ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανάπτυξη Διαδικτυακής Υπηρεσίας Ανάλυσης Συναισθήματος Δεδομένων Κοινωνικών Δικτύων με χρήση Γράφων ν-γραμμάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΑΡΙΣΤΟΤΕΛΗ Γ. ΒΛΑΧΟΥ

Επιβλέπουσα : Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Αθήνα, Μάρτιος 2017

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανάπτυξη Διαδικτυακής Υπηρεσίας Ανάλυσης Συναισθήματος Δεδομένων Κοινωνικών Δικτύων με χρήση Γράφων ν-γραμμάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΑΡΙΣΤΟΤΕΛΗ Γ. ΒΛΑΧΟΥ

Επιβλέπουσα : Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 27η Μαρτίου 2017.

.....	.....	.....
Θεοδώρα Βαρβαρίγου	Εμμανουήλ Βαρβαρίγος	Δημήτριος Ασκούνης
Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2017

.....

ΑΡΙΣΤΟΤΕΛΗΣ Γ. ΒΛΑΧΟΣ

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

Περίληψη

Η ραγδαία εξάπλωση του Διαδικτύου και των Μέσων Κοινωνικής Δικτύωσης κατά τα τελευταία χρόνια συνεπάγεται τη δημιουργία και τη διαρκή ενημέρωση μίας εκτενέστατης συλλογής πληροφοριών. Βασική πηγή της συλλογής είναι το περιεχόμενο που παράγεται από τους χρήστες μέσα από τις πρωτότυπες αναρτήσεις, τα σχόλια και το δημόσιο διαμοιρασμό εγγράφων και αρχείων σε Κοινωνικά Δίκτυα ή Ιστοτόπους. Οι συγκεκριμένες πληροφορίες συχνά αποτυπώνουν κυρίαρχες καταναλωτικές ή πολιτικές τάσεις. Κατά συνέπεια, προκύπτει η ανάγκη της μελέτης των διαθέσιμων δεδομένων και της εξαγωγής συμπερασμάτων ως προς τις προτιμήσεις που εκφράζουν με τρόπο αυτοματοποιημένο. Η Ανάλυση Συναισθήματος είναι η περιοχή της Επιστήμης που καλείται να αντιμετωπίσει το πρόβλημα που περιγράψαμε.

Η παρούσα εργασία μελετά την ανάπτυξη μίας διαδικτυακής υπηρεσίας Ανάλυσης Συναισθήματος δεδομένων προερχόμενων από Κοινωνικά Δίκτυα σε πραγματικό χρόνο. Αρχικά, γίνεται μία διεξοδική αναφορά στις μεθόδους Ανάλυσης Συναισθήματος που χρησιμοποιούνται διεθνώς, καθώς επίσης στις υπάρχουσες σχετικές διαδικτυακές υπηρεσίες. Στη συνέχεια, επεξηγείται το μοντέλο των γράφων ν-γραμμάτων που αποτελεί μία μέθοδο Ανάλυσης Συναισθήματος επιπλεπόμενης μάθησης ανεξάρτητης της γλώσσας και ανθεκτικής στο θόρυβο των Κοινωνικών Δικτύων. Ακολούθως, παρουσιάζεται το σύστημα που υλοποιεί τη μέθοδο των γράφων ν-γραμμάτων, καθώς και ο τρόπος εύρεσης των βέλτιστων παραμέτρων του συστήματος για την επίτευξη της μέγιστης ακρίβειας πρόβλεψης. Τέλος, αναλύεται η δομή και η επίδοση της διαδικτυακής υπηρεσίας Ανάλυσης Συναισθήματος που αναπτύξαμε. Η συγκεκριμένη υπηρεσία αξιοποιεί τη μέθοδο των γράφων ν-γραμμάτων για την ανάλυση δεδομένων Κοινωνικών Δικτύων σε πραγματικό χρόνο.

Λέξεις Κλειδιά

κοινωνικά δίκτυα, ανάλυση συναισθήματος, γράφοι ν-γραμμάτων, αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης, κατηγοριοποίηση πολικότητας, διαδικτυακή υπηρεσία ανάλυσης συναισθήματος

Abstract

Both the Internet and the Social Networks have grown rapidly over the recent years. Their increasing penetration to people of all ages and geographic areas results in vast amounts of data being created and stored each day. Social Network posts and shares as well as comments both in Social Networks and websites depict what users think of brands or politics and what affects them emotionally or motivates them. This makes all relevant information invaluable and poses a big challenge to Computer Science for processing and analyzing the available data. The mining of the sentiments hidden in those data is the subject of Sentiment Analysis.

The aim of the current thesis is to introduce an innovative web service for real time Sentiment Analysis of data originating from Social Networks. First, we thoroughtly present both the Sentiment Analysis methods and the relevant web services that are available today. Then, we decribe the n-gram graph model. This model is a supervised Machine Learning technique for the Sentiment Analysis of data regardless of language and the syntactic or semantic noise which exists in Social Networks. Subsequently, we outline a software implementing the n-gram graph model and we decribe how its optimal parameters are calculated in terms of accuracy. Last, we focus on the deployment and the performance of our web service which uses both the n-gram graph approach and the aforementioned software to conduct real time Sentiment Analysis.

Keywords

social networks, sentiment analysis, n-gram graphs, supervised machine learning algorithms, polarity classification, sentiment analysis web service

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την Καθηγήτρια Θεοδώρα Βαρβαρίγου που μου προσέφερε την ευκαιρία να εκπονήσω την παρούσα διπλωματική εργασία υπό την επίβλεψή της.

Επίσης, είμαι ιδιατέρως ευγνώμων στους επιβλέποντές μου, Δρ. Φώτη Αίσωπο και Δρ. Βρεττό Μουλό για την καθοδήγηση και την υποστήριξή τους καθόλη τη διάρκεια της συνεργασίας μας. Η εμπειρία τους και η βοήθειά τους ήταν καθοριστικής σημασίας για την ολοκλήρωση της διπλωματικής εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την υποστήριξη και τις όμορφες στιγμές που περάσαμε κατά τη διάρκεια των σπουδών μου στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσοβίου Πολυτεχνείου.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1

Εισαγωγή

Η ραγδαία ανάπτυξη του Διαδικτύου και η διείσδυσή του σε κάθε έκφανση της ανθρώπινης δραστηριότητας αλλάζει καθημερινά και ριζικά τον τρόπο με τον οποίο γίνεται αντιληπτή η αισθητή πραγματικότητα. Η δραστηριότητα των χρηστών στα κοινωνικά δίκτυα, η διάδοση πληροφοριών μέσα από παραδοσιακά μέσα ενημέρωσης ή προσωπικά ιστολόγια, οι συνεχείς δημόσιες συζητήσεις για θέματα που άπτονται της Πολιτικής, της Οικονομίας, της Επιστήμης και της Τεχνολογίας διαμορφώνουν καθοριστικά την αντίληψη που έχουμε για το σύγχρονο κόσμο, δημιουργούν καταναλωτικές τάσεις, εμπνέουν πολιτικά ρεύματα, ενδεχομένως καθορίζουν εκλογικές αναμετρήσεις, αποκαλύπτουν απόρρητες πληροφορίες που αφορούν θέματα προσωπικών δεδομένων και σεβασμού της ιδιωτικής ζωής, εγείρουν κρίσιμα ερωτήματα γύρω από ανθρωπιστικά προβλήματα και καθιστούν ολοένα πιο σαφή την έννοια της παγκόσμια κοινότητας.

Η άποψη της κοινής γνώμης αποτυπώνεται στα κοινωνικά δίκτυα. Το πλήθος των πληροφοριών που βρίσκονται αποθηκευμένες, αλλά και δημιουργούνται διαρκώς στα κοινωνικά δίκτυα είναι πολύτιμο. Εκ των πραγμάτων, προκύπτει η ανάγκη ανάλυσης του μεγάλου όγκου των δεδομένων με αυτοματοποιημένο τρόπο και δημιουργείται η γνωστική περιοχή της Ανάλυσης Συναισθήματος (Sentiment Analysis).

Η Ανάλυση Συναισθήματος σε δεδομένα κοινωνικών δικτύων κερδίζει διαρκώς έδαφος αφενός μεν στον ακαδημαϊκό χώρο λόγω των τεχνικών προκλήσεων που θέτει, αφετέρου δε στον επιχειρηματικό χώρο χάρη στις πολλά υποσχόμενες προοπτικές της. Εταιρείες που καινοτομούν στον τομέα δραστηριότητάς τους επενδύουν κεφάλαια στην εξόρυξη γνώμης από τα κοινωνικά μέσα δικτύωσης, χρησιμοποιώντας τεχνικές ανάλυσης συναισθήματος. Είναι ιδιαίτερα κρίσιμη η έγκαιρη ανάλυση της γνώμης των καταναλωτών στα κοινωνικά δίκτυα. Δίνει στις εταιρίες την ακριβή εικόνα που έχουν οι

καταναλωτές για τα προϊόντα τους, καθώς και τις ανάγκες που επιθυμούν να καλύψουν. Εξάλλου αυξάνεται συνεχώς ο αριθμός των ανθρώπων που πραγματοποποιούν τις αγορές τους διαδικτυακά ή στηρίζονται σε αξιολογήσεις άλλων καταναλωτών πριν λάβουν την τελική απόφαση αγοράς.

1.1 Διατύπωση Προβλήματος

Η Ανάλυση Συναισθήματος μπορεί να εφαρμοσθεί σε διάφορα επίπεδα ανάλυσης ανάλογα με το μέγεθος του κειμένου και τη ζητούμενη λεπτομέρεια της εξαγόμενης πληροφορίας:

- Επίπεδο Εγγράφου : γίνεται η παραδοχή ότι σε κάθε έγγραφο διατυπώνεται μία άποψη για ένα συγκεκριμένο αντικείμενο ή θέμα.
- Επίπεδο Πρότασης : διαχωρίζει κάθε έγγραφο σε προτάσεις υποθέτοντας ότι κάθε μία πρόταση εκφράζει μία μόνο άποψη.
- Επίπεδο Χαρακτηριστικού : διαχωρίζει κάθε έγγραφο ή πρόταση σε φράσεις οι οποίες αναφέρονται σε μία οντότητα συνολικά ή ξεχωριστά για κάθε ένα από τα χαρακτηριστικά της με διαφορετικό συναίσθημα.

Ο βαθμός λεπτομέρειας που εξετάζουμε και συναντάται κυρίως στη βιβλιογραφία εστιάζει στην πολικότητα του μηνύματος. Προσπαθεί να κατατάξει την άποψη που διατυπώνεται σε ένα κείμενο σε θετική,αρνητική ή ουδέτερη και για το λόγο αυτό η Ανάλυση Συναισθήματος αναφέρεται συχνά και με τον όρο Εξόρυξη Γνώμης. Πιο προχωρημένες προσεγγίσεις ασχολούνται με τον προσδιορισμό του συγκεκριμένου συναισθήματος που εκφράζει ο δημιουργός, για παράδειγμα χαρά, ευχαρίστηση, λύπη, αγανάκτηση.

Στην παρούσα εργασία μελετάμε την Ανάλυση Συναισθήματος σε επίπεδο εγγράφου αναλύοντας πολυγλωσσικά μηνύματα από το κοινωνικό δίκτυο Twitter. Σκοπός είναι ο προσδιορισμός της πολικότητας των κειμένων τα οποία αποτελούνται από μερικές προτάσεις. Ωστόσο, λόγω του περιορισμού του μήκους των μηνυμάτων που επιβάλλει το μέσο, το εύρος της ανάλυσης προσεγγίζει σε μεγάλο βαθμό το επίπεδο πρότασης.

Το πρόβλημα που εξετάζουμε είναι η ταξινόμηση πολυγλωσσικών κειμένων σε τρεις κατηγορίες (θετική, αρνητική ή ουδέτερη). Θεωρούμε πως κάθε μήνυμα ανήκει σε μόνο μία κατηγορία δηλαδή η πολικότητα είναι ενιαία σε όλο το σώμα του εγγράφου (single label classification). Παράλληλα, εστιάζουμε σε ένα σύνολο γλωσσών χωρίς ωστόσο να γνωρίζουμε εκ των προτέρων τη γλώσσα στην οποία κάθε μήνυμα έχει γραφτεί.

1.2 Προσέγγιση

Τα περισσότερα συστήματα Ανάλυσης Συναισθήματος ανιχνεύουν εκφραστικά μοτίβα και εφαρμόζουν τεχνικές που αξιολογούν τη σημασία και συναισθηματική χροιά συγκεκριμένων λέξεων και φράσεων (π.χ. SentiWordNet). Αν και αυτές οι προσεγγίσεις επιτυγχάνουν αρκετά υψηλή ακρίβεια όταν εφαρμοσθούν σε ένα συγκεκριμένο περιβάλλον με γνωστή θεματολογία, έχουν κατασκευασθεί με την παραδοχή ότι τα προς εξέταση δεδομένα είναι γραμμένα σε μία μόνο εκ των προτέρων γνωστή γλώσσα και δεν περιλαμβάνουν θορυβώδες περιεχόμενο, όπως νεολογισμούς, και ορθογραφικά λάθη, στοιχεία τα οποία αποτελούν εγγενή χαρακτηριστικά των δεδομένων από κοινωνικά δίκτυα. Σε αυτή την εργασία αντιμετωπίζουμε τους συγκεκριμένου περιορισμούς μελετώντας εκτενέστερα την εφαρμογή της μεθόδου γράφων ν-γραμμάτων σε μηνύματα του Twitter. Η εν λόγω προσέγγιση επιβλεπόμενης μάθησης που προτάθηκε από τους Aisopos, Papadakis, Tserpes και Varvarigou εμφανίζει ανεξαρτησία από τη γλώσσα (language neutrality) και υψηλή ανοχή στο θόρυβο (noise-tolerant) και περιγράφεται αναλυτικά στο Κεφάλαιο 3. Συνοπτικά, δημιουργεί ένα γράφο του οποίου οι κορυφές αντιστοιχούν σε χαρακτήρες ν-γραμμάτων ενός μηνύματος και τα βάρη των ακμών του αναφέρονται στη μέση απόσταση μεταξύ τους. Μηνύματα της ίδιας κατηγορίας πολικότητας συγχωνεύονται στον αντίστοιχο γράφο κλάσης και στη συνέχεια κάθε γράφος μηνύματος συγκρίνεται με τους γράφους κλάσεων για να προσδιορισθεί η πολικότητα του μηνύματος. Το μοντέλο γράφων ν-γραμμάτων εμφάνισε υψηλά ποσοστά ακρίβειας κατηγοριοποίησης εφαρμοζόμενο σε μεγάλο όγκο δεδομένων που παράχθηκε με αυτόματο τρόπο με την τεχνική της εξ αποστάσεως επίβλεψης (distant supervision).

1.3 Οργάνωση Κειμένου

Το κυρίως σώμα της παρούσας εργασίας διαρθώνεται στα ακόλουθα κεφάλαια:

- **Κεφάλαιο 2 :** Εκτίθεται σε μεγαλύτερο βαθμό το πρόβλημα της Ανάλυσης Συναισθήματος, οι εγγενείς δυσκολίες που εμφανίζονται ως συνάρτηση του εκάστοτε γλωσσικού περιβάλλοντος, οι ποικίλες προσεγγίσεις και οι περιορισμοί που τις συνοδεύουν, καθώς και οι υλοποιήσεις διαφόρων διαδικτυακών υπηρεσιών που προσφέρουν τη δυνατότητα Ανάλυσης Συναισθήματος ενός συνόλου δεδομένων.
- **Κεφάλαιο 3 :** Μελετάται η μέθοδος Ανάλυσης Συναισθήματος με γράφους ν-γραμμάτων και η άρση των γλωσσικών περιορισμών που επιφέρει. Επίσης, παρουσιάζονται οι αλγόριθμοι Μηχανικής Μάθησης που χρησιμοποιούνται στην εργασία.

- **Κεφάλαιο 4 :** Γίνεται εκτενής αναφορά στο σύστημα εκπαίδευσης των ταξινομητών που χρησιμοποιεί η υπηρεσία Ανάλυσης Συναισθήματος σε Πραγματικό Χρόνο. Στη συνέχεια του κεφαλαίου, παρουσιάζεται η υπηρεσία καθεαυτή.
- **Κεφάλαιο 5 :** Συνοψίζεται η συνεισφορά της παρούσας εργασίας στο γνωστικό αντικείμενο της Ανάλυσης Συναισθήματος και προτείνονται πιθανές μελλοντικές προεκτάσεις της υπηρεσίας Ανάλυσης Συναισθήματος σε Πραγματικό Χρόνο.

2

Ανάλυση Συναισθήματος

Τα τελευταία χρόνια η Ανάλυση Συναισθήματος προσελκύει όλο και περισσότερο το ενδιαφέρον της ακαδημαϊκής κοινότητας αλλά και της βιομηχανίας χάρη στις πιθανές εφαρμογές της, κυρίως στον τομέα της Επιχειρηματικής Ευφυίας (Business Intelligence). Στην προσπάθεια να αξιοποιήσουμε αποτελεσματικά τον τεράστιο όγκο δεδομένων που παράγονται καθημερινά από απλούς χρήστες (user-generated content) στα μέσα κοινωνικής δικτύωσης, έχουν πραγματοποιηθεί σημαντικές έρευνες εφαρμόζοντας διαφορετικές τεχνικές και προσεγγίσεις.

2.1 Το πρόβλημα της Ανάλυσης Συναισθήματος

Από τις πρώτες μελέτες στην περιοχή της Ανάλυσης Συναισθήματος [51] είχε ήδη γίνει αντιληπτό ότι η κατηγοριοποίηση συναισθήματος (sentiment classification) διαφέρει από το κλασικό πρόβλημα κατηγοριοποίησης κειμένου :

“Η κατηγοριοποίηση κείμενου - γνωστή και ως ταξινόμηση κειμένου ή ανίχνευση θέματος - αναφέρεται στην αντιστοίχηση κειμένων φυσικής γλώσσας σε θεματικές κατηγορίες ή κλάσεις οι οποίες ανήκουν σε ένα προκαθορισμένο σύνολο” [44]

Οι κατηγορίες καθορίζονται με βάση τα θέματα στόχους του εκάστοτε προβλήματος. Επομένως, διαφορετικά προβλήματα ταξινόμησης κειμένου βασίζονται σε διαφορετικά σύνολα κατηγοριών. Το πλήθος των κατηγοριών σε ένα σύνολο ποικίλει : μπορεί να εκτείνεται από ένα μικρό σύνολο δύο μόνο κατηγοριών έως σύνολα με δεκάδες κατηγορίες π.χ. οι κατηγορίες που απαιτούνται για την ταξινόμηση ενός άρθρου εφημερίδας με βάση τη θεματολογία που καλύπτει. Παράλληλα, ανάλογα με το πρόβλημα και το σύνολο κατηγοριών, ένα κείμενο μπορεί να ανήκει σε μία ή περισσότερες

επικαλυπτόμενες κατηγορίες π.χ. ένα άρθρο να αντιστοιχηθεί με τις κατηγορίες “πολιτική”, “οικονομία” και “επικαιρότητα”.

Αντίθετα, η Ανάλυση Συναισθήματος αναφέρεται σε ένα μικρό σύνολο κατηγοριών (π.χ. θετικό, αρνητικό, ουδέτερο - “1 αστέρι”, ...,“5 αστέρια”). Επειδή επικεντρώνεται στην κατάταξη ενός κειμένου ως προς την πολικότητα του, οι κατηγορίες είναι ανεξάρτητες της θεματολογίας του προβλήματος και μεταξύ τους αμοιβαία αποκλειόμενες.

Το πρόβλημα που προσπαθεί να επιλύσει η Ανάλυση Συναισθήματος είναι ένα από τα πιο απλά προβλήματα με τα οποία ασχολείται η Επεξεργασία Φυσικής Γλώσσας [22]. Ο υπολογιστής δε χρειάζεται να αντιλαμβάνεται πλήρως τη σημασιολογία της κάθε πρότασης αλλά θα πρέπει να εντοπίζει τη συνολική στάση του συγγραφέα και να την ταξινομεί ως προς την πολικότητά της. Οι απαιτήσεις απλοποιούν σε μεγάλο βαθμό το πρόβλημα της κατανόησης της φυσικής γλώσσας από τον υπολογιστή αλλά δεν παύει το πρόβλημα της ανίχνευσης της πολικότητας - στο οποίο εξειδικεύεται - να είναι αρκετές φορές δύσκολο ακόμη και για τον άνθρωπο.

2.2 Δυσκολίες και Προκλήσεις

Το εννοιολογικό πλαίσιο στο οποίο κινείται η Ανάλυση Συναισθήματος εξασφαλίζει την εφαρμογή των τεχνικών και μεθόδων της σε ένα μεγάλο εύρος θεμάτων χωρίς ιδιαίτερες τροποποιήσεις, επιτυγχάνοντας αρκετά ικανοποιητικά ποσοστά ακρίβειας. Στηριζόμενοι στην ανεξαρτησία του προβλήματος από τη θεματολογία, θα μπορούσαμε να ισχυριστούμε ότι η πολικότητα ενός κειμένου προκύπτει από την πολικότητα των μεμονωμένων λέξεων από τις οποίες απαρτίζεται. Συνεπώς, αναγνωρίζοντας ένα συγκεκριμένο σύνολο λέξεων-κλειδιών (keywords) θα μπορούσαμε να προσδιορίσουμε τη συνολική πολικότητα της άποψης που εκφράζεται στο κείμενο.

Η παραπάνω διαδικασία είναι μία από τις πρώτες μεθόδους που χρησιμοποιήθηκαν και υιοθετεί μία από τις πιο δημοφιλείς και αποτελεσματικές τεχνικές της ανίχνευσης θεματολογίας. Ωστόσο, η προσέγγιση μέσω λέξεων κλειδιών στο συγκεκριμένο πρόβλημα δεν εμφανίζει υψηλά ποσοστά ακρίβειας και έχει αποδεχθεί ελλιπής σε ορισμένες περιπτώσεις (“thwarted expectations” [34]).

Στο σημείο αυτό ανακύπτει το εξής ερώτημα : για ποιο λόγο το πρόβλημα της κατηγοριοποίησης συναισθήματος είναι πιο δύσκολο σε σχέση με την ανίχνευση θεματολογίας, αν λάβουμε υπόψη ότι οι κατηγορίες “θετικό”, “αρνητικό” και “ουδέτερο” είναι εννοιολογικά ξένες μεταξύ τους ;

Μία από τις πιο σημαντικές διαφορές με την κατηγοριοποίηση ως προς τη θεματολογία και τις δυσκολίες στην περιοχή της Ανάλυσης Συναισθήματος είναι ότι "το συναίσθημα/άποψη μπορεί πολλές φορές να εκφρασθεί με πιο λεπτό τρόπο χωρίς τη χρήση συναισθηματικά φορτισμένων (θετικά ή αρνητικά) λέξεων με αποτέλεσμα να είναι δύσκολο να αναγνωρισθεί από τους επιμέρους όρους του κειμένου όταν αυτοί εξετάζονται μεμονωμένα" [33].

Παράλληλα, πέρα από τον προσδιορισμό της πολικότητας όταν απουσιάζουν συναισθηματικά φορτισμένες λέξεις, ιδιαίτερα απαιτητικός είναι και ο διαχωρισμός των υποκειμενικών και αντικειμενικών λέξεων και φράσεων ενός κειμένου. Όπως αναφέρεται από τους Kim και Hovy στο [19] "πολλές φορές ακόμη και άνθρωποι διαφωνούν για το αν μία δήλωση αποτελεί ή όχι άποψη".

