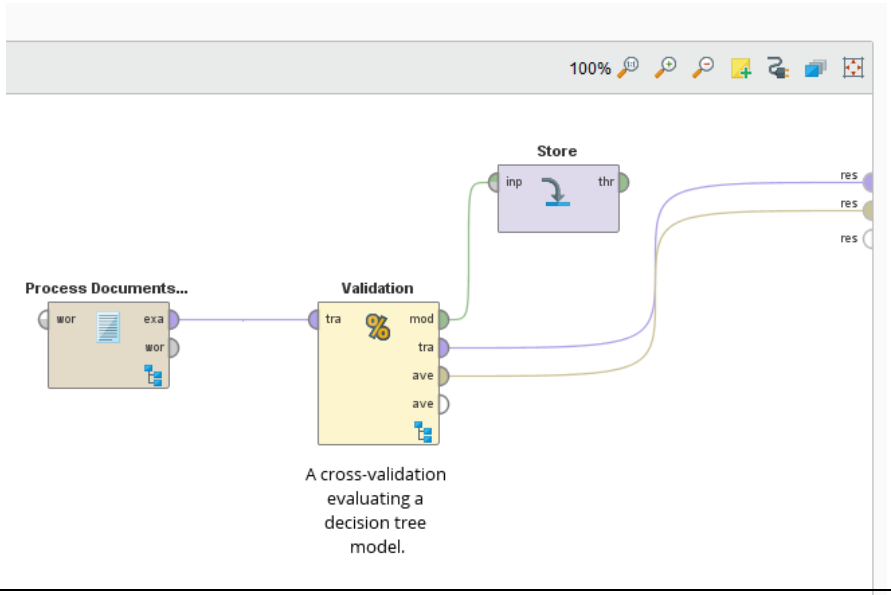
- Η αναγνώριση των συναισθημάτων από μηνύματα γίνεται με τη βοήθεια του RapidMiner 7.1 και του extension του Text Processing 7.1.1 που χρησιμοποιείται για την επεξεργασία κειμένου.
- Η εφαρμογή αναγνώρισης συναισθημάτων εκτελείται από έναν Java Application Server τον Tomcat ο οποίος αναλαμβάνει την εκτέλεση του classifier στη μεριά του εξυπηρετητή.

2.3 Rapid Miner

Σε αυτό το κεφάλαιο θα αναλύσουμε τον τρόπο με τον οποίο χρησιμοποιώντας το Rapid Miner και τα extensions του καταφέραμε να δημιουργήσουμε το μοντέλο που θα χρησιμοποιεί η εφαρμογή μας για να κάνει το classification για το unlabeled κείμενο που θα αποστέλλει ο χρήστης. Θα αναλύσουμε τον τρόπο με τον οποίο γίνεται το validation του μοντέλου και τα ποσοστά επιτυχίας που το μοντέλο αυτό επιτυγχάνει και τέλος θα περιγράψουμε τον τρόπο με τον οποίο έγινε η συλλογή των κειμένων που απαρτίζουν το training set από το οποίο γίνεται η εξόρυξη γνώσης.

2.3.1 Φάση 1η - Language Detection: Επιλογή Data Set / Training Set

Τα κείμενα που απαρτίζουν το training set αφορούν δεκατέσσερεις (14) διαφορετικές γλώσσες. Τα κείμενα επιλέχτηκαν από το Wikipedia.org καθώς η συγκεκριμένη σελίδα προσφέρει μία μεγάλη συλλογή από πληροφορίες γραμμένες σε πολλές γλώσσες απλώς εισάγοντας τον κώδικα της χώρας (Country Code CC). Έτσι για παράδειγμα για την Σουηδία με κωδικό SV αντλήσαμε κείμενα από την σελίδα http://sv.wikipedia.org . Προκειμένου να εξασφαλίσουμε ότι ακολουθώντας τα link θα εξακολουθούσαμε να παίρνουμε κείμενα στην γλώσσα που μας ενδιαφέρει ορίσαμε στον τελεστή crawl web να ακολουθάει μόνο link που ακολουθούσαν τον κανόνα http://CC.* όπου CC ο κωδικός της γλώσσας που θα θέλαμε να είναι γραμμένο το κείμενο (στο παράδειγμα μας http://sv.*). Από κάθε γλώσσα επιλέξαμε 10 κείμενα τα οποία αποθηκεύαμε σε φάκελο με όνομα το CC της γλώσσας (αυτοί οι φάκελοι όπως θα δούμε και αργότερα θα αποτελέσουν τις κλάσεις μας).



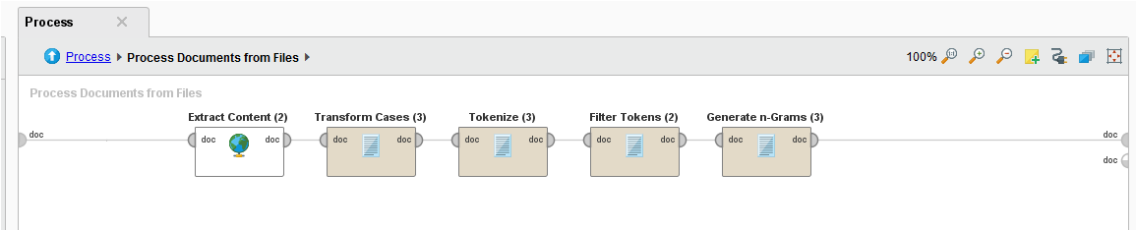
Rapid Miner Process

2.3.2 Φάση 1η - Language Detection: Δημιουργία Μοντέλου

Η δημιουργία μοντέλου περιλαμβάνει όλες εκείνες τις απαραίτητες δραστηριότητες που πρέπει να γίνουν προκειμένου τα κείμενα που αποτελούν την κάθε κλάση να μετατραπούν σε διάνυσμα, που θα περιέχει μόνο την ωφέλιμη πληροφορία. Στη συνέχεια το διάνυσμα αυτό θα περάσει από ένα validation με κάποιον αλγόριθμο και θα δημιουργηθεί ένα μοντέλο. Το μοντέλο αυτό θα δέχεται ένα άλλο διάνυσμα και θα το κατηγοριοποιεί ανάλογα με τις κλάσεις στις οποίες έχει εκπαιδευτεί.