Ένα άλλο ζήτημα που απασχολεί ιδιαίτερα την Ανάλυση Συναισθήματος είναι ο προσ- διορισμός του κατόχου - εκφραστή της άποψης (opinion holder) που διατυπώνεται στο κείμενο. Το συγκεκριμένο θέμα έχει μελετηθεί εκτενώς στη βιβλιογραφία, κυρίως σε αναλύσεις σε πολιτικά debates εξετάζοντας αν η γνώμη ανήκει στο συγγραφέα/δημιουργό ή στον σχολιαστή.

Όπως αναφέρθηκε στην Ενότητα 2.1, η γενικότερη αντίληψη της θετικής ή αρνητικής άποψης δεν εξαρτάται άμεσα από το εκάστοτε θέμα συζήτησης. Ωστόσο, το συναίσθημα και η υποκειμενικότητα ενός κειμένου εξαρτώνται από το σημασιολογικό πλαίσιο στο οποίο τοποθετείται [33]. Χαρακτηριστικό παράδειγμα : "πήγαινε διάβασε το βιβλίο". Η πρόταση εκφράζει θετική άποψη όταν αναφέρεται σε κριτική βιβλίου. Η ίδια πρόταση, όμως, εκφράζει εντελώς διαφορετική άποψη όταν χρησιμοποιείται σε κριτική ταινίας.

Άλλος ένας παράγοντας που επηρεάζει την πολικότητα είναι η σειρά των λέξεων και φράσεων στο κείμενο [33]. Οι ίδιες λέξεις με διαφορετική σειρά μπορεί να οδηγήσουν σε τελείως διαφορετική συνολική πολικότητα.

Τέλος, στις δυσκολίες που συναντά η Ανάλυση Συναισθήματος πρέπει να συμπεριληφθούν και οι προκλήσεις της ευρύτερης περιοχής της Επεξεργασίας Φυσικής Γλώσσας: αμφισημία, χειρισμός της άρνησης, ειρωνεία και σαρκασμός.

2.3 Ανάλυση Συναισθήματος σε Μικροϊστολόγια

Η Ανάλυση Συναισθήματος όταν εφαρμόζεται σε δεδομένα από μικροϊστολόγια και κοινωνικά δίκτυα καλείται να αντιμετωπίσει περαιτέρω δυσκολίες οι οποίες οφείλονται στην ιδιαίτερη φύση των κειμένων:

- Μήκος Κειμένου: τα μηνύματα είναι συνήθως σύντομα (π.χ. μέγιστο όριο 140 χαρακτήρες στο Twitter). Αν και ο περιορισμός μήκους μπορεί να οδηγήσει σε περιεκτικές και επί του θέματος τοποθετήσεις, πολλές φορές απουσιάζει το ευρύτερο εννοιολογικό πλαίσιο με αποτέλεσμα να μην είναι σαφής η πολικότητα του κειμένου [6].
- Λεξιλόγιο: τα περισσότερα κείμενα διατυπώνονται σε ανεπίσημη, καθομιλούμενη γλώσσα και εμφανίζουν πολύ μεγαλύτερη ποικιλομορφία σε σχέση με άλλα είδη κειμένου. Περιλαμβάνουν αργκό, νεολογισμούς, εσκεμμένες παραλλαγές λέξεων για έμφαση (επιμήκυνση φθόγγων, χρήση κεφαλαίων γραμμάτων), συντομογραφίες (π.χ. “gr8”-“great”) με αποτέλεσμα να μην είναι δυνατή η εφαρμογή λεκτικών αναλυτών ή άλλων εργαλείων που στηρίζονται στη γραπτή και πιο επίσημη μορφή της γλώσσας.
- Θόρυβος: οι πλατφόρμες κοινωνικής δικτύωσης επιτρέπουν μία αυθόρμητη επικοινωνία σε πραγματικό χρόνο όπου πολλές φορές οι χρήστες αναρτούν μηνύματα χωρίς να ελέγχουν για συντακτικά ή γραμματικά λάθη. Ένα μεγάλο ποσοστό από τα δεδομένα που παράγονται περιέχει ακούσια ορθογραφικά λάθη και ακατανόητες εκφράσεις τα οποία συνιστούν ουσιαστικά θόρυβο. Η αναγνώριση και αποκλεισμός τους αποτελεί ιδιαίτερη πρόκληση για τα σύγχρονα συστήματα ανίχνευσης συναισθήματος.
- Πολυγλωσσικό Περιεχόμενο: τα μέσα κοινωνικής δικτύωσης εξαπλώνονται σε μη αγγλόφωνες χώρες, αποκτώντας χρήστες που χρησιμοποιούν και γράφουν σε διαφορετικές γλώσσες, αρκετές φορές ακόμη και σε επίπεδο πρότασης ή μηνύματος. Το φαινόμενο αυτό έχει ως αποτέλεσμα ιδιαίτερα διαδεδομένες τεχνικές στοχευμένες σε συγκεκριμένες γλώσσες (language-specific) να καθίστανται πρακτικά μη εφαρμόσιμες.

2.4 Προσεγγίσεις

Έχουν προταθεί διάφορες τεχνικές που εξετάζουν την αναπαράσταση του κειμένου υπό διαφορετική οπτική γωνία, έχοντας, ωστόσο, κοινό στόχο: τον προσδιορισμό της πολικότητας. Οι πιο βασικές προσεγγίσεις συνοψίζονται στις εξής κατηγορίες:

- Λεξικό Συναισθήματος : βασίζεται στον ισχυρισμό ότι ο προσδιορισμός της πολικότητας ενός κειμένου βασίζεται στον σημασιολογικό προσανατολισμό (semantic orientation) των επιμέρους λέξεων και φράσεών του [51]. Δημιουργούνται, λοιπόν, λεξικά συναισθημάτων στα οποία περιέχονται λέξεις και φράσεις με την αντίστοιχη σημασιολογική πολικότητα και ισχύ τους (βασικά λήμματα). Στη συνέχεια, εμπλουτίζονται είτε αξιοποιώντας πληροφορίες από μεγάλα σώματα κειμένου (ανίχνευση συντακτικών μοτίβων, συχνότητα εμφάνισης λέξεων) (text-corpus based) ή χρησιμοποιώντας εξωτερικούς γλωσσολογικούς πόρους (θησαυρούς λέξεων,ερμηνευτικά λεξικά) (dictionary-based) για την επέκτασή τους με συνώνυμα, αντώνυμα και επιπλέον συντακτικές και σημασιολογικές πληροφορίες [16].
- Σχέσεις και Συνδέσεις : εξετάζει τις πιθανές σχέσεις και εξαρτήσεις μεταξύ των διάφορων χαρακτηριστικών του κειμένου. Μελετά τον τρόπο με τον οποίο συνδέονται μεταξύ τους οι παράγραφοι, οι προτάσεις και τα διάφορα μέρη του λόγου έτσι ώστε να ανιχνευτούν νοηματικές αντιθέσεις ή επικαλύψεις στις συνιστώσες του κειμένου και να προσδιορισθεί με μεγαλύτερη ακρίβεια η συνολική πολικότητα [36].
- Δομή του Λόγου : μελετά τη συντακτική δομή του κειμένου : κάθε λέξη εξετάζεται αν ανήκει στους κύριους όρους της πρότασης (υποκείμενο-ρήμα-αντικείμενο) και εντοπίζεται η θέση της μέσα στο κείμενο με σκοπό να προσδιορισθεί η τοπική της σημασιολογία (ενεργή ή παθητική συμμετοχή) και η βαρύτητά της στο συνολικό συναίσθημα. π.χ. στις κριτικές η συνολική στάση διατυπώνεται συνήθως προς το τέλος του κειμένου [34].
- Γλωσσικά Μοντέλα : δανείζεται αρκετά στοιχεία από την περιοχή της Αναγνώρισης Φωνής. Στηρίζεται στη στατιστική επεξεργασία του κειμένου με στόχο την κατασκευή ενός διανύσματος χαρακτηριστικών (feature vector) το οποίο θα χρησιμοποιηθεί στη συνέχεια για την κατηγοριοποίηση του συναισθήματος [32]. Ένα γλωσσικό μοντέλο αποτελεί την υπό συνθήκη κατανομή πιθανότητας της i-ιοστής λεκτικής μονάδας σε μία πρόταση , δηλαδή υποδηλώνει την πιθανότητα εμφάνισης της συγκεκριμένης λεκτικής μονάδας, γνωρίζοντας την κατηγορία όλων των προηγούμενων όρων της πρότασης. Τα πιο δημοφιλή μοντέλα βασίζονται στην αναπαράσταση του κειμένου με λέξεις ή χαρακτήρες ν-γραμμάτων. Η προσέγγιση αυτή διαφέρει σε σχέση με τις προαναφερθείσες καθώς απαιτεί ένα σύνολο από ήδη ταξινομημένα εκπαιδευτικά πρότυπα (training set) - τα οποία πρέπει να είναι

αντιπροσωπευτικά των κειμένων προς εξέταση (test set). Εφαρμόζοντας το μοντέλο στο σύνολο εκπαίδευσης, επιλέγουμε τα χαρακτηριστικά εκείνα που καθιστούν τα κείμενα διαφορετικών κατηγοριών διαχωρίσιμα.

2.5 Κατηγοριοποίηση

Η οπτική γωνία με την οποία προσεγγίζεται η δομή και αναπαράσταση ενός κειμένου καθορίζει και τον τρόπο κατηγοριοποίησης του. Οι τεχνικές που εφαρμόζονται χωρίζονται ανάλογα με το βαθμό παρέμβασης του ανθρώπου στη διαδικασία της μάθησης σε δύο βασικές κατηγορίες.

2.5.1 Μέθοδοι Επιβλεπόμενης Μάθησης

Αποτελεί την πιο δημοφιλή τεχνική κατηγοριοποίησης συναισθήματος. Στόχος είναι η δημιουργία ενός ταξινομητή (classifier) ο οποίος θα αντιστοιχίζει κείμενα με κατηγορίες (θετικά, αρνητικά, ουδέτερα) εφαρμόζοντας κάποιον αλγόριθμο.

Στην επιβλεπόμενη μάθηση, κάθε κείμενο αναπαρίσταται με ένα διάνυσμα χαρακτηριστικών έτσι, ώστε ο ταξινομητής να αναγνωρίσει και να μάθει τις πιο αντιπροσωπευτικές διαφορές ανάμεσα σε κείμενα που ανήκουν σε διαφορετικές κατηγορίες και για αυτό απαιτείται ένα σύνολο εκπαίδευσης. Οι αλγόριθμοι μηχανικής μάθησης εστιάζουν στη βελτιστοποίηση των εσωτερικών τους παραμέτρων ανάλογα με τη δυαδική τιμή ή βάρος ορισμένων χαρακτηριστικών ή στην κατασκευή επαγώμενων κανόνων ανάλογα με το ζεύγος χαρακτηριστικό-τιμή στο σύνολο εκπαίδευσης. Οι πιο γνωστοί αλγόριθμοι είναι: Naive Bayes, Multinomial Naive Bayes, C4.5 και Support Vector Machines (SVM).

Οι τεχνικές επιβλεπόμενης μάθησης επιτυγχάνουν υψηλά ποσοστά ακρίβειας και υπερτερούν των μη επιβλεπόμενων τεχνικών [34]. Ωστόσο, εμφανίζουν κάποια μειονεκτήματα. Απαιτείται αρκετός χρόνος και προσπάθεια για την κατασκευή ενός συνόλου εκπαίδευσης αλλά και για την εκπαίδευση του ταξινομητή μέχρις ότου βρεθούν οι βέλτιστες τιμές των παραμέτρων ή εξαχθούν οι απαραίτητοι κανόνες. Παράλληλα, η ακρίβεια ενός ταξινομητή εξαρτάται άμεσα από το σύνολο εκπαίδευσης. Συνεπώς, τα εκπαιδευτικά πρότυπα θα πρέπει να έχουν επιλεγεί κατάλληλα έτσι ώστε να είναι αντιπροσωπευτικά του συνολικού πληθυσμού των κειμένων.

2.5.1 Μέθοδοι Μη Επιβλεπόμενης Μάθησης

Η κατηγοριοποίηση συναισθήματος γίνεται με βάση το σημασιολογικό προσανατολισμό των λέξεων και φράσεων του κειμένου. Δεν απαιτείται σύνολο εκπαίδευσης για την εξαγωγή διανύσματος χαρακτηριστικών Αντίθετα, χρησιμοποιώντας προκατασκευασμένα λεξικά συναισθήματος, χαρακτηρίζονται οι διάφοροι όροι του κειμένου και προκύπτει η συνολική πολικότητα.

Αρκετές από τις μη επιβλεπόμενες μεθόδους επιτυγχάνουν ικανοποιητικά ποσοστά ακρίβειας σε συστηματική βάση όταν εφαρμόζονται σε γνωστά θεματικά πεδία όπου το λεξιλόγιο των κειμένων τους καλύπτεται από τα λεξικά συναισθήματος. Αποκτούν ιδιαίτερη δημοτικότητα καθώς δεν απαιτείται σύνολο εκπαίδευσης με αποτέλεσμα να μπορούν να εφαρμοστούν σε μεγαλύτερο εύρος θεμάτων σε σχέση με τις μεθόδους επιβλεπόμενης μάθησης [47]. Ωστόσο, παρουσιάζουν δύο σημαντικούς περιορισμούς [39]. Αρχικά, το πλήθος των λέξεων στα λεξικά είναι πεπερασμένο με αποτέλεσμα η τεχνική να μην μπορεί να εφαρμοστεί σε πολύ δυναμικά περιβάλλοντα όπως το Twitter όπου νεολογισμοί και συντομογραφίες συνεχώς εμφανίζονται. Επιπλέον, τα λεξικά συναισθήματος αναθέτουν συνήθως ένα σταθερό συναισθηματικό προσανατολισμό στις λέξεις χωρίς να εξετάζουν το ευρύτερο πλαίσιο στο οποίο χρησιμοποιούνται.

2.6 Σχετικές Εργασίες

Οι πρώτες μελέτες στην περιοχή της Ανάλυσης Συναισθήματος εξέταζαν κυρίως άρθρα εφημερίδων (κριτικές ταινιών και προϊόντων, πολιτικές και οικονομικές αναλύσεις). Τα τελευταία χρόνια, χάρη στην ανάπτυξη των μέσων κοινωνικής δικτύωσης, το ενδιαφέρον της ακαδημαϊκής κοινότητας, αλλά και της βιομηχανίας στράφηκε προς την επεξεργασία και ανάλυση των παραγόμενων δεδομένων. Αυτό είχε ως αποτέλεσμα να παραχθεί σημαντικό ερευνητικό έργο που εστιάζει αποκλειστικά σε δεδομένα από κοινωνικά δίκτυα. Ακολουθεί μία επισκόπηση των κύριων εργασιών σε δεδομένα από το Twitter, χωρισμένες σε ενότητες ανάλογα με το είδος της τεχνικής που εφαρμόζουν.

2.6.1 Εργασίες με χρήση Επιβλεπόμενης Μηχανικής Μάθησης

Οι Go et al. (2009) [14] υπήρξαν από τους πρώτους που μελέτησαν την ανάλυση συναισθήματος σε δεδομένα από το Twitter. Στη μελέτη τους ασχολούνται με τη δυαδική εκδοχή του προβλήματος κατηγοριοποίησης συναισθήματος, χαρακτηρίζοντας τα tweets ως θετικά ή αρνητικά. Λόγω της έλλειψης σε εκπαιδευτικά πρότυπα ήδη κατηγοριοποιημένα χειροκίνητα από άνθρωπο (manually annotated), εφαρμόζουν την τεχνική της εξ αποστάσεως επίβλεψης (distant supervision) για να εκπαιδεύσουν ένα ταξινομητή επιβλεπόμενης μηχανικής μάθησης. Μέσω του Twitter API, συλλέγουν ένα μεγάλο σύνολο από tweets τα οποία ταξινομούν αυτόματα σε κατηγορίες ανάλογα με τα emoticons (noisy labels) , διαγράφοντας tweets που περιέχουν emoticons και από τις δύο κατηγορίες. Το τελικό training set αποτελείται από 1.6 εκατομμύρια tweets, 800 χιλιάδες tweets από κάθε κατηγορία. Εφαρμόζουν στάδιο προεπεξεργασίας του αρχικού κειμένου των tweets όπου αφαιρούνται τα emoticons ενώ αναφορές σε χρήστες (@username) και υπερσύνδεσμοι αντικαθίστανται με κατάλληλες λέξεις-κλειδιά (placeholders). Για την κατηγοριοποίηση χρησιμοποιούν ως χαρακτηριστικά μονογράμματα, διγράμματα, συνδυασμό μονογραμμάτων και διγραμμάτων καθώς και επισημειώσεις για την ιδιότητα της κάθε λέξης (μέρος του λόγου), γνώρισμα που συναντάται στην βιβλιογραφία με τον όρο POS (part-of-speech) tags. Συγκρίνουν τους αλγορίθμους Naive Bayes, Maximum Entropy και Support Vector Machines (SVM). Η χρήση των SVM με μοναδικό χαρακτηριστικό τα μονογράμματα αποφέρει το καλύτερο αποτέλεσμα (82.9 %). Παρατηρούν πως η προσθήκη των διγραμμάτων στο διάνυσμα χαρακτηριστικών βελτιώνει την επίδοση των NaiveBayes και Maximum Entropy αλλά όχι των SVM. Τέλος, καταλήγουν στο συμπέρασμα ότι προσθέτοντας την άρνηση (negation) ως ξεχωριστό χαρακτηριστικό καθώς και τα POS tags δεν παρατηρείται

βελτίωση ενώ η χρήση μόνο των διγραμμάτων οδηγεί σε χειρότερα αποτελέσματα εξαιτίας του αραιού χώρου χαρακτηριστικών (feature space).

Οι Pak & Paroubek (2010) [31] χρησιμοποιούν επίσης θετικά και αρνητικά emoticons για να δημιουργήσουν ένα σύνολο εκπαίδευσης με 300 χιλιάδες tweets. Ωστόσο, συλλέγουν παράλληλα tweets από λογαριασμούς εφημερίδων στο Twitter για να τα χρησιμοποιήσουν ως ουδέτερα πρότυπα και να μελετήσουν το γενικότερο πρόβλημα κατηγοριοποίησης με τις τρεις κλάσεις. Στο στάδιο της προεπεξεργασίας τους, αφαιρούν τα ονόματα χρηστών, τα emoticons, τους υπερσυνδέσμους και τα άρθρα (a, an, the) (stopwords), οι λέξεις άρνησης (no, not) συνενώνονται με την προηγούμενη ή επόμενη λέξη και το κείμενο κατακερματίζεται στα κενά και τα σημεία στίξης (tokenization). Πειραματίζονται με μονογράμματα, διγράμματα και τριγράμματα. Κατασκευάζουν δύο εκδοχές του ταξινομητή Naive Bayes χρησιμοποιώντας διαφορετικά χαρακτηριστικά. Ο ένας στηρίζεται στην παρουσία ενός ν-γράμματος στο κείμενο (χαρακτηριστικό με δυαδική τιμή). Ο άλλος βασίζεται στην πληροφορία κατανομής των μερών του λόγου (POS) για να εκτιμήσει την παρουσία των POS tags και να υπολογίσει την εκ των υστέρων πιθανότητα (posterior probability) του μοντέλου Naive Bayes. Θεωρώντας πως τα δύο χαρακτηριστικά είναι υπό συνθήκη ανεξάρτητα και κατ' επέκταση και οι δύο ταξινομητές, η τελική ταξινόμηση γίνεται με βάση το λογάριθμο πιθανοφάνειας (log-likelihood). Παρατηρούν ότι το συγκεκριμένο μοντέλο υπερτερεί των Support Vector Machines (SVM) και των Conditional Random Fields (CRF) έχοντας καλύτερο αποτέλεσμα στον όχι και τόσο γνωστό δείκτη F0.5 = 0.63. Συμπεραίνουν πως τα διγράμματα πετυχαίνουν τη καλύτερη ακρίβεια γιατί "αποτελούν μία καλή ισορροπία ανάμεσα στην κάλυψη των εύρους (μονογράμματα) και στην ικανότητα αναγνώρισης συναισθηματικών μοτίβων έκφρασης (τριγράμματα)" [31].

Οι Barbosa & Feng (2010) [5] προτείνουν ένα ταξινομητή δύο φάσεων. Στην πρώτη φάση, τα tweets κατηγοριοποιούνται σε υποκειμενικά ή αντικειμενικά και στη συνέχεια τα υποκειμενικά διακρίνονται σε θετικά ή αρνητικά tweets. Ακολουθούν μία διαφορετική προσέγγιση για την κατασκευή του συνόλου εκπαίδευσης : χρησιμοποιούν ως noisy labels όχι τα emoticons αλλά τη "γνώμη" τριών εργαλείων ανίχνευσης συναισθήματος : Twendz2, Twitter Sentiment3 και TweetFeel2, διαγράφοντας τα tweets στα οποία δεν υπάρχει ομόφωνη απόφαση. Το τελικό σύνολο εκπαίδευσης περιλαμβάνει 200 χιλιάδες tweets για ανίχνευση υποκειμενικότητας και 71.046 θετικά και 79.628 αρνητικά tweets για κατηγοριοποίηση πολικότητας. Διαχωρίζουν τα χαρακτηριστικά τους σε δύο κατηγορίες : τα μέτα-χαρακτηριστικά (meta-features) και τα χαρακτηριστικά σύνταξης του tweets (tweet-syntax). Η πρώτη κατηγορία περιλαμβάνει χαρακτηριστικά όπως POS tags, ο βαθμός υποκειμενικότητας και πολικότητας της εκάστοτε λέξης όπως αυτά προσδιορίζονται στο λεξικό MPQA [54]. Η δοθείσα πολικότητα αντιστρέφεται από θετική σε αρνητική και αντίστροφα όταν μία άρνηση προηγείται της λέξης. Η δεύτερη κατηγορία περιλαμβάνει πιο ειδικά για το Twitter χαρακτηριστικά όπως η ύπαρξης retweets, hashtags, υπερσυνδέσμων, θαυμαστικών, ερωτηματικών, emoticons και κεφαλαίων

γραμμάτων. Η συχνότητα κάθε χαρακτηριστικού κανονικοποιείται διαιρώντας με το πλήθος των όρων του κάθε tweet. Συνολικά και από τις 2 κατηγορίες προκύπτουν 20 χαρακτηριστικά. Τα καλύτερα αποτελέσματα προκύπτουν χρησιμοποιώντας ως ταξινομητή τον SVM και στις δύο φάσεις επιτυγχάνουν δε ακρίβεια 81.9 % στην αναγνώριση υποκειμενικότητας και 81.3 % στην αναγνώριση πολικότητας ενώ ως βάση αναφοράς θεωρούνται τα μονογράμματα με ακρίβεια 72.4 % και 79.1 % αντίστοιχα στις δύο φάσεις. Παρατηρούν ότι τα μέτα-χαρακτηριστικά είναι πιο σημαντικά στη φάση προσδιορισμού της πολικότητας ενώ τα χαρακτηριστικά σύνταξης κατά την φάση αναγνώρισης της υποκειμενικότητας. Οι συγγραφείς καταλήγουν στο συμπέρασμα ότι επειδή χρησιμοποιούν μια πιο αφηρημένη αναπαράσταση των δεδομένων και όχι μεμονωμένους όρους του, η προσέγγισή τους εμφανίζει μεγαλύτερη ανοχή στο θόρυβο και μεροληψία (bias) του συνόλου εκπαίδευσης σε σχέση με άλλες μεθόδους ενώ εμφανίζει καλύτερη συμπεριφορά ως προς την ικανότητας γενίκευσης όταν χρησιμοποιούνται σχετικά λίγα εκπαιδευτικά πρότυπα.