Το διάνυσμα που αναφέρουμε δημιουργείται με το κριτήριο του TF-IDF το οποίο ουσιαστικά απεικονίζει τη βαρύτητα της κάθε λέξης (όπως αυτές δημιουργούνται από την επεξεργασία των κειμένων) σε κάθε ένα κείμενο με βάση τον αριθμό εμφάνισης (συχνότητα) της κάθε λέξης. Η προ-επεξεργασία των κειμένων έγκειται στην εξαγωγή του ωφέλιμου κειμένου από το πραγματικό κείμενο που είναι ένα κείμενο html. Η εξαγωγή του ωφέλιμου κειμένου γίνεται με μία σειρά από επεξεργασίες (τελεστές του Rapid Miner) οι οποίοι εντός του τελεστή process documents λαμβάνουν όλα τα αρχεία των δεκατεσσάρων (14) γλωσσών ,ως δεκατέσσερα (14) διαφορετικά labels (κλάσεις), και:

- Αφαιρούν όλα τα γνωστά html tags.
- Μετατρέπουν όλους τους χαρακτήρες σε lower case.
- Αφαιρούν τους ειδικούς χαρακτήρες.
- Αφαιρούν τις πολύ μικρές και τις πολύ μεγάλες λέξεις που κατά πάσα πιθανότητα δεν έχουν σημασία για το γενικό νόημα του κειμένου.
- Χωρίζουν το σύνολο των λέξεων σε συμβολοσειρές 3 χαρακτήρων.



Εξαγωγή ωφέλιμου κειμένου

Αφού λοιπόν γίνει η παραπάνω επεξεργασία και υπολογιστεί το διάνυσμα που περιγράφουμε παραπάνω το training set μας είναι έτοιμο να ελεγχθεί από έναν ten-fold nominal cross validation. Αυτό το validation επιτυγχάνεται χωρίζοντας το training set σε training set και test set (στο test set αποκρύπτεται η κλάση/label) και δοκιμάζεται ο αλγόριθμος που επιλέγουμε στο υποσύνολο του αρχικού training set. Χρησιμοποιούμε stratified sampling για πιο ομοιογενές δείγμα (δηλ. να περιέχονται δείγματα από όλες τις κλάσεις ομοιόμορφα).Ο αλγόριθμος του validation που επιλέξαμε για την δημιουργία του μοντέλου μας είναι ο k-NN (k πλησιέστεροι γείτονες) με παράμετρο 5. Ο αλγόριθμος αυτός μας έβγαλε τα καλύτερα αποτελέσματα από τους αλγορίθμους που χρησιμοποιήσαμε.

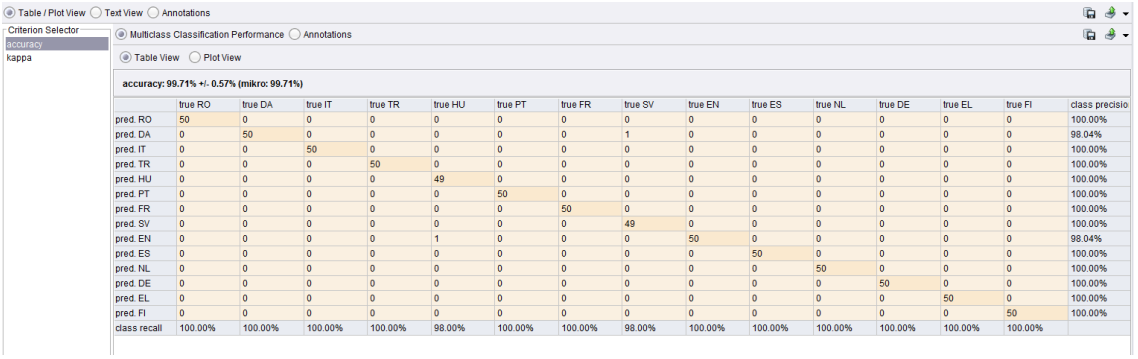


Table / Plot View ○ Text View ○ Annotations

Criterion Selector: accuracy, kappa

Multiclass Classification Performance ○ Annotations

Table View ○ Plot View

accuracy: 99.71% +/- 0.57% (mikro: 99.71%)

	true RO	true DA	true IT	true TR	true HU	true PT	true FR	true SV	true EN	true ES	true NL	true DE	true EL	true FI	class precisio
pred. RO	50	0	0	0	0	0	0	0	0	0	0	0	0	0	100.00%
pred. DA	0	50	0	0	0	0	0	1	0	0	0	0	0	0	98.04%
pred. IT	0	0	50	0	0	0	0	0	0	0	0	0	0	0	100.00%
pred. TR	0	0	0	50	0	0	0	0	0	0	0	0	0	0	100.00%
pred. HU	0	0	0	0	49	0	0	0	0	0	0	0	0	0	100.00%
pred. PT	0	0	0	0	0	50	0	0	0	0	0	0	0	0	100.00%
pred. FR	0	0	0	0	0	0	50	0	0	0	0	0	0	0	100.00%
pred. SV	0	0	0	0	0	0	0	49	0	0	0	0	0	0	100.00%
pred. EN	0	0	0	0	1	0	0	0	50	0	0	0	0	0	98.04%
pred. ES	0	0	0	0	0	0	0	0	0	50	0	0	0	0	100.00%
pred. NL	0	0	0	0	0	0	0	0	0	0	50	0	0	0	100.00%
pred. DE	0	0	0	0	0	0	0	0	0	0	0	50	0	0	100.00%
pred. EL	0	0	0	0	0	0	0	0	0	0	0	0	50	0	100.00%
pred. FI	0	0	0	0	0	0	0	0	0	0	0	0	0	50	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	98.00%	100.00%	100.00%	98.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	

k-NN Class Recall. Phase 1

Παρατηρούμε μεγάλο class recall (επιτυχίες πρόβλεψης) αλλά και μεγάλο class precision (μεγάλη ακρίβεια στις προβλέψεις).

Εκτός από τον παραπάνω αλγόριθμο χρησιμοποιήσαμε και:

- SVM – για πολλές κλάσεις (50 κείμενα ανά κλάση).
- Decision tree (50 κείμενα ανά κλάση).
- Neural Net (50 κείμενα ανά κλάση).

Οι παραπάνω αλγόριθμοι απορρίφθηκαν είτε λόγω χαμηλών ποσοστών είτε λόγω έλλειψης επεξεργαστικών πόρων για την ολοκλήρωση του μοντέλου. Ενδεικτικά το neural net για δείγμα 50 κειμένων για κάθε κλάση μετά από 3 μέρες εκτέλεσης (8Gb ram / 8 CPU cores) δεν κατάφερε να μας βγάλει αποτέλεσμα.