Οι Bermingham & Smeaton (2010) [6] εξετάζουν την επίδραση του μικρού μήκους των tweets στις συνήθεις τεχνικές επιβλεπόμενης μάθησης. Συλλέγουν tweets από τα δέκα πιο δημοφιλή θέματα (trending) σε πέντε κατηγορίες (ψυχαγωγία, προϊόντα & υπηρεσίες, αθλητικά, επικαιρότητα και εταιρείες) δημιουργώντας ένα σύνολο από χειροκίνητα κατηγοριοποιημένα πρότυπα (1.410 θετικά, 1.040 αρνητικά και 2.597 ουδέτερα tweets). Κατά την προεπεξεργασία των δεδομένων αντικαθιστούν τα ονόματα χρηστών, υπερσυνδέσμους και hashtags με προκατασκευασμένες λέξεις-κλειδιά. Ως χαρακτηριστικά αναπαράστασης του κειμένου χρησιμοποιούνται μονογράμματα, διγράμματα, τριγράμματα, POS tags και POS ν-γράμματα. Συγκρίνουν τα αποτελέσματα κατηγοριοποίησης από την εφαρμογή της μεθόδου σε tweets, κριτικές ταινιών και αναρτήσεις ιστολογίων. Παρατηρούν ότι ο Naive Bayes εμφανίζει υψηλότερα ποσοστά ακρίβειας σε σχέση με τα Support Vector Machines στη περίπτωση των tweets αλλά όχι σε μεγαλύτερου μήκους κείμενα (κριτικές και αναρτήσεις ιστολογίων). Χρησιμοποιώντας Naive Bayes και μονογράμματα επιτυγχάνουν ακρίβεια 74.85 % στο πρόβλημα της δυαδικής κατηγοριοποίησης και 61.3 % στο γενικότερο πρόβλημα των τριών κλάσεων. Η χρήση ν-γραμμάτων και POS tags βελτιώνει την ακρίβεια μόνο στην περίπτωση των μεγάλων κειμένων ενώ τα POS ν-γράμματα, η επίλυση συνωνύμων (stemming) και η αφαίρεση κοινών λέξεων (stopwording) δεν οδηγούν σε καλύτερα αποτελέσματα. Συμπεραίνουν ότι η ανάλυση συναισθήματος σε κείμενα μικρού μήκους όπως τα tweets είναι εν γένει πιο εύκολο πρόβλημα.

Οι Bifet & Frank (2010) [7] ασχολούνται με την ανάλυση συναισθήματος σε μεγάλη ροή δεδομένων (data stream) από το Twitter. Προτείνουν ένα Kappa στατιστικό δείκτη κυλιόμενου παραθύρου για να αξιολογήσουν την επίδοση κατηγοριοποίησης σε χρονομεταβλητές ροές δεδομένων. Χρησιμοποιούν το σύνολο δεδομένων (dataset) Stanford Twitter Sentiment των Go et al. [14] και το Edinburgh Twitter Corpus των Petrovic et al. [35], χρησιμοποιώντας τα emoticons ως δείκτες αυτόματης ταξινόμησης (noisy labels). Κατά την προεπεξεργασία των δεδομένων, αντικαθιστούν τα ονόματα

χρηστών και τους υπερσυνδέσμους με κατάλληλες λέξεις-κλειδιά ενώ ως χαρακτηριστικά χρησιμοποιούν μόνο τα μονογράμματα. Παράλληλα με την αξιολόγηση μέσω του προτεινόμενου δείκτη, αναφέρουν ως καλύτερα αποτελέσματα για το πρόβλημα της δυαδικής ταξινόμησης 82.45 % στο πρώτο dataset με χρήση Naive Bayes και 86.26 % στο δεύτερο σύνολο tweets με χρήση Στοχαστικής Κλίσης Καθόδου (Stochastic Gradient Descent SGD). Παρατηρούν πως οι Naive Bayes και SGD παρουσιάζουν παρόμοια ποσοστά ακρίβειας σε αντίθεση με τα δένδρα Hoeffing τα οποία υστερούν συστηματικά. Επομένως, συνιστούν να αποφεύγονται γενικά ταξινομητές δένδρα στην περίπτωση μεγάλης ροής δεδομένων και προτείνουν έναντι τη χρήση SGD καθώς προσαρμόζονται καλύτερα στις αλλαγές με τη πάροδο του χρόνου και παράλληλα οι αλλαγές στα βάρη των χαρακτηριστικών μπορούν να χρησιμοποιηθούν για να παρακολουθούνται αλλαγές στο συναίσθημα και απόψεις γύρω από συγκεκριμένα θέματα.

Οι Davidov et al. (2010) [9] ταξινομούν αυτόματα το σύνολο δεδομένων των O'Connor et al. [29] χρησιμοποιώντας ως δείκτες κατηγοριοποίησης (noisy labels) 50 hashtags και 15 emoticons. Το σύνολο χαρακτηριστικών περιλαμβάνει : λέξεις, ν-γράμματα (2-5), μήκος του κάθε tweet, πλήθος σημείων στίξης, θαυμαστικών, ερωτηματικών, εισαγωγικών, κεφαλαίων γραμμάτων και λέξεων καθώς και την ύπαρξη λέξεων με υψηλή συχνότητα εμφάνισης. Εφαρμόζοντας μία τεχνική παρόμοια με αυτή των k-Κοντινότερων Γειτόνων (k-Nearest Neighbours kNN) επιτυγχάνουν βέλτιστη ακρίβεια στο μέσο αρμονικό δείκτη F1 = 0.86 στην περίπτωση των emoticons και F1 = 0.8 στην περίπτωση των hashtags για τη δυαδική εκδοχή του προβλήματος μέσω 10-πλης σταυρωτής επικύρωσης (10-fold cross-validation). Στο γενικότερο πρόβλημα των τριών κλάσεων, η επίδοση ήταν αισθητά χαμηλότερη (0.64 και 0.31 αντίστοιχα). Παράλληλα, προτείνουν δύο διαφορετικές μεθόδους για την αυτόματη ανίχνευση της επικάλυψης συναισθήματος και των αλληλεξαρτήσεων ανάμεσα στις λέξεις του κειμένου. Παρατηρούν ότι οι λέξεις, τα σημεία στίξης και τα εκφραστικά μοτίβα είναι τα πιο σημαντικά χαρακτηριστικά ενώ τα ν-γράμματα οδηγούν σε οριακή βελτίωση.

Σε αντίθεση με τις περισσότερες μελέτες, οι Agarwal et al. (2011) [1] δεν περιορίζουν τη συλλογή tweets μόνο σε αυτά της αγγλικής γλώσσας μέσω του Twitter API αλλά χρησιμοποιούν την υπηρεσία Google Translate για μετάφρασή τους. Δημιουργούν ένα σύνολο από 8753 χειροκίνητα ταξινομημένα tweets στις τρεις κατηγορίες (θετικά, αρνητικά, ουδέτερα) αφού πρώτα διέγραψαν όσα περιείχαν λάθη λόγω μετάφρασης. Για την αξιολόγηση των μεθόδων τους, δημιουργούν ένα ισοζυγισμένο σύνολο δεδομένων περιλαμβάνοντας 1709 πρότυπα από κάθε κατηγορία, 5127 tweets συνολικά. Προτείνουν δύο νέες τεχνικές στο στάδιο της προεπεξεργασίας : δημιουργούν ένα λεξικό emoticons το οποίο περιέχει 170 emoticons όπως καταγράφονται στη Wikipedia χωρισμένα σε πέντε κατηγορίες ανάλογα με το συναίσθημα που εκφράζουν (υπερβολικά θετικό, θετικό, ουδέτερο, αρνητικό, υπερβολικά αρνητικό) και κατασκευάζουν ένα λεξικό με τη μετάφραση 5184 ακρωνύμιων. Έπειτα, αντικαθιστούν τα emoticons με την συναισθηματική πολικότητα στο λεξικό ,τα ακρωνύμια με την

κανονική τους μορφή καθώς και τα ονόματα χρηστών, τους υπερσυνδέσμους, τα hashtags και τις αρνήσεις με γνωστές λέξεις-κλειδιά και περιορίζουν τους επιμηκυμένους χαρακτήρες σε δύο π.χ. το cooooooooool σε coool. Αρκετά από τα χαρακτηριστικά που χρησιμοποιούν βασίζονται στη πρότερη πολικότητα των λέξεων την οποία προσδιορίζουν χρησιμοποιώντας το Dictionary of Affect in Language (DAL) [53] και το επεκτείνουν με συνώνυμα από το WordNet [10]. Στο σύνολο των χαρακτηριστικών περιλαμβάνονται το πλήθος, η συχνότητα και το ποσοστό των λέξεων, άρθρων (stopwords), αγγλικών λέξεων, σημείων στίξης, θαυμαστικών, tags, αρνήσεων και κεφαλαίων τα οποία υπολογίζονται για όλο και για το τελευταίο τρίτο του tweet. Συγκρίνουν πέντε διαφορετικά μοντέλα τα οποία στηρίζονται στα Support Vector Machines (SVM) και θέτουν ως βάση αναφοράς τα μονογράμματα. Παρατηρούν ότι ταξινομητές οι οποίοι στηρίζονται μόνο σε αφηρημένα γλωσσικά χαρακτηριστικά αποδίδουν εξίσου καλά με τα μονογράμματα τα οποία χρησιμοποιούν πολύ περισσότερα χαρακτηριστικά. Στο πρόβλημα της δυαδικής κατηγοριοποίησης επιτυγχάνουν βέλτιστη ακρίβεια 75.39 % με μοντέλο που συνδυάζει μονογράμματα και αφηρημένα γλωσσικά χαρακτηριστικά. Στο πρόβλημα των τριών κλάσεων το μοντέλο με τα καλύτερα ποσοστά ακρίβειας (60.83 %) συνδυάζει αφηρημένα γλωσσικά χαρακτηριστικά μαζί με μία ειδική δενδρική αναπαράσταση του κάθε όρου σε SVM με partial tree kernel [26]. Παρατηρούν ότι τα αφηρημένα γλωσσικά χαρακτηριστικά με τη περισσότερη πληροφορία είναι εκείνα τα οποία συνδυάζουν την πρότερη πολικότητα των λέξεων μαζί με τα POS tags. Καταλήγουν στο συμπέρασμα ότι η ανάλυση συναισθήματος σε δεδομένα από το Twitter δε διαφέρει από την ανάλυση συναισθήματος σε άλλα είδη κείμενου.

Οι Jiang et al. (2011) [18] εξετάζουν την ανάλυση συναισθήματος σε συγκεκριμένα θέματα στόχους (target-dependent) εφαρμόζοντας μία τεχνική τριών σταδίων. Όμοια με τους Barbosa και Feng [5], πρώτα ταξινομούν τα tweets σε υποκειμενικά και αντικειμενικά και στη συνέχεια (2η φάση) τα υποκειμενικά tweets σε θετικά και αρνητικά χρησιμοποιώντας δύο ξεχωριστούς ταξινομητές Support Vector Machines (SVM) με γραμμική συνάρτηση πυρήνα. Υποστηρίζουν ότι συνήθεις τεχνικές ([5, 14]) δεν επαρκούν καθώς όλα τα χαρακτηριστικά είναι ανεξάρτητα του στόχου. Στο τρίτο στάδιο προτείνουν μία μέθοδο η οποία βασίζεται σε γράφους με σκοπό να αυξήσουν την ακρίβεια : εξετάζουν το ευρύτερο εννοιολογικό πλαίσιο που τοποθετείται το κάθε tweet μέσω των συσχετιζόμενων με αυτό tweets όπως retweets, tweets που περιέχουν την ίδια οντότητα-στόχο και προέρχονται από τον ίδιο χρήστη καθώς και τις πιθανές απαντήσεις από ή στο εκάστοτε tweet. Μέσω του Twitter API συλλέγουν tweets που περιέχουν 5 δημοφιλείς οντότητες : Obama, Google, iPad, Lakers, Lady Gaga και τα ταξινομούν χειροκίνητα στις 3 κατηγορίες δημιουργώντας τελικά ένα σύνολο από 459 θετικά, 268 αρνητικά και 1212 ουδέτερα tweets. Κατά το στάδιο της προεπεξεργασίας με τη βοήθεια εξωτερικών εργαλείων εφαρμόζουν τεχνικές κανονικοποίησης των κειμένων (διόρθωση απλών ορθογραφικών λαθών και εμφατικής επιμήκυνσης λέξεων), επίλυσης συνωνύμων (stemming), γραμματικής αναγνώρισης (POS tagging) και συντακτικής ανάλυσης έτσι

ώστε να κατασκευάσουν χαρακτηριστικά ειδικά για τις εξεταζόμενες οντότητες. Παράλληλα, υπολογίζουν και χαρακτηριστικά ανεξάρτητα του στόχου μέσω μονογραμμάτων και του λεξικού General Inquirer 4. Τέλος, επιλύουν τις έμμεσες αναφορές αναζητώντας τις K πιο ισχυρά συσχετιζόμενες με τους στόχους λέξεις και φράσεις μέσω του δείκτη PMI (Pointwise Mutual Inforation). Παρατηρούν πως ο συνδυασμός των χαρακτηριστικών ειδικών του στόχου μαζί με άλλα χαρακτηριστικά οδηγεί σε καλύτερη επίδοση 68.2 % ξεπερνώντας κατά 7.9 % την δική τους υλοποίηση της εκδοχής των Barbosa και Feng. Αναφέρουν ως πιθανή αιτία το γεγονός ότι οι Barbosa και Feng χρησιμοποιούν πιο αφηρημένα χαρακτηριστικά ενώ η δική τους προσέγγιση στηρίζεται περισσότερο σε λεξιλογικά χαρακτηριστικά. Συμπεραίνουν πως τα χαρακτηριστικά εξαρτώμενα από τους στόχους συντελούν καθοριστικά ιδίως σε περιπτώσεις όπου το συναίσθημα δεν αναφέρεται στην πραγματικότητα στην οντότητα-στόχο.

Οι Kouloumpis et al. (2011) [20] ερευνούν την συμβολή των γλωσσολογικών χαρακτηριστικών στην αναγνώριση πολικότητας των tweets. Χρησιμοποιούν δύο γνωστά datasets και επιλέγουν διαφορετικό δείκτη αυτόματης κατηγοριοποίησης (noisy labels) για κάθε ένα. Εξετάζουν το Endinburgh Twitter Corpus των Petrovic et al. [35] μέσω hashtags ενδεικτικών του συναισθήματος (π.χ. #imthankfulfor, #ihate, #news) και το Stanford Twitter Sentiment Corpus των Go et al. [14] μέσω emoticons. Όμοια με τις περισσότερες μελέτες, στο στάδιο της προεπεξεργασίας αντικαθιστούν τα ονόματα χρηστών, τους υπερσυνδέσμους και τα hashtags με κατάλληλες λέξεις-κλειδιά, τις συντομογραφίες με την κανονικής τους μορφή και διορθώνουν την ορθογραφία των λέξεων από εμφατική επιμήκυνση και χρήση κεφαλαίων γραμμάτων. Αφαιρούν επίσης κοινές λέξεις και άρθρα (stopwording) και αναγνωρίζουν γραμματικά την κάθε λέξη (POS tagging). Χρησιμοποιούν ένα αρκετά μεγάλο σύνολο χαρακτηριστικών : μονογράμματα, διγράμματα, τα πρώτα 1000 μονογράμματα και διγράμματα με βάση το κέρδος πληροφορία κατά το δείκτη Chi-squared, τη πρότερη πολικότητα των λέξεων κατά το MPQA λεξικό [54], το πλήθος και ποσοστό των κυριότερων POS tags και δυαδικά χαρακτηριστικά μικροιστολογίων για την ύπαρξη εξειδικευμένων όρων (hashtags, emoticons). Παρατηρούν ότι η χρήση του ταξινομητή AdaBoost υπερτερεί των Support Vector Machines (SVM) έχοντας βέλτιστη επίδοση 75% στο πρόβλημα των τριών κλάσεων με χρήση όλων των χαρακτηριστικών εκτός του πλήθους των POS tags. Συμπεραίνουν ότι σε αντίθεση με τα χαρακτηριστικά μικροιστολογίων που ήταν τα πιο χρήσιμα, τα POS tags οδηγούν σε μείωση της ακρίβειας και δεν είναι μάλλον κατάλληλα για χρήση σε κείμενα από μικροϊστολόγια.

Οι Saif et al. (2011) [41] μελετούν το πρόβλημα της αραιότητας των δεδομένων λόγω του μικρού μήκους των μηνυμάτων του Twitter. Προτείνουν δύο διαφορετικές προσεγγίσεις της σημασιολογικής εξομάλυνσης (semantic smoothing) με σκοπό να εξάγουν σημασιολογικά κρυμμένες έννοιες από τα κείμενα και να τις χρησιμοποιήσουν ως επιπρόσθετα χαρακτηριστικά για την εκπαίδευση των ταξινομητών. Εξετάζουν ένα ισοζυγισμένο υποσύνολο 60 χιλιάδων tweets από το Standford Twitter

Sentiment Corpus των Go et al. [14] μαζί με το σύνολο εξέτασης με 177 αρνητικά και 182 θετικά χειροκίνητα ταξινομημένα tweets. Για την εξαγωγή των εννοιών χρησιμοποιούν την υπηρεσία AlchemyAPI5 όπου αναγνωρίζουν γνωσ τές οντότητες σ τα κείμενα των tweets. Στη πρώτη μέθοδο (shallow semantic smoothing) οι λέξεις αντικαθίστανται με τις αντίστοιχες σημασιολογικές τους έννοιες ενώ στη δεύτερη (interpolation) το γλωσσικό μοντέλο μονογράμματος παρεμβάλλεται μαζί ένα παραγωγικό μοντέλο λέξεων δοθέντων των σημασιολογικών εννοιών στον ταξινομητή Naive Bayes. Παρατηρούν πως ενώ η πρώτη μέθοδος οδηγεί σε μείωση της ακρίβειας κατά 5% σε σχέση με ένα ταξινομητή Naive Bayes με μόνο χαρακτηριστικό τα μονογράμματα, η μέθοδος της παρεμβολής οδηγεί σε οριακή βελτίωση επιτυγχάνοντας ακρίβεια 81.3% στη δυαδική κατηγοριοποίηση. Οι παραπάνω προσεγγίσεις βελτιώνονται στο [42] όπου προστίθεται προεπεξεργασία κειμένου: αντικατάσταση ονομάτων χρηστών, υπερσυνδέσμων με λέξεις-κλειδιά, διόρθωση της επιμήκυνσης φθόγγων, αφαίρεση hashtags, emoticons, μονών χαρακτήρων, ψηφίων και άλλων μη αλφαριθμητικών χαρακτήρων. Επεκτείνουν το αρχικό σύνολο εξέτασης σε 100 tweets και επιτυγχάνουν ακρίβεια 84% με βελτιωμένη έκδοση της μεθόδου παρεμβολής. Στο [43] οι συγγραφείς εξετάζουν τις μεθόδους σε δύο ακόμη σύνολα δεδομένων: HealthCare Reform (HCR) [46] και Obama McCain Debate [45] με καλύτερα ποσοστά ακρίβειας 79% και 69,15% αντίστοιχα. Καταλήγουν στο συμπέρασμα ότι η τεχνική της σημασιολογικής εξομάλυνσης εμφανίζει καλύτερα αποτελέσματα σε μεγάλα σύνολα δεδομένων που περιέχουν ποικίλη θεματολογία.

Οι Liu et al. (2012) [23] προτείνουν μία νέα προσέγγιση για τη συνένωση χειροκίνητα και αυτόματα μέσω noisy labels ταξινομημένων tweets χρησιμοποιώντας το γλωσσικό μοντέλο ESLM (Emoticon Smoothed Language Model). Αρχικά, εκπαιδεύουν ένα γλωσσικό μοντέλο με χειροκίνητα ταξινομημένα πρότυπα από το Sanders Corpus6 (570 θετικά, 654 αρνητικά, 2503 ουδέτερα tweets). Στη συνέχεια, μέσω του Twitter API συλλέγουν tweets που περιέχουν emoticons με σκοπό να εξομαλύνουν το γλωσσικό μοντέλο. Στο στάδιο της προεπεξεργασίας αντικαθιστούν ονόματα χρηστών, υπερσυνδέσμους και ψηφία με λέξεις-κλειδιά, διαγράφουν κοινές λέξεις (stopwording), αντικαθιστούν συνώνυμα (stemming) και κεφαλαία με πεζά γράμματα ενώ διαγράφουν retweets και διπλότυπα από το αρχικό dataset. Επίσης, ξεχωρίζουν υπερσυνδέσμους που αναφέρονται σε εικόνες/video από τα υπόλοιπα URLs. Παρατηρούν ότι το προτεινόμενο μοντέλο εμφανίζει πολύ καλύτερη συμπεριφορά συγκρινόμενο με ένα γλωσσικό μοντέλο πλήρως επιβλεπόμενο επιτυγχάνοντας ακρίβεια 82.5% και 79.5% στην αναγνώριση πολικότητας και υποκειμενικότητας αντίστοιχα. Τονίζουν τη σημασία των χειροκίνητα ταξινομημένων tweets επισημαίνοντας την προσθήκη τους ως κύριο λόγο βελτίωσης των αποτελεσμάτων.

Οι Mohammand et al. (2013) [24] ασχολούνται με την ανάλυση συναισθήματος σε επίπεδο μηνύματος και οντότητας σχεδιάζοντας μία εκδοχή ταξινομητή Support Vector Machines (SVM) για κάθε επίπεδο ανάλυσης. Εξετάζουν το πρόβλημα των τριών κλάσεων και αξιολογούν το μοντέλο τους

με δεδομένα από το διαγωνισμό SemEval2013 [27], καταλαμβάνοντας την 1η θέση και στα δύο υποπροβλήματα υποκειμενικότητας και πολικότητας. Το σύνολο εκπαίδευση περιέχει 3855 θετικά, 1624 αρνητικά και 4889 ουδέτερα χειροκίνητα ταξινομημένα tweets. Κατά την προεπεξεργασία των δεδομένων, αντικαθιστούν τα ονόματα χρηστών και τους υπερσυνδέσμους με κατάλληλες λέξεις-κλειδιά και κατακερματίζουν το κείμενο στους επιμέρους όρους, τους οποίους στη συνέχεια αναγνωρίζουν γραμματικά (POS tagging). Κάθε tweet αναπαρίσταται με ένα σύνολο χαρακτηριστικών : λέξεις και χαρακτήρες ν-γραμμάτων (3, 4, 5), πλήθος κεφαλαίων, POS tags, hashtags, emoticons, αρνήσεων και σημείων στίξης καθώς και λεξικογραφικών ιδιοτήτων που προσδιορίζονται με τη βοήθεια λεξικών. Παρατηρούν ότι ένας ταξινομητής Support Vector Machines (SVM) με όλα τα παραπάνω χαρακτηριστικά εμφανίζει πολύ καλύτερη συμπεριφορά σε σχέση με έναν απλό SVM ταξινομητή που χρησιμοποιεί μόνο μονογράμματα : F-score 69.02% έναντι 39.61% για το πρόβλημα της υποκειμενικότητας και 88.93% έναντι 80.28% στο πρόβλημα της πολικότητας. Συμπεραίνουν πως τα συναισθηματικά χαρακτηριστικά που προκύπτουν μέσω των λεξικών σε συνδυασμό με τα ν-γράμματα συνεισφέρουν το περισσότερο κέρδος στην αύξηση της ακρίβειας.

Οι Günther & Furrer (2013) [15] χρησιμοποιούν επίσης το SemEval2013 dataset [27] αλλά μελετούν την ανάλυση συναισθήματος μόνο σε επίπεδο μηνύματος-πρότασης. Κατά την προεπεξεργασία των δεδομένων, κανονικοποιούν τα κείμενα αντικαθιστώντας κεφαλαία με πεζά γράμματα, αφαιρώντας ψηφία και επαναλαμβανόμενους χαρακτήρες που προσδίδουν έμφαση. Εκτός από την ύπαρξη ή απουσία κανονικοποιημένων και συνώνυμων λέξεων, στο διάνυσμα των χαρακτηριστικών περιλαμβάνονται η χρήση άρνησης, η πρότερη πολικότητα κάθε όρου μέσω του SentiWordNet καθώς και η ύπαρξη/απουσία του εκάστοτε όρου σε clusters με λέξεις από το Twitter. Παρατηρούν ότι η κατασκευή ενός γραμμικού μοντέλου με συνάρτηση εκπαίδευσης Στοχαστική Κλίση Καθόδου (Stochastic Gradient Descent) υπερτερεί έναντι άλλων μεθόδων καταλαμβάνοντας τη 2η θέση στο διαγωνισμό με F1 = 0.65 και τονίζουν ότι η επιλογή ενός αλγορίθμου εκπαίδευσης είναι πιο σημαντική από την επιλογή των χαρακτηριστικών αυξανόμενου του πλήθους των προτύπων εκπαίδευσης.

Οι Aston et al. (2014) [4] μελετούν την ανάλυση συναισθήματος σε ροή δεδομένων από το Twitter όπου συνήθεις τεχνικής μάθησης δέσμης (batch learning) είναι αναποτελεσματικές. Εξετάζουν εναλλακτικούς αλγορίθμους μάθησης με περιορισμούς ως προς το χρόνο επεξεργασίας και χωρητικότητας διατηρώντας υψηλά ποσοστά ακρίβειας. Ασχολούνται με τα προβλήματα υποκειμενικότητας και πολικότητας ξεχωριστά, κατασκευάζοντας δύο εκδοχές από το σύνολο Sanders Corpus6. Αναπαριστούν το κείμενο μέσω ν-γραμμάτων αλλά παράλληλα επιτρέπουν την ύπαρξη ν-γραμμάτων διαφόρων μεγεθών στο σύνολο αναπαράστασης το οποίο αποκαλούν 1-ν γράμματα. Ο αριθμός των πιθανών ν-γραμμάτων αυξάνεται εκθετικά με την αύξηση του μεγέθους ν οπότε ο υπολογισμός όλων των πιθανών χαρακτηριστικών είναι πρακτικά ανέφικτος σε περιορισμένο χρόνο. Επιλέγουν, λοιπόν, τα N πρώτα χαρακτηριστικά ν-γραμμάτων όπως προκύπτουν μέσω 6

διαφορετικών αλγορίθμων αξιολόγησης της περιεχόμενης πληροφορίας (Chi-squared, Filtered Feature, Gain Ratio, Info Gain, OneR και Relief). Έπειτα, εξετάζουν 3 εκδοχές του ταξινομητή Perceptron (Simple, Best Learning Rate, Voted) καθώς και συνδυασμούς τους. Παρατηρούν ότι οι εκδοχές Best Learning Rate και Voted εμφανίζουν σταθερά και όμοια αποτελέσματα με την πρώτη, ωστόσο, να απαιτεί πολύ μεγαλύτερο χρόνο εκπαίδευσης. Συνδυάζοντας τις δύο αυτές τεχνικές επιτυγχάνουν καλύτερη ακρίβεια και στα δύο προβλήματα με F-score 85% και 78% στην ανίχνευση υποκειμενικότητας και πολικότητας αντίστοιχα. Συμπεραίνουν πως δε συντελούν με τον ίδιο βαθμό όλα τα χαρακτηριστικά στην κατηγοριοποίηση καθώς εξαιρώντας κάποια από αυτά, μειώνεται ο χρόνος εκτέλεσης χωρίς όμως να επηρεάζεται αρνητικά η επίδοση του ταξινομητή.

2.6.2 Εργασίες με χρήση Μη Επιβλεπόμενης Μηχανικής Μάθησης

Οι O'Connor et al. (2010) [29] εξετάζουν τη σύνδεση των δημοσκοπήσεων με την ανάλυση συναισθήματος σε tweets που αναφέρονται στον Πρόεδρο των ΗΠΑ Barrack Obama. Συλλέγουν μέσω του TwitterAPI 1 δις tweets αναρτήθηκαν στο διάστημα 2008 2009 χωρίς να ελέγχουν τα δημογραφικά χαρακτηριστικά των δημιουργών και τη γλώσσα γραφής. Κατηγοριοποιούν κάθε tweet μετρώντας αν περιέχει περισσότερες θετικές ή αρνητικές λέξεις, αναζητώντας την πολικότητα του κάθε όρου στο λεξικό συναισθημάτων MPQA [54]. Παρατηρούν πως αν και πρόκειται για μία απλή μέθοδο ανίχνευσης συναισθήματος, επιτυγχάνει να συλλέξει το συνολικό συναίσθημα της κοινής γνώμης και εμφανίζει υψηλή συσχέτιση με τα αποτελέσματα των δημοσκοπήσεων σε βαθμό μέχρι και 80%. Συμπεραίνουν πως οι απαιτητικές και χρονοβόρες τεχνικές εξόρυξης της κοινής γνώμης μέσω δημοσκοπήσεων μπορούν να ενισχυθούν και να συμπληρωθούν από την ανάλυση συναισθήματος στον τεράστιο όγκο εύκολα συλλέξιμων δεδομένων από τα κοινωνικά δίκτυα.

Οι Gayo-Avello et al. (2011) [11] ασχολούνται επίσης με την εξόρυξη κοινής γνώμης σε θέματα πολιτικής από tweets. Εξετάζουν την ικανότητα πρόβλεψης του τελικού αποτελέσματος εγκλογικών αναμετρήσεων εφαρμόζοντας δημοφιλείς τεχνικές σε tweets από τις εκλογές της Βουλής των Αντιπροσώπων των ΗΠΑ το 2010. Στηρίζονται στη μέθοδο των O'Connor et al.[29], χρησιμοποιούν το λεξικό MPQA [54] και εισάγουν κάποιες τροποποιήσεις έτσι ώστε το συνολικό μοντέλο να προσαρμόζεται στη φύση και τα ιδιαίτερα χαρακτηριστικά του κάθε εκλογικού συστήματος. Λαμβάνουν υπόψη tweets τα οποία περιέχουν ονόματα υποψηφίων από αντίπαλες παρατάξεις και δεν επιτρέπουν ένα tweet να έχει ταυτόχρονα δύο αντίθετες πολικότητες. Σε αντίθεση με άλλες μελέτες, δεν παρατηρούν άμεση συσχέτιση με τα αποτελέσματα δημοσκοπήσεων, εμφανίζοντας μέσο όρο σφάλματος 7.6% εκτός του αποδεκτού ορίου 2-3%. Καταλήγουν στο συμπέρασμα ότι η ακρίβεια των λεξικών συναισθήματος όταν εφαρμόζονται σε πολιτικές συζητήσεις είναι αρκετά χαμηλή και ως προσέγγιση ανεπαρκής και δηλώνουν ότι απαιτούνται πιο εξελιγμένες τεχνικές για να συλλάβουμε τη

δυναμική του πολιτικού λόγου στα κοινωνικά δίκτυα. Θεωρούν πως η αποτυχία της ανάλυσης συναισθήματος μέσω λεξικών είναι ως ένα βαθμό αναμενόμενη καθώς οι ακριβείς δημογραφικές πληροφορίες των χρηστών που συζητούν για τις εκλογές είναι ελάχιστες, η φύση και ποιότητα των online πολιτικών συζητήσεων αδιευκρίνιστη όπως επίσης και ο τρόπος με τον οποίο ομάδες με διαφορετικές ιδεολογίες συμμετέχουν και ασκούν επιρροή μέσω των κοινωνικών δικτύων.

Οι Thelwall et al. (2010) [49] προτείνουν ένα νέο αλγόριθμο βασισμένο σε λεξικό συναισθήματος τον οποίο αποκαλούν SentiStrength χρησιμοποιώντας επίσης μη λεξικογραφικές γλωσσολογικές πληροφορίες και κανόνες. Εκτός από την πολικότητα κάθε κειμένου (θετικό/αρνητικό) υπολογίζουν και την αντίστοιχη ισχύ του συναισθήματος με εύρος τιμών 1 έως 5. Αρχικά, χρησιμοποιούν ένα σύνολο από 2600 σχόλια από το MySpace και κατασκευάζουν μία λίστα με 298 θετικούς και 465 αρνητικούς όρους ταξινομημένους ως προς την πολικότητά τους μαζί με την αντίστοιχη ισχύ τους. Έπειτα, επεκτείνουν το μοντέλο με λίστες από emoticons, όρους άρνησης, λέξεις που αυξάνουν ή μειώνουν τη ισχύ του συναισθήματος των συμφραζόμενων όρων (booster words). Παράλληλα, στο στάδιο της προεπεξεργασίας, διορθώνονται απλά ορθογραφικά λάθη και αντιμετωπίζονται φαινόμενα εμφατικής επιμήκυνσης (επαναλαμβανόμενα γράμματα, φθόγγοι και σημεία στίξης). Συγκρίνοντας το μοντέλο σε σχέση με διάφορους ταξινομητές επιβλεπόμενης μηχανικής μάθησης παρατηρούν ότι συμπεριφέρεται καλύτερα στην αναγνώριση των αρνητικών αλλά όχι των θετικών σχολίων. Βελτιωμένη έκδοση του αλγορίθμου παρουσιάζεται στο [48] όπου οι συγγραφείς αυξάνουν τους όρους από 693 σε 2310, εισάγουν μία λίστα με ιδιώματα καθώς και την έννοια της ενίσχυσης της πολικότητας λόγω εμφατικής επιμήκυνσης. Συγκρίνουν πάλι το μοντέλο με διαφορετικούς αλγορίθμους μάθησης σε διαφορετικά σύνολα δεδομένων και από το Twitterκαι παρατηρούν πως σε γενικές γραμμές εμφανίζει ικανοποιητικά ποσοστά ακρίβειας και μόνο ο ταξινομήτης Linear Regression υπερτερεί συστηματικά. Καταλήγουν ότι η ανάλυση συναισθήματος που βασίζεται σε λεξικά συναισθήματος και κανόνες έχει σταθερή συμπεριφορά και είναι ανεξάρτητη του πεδίου εφαρμογής.

Οι Zhang et al. (2011) [56] προτείνουν ένα μοντέλο βασισμένο σε κανόνες για την ανάλυση συναισθήματος σε επίπεδο οντότητας σε δεδομένα που συλλέγουν από το Twitter. Προεπεξεργάζονται το σύνολο δεδομένων : διαγράφουν διπλότυπα, αφαιρούν ονόματα χρηστών και υπερσυνδέσμους, αντικαθιστούν συντομογραφίες με την κανονική τους μορφή και αναγνωρίζουν γραμματικά τους επιμέρους όρους των μηνυμάτων (POS tagging). Έπειτα, υπολογίζουν τη συναισθηματική τιμή κάθε όρου με βάση την ομοιότητά του με λέξεις από το λεξικό συναισθημάτων και επιλύουν τις απλές αναφορές αντιστοιχίζοντας αντωνυμίες με την πιο κοντινή οντότητα του κειμένου. Εφαρμόζοντας το σύνολο κανόνων ο αλγόριθμος διαχωρίζει τις προτάσεις σε δηλωτικές, προστακτικές και ερωτηματικές ενώ παράλληλα μπορεί να αναγνωρίσει συγκρίσεις, αρνήσεις και αντιθετικές περιόδους. Τονίζουν ότι αυτή η μέθοδος εμφανίζει αρκετά καλή ακρίβεια (precision)

αλλά χαμηλή ανάκληση (recall). Εκπαιδεύουν, λοιπόν, ένα δυαδικό ταξινομητή Support Vector Machines (SVM) με πρότυπα που προκύπτουν από την παραπάνω μη-επιβλεπόμενη διαδικασία ο οποίος ταξινομεί τα tweets στις τελικές τους κατηγορίες. Παρατηρούν πως η προσθήκη του ταξινομητή βελτιώνει δραματικά την ανάκληση και το F-score και παράλληλα ξεπερνά αρκετούς από τους state-of-the-art τεχνικές-σημεία αναφοράς.

Οι Kumar & Sebastian (2012) [21] ασχολούνται με την εξαγωγή γνώμης από tweets προτείνοντας μία μη επιβλεπόμενη υβριδική μέθοδο που συνδυάζει μεγάλα σώματα κειμένου (corpus) και λεξικά για να προσδιορίσει τον σημασιολογικό προσανατολισμό των όρων του κειμένου. Προτείνουν ένα τρόπο υπολογισμού της τιμής του συναισθήματος ο οποίος εκτός από τις πολικότητες του λεξικού λαμβάνει υπόψη και το πλήθος των emoticons, επαναλαμβανόμενων γραμμάτων, θαυμαστικών και κεφαλαίων, χαρακτηριστικά που χρησιμοποιούνται συνήθως από τις επιβλεπόμενες τεχνικές.

Οι Hu et al. (2013) [17] μελετούν τη μη-επιβλεπόμενη ανάλυση συναισθήματος στα κοινωνικά δίκτυα με τη βοήθεια "σημάτων συναισθήματος" , δηλαδή οποιαδήποτε πληροφορία η οποία μπορεί να συσχετισθεί με συναισθηματική πολικότητα. Εξετάζουν την επίδραση των emoticons και των κοινών τους εμφανίσεων στην κατηγοριοποίηση των δεδομένων από τα σύνολα Stanford Twitter Sentiment Corpus [14] και το Obama McCain Debate Corpus [45]. Στο στάδιο της προεπεξεργασίας αναπαριστούν τα κείμενα μέσω μονογραμμάτων και επιλέγουν την εμφάνιση των όρων (term presence) ως χαρακτηριστικό το οποίο σε συνδυασμό με το λεξικό MPQA [54] χρησιμοποιούνται για να υπολογίσουν τους δείκτες ένδειξης (indication) και συσχέτισης (correlation) συναισθήματος των tweets. Παρατηρούν ότι το μοντέλο εμφανίζει καλύτερη συμπεριφορά σε σχέση με άλλες μη-επιβλεπόμενες τεχνικές με βέλτιστη επίδοση 74.2% και 70.97% αντίστοιχα στα 2 σύνολα δεδομένων. Συμπεραίνουν πως η χρήση σημάτων συναισθήματος βελτιώνει την ακρίβεια με το δείκτη ένδειξης να έχει τη μεγαλύτερη συνεισφορά.

Οι Ortega et al. (2013) [30] προτείνουν ένα μη-επιβλεπόμενο σύστημα ανάλυσης συναισθήματος για το πρόβλημα της γενικής κατηγοριοποίησης των tweets. Αρχικά, προεπεξεργάζονται τα μηνύματα : κατακερματίζουν τις προτάσεις σε όρους, αφαιρούν retweets, ονόματα χρηστών, υπερσυνδέσμους, τον χαρακτήρα "#" από τα hashtags. Έπειτα, αντικαθιστούν τα emoticons με λέξεις συναισθήματος από ένα χειροκίνητα κατασκευασμένο λεξικό emoticons μέσω της Wikipedia και τις συντομογραφίες με την πλήρη μορφή τους και, τέλος, αφαιρούν συνήθεις λέξεις, λημματοποιούν και αναγνωρίζουν γραμματικά (POS tagging) τους όρους του κειμένου. Στη συνέχεια, υπολογίζουν τη συναισθηματική πολικότητα κάθε λέξης λαμβάνοντας υπόψη και το ευρύτερο εννοιολογικό πλαίσιο στο οποίο εμφανίζεται, αποσαφηνίζοντας την έννοιά της μέσω του WordNet και του SentiWordNet. Η τελική κατηγοριοποίηση του κάθε tweet πραγματοποιείται μέσω ενός μοντέλου κανόνων. Εξετάζουν την ακρίβεια του συστήματος σε δεδομένα από το διαγωνισμό SemEval2013 [27] όπου καταλαμβάνουν

τη 25η θέση με επίδοση F1=51.17%. Τονίζουν πως τα αποτελέσματα είναι αρκετά ικανοποιητικά δεδομένης της δυσκολίας του προβλήματος και του γεγονότος ότι δε χρησιμοποιούν κανένα σύνολο εκπαίδευσης, είναι μία καθαρά μη επιβλεπόμενη προσέγγιση.

Οι Saif et al. (2014) [40] παρουσιάζουν την μη-επιβλεπόμενη μέθοδο SentiCircle η οποία βασίζεται σε λεξικό συναισθήματος. Θεωρούν πως το συναίσθημα ενός όρου δεν είναι στατικό αλλά εξαρτάται από το εννοιολογικό πλαίσιο στο οποίο ανήκει καθώς και από τα συμφραζόμενα. Προτείνουν, λοιπόν, το δείκτη TDOC (Term Degree of Correlation) για να υπολογίσουν τη σχέση ανάμεσα σε μία λέξη w και τους συμφραζόμενους όρους ci με την ίδια σημασιολογική χροιά. Έπειτα, αναπαριστούν τη λέξη w και τους όρους ci σε πολικό σύστημα συντεταγμένων με κέντρο τη λέξη w , ακτίνα τον δείκτη TDOC και γωνία την πρότερη πολικότητα κάθε όρου όπως προσδιορίζεται από λεξικό συναισθήματος. Αξιοποιώντας τις τριγωνομετρικές ιδιότητες της αναπαράστασης υπολογίζουν τον συναισθηματικό προσανατολισμό και την ισχύ της λέξης w. Εξετάζουν 3 διαφορετικά σύνολα δεδομένων: Stanford Twitter Sentiment Corpus [14], Obama McCain Debate Corpus και Health Care Reform και παρατηρούν ότι η μέθοδος SentiCircle συναγωνίζεται την state-of-the-art μέθοδο SentiStrength έχοντας μέση ακρίβεια 72.39% έναντι 71.7%.

2.7 Διαδικτυακές Υπηρεσίες

Στην παρούσα ενότητα παραθέτουμε παραδείγματα διαδικτυακών υπηρεσιών ανάλυσης συναισθήματος και συνοψίζουμε τα χαρακτηριστικά που διαθέτουν.

Repustate[1]

Η υπηρεσία Repustate είναι μία μηχανή ανάλυσης συναισθήματος που υποστηρίζει πολλαπλές γλώσσες όπως Αραβικά, Κινέζικα, Αγγλικά, Γαλλικά, Γερμανικά, Εβραϊκά, Ισπανικά και Ταϊλανδέζικα. Ο χρήστης καλείται να εισάγει ένα κείμενο το πολύ 2048 χαρακτήρων, καθώς επίσης τη γλώσσα που είναι γραμμένο το κείμενο. Η υπηρεσία μπορεί να συμπεράνει το συνολικό συναίσθημα του κειμένου ή να διακρίνει τα θέματα που θίγονται στο κείμενο και το συναίσθημα που περιβάλλει το κάθε θέμα. Η διάκριση των θεμάτων επιτυγχάνεται με χρήση τεχνικών επεξεργασίας Φυσικής Γλώσσας και ο υπολογισμός του συναισθήματος γίνεται στην κλίμακα [-1,1]. Το κείμενο μπορεί να προέρχεται από κάποιο κοινωνικό δίκτυο. Παραδείγματα χρήσης της υπηρεσίας είναι ο έλεγχος ικανοποίησης των πελατών μίας επιχείρησης, η έρευνα προτιμήσεων των καταναλωτών μέσω

[1] www.repustate.com

ενημερωτικών δελτίων ή η διατήρηση της εμπιστοσύνης των πελατών με την παροχή κατάλληλων προσφορών.

Lexalytics[2]

Η υπηρεσία Lexalytics είναι μία μηχανή ανάλυσης συναισθήματος που υποστηρίζει πολλαπλές γλώσσες. Αναλύει ιστοσελίδες, περιεχόμενο κοινωνικών δικτύων ή απλό κείμενο και ανιχνεύει το συναίσθημα σε επίπεδο συλλογής εγγράφων, απλού εγγράφου, παραγράφου, πρότασης ή οντότητας. Σε επίπεδο κειμένων, ανιχνεύει τα αντικείμενα που σχολιάζονται, την έκταση που καταλαμβάνει ο σχολιασμός τους, το συναίσθημα που αφενός συνοδεύει τα αντικείμενα και αφετέρου διέπει το κείμενο, καθώς και τις συγκεκριμένες λέξεις που υποδεικνύουν το συναίσθημα. Ο υπολογισμός του συναισθήματος γίνεται στην κλίμακα [-1,1]. Η ανάλυση γίνεται με τη βοήθεια λεξικού που διαθέτει το σύστημα. Ο χρήστης έχει τη δυνατότητα να προσθέσει ένα λεξικό της αρεσκείας του και να διεξάγει την ανάλυση βάσει αυτού.

Sentiment140[3]

Η υπηρεσία Sentiment140 αποτελεί ένα πρόγραμμα του πανεπιστημίου του Stanford. Σκοπός της είναι η ανίχνευση του συναισθήματος για ένα προϊόν, κάποια εταιρεία ή κάποιο δημόσιο πρόσωπο μέσω της ανάλυσης διαφόρων Tweets. Η υπηρεσία συγκεντρώνει μία συλλογή από Tweets που περιέχουν την προς εξέταση λέξη, διακρίνει το συναίσθημα κάθε Tweet σε θετικό, αρνητικό ή ουδέτερο και υπολογίζει το ποσοστό των θετικών και των αρνητικών Tweets.

Η μέθοδος που επιστρατεύεται εντοπίζει και αφαιρεί τους ειδικούς χαρακτήρες κάθε Tweet. Κατόπιν, επεξεργάζεται το κείμενο με μονογράμματα ή διγράμματα και χρησιμοποιεί διαφόρους αλγορίθμους μηχανικής μάθησης, όπως Naive Bayes, Maximum Entropy και Support Vector Machines, για την εκπαίδευση των ταξινομητών. Υπάρχει δυνατότητα ανάλυσης Tweets γραμμένων σε Αγγλικά ή Ισπανικά.

SentimentAnalyzer[4]

Η υπηρεσία SentimentAnalyzer είναι μία απλή διαδικτυακή εφαρμογή που εντοπίζει το συναίσθημα Αγγλικών, Γερμανικών και Γαλλικών κειμένων. Συγκεκριμένα, διακρίνει την πολικότητα του

[2] www.lexalytics.com
[3] www.sentiment140.com
[4] sentimentanalyzer.appspot.com

συναισθήματος του κειμένου σε θετική, αρνητική ή ουδέτερη και υπολογίζει την ακριβή συναισθηματική τοποθέτηση στην κλίμακα [-1,1].

3

Ανάλυση Συναισθήματος με Γράφους ν-γραμμάτων

Οι περισσότερες μέθοδοι Ανάλυσης Συναισθήματος εντοπίζουν εκφραστικά μοτίβα σε ένα προκαθορισμένο σύνολο φυσικών γλώσσών. Κατά συνέπεια, οι μέθοδοι δεν είναι εφαρμόσιμες σε πολυγλωσσικά σύνολα δεδομένων, καθώς επίσης σε γλωσσικά περιβάλλοντα που περιέχουν ιδιωματισμούς, συντμήσεις και νεολογισμούς. Ωστόσο, η ευρεία χρήση και η συνεχώς αυξανόμενη διείσδυση των Κοινωνικών Δικτύων στο σύγχρονο κόσμο, δημιουργούν καθημερινά ένα εν πολλοίς πολύτιμο περιεχόμενο που αποτυπώνει τις απόψεις των χρηστών για ποικίλα θέματα, διαγράφει τις καταναλωτικές ή ιδεολογικές τάσεις που διέπουν επιμέρους κοινωνίες ή ευρύτερα σύνολα χρηστών, δίνει βήμα σε χρήστες με αυξημένη πολιτική, ιδεολογική ή επιχειρηματική επιρροή, αλλά και καταγράφει σχολιασμούς απλών καθημερινών γεγονότων από πλήθος χρηστών. Η μελέτη του περιεχομένου των Κοινωνικών Δικτύων και η συνακόλουθη εξαγωγή συμπερασμάτων επιβάλλουν την εφαρμογή μεθόδων Ανάλυσης Συναισθήματος ανεξάρτητων του εκάστοτε γλωσσικού περιβάλλοντος. Για την υπέρβαση των περιορισμών που θέτει το κάθε περιβάλλον, στην παρούσα εργασία υιοθετούμε την προσέγγιση των Aisopos et al. [3] για την Ανάλυση Συναισθήματος δεδομένων του κοινωνικού δικτύου Twitter με τη βοήθεια γράφων ν-γραμμάτων. Στις ενότητες που ακολουθούν επεξηγείται η εν λόγω μέθοδος και παρουσιάζονται οι αλγόριθμοι Μηχανικής Μάθησης που χρησιμοποιούνται στην υλοποίηση και εφαρμογή της προσέγγισης.