2.3.3 Φάση 1η - Language Detection: Εκτέλεση Μοντέλου

```
//Rapid Miner run from Java
RapidMiner.setExecutionMode(ExecutionMode.COMMAND_LINE);
RapidMiner.init();
File x=new File("C:\\Users\\John\\.RapidMiner\\repositories\\Local
Repository\\Project_Prediction.rmp");
com.rapidminer.Process process = new  com.rapidminer.Process(x);
IOContainer ioResult=process.run();

ExampleSet resultSet=(ExampleSet)ioResult.getElementAt(0);
ExampleTable mytable=resultSet.getExampleTable();

Example example=resultSet.getExample(0);
Attribute predict=example.getAttributes().get("prediction(label)");

String resultString=example.getValueAsString(predict);
```

Εκτέλεση του Rapid Miner στον εξυπηρετητή μέσω java

2.3.4 Φάση 2η - Sentiment Classification: Επιλογή Data Set / Training Set

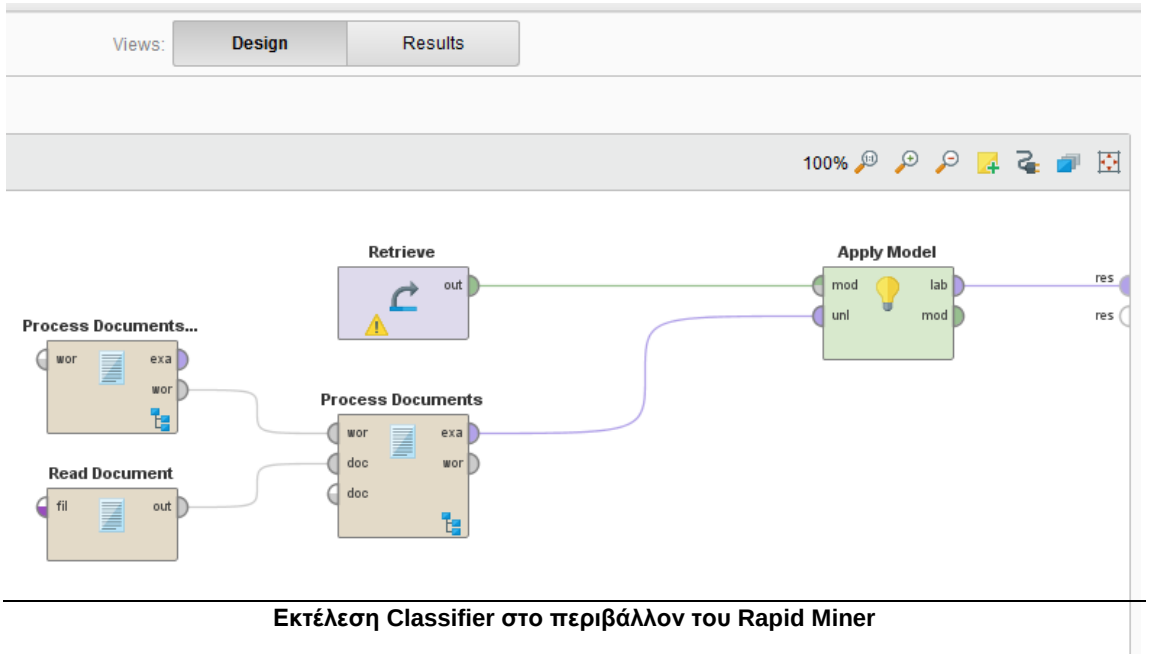
Η επιλογή των κειμένων έγινε από την ερευνητική εργασία "Sentiment Analysis of Greek Tweets and Hashtags using a Sentiment Lexicon" των Georgios Kalamatianos et al. (Arampatzis, 2015) οι οποίοι παρέχουν ένα έγγραφο με rated Tweets στην ελληνική γλώσσα. Σε αντίθεση με τη πρώτη φάση στην οποία χρησιμοποιήσαμε web crawler για τη συλλογή του δείγματος στη παρούσα φάση κληθήκαμε να προ-επεξεργαστούμε τα δεδομένα ώστε να τα εισάγουμε στο Rapid Miner στην ίδια μορφή με τη πρώτη φάση. Για το λόγο αυτό δημιουργήσαμε μία εφαρμογή σε JAVA η οποία ήταν υπεύθυνη για το data sanitization και distribution στις διαθέσιμες κλάσεις συναισθημάτων. Πιο συγκεκριμένα η εν λόγω εφαρμογή χρησιμοποιούσε σαν είσοδο το Excel που περιέχει τα rated Tweets και δημιουργούσε τα αρχεία κειμένου που είναι απαραίτητα για την διαδικασία της δημιουργίας του μοντέλου.

```
//Training Set Normalization
File file = new File("D:\\Dropbox\\2.MSc University of
Piraeus\\THESIS\\Sentiments\\RatedTweets.xlsx");
Workbook wb = new XSSFWorkbook(new FileInputStream(file));

for (int i = 0; i < wb.getNumberOfSheets(); i++) {
    Sheet sheet = wb.getSheetAt(i);
    int j = 0;
    for (Row row : sheet) {
        try {
        PrintWriter writer = new PrintWriter("D:\\Dropbox\\2.MSc
University of Piraeus\\
        THESIS\\Sentiments\\"+sentiment_classes[max_index]+"\\
        "+sentiment_classes[max_index]
        +"_file_"+j+".txt", "UTF-8");
            writer.println(row.getCell(14));
            writer.close();

            j++;
        } catch (Exception ex) {
            System.out.println(ex.toString());
        }
    }
}
```