3.1 Αναπαράσταση Δεδομένων

3.1.1 Βασικές Έννοιες και Ορισμοί

Στην Επεξεργασίας Φυσικής Γλώσσας, η χρήση ν-γραμμάτων χρησιμοποιείται συχνά για τη διόρθωση ορθογραφικών λαθών, το φιλτράρισμα ανεπιθύμητης αλληλογραφίας και την ανίχνευση λογοκλοπής. Ένα ν-γραμμα είναι ένα διατεταγμένο σύνολο ν λέξεων ή χαρακτήρων. Στην Ανάλυση Συναισθήματος τα ν-γράμματα αξιοποιούνται ως γλωσσικό μοντέλο για την εξαγωγή των τιμών του διανύσματος χαρακτηριστικών.

Παράδειγμα 3.1.1 Παραδείγματα ν-γραμμάτων της πρότασης test texts είναι τα εξής:

- μονογράμματα λέξεων: test, texts
- διγράμματα χαρακτήρων: te, es, st, t_, _t, te, ex, xt, ts
- 3-γράμματα χαρακτήρων: tes, est, st_, t_t, _te, tex, ext, xts

Στη συνέχεια, με το γενικό όρο στοιχείο θα αναφερόμαστε είτε σε λέξη είτε σε χαρακτήρα.

Ορισμός 3.1.1 Έστω ακολουθία στοιχείων $T^l \equiv c_1 c_2 \dots c_l$ μήκους l και ν ένας κατάλληλος θετικός ακέραιος. Ένα ν-γραμμα στοιχείων $S^\nu \equiv s_1 s_2 \dots s_\nu$ είναι μία υπακολουθία μήκους ν της T^l. Ένα ν-γραμμα που εκτείνεται από το στοιχείο i έως το k θα σημειώνεται ως $S_{i,k}$, ενώ ένα ν-γραμμα μήκους ν ως S^ν, σε αντιδιαστολή με το σύμβολο S_j που θα δηλώνει απαρίθμηση ενός ν-γράμματος. Το μήκος ενός ν-γράμματος ονομάζεται επίσης βαθμός του ν-γράμματος.

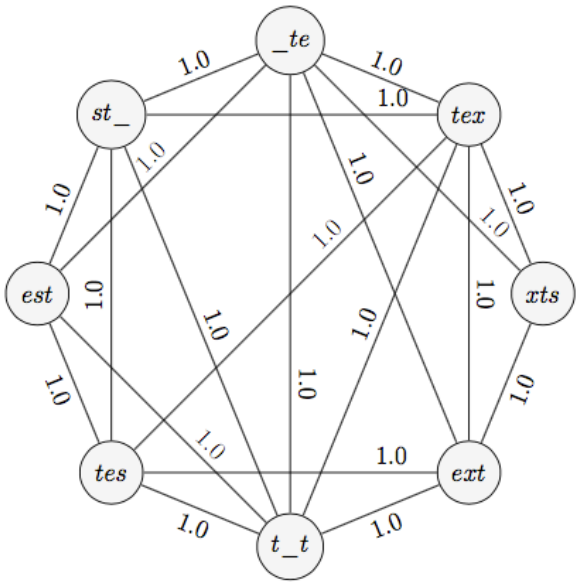
Η συνολική αναπαράσταση ενός κειμένου επιτυγχάνεται με τη βοήθεια του γράφου ν-γραμμάτων. Το κείμενο οργανώνεται σε ν-γράμματα και τα ν-γράμματα συσχετίζονται μέσω ακμών κατάλληλου βάρους για τη δημιουργία του γράφου ν-γραμμάτων.

Ένας γράφος ν-γραμμάτων είναι ένας γράφος $G = \{V^G, E^G, L, W\}$ όπου V^G είναι το σύνολο κορυφών, E^G το σύνολο ακμών, L μία συνάρτηση που αποδίδει ένα ν-γραμμα ως επισημείωση σε κάθε κορυφή και W μία συνάρτηση που αποδίδει ένα βάρος σε κάθε ακμή. Κάθε ακμή $uv \in E^G$ υποδηλώνει την εγγύτητα των ν-γραμμάτων που αντιστοιχούν στις κορυφές u και v. Το βάρος της ακμής uv προκύπτει από την απόσταση μεταξύ των ν-γραμμάτων των κορυφών u και v στο κείμενο. Για τον

καθορισμό τόσο της εγγύτητας όσο και της απόστασης δύο ν-γραμμάτων εντός ενός κειμένου χρησιμοποιείται ως αναφορά ένα παράθυρο μήκους D_{win}.

Ορισμός 3.1.2 Έστω $\Sigma = \{S_1, S_2, \ldots, S_m\}$ είναι το σύνολο όλων ανεξαιρέτως των ν-γραμμάτων ενός κειμένου. Τότε $G = \{V^G, E^G, L, W\}$ είναι ο γράφος των ν-γραμμάτων του κειμένου όπου V^G είναι το σύνολο κορυφών, E^G είναι το σύνολο των ακμών, $L: V^G \rightarrow \Lambda$ είναι η συνάρτηση επισημείωσης ενός ν-γράμματος σε κάθε κορυφή και $W: E^G \rightarrow R$ είναι η συνάρτηση ανάθεσης βάρους σε κάθε ακμή. Δύο κορυφές u και v ενώνονται μέσω της ακμής uv αν και μόνο αν τα ν-γράμματα που αντιστοιχούν στις κορυφές u και v βρίσκονται στο κείμενο εντός παραθύρου μήκους D_{win} .

Στο σχήμα που ακολουθεί απεικονίζεται ο γράφος τριγραμμάτων για τη συμβολοσειρά “test_texts”. Παρατηρούμε ότι η πυκνότητα πληροφορίας της αναπαράστασης με γράφο είναι σαφώς μεγαλύτερη συγκρινόμενη με την απλή παράθεση των τριγραμμάτων στο παράδειγμα 3.1.1.



Σχήμα 3.1: Πάραδειγμα Γράφου 3-γραμμάτων

Το βάρος της ακμής uv είναι μοναδιαίο αν και μόνο αν τα τριγράμματα u και v γειτνιάζουν το πολύ σε απόσταση τριών χαρακτήρων. Διαφορετικά, το βάρος είναι μηδέν και δε σχηματίζεται ακμή. Στη συγκεκριμένη περίπτωση, το παράθυρο έχει μήκος τρία.

Συνήθως χρησιμοποιείται ένα σταθερού μήκους παράθυρο D_{win} χαρακτήρων ή λέξεων γύρω από ένα συγκεκριμένο ν-γράμμα S με όλους τους χαρακτήρες ή λέξεις εντός του παραθύρου να θεωρούνται γείτονες του S.

Έχει ιδιαίτερη σημασία να επιλεγεί κατάλληλο μήκος παραθύρου, καθώς "δεν είναι όλες οι αποστάσεις το ίδιο σημαντικές και επομένως δύο ν-γράμματα σε απόσταση 150 χαρακτήρων δεν έχουν μάλλον ουσιαστική σύνδεση και εξάρτηση."[12].

Επίσης, καθοριστικό ρόλο έχει ο τρόπος υπολογισμού της γειτνίασης δύο ν-γραμμάτων. Μπορεί κανείς να λάβει υπόψη :

- μόνο τους προηγούμενους χαρακτήρες του S κατά την κύλιση του παραθύρου D_{win} στο κείμενο (ασύμμετρη προσέγγιση),
- και τους επόμενους χαρακτήρες (συμμετρική προσέγγιση),
- και τους επόμενους χαρακτήρες αλλά και την πραγματική απόσταση μεταξύ του S και του εκάστοτε ν-γράμματος (κανονικοποιημένη κατά Gauss συμμετρική προσέγγιση).

3.1.2 Αναπαράσταση Δεδομένων με Γράφους ν-γραμμάτων

Στην παρούσα ενότητα περιγράφεται η εφαρμογή της μεθόδου αναπαράστασης μέσω γράφων ν-γραμμάτων στην ανάλυση συναισθήματος σε δεδομένα από το κοινωνικό δίκτυο Twitter. Η περιγραφή και τα συμπεράσματα προέρχονται από τους Aisopos et al. (2011) στο [3].

Κάθε tweet t_i αναπαρίσταται με ένα γράφο ν-γραμμάτων που ονομάζεται γράφος μηνύματος, tweet graph, και συμβολίζεται με G_{t_i}. Για την κατασκευή του γράφου χρησιμοποιείται ένα κυλιόμενο παράθυρο D_{win} μήκους ν όπου το κείμενο του tweet αναλύεται σε επικαλυπτόμενες συμβολοακολουθίες μήκους ν, δηλαδή ν-γράμματα χαρακτήρων. Μία ακμή που συνδέει ένα ζεύγος ν-γραμμάτων υποδηλώνει ότι τα ν-γράμματα γειτνιάζουν στο κείμενο σε απόσταση το πολύ ν χαρακτήρων.

Γράφοι μηνυμάτων με όμοια πολικότητα συνθέτουν το γράφο G^{T^P} της συγκεκριμένης κλάσης πολικότητας T^P. Ο γράφος G^T δημιουργείται από τα μηνύματα της πολικότητας T^P του συνόλου εκπαίδευσης. Το μήνυμα t_i της T^P μετασχηματίζεται σε ένα γράφο G_{t_i} που συνενώνεται με το γράφο της κλάσης πολικότητας και δημιουργείται ο γράφος $G_i^{T^P}$. Αρχικά, ο γράφος της κλάσης είναι κενός. Μόλις ολοκληρωθεί η διαδικασία, προκύπτει ο γράφος G^{T^P}.

Επομένως, $G_i^{T^P} = \{V^i, E^i, W^i\}$ όπου $V^i = V^{i-1} \cup V^{G_{t_i}}$, $E^i = E^{i-1} \cup E^{G_{t_i}}$ και $W^i(e) = W^{i-1}(e) + \frac{W^{G_{t_i}}(e) - W^{i-1}(e)}{i}$.

Όπως εξηγείται στο [13], η διαίρεση με i εξασφαλίζει ότι το αθροιζόμενο βάρος συγκλίνει στη μέση τιμή των αντίστοιχων βαρών των ακμών ανάμεσα σε όλους τους γράφους μηνύματος G_{t_i} έτσι, ώστε η ενημέρωση να είναι ανεξάρτητη της σειράς με την οποία συγχωνεύονται τα tweets. Μετά τη συγχώνευση όλων των tweets της T^P στο γράφο G^{T^P}, οι ακμές $E^{G^{T^P}}$, αποτυπώνουν τα πιο χαρακτηριστικά εκφραστικά μοτίβα που εμφανίζουν τα μηνύματα της εκάστοτε κλάσης, όπως επαναλαμβανόμενες και γειτονικές συμβολοακολουθίες, ειδικοί χαρακτήρες και ψηφία.

Για να εκτιμήσουμε την ομοιότητα μεταξύ ενός γράφου μηνύματος G_{t_i} και ενός γράφου κλάσης G^{T^P} χρησιμοποιούνται τρεις διαφορετικοί δείκτες - μετρικές ομοιότητας [12]:

I. Ομοιότητα Συνοχής (Containment Similarity - CS): εκφράζει το ποσοστό των ακμών του γράφου G_{t_i} οι οποίες περιέχονται και στο γράφο G^{T^P}. Αν G είναι ένας γράφος ν-γραμμάτων και e μία ακμή ενός γράφου ν-γραμμάτων, τότε ορίζουμε τη συνάρτηση μ όπου $\mu(e, G) = 1$ αν και μόνο αν $e \in G$ και 0 αλλιώς.

$$CS(G_{t_i}, G^{T^P}) = \sum_{e \in G_{t_i}} \frac{\mu(e, G^{T^P})}{min(|E_{G_{t_i}}|, |E_{G^{T^P}}|)}$$

II. Ομοιότητα Τιμής (Value Similarity - VS): εκφράζει το πλήθος των ακμών του γράφου G_{t_i} οι οποίες περιέχονται και στο G^{T^P}, λαμβάνοντας υπόψη τα βάρη τους. Κάθε κοινή ακμή e έχει βάρη $w^{t_i}(e)$ και $w^{T^P}(e)$ στους γράφους G_{t_i} και G^{T^P} συνεισφέροντας $\frac{VR(e)}{max(|E_{G_{t_i}}|, |E_{G^{T^P}}|)}$ στο δείκτη VS. Ο όρος VR, Value Ratio, είναι ένας συμμετρικός συντελεστής κλίμακας $VR: [0,1] \rightarrow [0,1]$ με $VR(e) = \frac{min(|E_{G_{t_i}}|, |E_{G^{T^P}}|)}{max(|E_{G_{t_i}}|, |E_{G^{T^P}}|))}$ οπότε η Ομοιότητα Τιμής υπολογίζεται από τη σχέση:

$$VS(G_{t_i}, G^{T^P}) = \sum_{e \in G_{t_i}} \frac{VR(e)}{max(|E_{G_{t_i}}|, |E_{G^{T^P}}|)}$$

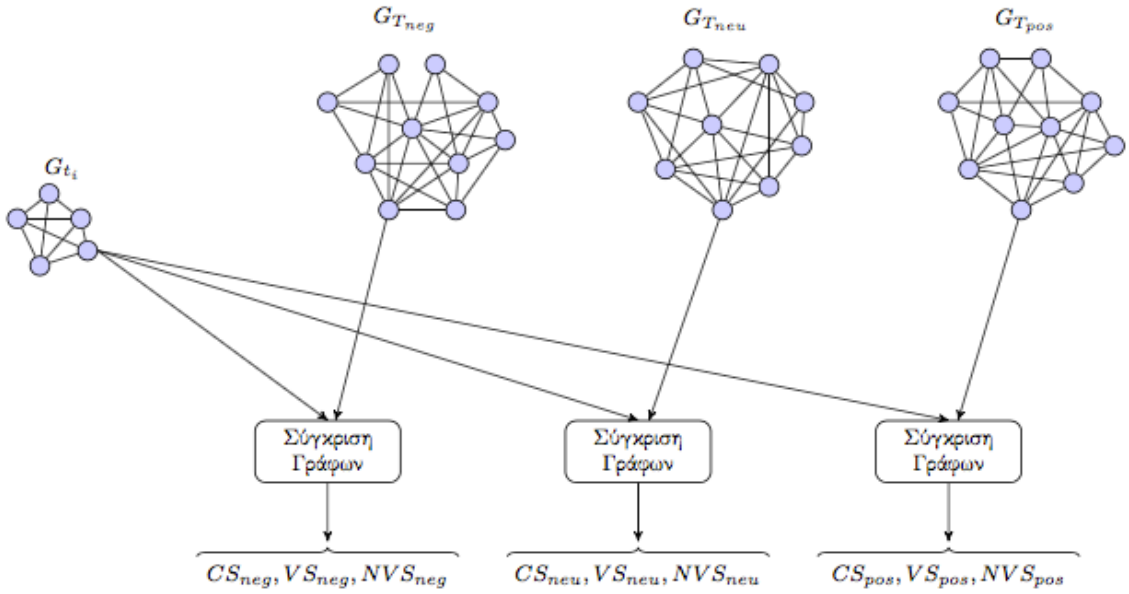
Ο δείκτης VS συγκλίνει στο 1 για γράφους G_{t_i} και G^{T^P} που μοιράζονται ακμές και παρόμοια βάρη με την τιμή VS = 1 να δηλώνει το τέλειο ταίριασμα μεταξύ των δύο συγκρινόμενων γράφων.

III. Κανονικοποιημένη Ομοιότητα Τιμής (Normalized Value Similarity - NVS): αποσυνδέει το δείκτη Ομοιότητας Τιμής VS από την επίδραση του μεγέθους του μεγαλύτερου γράφου διαιρώντας με το δείκτη Ομοιότητας Μεγέθους (Size Similarity - SS).

$$NVS(G_{t_i}, G^{T^P}) = \frac{VS(G_{t_i}, G^{T^P})}{SS(G_{t_i}, G^{T^P})}$$

όπου $SS(G_{t_i}, G^{T^P})) = \frac{min(|G_{t_i}|,|G^{T^P}|)}{max(|G_{t_i}|,|G^{T^P}|)}$.

Ο προσδιορισμός του συναισθήματος ενός tweet t_i ξεκινάει με τη σύγκριση του γράφου μηνύματος G_{t_i} με τους γράφους $G^{T^{POS}}$, $G^{T^{NEG}}$ και $G^{T^{NEUT}}$ και τον προσδιορισμό της εγγύτητάς του με κάθε κλάση. Δηλαδή υπολογίζονται οι 3 δείκτες ομοιότητας (CS, VS, NVS) για κάθε κλάση και τοποθετούνται στο διάνυσμα χαρακτηριστικών που δίνεται ως είσοδος στον ταξινομητή. Η διαδικασία απεικονίζεται στο σχήμα που ακολουθεί.



Σχήμα 3.2: Διαδικασία υπολογισμού διανύσματος χαρακτηριστικών με το μοντέλο γράφων ν-γραμμάτων

Κατόπιν, ο ταξινομητής εξετάζει το διάνυσμα χαρακτηριστικών και αποφαίνεται την πιθανότερη πολικότητα του tweet.

Οι Aisopos et al. [2] χρησιμοποιούν γράφους ν-γραμμάτων με χαρακτήρες και όχι λέξεις, καθώς η εξαγωγή των λέξεων από το εκάστοτε κείμενο απαιτεί τεχνικές εξειδικευμένες σε κάθε γλώσσα και

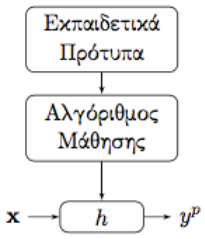
αίρει τη γλωσσική ανεξαρτησία της προσέγγισης. Επίσης, οι μέθοδοι που στηρίζονται στην εξαγωγή λέξεων δεν είναι αποτελεσματικές στην ανάλυση δεδομένων κοινωνικών δικτύων λόγω της πολυγλωσσίας και των ποικίλων ιδιωματισμών των δικτύων. Ως αποτέλεσμα, όροι σημασιολογικά ταυτόσημοι, αλλά διαφορετικής συντακτικής μορφής γίνονται αντιληπτοί ως διακριτές έννοιες. Κατά συνέπεια, το μέγεθος του διανύσματος χαρακτηριστικών αυξάνεται δραματικά και επιβαρύνει σημαντικά τη χωρική και χρονική πολυπλοκότητα των αλγορίθμων μάθησης και ταξινόμησης. Αντιθέτως, το μέγεθος του διανύσματος χαρακτηριστικών στο μοντέλο ν-γραμμάτων με χαρακτήρες εξαρτάται αποκλειστικά από τον αριθμό των κλάσεων κατηγοριοποίησης.

3.2 Αλγόριθμοι κατηγοριοποίησης

Το πρόβλημα της κατηγοριοποίησης δεδομένων συνίσταται στην πρόβλεψη της τιμής μίας μεταβλητής κλάσης y με σύνολο τιμών $c_1, c_2, \ldots, c_k$ λαμβάνοντας υπόψη ένα σύνολο χαρακτηριστικών $x = \{x_1, x_2, \ldots, x_n\}$. Το πρόβλημα αντιμετωπίζεται σε δύο βήματα, της μάθησης και της κατηγοριοποίησης.

Κατά τη μάθηση, ο ταξινομητής εκπαιδεύεται σύμφωνα με κάποιο αλγόριθμο μάθησης και κατασκευάζει μία συνάρτηση κατηγοριοποίησης $h : X \rightarrow Y$ μελετώντας m εκπαιδευτικά πρότυπα της μορφής $(x^{(i)}, y^{(i)})$όπου $y^{(i)}$ είναι η κλάση που ανήκει το πρότυπο i.

Κατά την κατηγοριοποίηση, ο ταξινομητής δέχεται ως είσοδο κάποιο άγνωστο στοιχείο **x** που ονομάζεται διάνυσμα χαρακτηριστικών και απαντάει με την πιθανότερη κλάση $h(x) = y^p$ που ανήκει το **x**. Παρατηρούμε ότι η ακρίβεια του ταξινομητή εξαρτάται από τον αλγόριθμο μάθησης που έχει χρησιμοποιηθεί κατά τη μάθηση. Στο σχήμα που ακολουθεί περιγράφεται συνοπτικά η διαδικασία της κατηγοριοποίησης.



Σχήμα 3.3: Διαδικασία Κατηγοριοποίησης

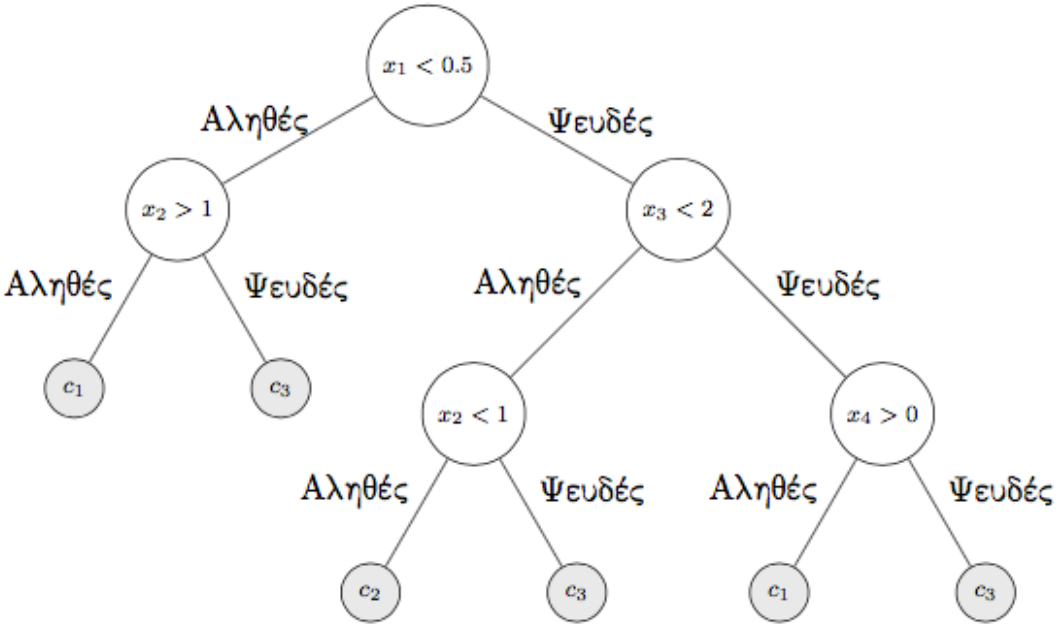
Στη συνέχεια της παρούσας ενότητας, παραθέτουμε τους αλγορίθμους μάθησης που χρησιμοποιήθηκαν για την κατασκευή των ταξινομητών της υπηρεσίας.

3.2.1 Δένδρα αποφάσεων

Δένδρο απόφασης είναι ένας ταξινομητής δενδρικής μορφής. Κάθε εσωτερικός κόμβος του δένδρου αντιστοιχεί σε μία κατάσταση που διαμερίζει τα δεδομένα σε διακριτές ομάδες σύμφωνα με κάποιο

χαρακτηριστικό. Κάθε φύλλο αντιστοιχεί σε μία μοναδική κλάση που ορίζεται από το μονοπάτι με αρχή τη ρίζα του δένδρου και πέρας το φύλλο.