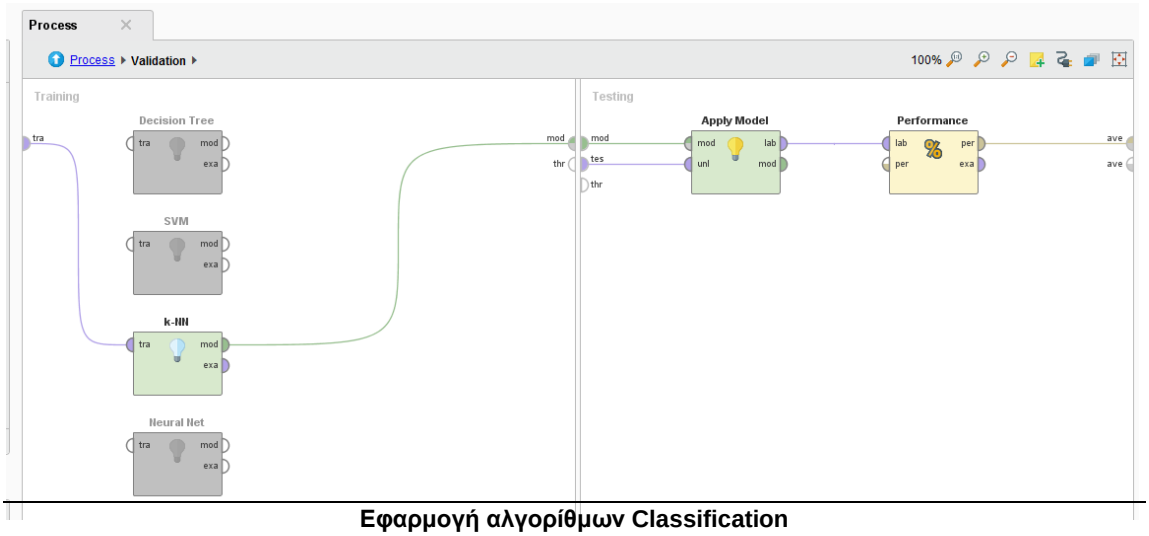
Κανονικοποίηση του Training Set

2.3.5 Φάση 2η - Sentiment Classification: Δημιουργία Μοντέλου

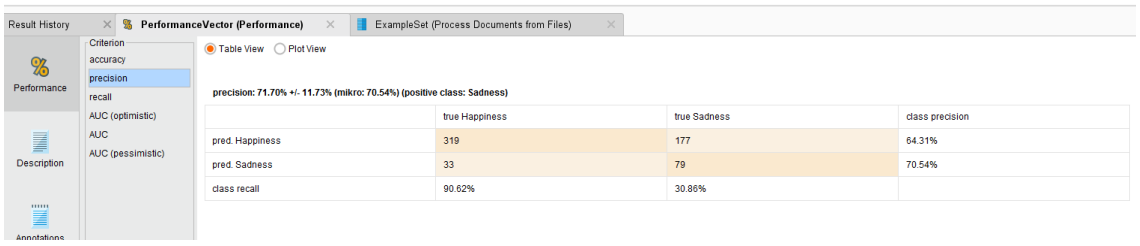
Η λογική δημιουργίας του πρώτου classifier ακολουθήθηκε και στη δημιουργία του δεύτερου μοντέλου. Αλλάζοντας κάποιες παραμέτρους στη διαδικασία επεξεργασίας των κειμένων όπως το vector creation και το prune method καταφέραμε από αυτό το μικρό data set να επιτύχουμε class precision μεγαλύτερη από 70% (71.9%).



Εκτέλεση Classifier στο περιβάλλον του Rapid Miner



Εφαρμογή αλγορίθμων Classification



K-NN Class Recall. Phase 2

2.3.6 Φάση 2η - Sentiment Classification: Εκτέλεση Μοντέλου

```
//Rapid Miner run from Java
RapidMiner.setExecutionMode(ExecutionMode.COMMAND_LINE);
RapidMiner.init();
File x=new File("C:\\Users\\John\\.RapidMiner\\repositories\\Local
Repository\\Sentiment_MODEL.rmp");
com.rapidminer.Process process = new  com.rapidminer.Process(x);
IOContainer ioResult=process.run();

ExampleSet resultSet=(ExampleSet)ioResult.getElementAt(0);
ExampleTable mytable=resultSet.getExampleTable();

Example example=resultSet.getExample(0);
Attribute predict=example.getAttributes().get("prediction(label)");

String resultString=example.getValueAsString(predict);
```

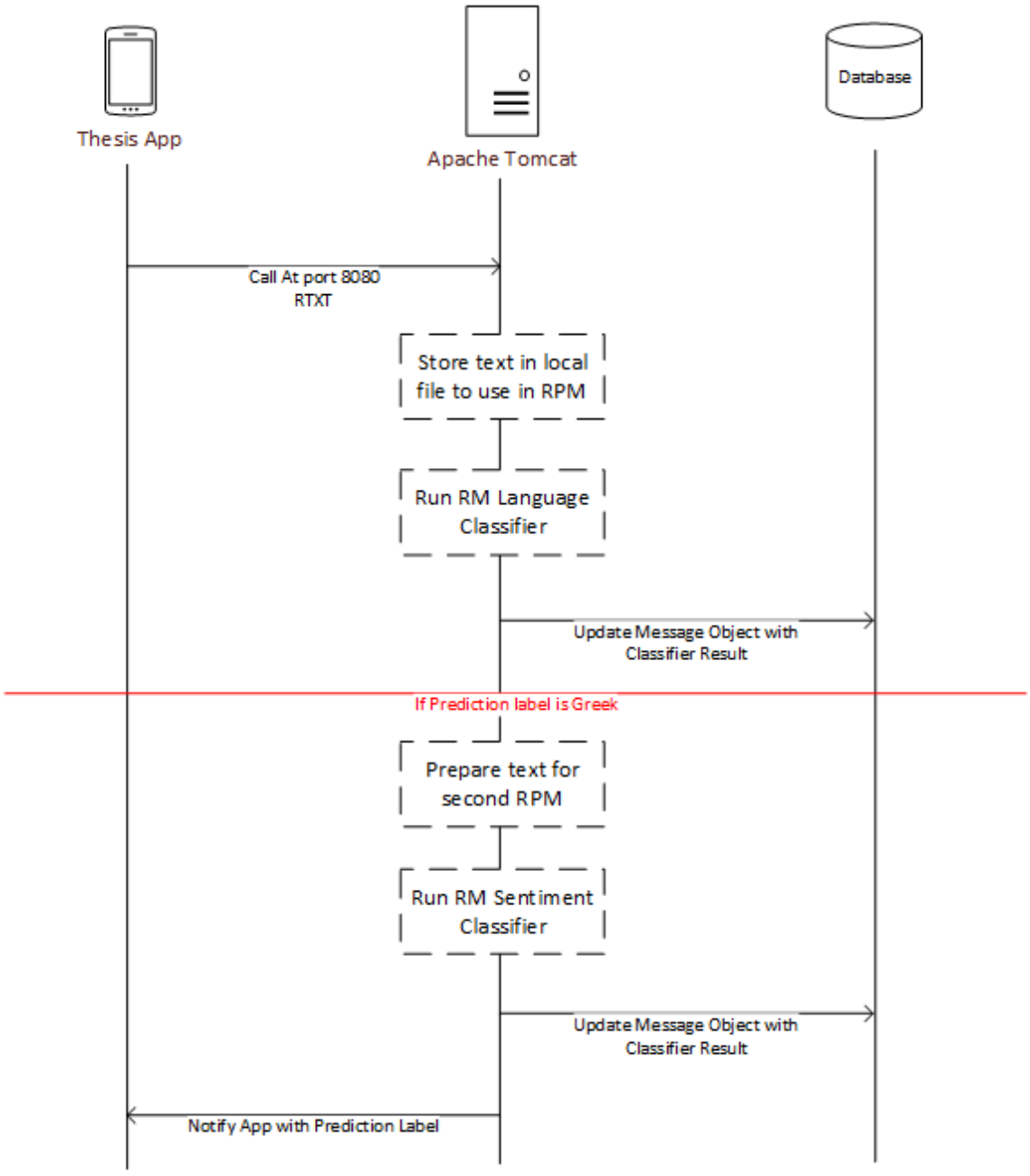
Εκτέλεση του Sentiment Classifier στον εξυπηρετητή μέσω java