Το δένδρο κατασκευάζεται με αναδρομική διάσπαση των υποσυνόλων των δεδομένων μάθησης σύμφωνα με την επιλογή χαρακτηριστικών και τις συνθήκες ελέγχου. Η επιλογή των χαρακτηριστικών γίνεται με μία συνάρτηση αξιολόγησης, συνήθως του Κέρδους Πληροφορίας (Information Game) της Εντροπίας Πληροφορίας (Information Entropy). Ευρέως γνωστοί αλγόριθμοι μηχανικής μάθησης με χρήση δένδρων είναι οι ID3, C4.5 και CART.



Σχήμα 3.4: Παράδειγμα Δένδρου Αποφάσεων

Ορισμός 3.2.1 Θεωρούμε το δένδρο απόφασης T, τις k κλάσεις c_i της μεταβλητής y και το σύνολο S των δεδομένων εκπαίδευσης. Ορίζουμε την εντροπία του S ως

$$E(S) = -\sum_{i=1}^{k} P(y = c_i|S) \cdot log_2 P(y = c_i|S)$$

όπου $P(y = c_i|S)$ είναι το ποσοστό των προτύπων του S που ανήκουν στην κατηγορία c_i.

Η εντροπία είναι μέτρο της ομογενότητα της μεταβλητής κλάσης y στο χώρο S. Στην περίπτωση που ο χώρος S αντιστοιχεί στη ρίζα του δένδρου, τότε η εντροπία υπολογίζεται για όλο το σύνολο των δεδομένων.

Ορισμός 3.2.2 Θεωρούμε το δένδρο απόφασης T, το σύνολο S των δεδομένων εκπαίδευσης και μία ανεξάρτητη μεταβλητή A. Ορίζουμε το κέρδος πληροφορίας ως

$$G(S, A) = E(S) - \sum_{u} \frac{|S_u|}{|S|} \cdot E(S_u)$$

όπου $E(S)$ είναι η εντροπία πληροφορίας του υπό εξέταση κόμβου, u μία από τις δυνατές τιμές του A, S_u το πλήθος των δεδομένων με $A = u$ και $E(S_u)$ η εντροπία πληροφορίας του υπό εξέταση κόμβου ως προς την τιμή $A = u$.

Το κέρδος πληροφορίας αναπαριστά τη μείωση της εντροπίας του συνόλου εκπαίδευσης αν επιλεγεί ως παράμετρος διαχωρισμού η μεταβλητή A.

Όταν μειώνεται η εντροπία πληροφορίας, αυξάνεται η πυκνότητα πληροφορίας και η περιγραφή γίνεται πιο συμπαγής.

Τα δένδρα αποφάσεων απαιτούν λίγη προεπεξεργασία δεδομένων, μπορούν να διαχειριστούν μεγάλα σύνολα δεδομένων σε σύντομο χρόνο και γενικά αποτελούν ένα γρήγορο ταξινομητή "ανοικτού τύπου" (white box model), καθώς είναι εμφανή τα χαρακτηριστικά στα οποία δίνεται η μεγαλύτερη βαρύτητα και το τελικό αποτέλεσμα είναι επαληθεύσιμο.

Η κατασκευή του βέλτιστου δένδρου αποφάσεων είναι πρόβλημα NP-Complete. Κατά συνέπεια, οι αλγόριθμοι μάθησης εφαρμόζουν ευριστικές μεθόδους. Για παράδειγμα, ο αλγόριθμος ID3 στηρίζεται στο κριτήριο της άπληστης επιλογής με υπολογισμό τοπικά βέλτιστων αποφάσεων σε κάθε κόμβο και χωρίς εγγύηση εξαγωγής του ολικά βέλτιστου δένδρου απόφασης. Ένα πρόβλημα των συγκεκριμένων αλγορίθμων είναι ότι για μικρές διακυμάνσεις στα δεδομένα παράγουν πολύ διαφορετικά δένδρα αποφάσεων.

3.2.2 Λογιστική Παλινδρόμηση

Η Λογιστική Παλινδρόμηση, Logistic Regression, ταξινομεί τα δεδομένα υπολογίζοντας για κάθε μία από τις δυνατές κλάσεις την εκ των υστέρων, posteriori, πιθανότητα το δεδομένο εισόδου να ανήκει στην εκάστοτε κλάση c_i δοθέντος του διανύσματος χαρακτηριστικών **x** και επιλέγει ως αναμενόμενη κλάση y^p εκείνη με τη μέγιστη πιθανότητα. Για τον υπολογισμό της πιθανότητας p_{c_i} που αντιστοιχεί σε κάθε κλάση εφαρμόζει στα δεδομένα εκπαίδευσης ένα γραμμικό μοντέλο f(**x**,**w**) όπου **w** το διάνυσμα συντελεστών ξεχωριστό για κάθε κλάση. Ωστόσο, η πιθανότητα p_{c_i} εξ ορισμού έχει πεδίο τιμών το $[0,1]$, ενώ οι γραμμικές συναρτήσεις είναι μη φραγμένες. Επομένως, για να αντιστοιχίζουμε την p_{c_i} σε μη φραγμένο πεδίο χρησιμοποιούμε το λογιστικό μετασχηματισμό:

$$logit(p_{c_i}) = log(\frac{p_{c_i}}{1 - p_{c_i}})$$

Στην περίπτωση των δύο κλάσεων, οι πιθανότητες p_{c_1} και p_{c_2} είναι συμπληρωματικές, καθώς $p_{c_1} + p_{c_2} = 1$, οπότε προκύπτει:

$$logit(p_{c_i}(x)) = log(\frac{p_{c_i}(x)}{1 - p_{c_i}(x)}) = \beta_0 + w_1x_1 + \ldots + w_nx_n = \beta_0 + wx$$

όπου β_0 ένας βαθμωτός όρος.
Επιλύοντας ως προς $p_{c_i}(x)$ λαμβάνουμε:

$$p_{c_i}(x) = \frac{1}{1 + e^{(1-(\beta_0 + wx))}}$$

Το σύνορο απόφασης που διαχωρίζει τις δύο κλάσεις προκύπτει από τη λύση της εξίσωσης $\beta_0 + w_1x_1 + \ldots + w_nx_n = \beta_0 + wx = 0$ και δεν απαιτείται ξεχωριστό διάνυσμα συντελεστών και βαθμωτός όρος για τη δεύτερη κλάση. Ο βαθμωτός όρος β_0 και το διάνυσμα συντελεστών w προσδιορίζονται αναζητώντας τις τιμές εκείνες που μεγιστοποιούν την πιθανότητα στο σύνολο εκπαίδευσης.

Στην περίπτωση όπου οι κλάσεις είναι περισσότερες από δύο, έστω $k > 2$, τότε προκύπτει η Πολυωνυμική Λογιστική Παλινδρόμηση, Multinomial Logistic Regression ή Maximum Entropy. Για κάθε κλάση $c \in C^k$ απαιτείται ξεχωριστός βαθμωτός όρος ${\beta_0}^{(c)}$ και αντίστοιχο διάνυσμα συντελεστών $w^{(c)}$ και οι υπό συνθήκη πιθανότητες υπολογίζονται από τη σχέση:

$$P(y = c|\bar{x} = x) = \frac{e^{({\beta_0}^{(c)} + w^{(c)}x)}}{\sum_c e^{({\beta_0}^{(c)} + w^{(c)}x)}}$$

3.2.3 Naive Bayes

Ο Naive Bayes είναι ένας πιθανοτικός ταξινομητής ο οποίος χρησιμοποιεί το Θεώρημα του Bayes για να υπολογίσει την εκ των υστέρων, posteriori, πιθανότητα $P(x\,|y)$ για κάθε κλάση c του C^k δοθέντος του διανύσματος χαρακτηριστικών:

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

Όμοια με την περίπτωση της Λογιστικής Παλινδρόμησης, η αναμενόμενη κλάση $y^p = c_i$ είναι εκείνη με τη βέλτιστη πιθανότητα $P(y = c_i|x)$. Ωστόσο, ο Naive Bayes δεν υπολογίζει απευθείας την εκ των υστέρων πιθανότητα, αλλά βασίζεται στην πιθανότητα $P(x\,|y)$. Η συγκεκριμένη πιθανότητα, γνωστή και ως likelihood, αναφέρεται στο πόσο πιθανό είναι να παραχθεί ένα δεδομένο με διάνυσμα χαρακτηριστικών x θεωρώντας δεδομένη την κλάση y όπου ανήκει.

Η πιθανότητα $P(x)$ είναι ανεξάρτητη της μεταβλητής y και σταθερή για όλες τις κλάσεις. Για την εκ των υστέρων πιθανότητα αληθεύει:

$$P(y|\,x) \sim P(x\mid y) \cdot P(y)$$

Η πιθανότητα εμφάνισης κάθε κλάσης $P(y)$ μπορεί να υπολογιστεί από το σύνολο δεδομένων. Ωστόσο, η $P(x\mid y)$ εξαρτάται από τη συνδυασμένη κατανομή πιθανότητας των x και y και ο υπολογισμός της είναι ιδιαίτερα απαιτητικός ακόμη και σε μικρά σύνολα δεδομένων, καθώς το □ είναι μία πολυδιάστατη τυχαία μεταβλητή.

Ο ταξινομητής Naive Bayes κάνει την παραδοχή ότι οποιεσδήποτε δύο μεταβλητές x_i, x_j με $i \neq j$είναι ανεξάρτητες μεταξύ τους δοθείσης της κλάσης y οπότε $P(x_i|x_j, y) = P(x_i|y)$ για κάθε ζεύγος $i, j \in [1, n]$. Με εφαρμογή του κανόνα της αλυσίδας και της παραδοχής, προκύπτει ότι:

$$P(x|y) = P(x_1, \ldots x_n|y) = \prod_{i=1}^{n} P(x_i|y)$$

Επομένως, για την εκ των υστέρων πιθανότητα αληθεύει ότι

$$P(y|x) \sim P(y) \prod_{i=1}^{n} P(x_i|y)$$

Η αναμενόμενη κλάση y^p είναι εκείνη που μεγιστοποιεί την ποσότητα $P(y)\prod_{i=1}^{n}P(x_i|y)$.

Με άλλα λόγια

$$y^p = argmax_c P(y=c)\prod_{i=1}^{n}P(x_i|y=c)$$

Ο Naive Bayes αναφέρεται γενικά στην υπό συνθήκη ανεξαρτησία των μεταβλητών - χαρακτηριστικών x_i αφήνοντας απροσδιόριστη την κατανομή πιθανότητάς τους. Αν θεωρήσουμε πως κάθε χαρακτηριστικό x_i ακολουθεί πολυωνυμική κατανομή, τότε προκύπτει ο Πολυωνυμικός Naive Bayes, Multinomial Naive Bayes, όπου :

$$P(x|y) = \frac{(\sum_i x_i)!}{(\prod_i x_i)!}\cdot\prod_{i=1}^{n}{p_i}^{(x_i)}$$

3.2.4 Πολυεπίπεδο Perceptron

Τα Τεχνητά Νευρωνικά Δίκτυα είναι μαθηματικά μοντέλα επεξεργασίας δεδομένων που αποτελούνται από ένα πλήθος τεχνητών νευρώνων οργανωμένων σε δομές παρόμοιες με αυτές των βιολογικών νευρωνικών δικτύων όπως ο ανθρώπινος εγκέφαλος.

Ένας από τους πιο διαδεδομένους τύπους τεχνητών νευρωνικών δικτύων είναι το πολυεπίπεδο νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης, multilayer feedforward network, ή πολυεπίπεδο perceptron, multilayer perceptron.

Στο πολυεπίπεδο perceptro οι τεχνητοί νευρώνες είναι οργανωμένοι σε μία σειρά από στρώματα ή επίπεδα, layers. Το πρώτο από αυτά τα επίπεδα ονομάζεται επίπεδο εισόδου, input layer, και χρησιμοποιείται για την εισαγωγή των δεδομένων. Τα στοιχεία του κατ'ουσίαν δεν είναι νευρώνες, καθώς δεν εκτελούν κάποιο υπολογισμό. Στη συνέχεια, ενδέχεται να ακολουθούν, προαιρετικά, ένα ή περισσότερα ενδιάμεσα ή κρυφά επίπεδα (hidden layers), ενώ στο τέλος υπάρχει το επίπεδο εξόδου(output layer).

Το συγκεκριμένο νευρωνικό δίκτυο ονομάζεται εμπρόσθιας τροφοδότησης, διότι επιτρέπονται συνδέσεις μόνο μεταξύ νευρώνων διαδοχικών στρωμάτων. Ως αποτέλεσμα, η ροή πληροφορίας είναι πρόσθια. Παράλληλα, τα στρώματα είναι πλήρως συνδεδεμένα, καθώς κάθε νευρώνας σε ένα επίπεδο συνδέεται με όλους τους νευρώνες του επόμενου επιπέδου.

Οι McCulloch and Pitts (1943) πρότειναν την ιδέα ενός τεχνητού νευρώνα j ο οποίος υπολογίζει μία συνάρτηση g ως σταθμισμένο άθροισμα των n εισόδων :

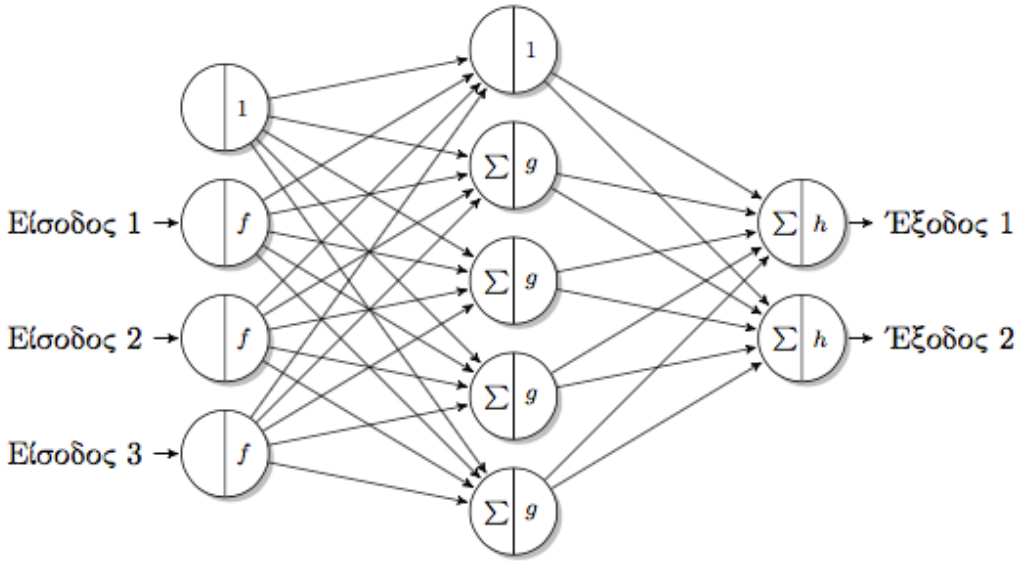
$$y_j(x) = g(\sum_{i=0}^{n} w_i x_i)$$

όπου $(w_1, w_2, \ldots, w_n)$ είναι οι συντελεστές βαρύτητας ή βάρη που εφαρμόζονται στις εισόδους $(x_1, x_2, \ldots, x_n)$

Σε ένα πολυεπίπεδο νευρωνικό δίκτυο, η έξοδος y_j δημιουργεί μέρος της εισόδου που θα δοθεί στους νευρώνες του επόμενου επιπέδου.

Η συνάρτηση ενεργοποίησης g είναι συνήθως μία από τις παρακάτω :

- βηματική συνάρτηση (step function) : ανάλογα με την τιμή του αθροίσματος και την παράμετρο κατωφλίου T_{thres} προκύπτει $y_j \in \{0,1\}$.
- συνάρτηση προσήμου (sign function) : ανάλογα με την τιμή του αθροίσματος και την παράμετρο κατωφλίου T_{thres} προκύπτει $y_j \in [-1,1]$.
- λογιστική συνάρτηση (logistic function) : ανάλογα με την τιμή του αθροίσματος και την μορφή της σιγμοειδούς συνάρτησης h_{sig} προκύπτει $y_j \in [0,1]$.



Σχήμα 3.5: Πολυεπίπεδο Perceptron εμπρόσθιας τροφοδότησης με ένα κρύφο επίπεδο και κόμβους πόλωσης

Το πολυεπίπεδο νευρωνικό δίκτυο Perceptron ταξινομεί τα δεδομένα υλοποιώντας μία συνάρτηση μεταφοράς Τ. Η συνάρτηση μεταφοράς Τ συνδέει την είσοδο, δηλαδή το διάνυσμα χαρακτηριστικών, με την έξοδο, δηλαδή την κλάση στην οποία ανήκει το εκάστοτε δεδομένο.

Κατά τη διάρκεια της εκπαίδευσης, μέσω της μεθόδου οπισθοδιάδοσης, backpropagation, οι παράμετροι των συναρτήσεων ενεργοποίησης, σε συνδυασμό με τους συντελεστές βαρύτητας των νευρώνων, αναπροσαρμόζονται επαναληπτικά έτσι, ώστε να βελτιστοποιηθεί η συνάρτηση μεταφοράς Τ.

Για να απλοποιηθεί η διαδικασία, προστίθενται κόμβοι πόλωσης, bias nodes, με σταθερή τιμή εξόδου 1 σε κάθε επίπεδο πλην της εξόδου έτσι, ώστε να αποπλεχθεί η έξοδος από τις παραμέτρους των συναρτήσεων ενεργοποίησης και η αναπροσαρμογή του δικτύου να εξαρτάται μόνο από την ενημέρωση των συντελεστών βαρύτητας.

Η έξοδος του ταξινομητή προκύπτει στο στάδιο εκτέλεσης όπου οι τιμές του διανύσματος χαρακτηριστικών διαδίδονται από το επίπεδο εισόδου σε όλο το υπόλοιπο δίκτυο.

Αποδεικνύεται πως ένα νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης με ένα κρυφό επίπεδο μπορεί να προσεγγίσει οποιαδήποτε μη γραμμική συνάρτηση.

3.2.5 Μηχανές Διανσμάτων Υποστήριξης

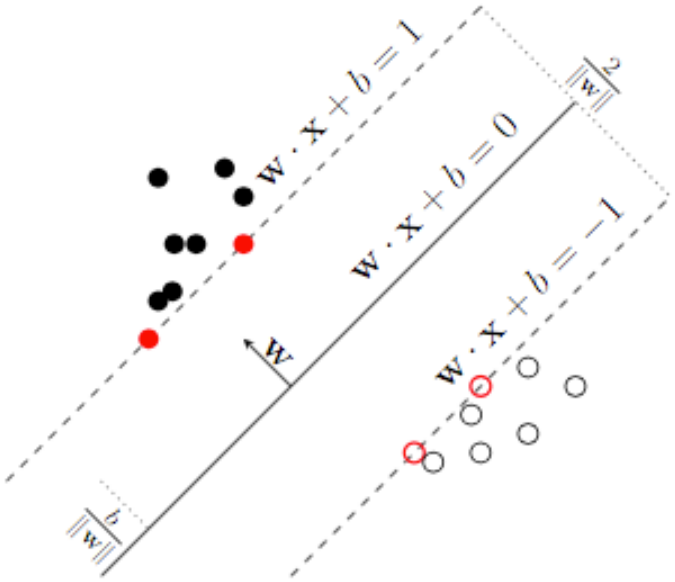
Οι Μηχανές Διανυσμάτων Υποστήριξης, Support Vector Machines ή SVMs, είναι μη πιθανοτικοί γραμμικοί δυαδικοί ταξινομητές. Προτάθηκαν το 1992 από τους Vapnik et al. ως μία νέα μέθοδος μάθησης, εντούτοις η γενικότερη ιδέα στην οποία στηρίζονται είχε προταθεί ήδη από τη δεκατία του 1960. Συνδυάζουν στοιχεία από τη Θεωρία Στατιστικής Μάθησης και τα Νευρωνικά Δίκτυα τύπου Perceptron.

Για τον ορισμό τους, ας θεωρήσουμε ένα σύνολο εκπαίδευσης D με n εκπαιδευτικά πρότυπα. Το σύνολο D έχει τη μορφή $D = \{(x_i, y_i) \mid x_i \in R^p, y_i \in \{-1,1\}, 1 \leq i \leq n\}$ και x_i είναι το διάνυσμα χαρακτηριστικών p-διαστάσεων του δείγματος i και y_i η κλάση στην οποία ανήκει.

Μία Μηχανή Διανυσμάτων Υποστήριξης προσπαθεί να προσδιορίσει το βέλτιστο υπερεπίπεδο, hyperplane, τέτοιο, ώστε $w \cdot x - b = 0$ όπου w είναι το κάθετο διάνυσμα του υπερεπιπέδου που ορίζεται από το χώρο χαρακτηριστικών R^p. Βέλτιστο είναι εκείνο το υπερεπίπεδο που μεγιστοποιεί

την απόσταση μεταξύ των αρνητικών και των θετικών δειγμάτων του συνόλου εκπαίδευσης, maximum margin hypersurface.

Θεωρούμε τη διαδικασία μάθησης ενός SVM ως ένα πρόβλημα βελτιστοποίησης. Υπό την προϋπόθεση ότι τα δεδομένα είναι γραμμικώς διαχωρίσιμα, τα δύο υπερεπίπεδα περιγράφονται από τις σχέσεις $w \cdot x - b = 1$ και $w \cdot x - b = -1$. Η μεταξύ τους απόσταση είναι ίση με $\frac{2}{||w||}$. Ελαχιστοποιώντας το $||w||$, λαμβάνουμε το βέλτιστο υπερεπίπεδο.



Σχήμα 3.6: Μηχανή Διανυσμάτων Υποστήριξης για δεδομένα δύο κλάσεων

Προσθέτοντας περιορισμούς έτσι, ώστε να μην απεικονίζονται δεδομένα στο διάστημα μεταξύ των υπερεπιπέδων και αξιοποιώντας τους πολλαπλασιαστές Lagrange λ_i, το πρόβλημα βελτιστοποίησης λαμβάνει την εξής μορφή :

$$L(w, b, \lambda_1, \ldots, \lambda_n) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \lambda_i (y_i(w \cdot x - b) - 1)$$

Η αντικεμενική συνάρτηση L πρέπει να ελαχιστοποιηθεί ως προς w και b και να μεγιστοποιηθεί ως προς λ_i.

Με εφαρμογή των συνθηκών Karush-Kuhn-Tucker, το βέλτιστο υπερεπίπεδο προκύπτει από τη σχέση:

$$g^*(x) = w \cdot x + b = \sum_{i=1}^{n} \lambda_i y_i x_i \cdot x + b$$

Μία Μηχανή Διανυσμάτων Υποστήριξης μπορεί να ταξινομεί περιπτώσεις που είναι παρόμοιες, αλλά όχι πανομοιότυπες, με κάποιο πρότυπο εκπαίδευσης. Τελικά, το αποτέλεσμα εξόδου είναι μία αριθμητική τιμή στο διάστημα [−1,1] και όχι κάποια πιθανότητα όπως σε άλλους ταξινομητές.

Για την επίλυση του γενικότερου προβλήματος της κατηγοριοποίησης δεδομένων σε περισσότερες από δύο κλάσεις έχει προταθεί η κατασκευή ενός συνόλου δυαδικών ταξινομητών SVM με δύο εκδοχές :

- one-versus-all : κάθε ταξινομητής διαχωρίζει ανάμεσα σε μία κλάση και όλες τις υπόλοιπες. Η τελική κλάση είναι αυτή με τη μεγαλύτερη τιμή εξόδου (winner-takes-all strategy).