Πρωτόκολλο επικοινωνίας Χρήστη – Classifier

Το πρωτόκολλο επικοινωνίας ανάμεσα στο χρήστη και τον Classifier παρουσιάζεται στο σχήμα 27 και ξεκινάει με την αποστολή ενός μηνύματος μέσω της εφαρμογής του χρήστη και η σύνδεση γίνεται στη θύρα 8080 του εξυπηρετητή. Με την εγκαθίδρυση της σύνδεσης η εφαρμογή του χρήστη αποστέλλει το μήνυμα κειμένου στον εξυπηρετητή για να περάσει από τους δύο classifiers. Αρχικά το μήνυμα περνάει από το πρώτο μοντέλο το οποίο αναγνωρίζει τη γλώσσα του εν λόγω κειμένου. Στη περίπτωση που η γλώσσα που αναγνωρίσει ο πρώτος classifier είναι η ελληνική τότε εκτελείται το δεύτερο μοντέλο προκειμένου να προχωρήσει σε sentiment classification του εν λόγω μηνύματος. Ο λόγος για τον οποίο εκτελείται η δεύτερη διαδικασία μόνο για ελληνικά κείμενα έγκειται στα διαθέσιμα datasets για train του μοντέλου όπως έχουμε αναφέρει και στο κεφάλαιο 2.3.2. Τα αποτελέσματα των δύο classifiers αποθηκεύονται στη βάση δεδομένων ώστε σε μελλοντική εργασία να συνδεθούν με τα υπόλοιπα δεδομένα που λαμβάνουμε κατά την αποστολή των μηνυμάτων (δεδομένα key-logging και δεδομένα από τους διαθέσιμους αισθητήρες) για τη καλύτερη κατανόηση των μηνυμάτων και εξαγωγή χρήσιμων συμπερασμάτων.

```
//Apache Tomcat Restart
try {
    Process child = Runtime.getRuntime().exec(/catalina.sh stop");
    System.out.println("Server stopped");
        Process child = Runtime.getRuntime().exec(/catalina.sh
start");
    System.out.println("Server Started");
} catch (IOException ex) {

Logger.getLogger(Installation.class.getName()).log(Level.SEVERE,
null, ex);
     System.out.println("Error in restarting Server");
}
```

Επανεκκίνηση του Apache Tomcat μέσω του εξυπηρετητή σε γλώσσα java



Πρωτόκολλο επικοινωνίας χρήστη – classifier

Σχετική Βιβλιογραφία

4.1 Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural *Language Processing Techniques*

Το 2003 οι Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu και Wayne Niblacki *(Jeonghee Yi, 2003)* δημοσιεύουν τη μελέτη τους σχετικά με τον "Αναλυτή Συναισθημάτων" (Sentiment Analyzer) που είχαν δημιουργήσει. Πρόκειται για μία από τις πρώτες ερευνητικές αναφορές στην επιστήμη του Sentiment Analysis και Opinion Mining. Ο εν λόγω αναλυτής κατηγοριοποιούσε τα συναισθήματα ανάλογα με το εκάστοτε θέμα ενός κειμένου χρησιμοποιώντας τεχνολογίες επεξεργασίας φυσικής γλώσσας NLP. Η εργασία τους χωριζόταν σε τρία επιμέρους συστήματα τα οποία ήταν 1) Ο αναλυτής της θεματικής ενότητας, 2) Το σύστημα εξόρυξης συναισθημάτων και 3) Ο αναλυτής συναισθημάτων που συνδύαζε τα 2 προαναφερθέντα υποσυστήματα. Ο εν λόγω αναλυτής εφαρμόστηκε σε κείμενα αξιολόγησης προϊόντων. Τα αποτελέσματα της έρευνας ήταν ιδιαίτερα θετικά και ενθαρρυντικά για την ερευνητική κοινότητα καθώς επιτεύχθηκε ποσοστό ακρίβειας που άγγιζε το 91.0% με 93.0%. Κλείνοντας τη δημοσίευση τους οι Nasukawa et al. σημειώνουν τη σημασία του ανθρώπινου παράγοντα (expert) στη διαδικασία της επικύρωσης των αποτελεσμάτων (validation process) και αναφέρουν πως ένα από τα θέματα που θα μπορούσε να αποτελέσει μελλοντική εργασία είναι η περεταίρω αυτοματοποίηση αυτής της διαδικασίας.

4.2 NRC-Canada: Building the State-of-the-Art *in Sentiment Analysis of Tweets*

Το 2013 οι Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu *(Saif M. Mohammad, 2013)* παρουσιάζουν τη μελέτη τους στην οποία περιγράφουν δύο SVM classifiers που χρησιμοποιήθηκαν για την ανάλυση συναισθημάτων σε σύντομα κείμενα όπως tweets και SMS (σε επίπεδο μηνύματος - message level task) αλλά και των συναισθημάτων που κρύβονται στους διάφορους όρους που περιλαμβάνονται μέσα στα εν λόγω μηνύματα αντίστοιχα (σε επίπεδο όρων - term level task). Αποτέλεσμα της μελέτης ήταν επίσης και η δημιουργία δύο μεγάλων λεξικών συσχέτισης ανάμεσα σε όρους και συναισθήματα. Σε αυτό το σημείο πρέπει να τονίσουμε πως η χρήση των λεξικών (lexicons) είναι μία από τις πιο διαδεδομένες μεθόδους ανάλυσης φυσικής γλώσσας και χρησιμοποιείται ευρέως για την εξόρυξη συναισθημάτων από κείμενο. Οι classifiers που δημιουργήθηκαν στα πλαίσια της εν λόγω μελέτης, οι οποίοι δοκιμάστηκαν στο Conference on Semantic Evaluation Exercises (SemEval-2013), μπορούσαν να κατηγοριοποιήσουν τα κείμενα που τροφοδοτήθηκαν στα μοντέλα σε τρεις κλάσεις ως προς τα συναισθήματα των χρηστών (θετικά, αρνητικά, ουδέτερα). Στις δοκιμές οι οποίες πραγματοποιήθηκαν στο προαναφερθέν συνέδριο οι classifiers της εν λόγω μελέτης κατέλαβαν τη πρώτη και τη δεύτερη θέση στην αναγνώριση συναισθημάτων από SMS και αναγνώριση συναισθημάτων από επιλεγμένους όρους αντίστοιχα.

4.3 Comparing and Combining Sentiment Analysis Methods

Το 2013 οι Pollyanna Gonçalves et al. *(Pollyanna Gonçalves, 2013)* δημοσιεύουν τη μελέτη τους που αφορά τη συγκριτική ανάλυση οχτώ (8) συστημάτων ανάλυσης συναισθημάτων από κείμενα. Οι συγγραφείς της εν λόγω δημοσίευσης αντιλαμβανόμενοι αρχικά τη σημασία της ανάλυσης των συναισθημάτων και των απόψεων των χρηστών των διαδικτυακών κοινωνικών δικτύων αλλά και την έλλειψη αντίστοιχης μελέτης αποφάσισαν να δημιουργήσουν μία νέα μέθοδο που συνδυάζει υπάρχουσες προσεγγίσεις και παρέχει καλύτερα αποτελέσματα κάλυψης (coverage) και σύγκλισης. Ακόμα αποτέλεσμα της εν λόγω μελέτης αποτελεί και το API iFeel που αποτελεί ένα "ανοιχτό" σύστημα αξιολόγησης και σύγκρισης μεθόδων ανάλυσης συναισθημάτων από κείμενο. Στο κύριο μέρος της μελέτης τους οι Pollyanna Gonçalves et al. παρουσιάζουν τα βήματα που ακολουθήσανε και τα εργαλεία που χρησιμοποιήσαν προκειμένου να συγκρίνουν τους διάφορους αναλυτές αλλά και τα συνδυαστικά συστήματα που προέκυψαν από τη χρήση δύο ή περισσότερων από τα προς αξιολόγηση συστήματα. Τέλος οι συγγραφείς αφού απεικονίσουν όλα τους τα ευρήματα με γραφικές παραστάσεις και συγκριτικούς πίνακες παρουσιάζουν το iFeel web System το οποίο όπως αναφέρουν "Επιτρέπει σε οποιονδήποτε να δοκιμάσει διάφορα συστήματα και μεθόδους ανάλυσης σε κείμενα της επιλογής τους".

4.4 From Tweets to Polls: Linking Text Sentiment *to Public Opinion Time Series*

Το 2010 οι Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge και Noah A. Smith (*Brendan O'Connor, 2010)* δημοσιεύουν τη μελέτη τους στην οποία συνδυάζουν αποτελέσματα εκλογών με την επεξεργασία και ανάλυση κειμένων από το Twitter και άλλες micro-blogging εφαρμογές. Στην εν λόγω δημοσίευση οι ερευνητές χρησιμοποιούν αποτελέσματα τηλεφωνικών και άλλων ερευνών σχετικά με πολιτικά θέματα όπως την εκλογή του προέδρου των Ηνωμένων πολιτειών (εκλογές 2009 ανάμεσα σε Obama και McCain) αλλά και μηνυμάτων χρηστών στο Twitter και σε άλλα παρεμφερή μέσα σε συνδυασμό με τα δημογραφικά στοιχεία που τα εν λόγω συστήματα προσφέρουν. Τα αποτελέσματα αυτής της πρωτότυπης έρευνας τόσο σε ανάλυση κειμένου όσο και στην ανάλυση προβλέψεων αναδεικνύουν την ανάγκη για τη δημιουργία συστημάτων ανάλυσης συναισθημάτων που απαντούν σε συγκεκριμένες ερωτήσεις. Τέλος οι συγγραφείς τονίζουν τη σημασία των στοιχείων που παρέχουν οι, κάθε είδους έρευνες (surveys), στην εξέλιξη του τομέα της ανάλυσης συναισθημάτων και στη δημιουργία πιο εκλεπτυσμένων εφαρμογών.

4.5 Lexical Normalization for Social Media Text

Το 2013 οι Han, B., Cook, P., and Baldwin, T. *(Han, 2013)* δημοσιεύουν τη μελέτη τους σχετικά με τις μεθόδους κανονικοποίησης των λεξικών και την αναγνώριση λέξεων και παραλλαγών που θα μπορούσαν να παρεμποδίσουν τη λειτουργικότητα των τεχνικών της επεξεργασίας φυσικής γλώσσας. Η κανονικοποίηση του κειμένου είναι μία πολύπλοκη διαδικασία και μοιάζει με εκείνη του ορθογραφικού ελέγχου αλλά διαφέρει στο γεγονός ότι συνήθως γίνεται σκόπιμα για λόγους ευκολίας γραφής ή περιορισμών του συστήματος (140 χαρακτήρες όριο για κάθε Tweet). Η διαδικασία της σύνθεσης της λέξης από συντομογραφία (deabbreviation - π.χ. b4 σε before), της ανάκτησης της λέξης ύστερα από ελεύθερη γραφή (π.χ. gooooood σε good) και η αναγνώριση παραλλαγών σε λέξεις που δεν ανήκουν στα υπάρχοντα λεξικά καθιστούν το αντικείμενο αυτής της έρευνας ακόμα πιο δύσκολο. Στη μακροσκελή δημοσίευση τους οι Han et al. παρουσιάζουν τα αποτελέσματα της δοκιμής διάφορων τεχνικών κανονικοποίησης αλλά και μία συγκριτική μελέτη ανάμεσα σε υπάρχοντα συστήματα και στην προτεινομένη προσέγγιση. Τα αποτελέσματα της μελέτης είναι ιδιαίτερα ενθαρρυντικά καθώς η προτεινόμενη λύση υπερτερεί όλων των άλλων λύσεων σε όλους τους τομείς της συγκριτικής αξιολόγησης.