- one-versus-one : κάθε ταξινομητής διαχωρίζει ανάμεσα σε ένα ζεύγος κλάσεων. Για την κατηγοριοποίηση των δεδομένων, πραγματοποιείται μία διαδικασία ψηφοφορίας όπου κάθε ταξινομητής αντιστοιχεί το εκάστοτε δεδομένο εισόδου σε μία από τις δύο κλάσεις. Η κλάση με τις περισσότερες ψήφους αναδεικνύεται νικήτρια (max-wins voting strategy).

Το κυρίαρχο πλεονέκτημα των Μηχανών Διανυσμάτων Υποστήριξης έναντι των Νευρωνικών Δικτύων τύπου Perceptron είναι η παραγωγή πιο σύνθετων υπερεπιφανειών. Κατά συνέπεια, υπερβαίνουν τα προβλήματα των τοπικών ελαχίστων και της διασποράς των λύσεων στο χώρο αναζήτησης. Ο λόγος είναι η ενσωμάτωση μετασχηματισμών και συνδυασμών των αρχικών μεταβλητών σύμφωνα με τις απαιτήσεις του εκάστοτε προβλήματος μέσω της χρήσης ενός πεπερασμένου αριθμού υποσυνόλων του συνόλου εκπαίδευσης (τα διανύσματα υποστήριξης) καθώς και συναρτήσεις πυρήνα (kernel functions) προκειμένου να μετασχηματίσουν τον αρχικό χώρο υποθέσεων και να βρουν τη βέλτιστη μη γραμμική υπερεπιφάνεια που ελαχιστοποιεί το σφάλμα αναζήτησης.[57]

3.2.6 κ-Κοντινότεροι Γείτονες

Ο ταξινομήτης των k-Κοντινότερων Γειτόνων, k-Nearest Neighbors ή k-NN, ανήκει στην κατηγορία των αλγορίθμων μηχανικής μάθησης κατά περίπτωση, instance-based learning. Σε αντίθεση με τις προαναφερθείσες μεθόδους που μετασχηματίζουν τα πρότυπα εκπαίδευσης σε συμπαγή δεδομένα, στη μάθηση κατά περίπτωση τα δεδομένα διατηρούνται αυτούσια.

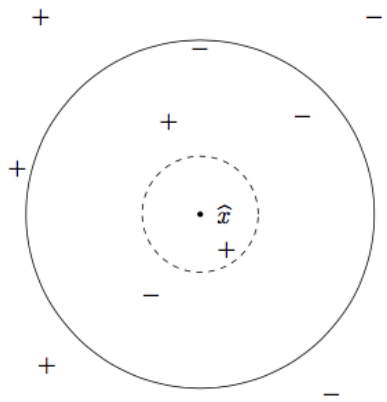
Όταν ένα σύστημα που εφαρμόζει μηχανική μάθηση κατά περίπτωση κληθεί να αποφανθεί για την κατηγορία στην οποία ανήκει ένα νέο δεδομένο εισόδου, το σύστημα εξετάζει εκείνη τη στιγμή τη σχέση του δεδομένου εισόδου με τα ήδη αποθηκευμένα παραδείγματα. Η μέθοδος που περιγράφουμε αναβάλλει τη μάθηση μέχρι τη στιγμή που θα εμφανιστεί κάποιο νέο στιγμιότυπο. Για το λόγο αυτό, ονομάζεται οκνηρή μάθηση, lazy learning, σε αντιδιαστολή με τις υπόλοιπες μεθόδους μάθησης που χαρακτηρίζονται ως πρόθυμες μέθοδοι μάθησης, eager learners. Αντιθέτως, οι πρόθυμες μέθοδοι μάθησης κατασκευάζουν άμεσα το μοντέλο μάθησης από το σύνολο εκπαίδευσης και χωρίς να περιμένουν για την άφιξη μίας νέας περίπτωσης. [57]

Στον αλγόριθμο των k-Κοντινότερων Γειτόνων γίνεται η παραδοχή ότι τα διάφορα πρότυπα δύνανται να αναπαρασταθούν ως σημεία ενός ν-διάστατου Ευκλείδειου χώρου R^v όπου ν είναι το πλήθος των χαρακτηριστικών εισόδου. Κάθε νέο στιγμιότυπο τοποθετείται στο ν-διάστατο Ευκλείδιο χώρο ως νέο σημείο. Η κλάση στην οποία ανήκει το νέο σημείο προσδιορίζεται λαμβάνοντας υπόψη την πλειοψηφία των αποφάσεων των k πλησιέστερων σημείων που προέρχονται από τα πρότυπα εκπαίδευσης του συνόλου εκπαίδευσης D. Στις περισσότερες περιπτώσεις, η μετρική που χρησιμοποιείται για τον καθορισμό των k κοντινότερων γείτονων ενός σημείου είναι η Ευκλείδεια απόσταση.

Λαμβάνοντας υπόψη τα παραπάνω, θεωρούμε ένα νέο στιγμιότυπο $\hat{\hat{x}}$ με σύνολο χαρακτηριστικών $\{a_1(\hat{x}), a_2(\hat{x}), \ldots, a_n(\hat{x})\}$ ως επίσης ένα αποθηκευμένο πρότυπο x με σύνολο χαρακτηριστικών $\{a_1(x), a_2(x), \ldots, a_n(x)\}$. Το τετράγωνο της απόστασης των στιγμιοτύπων $\hat{x}$ και x είναι ίσο με το άθροισμα των τετραγώνων της διαφοράς των χαρακτηριστικών $a_i(\hat{x})$ και $a_i(x)$, δηλαδή

$$distance(\hat{x}, x) = \sqrt{\sum_{i=1}^{v} (a_i(\hat{x}) - a_i(x))^2}$$

Στο σχήμα που ακολουθεί, αναπαρίστανται τα πρότυπα δύο κλάσεων, καθώς επίσης ένα νέο στιγμιότυπο $\hat{\hat{x}}$. Στην περίπτωση που ληφθεί υπόψη μόνο ο πλησιέστερος γείτονας, το νέο στιγμιότυπο χαρακτηρίζεται ως θετικό. Αντιθέτως, εάν ληφθούν υπόψη οι πέντε πλησιέστεροι γείτονες, τότε το νέο στιγμιότυπο χαρακτηρίζεται ως αρνητικό, επειδή οι τρεις από τους πέντε γείτονες έχουν αρνητικό πρόσημο.



Σχήμα 3.7: Προσδιορισμός κατηγορίας με βάση τον 1 και τους 5 κοντινότερους γείτονες

4

Διαδικτυακή Υπηρεσία Ανάλυσης Συναισθήματος

Ο σύγχρονος επιχειρηματικός κόσμος χρησιμοποιεί κυρίως δύο τεχνολογίες για τη διαχείριση δεδομένων, τις επιχειρησιακές βάσεις δεδομένων (operational databases) και τις αποθήκες δεδομένων (data warehouses). Οι επιχειρησιακές βάσεις δεδομένων εγγυώνται τη συνέπεια και τη συνοχή δεδομένων που υφίστανται τακτικές ενημερώσεις. Τα εχέγγυα συνέπειας που προσφέρουν είναι γνωστά ως ιδιότητες ACID (ACID properties) ή εγγυήσεις συναλλαγής (transactional guarantees). Οι επιχειρησιακές βάσεις δεδομένων είναι On Line Transactional Processing (OLTP) συστήματα ικανά να επεξεργαστούν σύντομες συναλλαγές, στην πλειοψηφία τους ενημερώσεις δεδομένων. Οι αποθήκες δεδομένων χρησιμοποιούνται για στατιστική ανάλυση επιχειρησιακών δεδομένων (business analytics). Διαθέτουν τη δυνατότητα online απάντησης ερωτήσεων που αφορούν δεδομένα μεγάλου όγκου. Για αυτό το λόγο χαρακτηρίζονται ως On Line Analytical Processing (OLAP) συστήματα.

Ωστόσο, τόσο τα OLTP όσο και τα OLAP συστήματα εμφανίζουν κακή επίδοση όταν καλούνται να εκτελέσουν το ένα την εργασία του άλλου. Ως εκ τούτου, οι επιχειρήσεις διατηρούν και τα δύο είδη βάσεων δεδομένων. Έτσι, υπάρχει η διαρκής ανάγκη μεταφοράς δεδομένων από τις επιχειρησιακές βάσεις δεδομένων στις αποθήκες δεδομένων. Η διαδικασία ονομάζεται Extract-Transform-Load (ETL) και ευθύνεται για το 80% του κόστους της στατιστικής ανάλυσης επιχειρησιακών δεδομένων.

Μία πλατφόρμα που επιχειρεί να λύσει το συγκεκριμένο πρόβλημα είναι η LeanBigData [59]. Η πλατφόρμα αναπτύχθηκε με κύριο στόχο τη μείωση του κόστους της στατιστικής ανάλυσης επιχειρησιακών δεδομένων αποτελώντας μία υπηρεσία διαχείρισης δεδομένων μεγάλου όγκου (big data) σε πραγματικό χρόνο που προσφέρει παράλληλα τις λειτουργικότητες OLTP και OLAP.

Η λειτουργία της πλατφόρμας LeanBigData έχει επικυρωθεί από διάφορες περιπτώσεις χρήσης (use cases) συμπεριλαμβανομένης της Ανάλυσης Συναισθήματος σε Κοινωνικά Δίκτυα. Η συγκεκριμένη περίπτωση χρήσης περιλαμβάνει μία συνιστώσα υπηρεσία Ανάλυσης Συναισθήματος σε πραγματικό χρόνο.

Ερευνητικά εγχειρήματα όπως η πλατφόρμα LeanBigData υποδεικνύουν την ανάγκη και τη σημασία ανάπτυξης αποδοτικών και αξιόπιστων διαδικτυακών υπηρεσιών που προσφέρουν Ανάλυση Συναισθήματος σε πραγματικό χρόνο. Το γεγονός αυτό αποτέλεσε την αφορμή ενασχόλησής μας με την ανάπτυξη μίας διαδικτυακής υπηρεσίας Ανάλυσης Συναισθήματος που παρουσιάζουμε στην παρούσα ενότητα.

4.1 Αρχιτεκτονική Representational State Transfer - REST

Η αρχιτεκτονική REST, Representational State Transfer, διαδικτυακών υπηρεσιών προσφέρει διαλειτουργικότητα μεταξύ υπολογιστικών συστημάτων στο Διαδίκτυο. Οι υπηρεσίες που είναι συμβατές με την αρχιτεκτονική REST επιτρέπουν στα συστήματα-πελάτες την πρόσβαση και τροποποίηση πόρων, συνήθως σε μορφή κειμένου, χρησιμοποιώντας ένα προκαθορισμένο σύνολο ακαταστασικών (stateless) λειτουργιών. Σύμφωνα με τον αρχικό ορισμό τους στο World Wide Web, οι διαδικτυακοί πόροι είναι έγγραφα ή αρχεία που προσδιορίζονται μέσω των URLs. Σήμερα, ο ορισμός των διαδικτυακών πόρων είναι πιο γενικός και αφηρημένος και περιλαμβάνει οποιαδήποτε οντότητα μπορεί να προσδιοριστεί, να ονομαστεί, να διευθυνσιοδοτηθεί ή να τροποποιηθεί με κάποιο τρόπο στο Διαδίκτυο.

Σε μία RESTful διαδικτυακή υπηρεσία, τα αιτήματα για κάποιο πόρο ακολουθούνται από μία απάντηση συνήθως σε μορφή XML, HTML ή JSON. Η απάντηση αναφέρει πιθανή αλλαγή που έχει γίνει στον αιτούμενο πόρο μέσω κάποιας λειτουργίας τροποποίησης. Επειδή η αρχιτεκτονική REST χρησιμοποιεί το πρωτόκολλο HTTP για τη μετάδοση και λήψη δεδομένων, οι λειτουργίες που υποστηρίζονται είναι οι προκαθορισμένες του HTTP, όπως οι GET, POST, PUT και DELETE.

Η αξιοποίηση του ακαταστασιακού πρωτοκόλλου και των προκαθορισμένων μεθόδων προσπέλασης πόρων προσφέρει στις υπηρεσίες REST υψηλή επίδοση, αξιοπιστία, δυνατότητα κλιμάκωσης και ενημέρωση του συστήματος κατά τη διάρκεια λειτουργίας των υπηρεσιών.

Πόροι

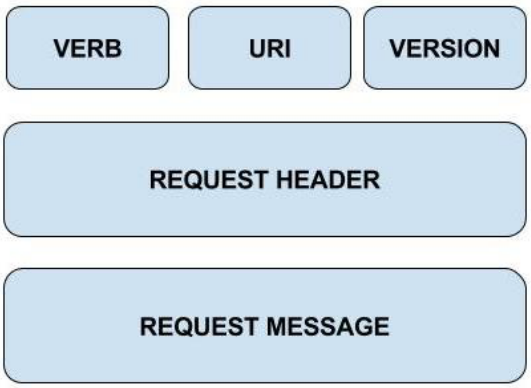
Στην αρχιτεκτονική REST η έννοια του πόρου είναι αντίστοιχη της έννοιας του αντικειμένου στον Αντικειμενοστραφή προγραμματισμό. Η αρχιτεκτονική REST αντιμετωπίζει οποιαδήποτε πληροφορία ως πόρο. Οι πόροι ενδέχεται να είναι αρχεία κειμένου, σελίδες HTML, εικόνες, βίντεο ή δυναμικά επιχειρησιακά δεδομένα. Κάθε πόρος προσδιορίζεται κατά μοναδικό τρόπο με URIs. Οι

πόροι κωδικοποιούνται συνήθως ως Text, JSON ή XML. Κάθε εξυπηρετητής παρέχει πρόσβαση στους πόρους, ενώ κάθε πελάτης ζητάει, προσπελαύνει και τροποποιεί ή αναπαριστά τους πόρους.

Μηνύματα

Η αρχιτεκτονική REST χρησιμοποιεί το πρωτόκολλο HTTP για την επικοινωνία μεταξύ πελάτη και εξυπηρετητή. Ο πελάτης στέλνει ένα μήνυμα ως αίτηση HTTP (HTTP Request) και ο εξυπηρετητής αποκρίνεται με μία απάντηση HTTP (HTTP Responce). Τα μηνύματα περιέχουν δεδομένα και μεταδεδομένα, δηλαδή πληροφορίες για το ίδιο το μήνυμα.

Αίτηση HTTP

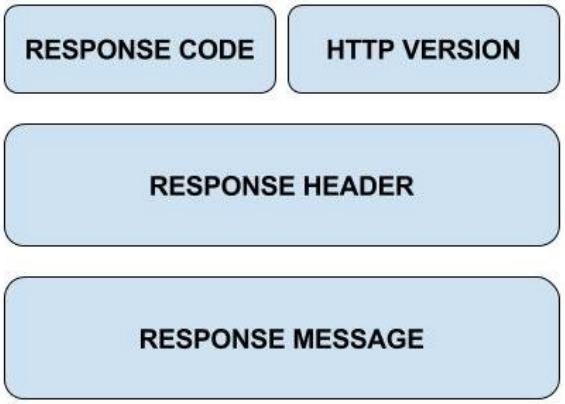


Σχήμα 4.1: Δομή αίτησης HTTP

Μία αίτηση HTTP έχει 5 μέρη:

- Verb: Δηλώνει την HTTP μέθοδο που θα χρησιμοποιηθεί, δηλαδή GET, POST, DELETE, PUT κλπ.
- URI: Προσδιορίζει τον πόρο στον εξυπηρετητή.
- Version: Προσδιορίζει την έκδοση του HTTP, για παράδειγμα HTTP v2.0.
- Request Header: Περιέχει μεταδεδομένα της αίτησης HTTP στη μορφή ζευγών κλειδιού - τιμής.
- Request Message: Το περιεχόμενο της αίτησης ή την αναπαράσταση του πόρου.

Απόκριση HTTP



Σχήμα 4.2: Δομή απόκρισης HTTP

Μία απόκριση HTTP έχει 4 μέρη:

- Response Code: Δηλώνει τη διαθεσιμότητα του πόρου που ζητήθηκε.
- HTTP Version: Προσδιορίζει την έκδοση του HTTP, για παράδειγμα HTTP v2.0.
- Response Header: Περιέχει μεταδεδομένα της απόκρισης HTTP στη μορφή ζευγών κλειδιού - τιμής. Για παράδειγμα, το μήκος περιεχομένου, το είδος του περιεχομένου και την ημερομηνία της απόκρισης.
- Response Message: Το περιεχόμενο της απόκρισης ή την αναπαράσταση του πόρου.

Διευθυνσιοδότηση

Στην αρχιτεκτονική REST, κάθε πόρος προσδιορίζεται με ένα URI της μορφής

<protocol>://<service-name>/<ResourceType>/<ResourceID>

Μέθοδοι

Οι HTTP μέθοδοι επικοινωνίας που χρησιμοποιούνται στην αρχιτεκτονική REST είναι οι εξής:

- GET - για ανάγνωση μόνο ενός πόρου,
- PUT - για δημιουργία ενός πόρου,
- DELETE - για διαγραφή ενός πόρου,
- POST - για ενημέρωση ενός υπάρχοντος πόρου ή δημιουργία νέου,
- OPTIONS - για την ανάγνωση των υποστηριζόμενων λειτουργιών σε ένα πόρο.

Ακαταστασικότητα

Σύμφωνα με την αρχιτεκτονική REST, η υπηρεσία δεν πρέπει να διατηρεί στοιχεία της κατάστασης του πελάτη στον εξυπηρετητή. Είναι ευθύνη του πελάτη να προσδιορίζει τα στοιχεία του στον εξυπηρετητή. Ακολούθως, ο εξυπηρετητής μπορεί να αποθηκεύει τα στοιχεία έτσι, ώστε να επεξεργάζεται το αίτημα του πελάτη. Για παράδειγμα, ο πελάτης προσδιορίζει το αναγνωριστικό της συνεδρίας στον εξυπηρετητή. Η συγκεκριμένη ιδιότητα της αρχιτεκτονικής ονομάζεται ακαταστασικότητα, statelessness.

Βασικά πλεονεκτήματα της ακαταστασικότητας είναι:

- Η ανεξαρτησία χειρισμού εκ μέρους του εξυπηρετητή κάθε αίτησης.
- Η απλοποίηση του σχεδιασμού της εφαρμογής.
- Η συμβατότητα των RESTful υπηρεσιών με το επίσης ακαταστασικό πρωτόκολλο HTTP.

Μειονέκτημα της ακαταστασικότητας είναι η ανάγκη για μεταφορά επιπλέον πληροφοριών από τον πελάτη στον εξυπηρετητή έτσι, ώστε το περιεχόμενο να είναι συνάρτηση της ταυτότητας του χρήστη.

Προσωρινή Αποθήκευση

Προσωρινή Αποθήκευση ή Caching είναι η αποθήκευση της απόκρισης του εξυπηρετητή στον πελάτη έτσι, ώστε να αποφευχθούν πολλαπλά αιτήματα από ένα πελάτη για τον ίδιο πόρο. Η απόκριση του εξυπηρετητή πρέπει να περιέχει οδηγίες για την υλοποίηση του caching προκειμένου ο πελάτης να γνωρίζει εάν και για πόσο χρόνο μπορεί να αποθηκεύσει το περιεχόμενο της απόκρισης. Οι επικεφαλίδες της απόκρισης που περιέχουν τις σχετικές οδηγίες μπορεί να είναι οι εξής:

- Ημερομηνία της δημιουργίας του πόρου.
- Ημερομηνία τελευταίας τροποποίησης του πόρου.
- Cache-control που υποδηλώνει εάν, πως και από ποιους επιτρέπεται το caching.
    - Ιδιωτικό: μόνο κάθε πελάτης μπορεί να κάνει caching.
    - Δημόσιο: κάθε ενδιάμεσος μπορεί να κάνει caching.
    - no-cache/no-store: δεν επιτρέπεται το caching.
    - Μέγιστη διάρκεια: το χρονικό διάστημα που μπορεί να διατηρηθεί ο πόρος cached. Αφού παρέλθει το διάστημα, πρέπει να πραγματοποιηθεί εκ νέου αίτηση στον εξυπηρετητή.
    - Must-revalidate: Υπόδειξη στον εξυπηρετητή ότι πρέπει να επανεπικυρώσει τον πόρο εφόσον έχει παρέλθει η μέγιστη διάρκεια του caching.
- Ημερομηνία λήξης του caching.
- Χρονικό διάστημα από τη στιγμή που ο πόρος ανακτήθηκε από τον εξυπηρετητή.

Ασφάλεια

Οι RESTful διαδικτυακές υπηρεσίες λειτουργούν με HTTP URL Paths. Επομένως, είναι πολύ σημαντικό οι υπηρεσίες να προστατεύονται με τον ίδιο τρόπο που προστατεύεται ένας ιστότοπος (website). Κάποιες βέλτιστες πρακτικές για την επίτευξη ενός στοιχειώδους επιπέδου ασφάλειας είναι οι ακόλουθες:

- Επικύρωση: Πρέπει να γίνεται επικύρωση όλων των αιτημάτων στον εξυπηρετητή έτσι, ώστε να αποτρέπονται επιθέσεις τύπου έγχυσης (injection) SQL ή NoSQL.
- Πιστοποίηση ανά συνεδρία: Πρέπει να γίνεται πιστοποίηση του χρήστη για κάθε συνεδρία όταν υπάρχει αίτηση για κάποια μέθοδο της διαδικτυακής υπηρεσίας.
- Μη ενσωμάτωση ευαίσθητων δεδομένων στο URL: Σε καμία περίπτωση δεν πρέπει να αποτυπώνεται το όνομα του χρήστη, το συνθηματικό εισόδου ή το αναγνωριστικό συνεδρίας στο URL. Οι όποιες ευαίσθητες πληροφορίες πρέπει να μεταδίδονται κρυπτογραφημένες στη διαδικτυακή υπηρεσία μέσω της μεθόδου POST.
- Περιορισμοί στην εκτέλεση μεθόδων: Μέθοδοι όπως οι GET, POST και DELETE πρέπει να εκτελούνται εφόσον ο αιτών διαθέτει τα κατάλληλα δικαιώματα. Σε καμία περίπτωση δεν πρέπει η μέθοδος POST να μπορεί να διαγράψει δεδομένα.
- Έλεγχος παραμορφωμένων XML ή JSON αρχείων: Κάθε είσοδος της μορφής XML ή JSON στην υπηρεσία πρέπει να έχει ορθή συντακτική μορφή. Επομένως, πρέπει πάντα να γίνεται ο κατάλληλος έλεγχος των XML ή JSON αρχείων που δίνονται ως είσοδοι.
- Εμφάνιση κατάλληλων κωδικών κατάστασης: Μία διαδικτυακή υπηρεσία οφείλει να απαντάει με τον εκάστοτε κατάλληλο κωδικό κατάστασης κατόπιν αποδοχής μίας αίτησης από κάποιο πελάτη. Για λόγους πληρότητας, παραθέτουμε τους κωδικούς κατάστασης που οφείλει να παρέχει η υπηρεσία στον ακόλουθο πίνακα.