4.6 Sentiment Analysis of Short Informal Texts

Το 2014 οι Svetlana Kiritchenko, Xiaodan Zhu και Saif M. Mohammad (*Svetlana Kiritchenko, 2014)* παρουσιάζουν τη μελέτη τους στην οποία περιγράφουν ένα state-of-the-art σύστημα ανάλυσης συναισθημάτων σε σύντομα μηνύματα όπως SMSs και Tweets τόσο σε επίπεδο μηνύματος όσο και σε επίπεδο όρων (message-level task και term-level task). Το εν λόγω σύστημα βασίζεται σε στατιστική κατηγοριοποίηση με επίβλεψη (supervised statistical text classification) το οποίο εκμεταλλεύεται μία πληθώρα σημασιολογικών και συναισθηματικών λειτουργιών με τη χρήση νέων, ειδικών λεξικών μεγάλης κάλυψης. Τα εν λόγω λεξικά δημιουργούνται αυτόματα από τα Tweets και μπορούν να περιέχουν hashtags (θεματικές ενότητες) και emoticons. Το σύστημα που προέκυψε από αυτή τη μελέτη και διακρίθηκε στο Conference on Semantic Evaluation Exercises (SemEval-2013) *(Saif M. Mohammad, 2013)* αποδεικνύει ότι τεχνικές αντιστροφής της πολικότητας των λέξεων δεν είναι πάντα ακριβείς ούτε κατάλληλες για τη διαδικασία της ανάλυσης των συναισθημάτων. Πιο συγκεκριμένα η εν λόγω μελέτη αποδεικνύει ότι όταν οι θετικοί όροι αναιρούνται (αντιστρέφονται), τείνουν να μεταφέρουν ένα αρνητικό συναίσθημα. Αντίθετα, όταν οι αρνητικοί όροι αναιρούνται (αντιστρέφονται), τείνουν να εξακολουθούν να μεταφέρουν ένα αρνητικό συναίσθημα. Επιπλέον, η ένταση αξιολόγησης για τόσο τους θετικούς όσο και για τους αρνητικούς όρους αλλάζει στο πλαίσιο τις αντιστροφής, και το ποσό της μεταβολής κυμαίνεται από όρο σε όρο. Προκειμένου να καταλάβουμε επαρκώς τις επιπτώσεις της αντιστροφής για μεμονωμένους όρους, οι συγγραφείς προτείναν της εμπειρική εκτίμηση των αποτελεσμάτων της ανάλυσης και τη δημιουργία δύο λεξικών για όρους σε αρνητικό πλαίσιο και σε θετικό πλαίσιο αντίστοιχα. Το εν λόγω σύστημα μπορεί να διαχειριστεί 100 Tweets ανά δευτερόλεπτο ενώ έχει καταφέρει να αξιολογήσει 135 εκατομμύρια Tweets σε ένα cluster 50 υπολογιστών σε 11 ώρες λειτουργίας.

4.7 Sentiment Analysis of Greek Tweets and Has*htags using a Sentiment Lexicon*

Το 2015 οι Georgios Kalamatianos, Dimitrios Mallis, Symeon Symeonidis, Avi Arampatzis *(Arampatzis, 2015)* δημοσιεύουν τη μελέτη τους στην οποία περιγράφουν τη διαδικασία ανάλυσης συναισθημάτων σε Tweets γραμμένα στην Ελληνική γλώσσα. Στην εν λόγω δημοσίευση οι συγγραφείς τονίζουν το κενό που παρατήρησαν στο τομέα της ανάλυσης συναισθημάτων στην Ελληνική γλώσσα εξαιτίας της έλλειψης κατάλληλων συνόλων δεδομένων. Στόχος των Georgios Kalamatianos et al. ήταν η δημιουργία ενός ολοκληρωμένου συνόλου δεδομένων που αφορά την ελληνική γλώσσα και περιέχει κατηγοριοποιημένα Tweets (υπό ανθρώπινη επίβλεψη), τη δημιουργία ενός αυτόματου συστήματος κατηγοριοποίησης ελληνικών Tweet για τις ακόλουθες έξι (6) κατηγορίες, "Θυμό", "Απέχθεια", "Φόβο", "Ευτυχία", "Λύπη" και "Έκπληξη", τη δημιουργία ενός αυτόματου συστήματος κατηγοριοποίησης ελληνικών Hashtags για τις παραπάνω έξι (6) κατηγορίες και τέλος την εξέταση των πτυχών των συναισθημάτων που εξαρτώνται από το χρόνο, όπως οι μεταβολές στην ένταση τους, για ορισμένα hashtags την πάροδο του χρόνου. Τα αποτελέσματα της εν λόγω μελέτης βοήθησαν σε μεγάλο βαθμό την εκπόνηση της παρούσας διπλωματικής εργασίας. Τα σύνολα δεδομένων όπως παρουσιάστηκαν από τους συγγραφείς χρησιμοποιήθηκαν ως σύνολο δοκιμών για το μοντέλο κατηγοριοποίησης που παρουσιάζεται σε αυτή τη διπλωματική εργασία.

4.8 Twitter as a Corpus for Sentiment Analysis *and Opinion Mining*

Το 2010 οι Alexander Pak, Patrick Paroubek *(Alexander Pak, 2010)* δημοσιεύουν τη μελέτη τους στην οποία περιγράφουν τη διαδικασία της αυτοματοποιημένης συλλογής των Tweets προκειμένου να δημιουργήσουν ένα σώμα δεδομένων (corpus) για την ανάλυση συναισθήματος και τους σκοπούς της εξόρυξης γνώμης. Στη μελέτη αυτή οι συγγραφείς πραγματοποιούν γλωσσολογική ανάλυση των συλλεχθέντων δεδομένων και εξηγούν τις ανακαλύψεις τους. Χρησιμοποιώντας το σώμα (corpus), οι συγγραφείς δημιούργησαν ένα ταξινομητή συναισθημάτων, που είναι σε θέση να προσδιορίσει θετικά, αρνητικά και ουδέτερα συναισθήματα για ένα κείμενο. Πειραματικές αξιολογήσεις και δοκιμές δείχνουν ότι οι προτεινόμενες τεχνικές που παρουσιάζονται στην εν λόγω μελέτη είναι αποδοτικές και μάλιστα πιο αποδοτικές από παλαιότερες μεθόδους. Στην παρούσα έρευνά η επιλεγμένη γλώσσα ήταν τα αγγλικά , ωστόσο, η προτεινόμενη τεχνική μπορεί να χρησιμοποιηθεί με οποιαδήποτε άλλη γλώσσα.

4.9 Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums

Το 2007 οι Ahmed Abbasi, Hsinchun Chen, και Arab Salem *(Ahmed Abbasi, 2007)* προτείνουν την εφαρμογή τεχνικών ανάλυσης συναισθήματος για αναγνώριση εξτρεμιστικών φόρουμ. Η ανάλυσή που παρουσιάζεται περιλαμβάνει την ταξινόμηση των συναισθημάτων σε μία κριτική ταινίας και δύο φόρουμ: ένα ρατσιστών των ΗΠΑ και ένα εξτρεμιστικής ομάδα της Μέσης Ανατολής. Στη μελέτη οι συγγραφείς αξιολογούν τα διαφορετικά σύνολα που αποτελείται από διαφορετικά συντακτικά και τεχνοτροπικά χαρακτηριστικά (ως τεχνοτροπικά χαρακτηριστικά παρουσιάζονται η κατανομή του μήκους των λέξεων, ο πλούτος του λεξιλογίου, διάφορα λεξιλογικά χαρακτηριστικά και συχνότητα των ειδικών χαρακτήρων). Παρουσιάζεται επίσης η ανάπτυξη ενός Γενετικού αλγορίθμου σταθμισμένης εντροπίας (EWGA) για την επιλογή των χαρακτηριστικών. Τα προαναφερθέντα χαρακτηριστικά και οι τεχνικές οδήγησαν στη δημιουργία μιας προσέγγισης ανάλυσης συναισθήματος προσανατολισμένη στην κατάταξη των ιστοσελίδων ανά συναισθήματα σε διάφορες γλώσσες. Τα αποτελέσματα χρησιμοποιώντας Support Vector Machine (SVM) δείχνουν ένα υψηλό επίπεδο ακρίβειας ταξινόμησης, αποδεικνύοντας την αποτελεσματικότητα αυτής της προσέγγισης για την ταξινόμηση και την ανάλυση των συναισθημάτων σε εξτρεμιστικά φόρουμ τόσο στην αγγλική γλώσσα όσο και στην αραβική. Τα αποτελέσματα της μελέτης αυτής είναι ιδιαίτερα ενδιαφέροντα καθώς με την χρήση αυτών των τεχνικών ανάλυσης μπορούμε να εντοπίσουμε εξτρεμιστικές ομάδες και να εμποδίσουμε τη διάδοση πληροφοριών και προπαγάνδας

Αποτελέσματα και συμπεράσματα

Τα συγκεντρωτικά αποτελέσματα απόδοσης της πλατφόρμας Thesis είναι:

- Η αναγνώρισης της γλώσσας των μηνυμάτων όπως υλοποιήθηκε στο RapidMiner μας έδωσε τα εξής ποσοστά ακρίβειας, 100% σε μικρό δείγμα 5 κλάσεων (20 κείμενα ανά κλάση) υπογραφών και περίπου 99.7% σε μεγάλο δείγμα με μεγάλη ομοιότητα 14 κλάσεων. Οι γλώσσες που μελετήθηκαν είναι οι εξής (2 letter ISO code): RO, DA, IT, HU, PT, FR, SV, EN, ES, NL, DE, EL, FI.
- Η αναγνώριση των συναισθημάτων των μηνυμάτων όπως υλοποιήθηκε στο RapidMiner μάς έδωσε ποσοστά ακρίβειας της τάξης του 71.70% σε ένα δείγμα 608 Tweets όπως προέκυψαν από την εκκαθάριση του δείγματος. Τα συναισθήματα που αποτέλεσαν τις κλάσεις μάς είναι αυτά της Ευτυχίας και της Λύπης (Happiness - Sadness). Παρατηρήσαμε μεγάλη απόκλιση στην αναγνώριση της Λύπης καθώς ενώ στην Ευτυχία επιτύχαμε class recall της τάξης του 90.62% δε συνέβη το ίδιο και για την αναγνώριση της λύπης.

Περισσότερα αποτελέσματα και συμπεράσματα παρουσιάζουμε στο δεύτερο μέρος (Μέρος Β) της παρούσας διπλωματικής εργασίας

Ακρωνύμια

OSN: Online Social Network
OSNs: Online Social Networks
NLP: Natural Language Processing
API: Application Programming Interface
GCM: Google Cloud Messaging
LAMP: Linux - Apache - PHP - MySQL

Γλωσσάρι

NLP: Με τον όρο NLP ή Natural Language Processing (επεξεργασία φυσικής γλώσσας) εννοούμε το υποπεδίο των σχετικών με τη χρήση υπολογιστικών τεχνικών προκειμένου υπολογιστικά συστήματα να μάθουν, να κατανοήσουν και να παράγουν περιεχόμενο σε ανθρώπινη γλώσσα, γνωστή και ως υπολογιστική γλωσσολογία

Classifier: Πρόκειται για μία υπολογιστική/στατιστική διαδικασία της μηχανικής μάθησης κατά την οποία ένα δείγμα κατατάσσεται σε ένα από τα δεδομένα σύνολα ή κλάσεις που είναι γνωστές στο υπολογιστικό σύστημα

10-fold-cross-validation: Πρόκειται για τη διαδικασία αξιολόγησης ενός classifier. Η διαδικασία έχει ως εξής.

- Βήμα 1ο: Από ένα ολοκληρωμένο και κατηγοριοποιημένο δείγμα επιλέγουμε το 10% των δειγμάτων (η δειγματοληψία είναι στρωματοποιημένη - stratified sampling) το οποίο αφαιρούμε από το training set.
- Βήμα 2ο: Εκπαιδεύουμε το μοντέλο με το 90% των δειγμάτων.
- Βήμα 3ο: Εκτελούμε το μοντέλο σε κάθε μία από τις εγγραφές που έχουμε παραλείψει και συγκρίνουμε το predicted label με το αρχικό label.
- Βήμα 4ο: Συγκεντρώνουμε τα αποτελέσματα και επαναλαμβάνουμε 10 φορές.

Bibliography

Ahmed Abbasi, H. C. (2007). Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. *2007 ACM 1073-0516/01/0300-0034.*

Alexander Pak, P. P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Conference: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010.*

Arampatzis, G. K. (2015). Sentiment Analysis of Greek Tweets and Hashtags using a Sentiment Lexicon. *PCI 2015, October 01-03, 2015, Athens, Greece.*

Brendan O'Connor, R. B. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Association for the Advancement of Artificial Intelligence.*

Han, B. C. (2013). Lexical normalization for social media text. *ACM Trans. Intell. Syst. Technol. 4, 1, Article 5.*

Jeonghee Yi, T. N. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing technique. *Proceedings of the Third IEEE International Conference on Data Mining*, (σσ. 427-434).

M. Cha, H. H. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *International AAAI Conference on Weblogs and Social Media (ICWSM).*

Pollyanna Gonçalves, M. A. (2013). Comparing and Combining Sentiment Analysis Methods. *COSN'13, October 07–08, 2013, Boston, MA, USA.*

Saif M. Mohammad, S. K. (2013). NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. *National Research Council Canada Ottawa, Ontario, Canada K1A 0R6.*

Svetlana Kiritchenko, X. Z. (2014). Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research 50.* (2014). *Journal of Artificial Intelligence Research 50.*

Wickre, K. (2013). Celebrating Twitter7.

Βιογραφικό Σημείωμα

Ο Παρασκάκης Ιωάννης γεννήθηκε στην Αθήνα το 1989. Είναι απόφοιτος του τμήματος Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων του Πανεπιστημίου Αιγαίου και μεταπτυχιακός φοιτητής του τμήματος Πληροφορικής του Πανεπιστημίου Πειραιώς. Εργάζεται ως προγραμματιστής στην Αθήνα τα τελευταία τρία χρόνια.