#	Κωδικός HTTP	Περιγραφή
1	200	OK − Επιτυχία.
2	201	CREATED − Στην περίπτωση που ο πόρος δημιουργήθηκε επιτυχώς με αίτημα POST ή PUT. Επιστρέφει το σύνδεσμο στο νέο πόρο.
3	204	NO CONTENT − Όταν το σώμα της απόκρισης είναι κενό. Για παράδειγμα, κατόπιν μίας αίτησης DELETE.
4	304	NOT MODIFIED − Χρησιμοποιείται για τον περιορισμό εύρους ζώνης σε υπό συνθήκη αιτήματα GET. Το σώμα της απάντησης πρέπει να είναι κενό. Οι κεφαλίδες πρέπει να περιέχουν στοιχεία ημερομηνίας, τοποθεσίας κλπ.
5	400	BAD REQUEST − Δηλώνει ότι η είσοδος είναι άκυρη. Για παράδειγμα, σφάλμα επικύρωσης δεδομένων.
6	401	UNAUTHORIZED − Δηλώνει ότι ο χρήστης εισάγει άκυρα ή λάθος στοιχεία πιστοποίησης.
7	403	FORBIDDEN − Δηλώνει ότι ο χρήστης δεν έχει πρόσβαση στη μέθοδο. Για παράδειγμα, αίτημα DELETE χωρίς δικαιώματα διαχειριστή.
8	404	NOT FOUND − Δηλώνει ότι η μέθοδος δεν είναι διαθέσιμη.
9	409	CONFLICT − Δηλώνειώ σύγκρουση κατά την εκτέλεση μεθόδου. Για παράδειγμα, προσθήκη διπλής εγγραφής.
10	500	INTERNAL SERVER ERROR − Δηλώνει ότι ο εξυπηρετητής έδωσε εξαίρεση κατά την εκτέλεση κάποιας μεθόδου.

Πίνακας 4.1: Κωδικοί κατάστασης HTTP διαδικτυακής υπηρεσία αρχιτεκτονικής REST

Java (JAX-RS)

Η JAX-RS είναι μία προγραμματιστική διεπιφάνεια Java που χρησιμοποιεί τις παρακάτω επισημειώσεις προκειμένου να απλοποιήσει την ανάπτυξη διαδικτυακών υπηρεσιών.

#	Επισημείωση	Περιγραφή
1	@Path	Σχετικό μονοπάτι της κλάσης ή της μεθόδου.
2	@GET	HTTP Get αίτηση για την απόκτηση κάποιου πόρου.
3	@PUT	HTTP Put αίτηση για τη δημιουργία ενός πόρου.
4	@POST	HTTP Post αίτηση για τη δημιουργία ή την ενημέρωση ενός πόρου.
5	@DELETE	HTTP Delete αίτηση για τη διαγραφή ενός πόρου.
6	@HEAD	HTTP Head αίτηση για τον έλεγχο διαθεσιμότητας μεθόδου.
7	@Produces	Δηλώνει τί απόκριση HTTP παράγει η διαδικτυακή υπηρεσία. Για παράδειγμα, APPLICATION/XML, TEXT/HTML, APPLICATION/JSON etc.
8	@Consumes	Δηλώνει το είδος του αιτήματος της αίτησης HTTP Request. Για παράδειγμα, application/x-www-formurlencoded για την αποδοχή δεδομένων από το σώμα τηε αίτησης HTTP με τη μέθοδο POST.
9	@PathParam	Ενσωματώνει την παράμετρο που περνάει στη μέθοδο σε μία τιμή στο μονοπάτι.
10	@QueryParam	Ενσωματώνει την παράμετρο που περνάει στη μέθοδο σε μία παράμετρο ερωτήματος στο μονοπάτι.
11	@MatrixParam	Ενσωματώνει την παράμετρο που περνάει στη μέθοδο σε μία παράμετρο τύπου πίνακα HTTP στο μονοπάτι.
12	@HeaderParam	Ενσωματώνει την παράμετρο που περνάει στη μέθοδο σε μία κεφαλίδα HTTP.
13	@CookieParam	Ενσωματώνει την παράμετρο που περνάει στη μέθοδο σε ένα Cookie.
14	@FormParam	Ενσωματώνει την παράμετρο που περνάει στη μέθοδο σε μία φόρμα.
15	@DefaultValue	Αναθέτει μία προκαθορισμένη τιμή στην παράμετρο που περνάει στη μέθοδο.
16	@Context	Το περιεχόμενο του πόρου. Για παράδειγμα, αίτηση HTTP ως περιεχόμενο.

Πίνακας 4.2: Επισημειώσεις της προγραμματιστικής διεπιφάνειας JAX-RS

4.2 Εκπαίδευση και αξιολόγηση των ταξινομητών

Η διαδικτυακή υπηρεσία που αναπτύξαμε αποσκοπεί στην ανάλυση πολυγλωσσικών συλλογών μηνυμάτων από Κοινωνικά Δίκτυα. Όπως έχουμε ήδη αναφέρει στην αρχή του παρόντος κεφαλαίου, αυτός είναι ο λόγος που υιοθετείται η προσέγγιση των Aisopos et al. [3] για την Ανάλυση Συναισθήματος με τη βοήθεια γράφων ν-γραμμάτων, όπως περιγράψαμε στο Κεφάλαιο 3.

Η μέθοδος των γράφων ν-γραμμάτων υλοποιείται με τη βοήθεια του λογισμικού εξόρυξης δεδομένων Weka για την εκπαίδευση των ταξινομητών C4.5 tree, Logistic Regression LReg, Naive Bayes NB, Naive Bayes Multinomial NBM, Multilayer Perceptron MLP, Support Vector Machines SVM και k-Nearest Neighbours k-NN.

Στο σημείο αυτό αναφέρουμε ότι τα αποτελέσματα, οι πίνακες, τα σχήματα και τα συμπεράσματα που αναφέρονται στο υπόλοιπο της παρούσας υποενότητας βρίσκονται στη σχετική εργασία του Δ. Τζαννέτου [58] και παρατίθενται για λόγους σύνοψης.

Η αξιολόγηση της ακρίβειας των ταξινομητών πραγματοποιείται με κριτήριο την ευρέως διαδεδομένη μετρική απόδοσης ακρίβεια κατηγοριοποίησης □ που ορίζεται ως εξής:

$$a = \frac{\sum_{c \in C} |ορθώς\ ταξινομημένα\ δεδομένα\ c|}{|δεδομένα\ εξέτασης|}$$

όπου $C = \{c_{neg}, c_{pos}, c_{neu}\}$ είναι το σύνολο των κατηγοριών πολικότητας.

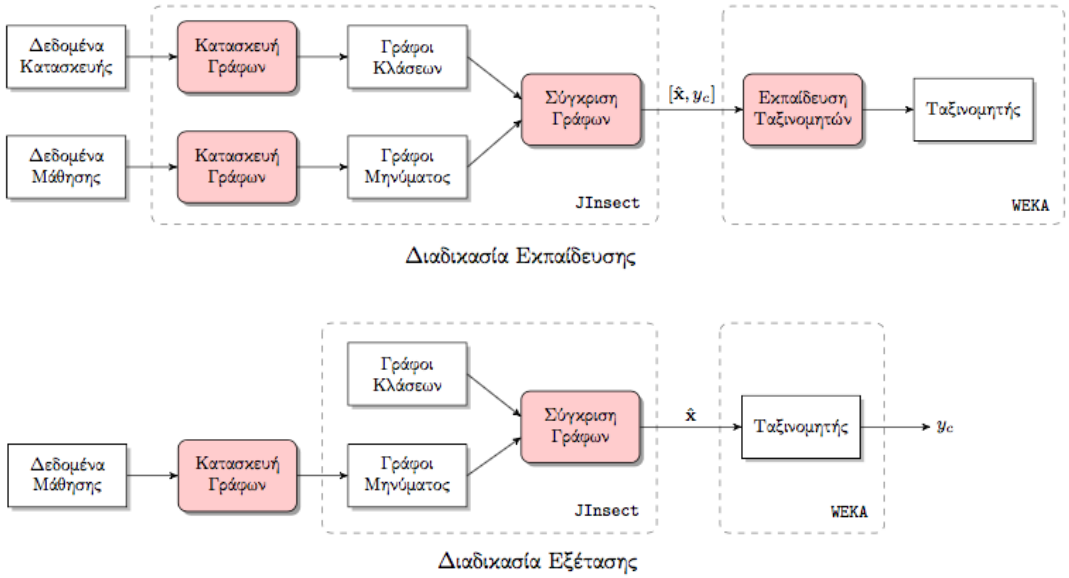
Η αξιολόγηση της διακριτικής ικανότητας των ταξινομητών ως προς τις κατηγορίες πραγματοποιείται με τους δείκτες ακρίβειας (precision), ανάκλησης (recall) και του αρμονικού τους μέσου όρου □$_1$ που ορίζονται ως εξής:

$$Precision = \frac{\sum_{c \in C} \frac{|ορθώς\ ταξινομημένα\ δεδομένα\ ως\ c|}{|δεδομένα\ που\ ταξινομήθηκαν\ ως\ c|}}{|C|}$$

$$Recall = \frac{\sum_{c \in C} \frac{|ορθώς\ ταξινομημένα\ δεδομένα\ ως\ c|}{|δεδομένα\ που\ ανήκουν\ στην\ κατηγορία\ c|}}{|C|}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Για κατασκευή και την αξιολόγηση της επίδοσης των ταξινομητών, το λογισμικό ακολουθεί διαδοχικά τις διαδικασίες εκπαίδευσης και εξέτασης. Η ροή εκτέλεσης περιγράφεται στα ακόλουθα σχήματα.



Σχήμα 4.3: Λειτουργικότητες εκπαίδευσης και εξέτασης

Τα στάδια είναι τα εξής:

I. Κατασκευή Γράφων Κλάσης Πολικότητας: Κατασκευάζονται οι γράφοι συναισθήματος από τα μηνύματα (tweets) της αντίστοιχης κλάσης των δεδομένων κατασκευής. Κάθε μήνυμα διαθέτει μία επισήμανση για την κατηγορία συναισθήματος στην οποία υπάγεται.

II. Εκπαίδευση Ταξινομητών: Για κάθε μήνυμα (tweet) του συνόλου μάθησης δημιουργείται ο γράφος μηνύματος. Κάθε γράφος μηνύματος συγκρίνεται με κάθε γράφο κλάσης πολικότητας και υπολογίζονται οι δείκτες ομοιότητας. Οι δείκτες ομοιότητας κάθε μηνύματος και η κλάση που ανήκει δίνονται ως είσοδοι στο βήμα εκπαίδευσης των ταξινομητών.

III. Εξέταση Ταξινομητών: Για κάθε μήνυμα (tweet) του συνόλου εξέτασης δημιουργείται ο γράφος μηνύματος. Κάθε γράφος μηνύματος συγκρίνεται με κάθε γράφο κλάσης πολικότητας και υπολογίζονται οι δείκτες ομοιότητας. Οι δείκτες ομοιότητας κάθε μηνύματος και η κλάση που ανήκει δίνονται ως είσοδοι στο βήμα εξέτασης των ταξινομητών. Οι εκπαιδευμένοι ταξινομητές προβλέπουν την πολικότητα κάθε μηνύματος. Μία ταξινόμηση θεωρείται επιτυχής στην περίπτωση που η προβλεπόμενη πολικότητα συμπίπτει με την πραγματική. Διαφορετικά, η ταξινόμηση θεωρείται ανεπιτυχής.

Η κατασκευή των γράφων και ο υπολογισμός των δεικτών ομοιότητας πραγματοποιείται με τις μεθόδους που προσφέρει η βιβλιοθήκη JInsect, ενώ οι αλγόριθμοι κατηγοριοποίησης υλοποιούνται στη βιβλιοθήκη Weka.

Οι κλάσεις ταξινομητών της βιβλιοθήκης Weka που χρησιμοποιήθηκαν είναι:

I. J48: Υλοποίηση του αλγορίθμου C4.5 με ενεργοποιημένη τη λειτουργία κλαδέματος (pruning).
II. Logistic: Η πολυωνυμική εκδοχή της Λογιστικής Παλινδρόμησης που ονομάζεται Maximum Entropy.
III. NaiveBayes: Ο αλγόριθμος Naive Bayes χωρίς τη δυνατότητας ενημέρωσης.
IV. NaiveBayesMultinomial: Η πολυωνυμική εκδοχή του αλγορίθμου Naive Bayes.
V. Multilayer Perceptron: Το μοντέλο του Πολυεπίπεδου Perceptron με αριθμό κρυφών επιπέδων ίσο με $\frac{1}{2} \cdot (\sigma\acute{\upsilon}\nu o\lambda o\ \chi\alpha\rho\alpha\kappa\tau\eta\rho\iota\sigma\tau\iota\kappa\acute{\omega}\nu + \sigma\acute{\upsilon}\nu o\lambda o\ \kappa\lambda\acute{\alpha}\sigma\epsilon\omega\nu) = 6$.Η συγκεκριμένη τιμή είναι η προκαθορισμένη.
VI. SMO: Η υλοποίηση της Μηχανής Διανυσμάτων Υποστήριξης με χρήση της πολυωνυμικής συνάρτησης πυρήνα PolyKernel και στρατηγική 1-vs-1.
VII. IBk: Η υλοποίηση του αλγορίθμου k-Κοντινότερων γειτόνων με $k = 3$.

Το σύνολο των μηνυμάτων που χρησιμοποιείται για την εκπαίδευση και την αξιολόγηση των αλγορίθμων κατηγοριοποίησης συνίσταται από σύνολα δεδομένων ταξινομημένων με χειροκίνητο τρόπο. Τα ταξινομημένα δεδομένα είναι διαθέσιμα ελεύθερα στην ερευνητική κοινότητα και έχουν χρησιμοποιηθεί σε πλήθος σχετικών ερευνητικών εργασιών. Η χρήση δεδομένων ταξινομημένων με χειροκίνητο τρόπο πλεονεκτεί έναντι άλλων μεθόδων σε ακρίβεια, αλλά περιορίζει σημαντικά τον όγκο των εξεταζόμενων δεδομένων.

Η Πολιτική Χρήσης και Προστασίας του Προσωπικού Απορρήτου του Twitter δεν επιτρέπει τη δημοσίευση του περιεχομένου των μηνυμάτων, αλλά μόνο των αναγνωριστικών (tweetIDs) των μηνυμάτων. Ωστόσο, τα αναγνωριστικά μπορούν να χρησιμοποιηθούν για να ανακτηθεί το αρχικό κείμενο μέσω του TwitterAPI. Επειδή όμως αρκετοί από τους χρήστες - δημιουργούς των μηνυμάτων επιλέγουν να διαγράψουν ή να κλειδώσουν το λογαριασμό τους κατόπιν της εξόρυξης των αναγνωριστικών, μέρος της συλλογής των δεδομένων ενδέχεται να μην είναι πλέον ανακτήσιμο.

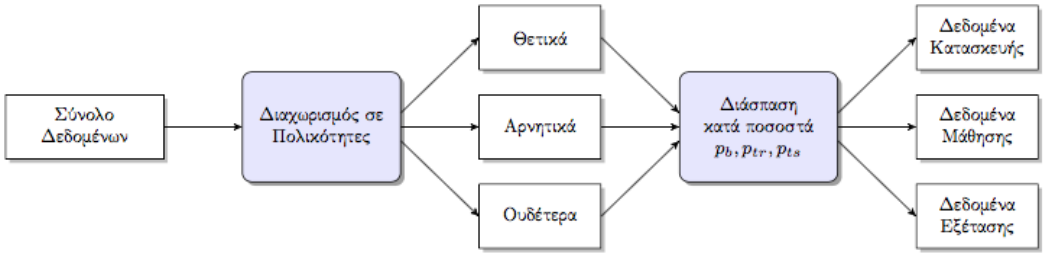
Βασικές πληροφορίες για τα σύνολα δεδομένων που χρησιμοποιήσαμε παρατίθενται στον ακόλουθο πίνακα.

Dataset	Δημιουργός	Γλώσσα	Πλήθος	Θετικά	Αρνητικά	Ουδέτερα
Arabic Twitter Corpus (RRArabic)	Refaee et al.[37]	Αραβικά	5821	767	1670	3384
HealthCare Reform (HCR)	Speriosu et al.[46]	Αγγλικά	2392	541	1381	470
Obama-McCain Debate (OMD)	Shamma et al.[45]	Αγγλικά	1904	709	1195	-
Multi-DaiLabor (MDL)	Narr et al.[28]	Αγγλικά	10594	2334	1486	6774
		Γαλλικά	2155	500	481	1174
		Γερμανικά	2637	496	334	1807
		Πορτογαλικά	2395	923	627	845
Manual Groundtruth (NTUA)	Aisopos et al.	Αγγλικά	500	159	119	222
Sanders Twitter	Sanders	Αγγλικά	3152	473	513	2166
SemEval-2014 Task (SEM)	Rosenthal et al.[38]	Αγγλικά	12838	4855	1986	5997
SentiTuites-PT	Moreira et al.	Πορτογαλικά	10075	1290	5413	3372
Sentiment Strength (SSTweet)	Thelwall et al.[48]	Αγγλικά	4242	1252	1037	1953
Tromp MultiLingual (Tromp)	Tromp [50]	Αγγλικά	12128	4885	3691	3552
		Ολλανδικά	5086	1277	1601	2208
TASS 2013 Corpus (TASS)	Villena et al.[52]	Ισπανικά	45482	24589	18161	2732
Σύνολο			**121401**	**45050**	**39695**	**36656**

Πίνακας 4.3: Σύνολα χειροκίνητα ταξινομημένων δεδομένων

Για την ανάκτηση του περιεχομένου σε σύνολα δεδομένων που δεν περιείχαν το αρχικό κείμενο χρησιμοποιήθηκε η βιβλιοθήκη twitter4j της JAVA.

Για τη λειτουργία του συστήματος Ανάλυσης Συναισθήματος απαιτείται η διαμέριση του αρχικού συνόλου δεδομένων σε δεδομένα κατασκευής, μάθησης και εξέτασης. Για τη δημιουργία της διαμέρισης, αρχικά αναδιατάσουμε με τυχαίο τρόπο τα μηνύματα του αρχικού συνόλου δεδομένων. Κατόπιν, επιλέγουμε πόσα μηνύματα θα εντάξουμε σε κάθε ένα από τα σύνολα κατασκευής, μάθησης και εξέτασης και τελικά δημιουργούμε τη διαμέριση. Η διαδικασία της διαμέρισης σκιαγραφείται στο ακόλουθο σχήμα.



Σχήμα 4.4: Διαδικασία διάσπασης συνόλου δεδομένων

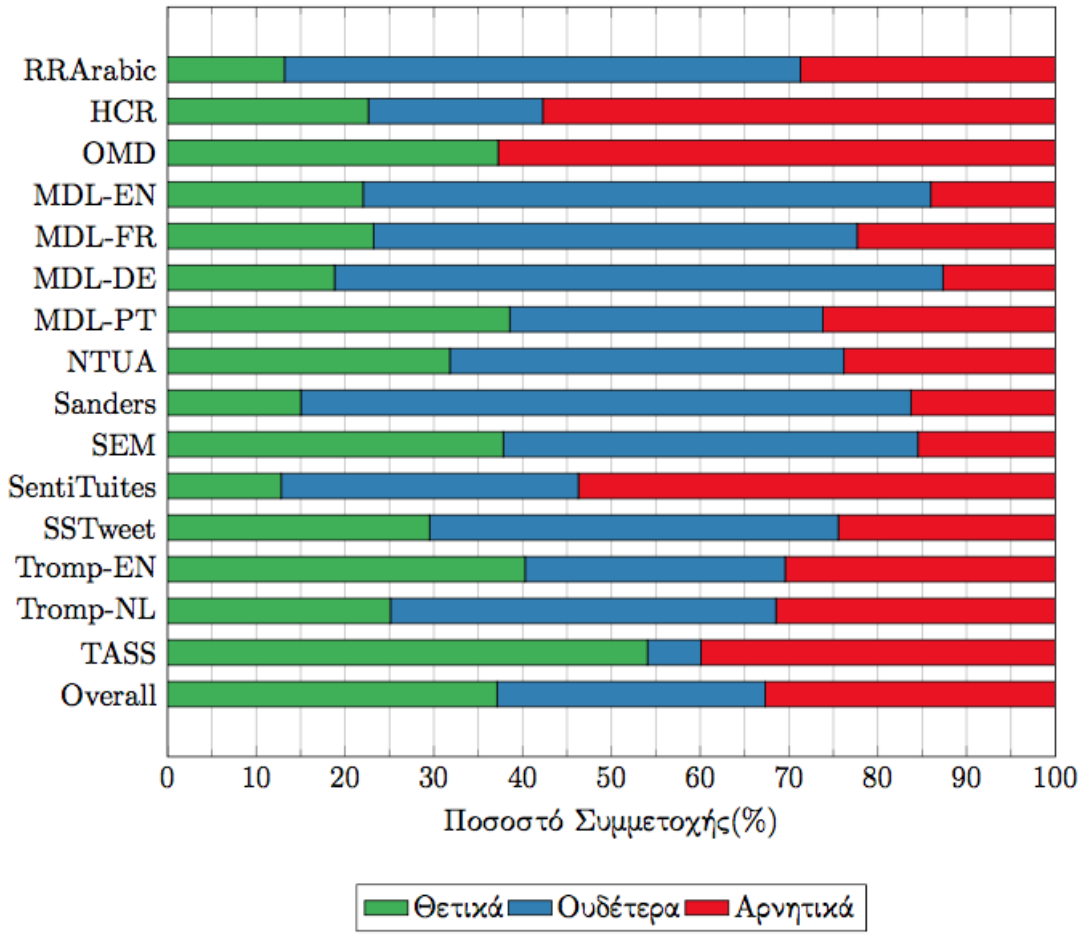
Η επίδοση ενός συστήματος επιβλεπόμενης μάθησης στο σύνολο εξέτασης εξαρτάται από την αντιπροσωπευτικότητα των δεδομένων που χρησιμοποιούνται κατά τη διαδικασία της εκπαίδευσης, στην περίπτωσή μας στις διαδικασίες κατασκευής και μάθησης.

Μέσω της διαδικασίας διαμέρισης και της κατάλληλης επιλογής των ποσοστών κατασκευής p_b, εκπαίδευσης p_{tr} και ελέγχου p_{ts}, κάθε σύνολο δεδομένων συμμετέχει αναλογικά σε κάθε στάδιο λειτουργίας του συστήματος ενώ παράλληλα διατηρείται η κατανομή της πολικότητας του συναισθήματος με τρόπο που εξασφαλίζει την αντιπροσωπευτικότητα του συνόλου εκπαίδευσης. Ως αποτέλεσμα, αποκλείεται η περίπτωση υπερεξειδίκευσης του συστήματος σε μία γλώσσα ή κατηγορία πολικότητας που δε συναντάται σε τόσο μεγάλο βαθμό στο σύνολο εξέτασης.

Το ερώτημα που θα απαντήσουμε στη συνέχεια είναι το κριτήριο επιλογής των ποσοστών κατασκευής p_b, εκπαίδευσης p_{tr} και ελέγχου p_{ts}.

Για να απαντήσουμε το εν λόγω ερώτημα, πρέπει να προχωρήσουμε σε κάποιες διαπιστώσεις αναφορικά με τη σύνθεση των επιμέρους συνόλων μηνυμάτων, καθώς και του ενοποιημένου συνόλου δεδομένων. Δεν παραθέτουμε τους υπολογισμούς που αποδίδουν τους αριθμούς που αναφέρουμε στην αμέσως επόμενη παράγραφο. Εντούτοις, παραθέτουμε κάποια διαγράμματα που συνοψίζουν με ευπαρουσίαστο τρόπο κάποιες διαπιστώσεις.

Στο ενοποιημένο σύνολο δεδομένων οι κλάσεις πολικότητας είναι σχετικά ισοκατανεμημένες. Συγκεκριμένα, στη θετική πολικότητα ανήκει το 37.1% των μηνυμάτων, στην αρνητική το 32.7% και στην ουδέτερη το 30.2%. Ωστόσο, παρατηρούμε ότι σε κάθε υποσύνολο που συνθέτει το ενοποιημένο σύνολο δεδομένων κυριαρχεί σε ποσοστό μία κλάση πολικότητας. Η συγκεκριμένη διαπίστωση αναπαρίσταται στο διάγραμμα ποσοστών που ακολουθεί.

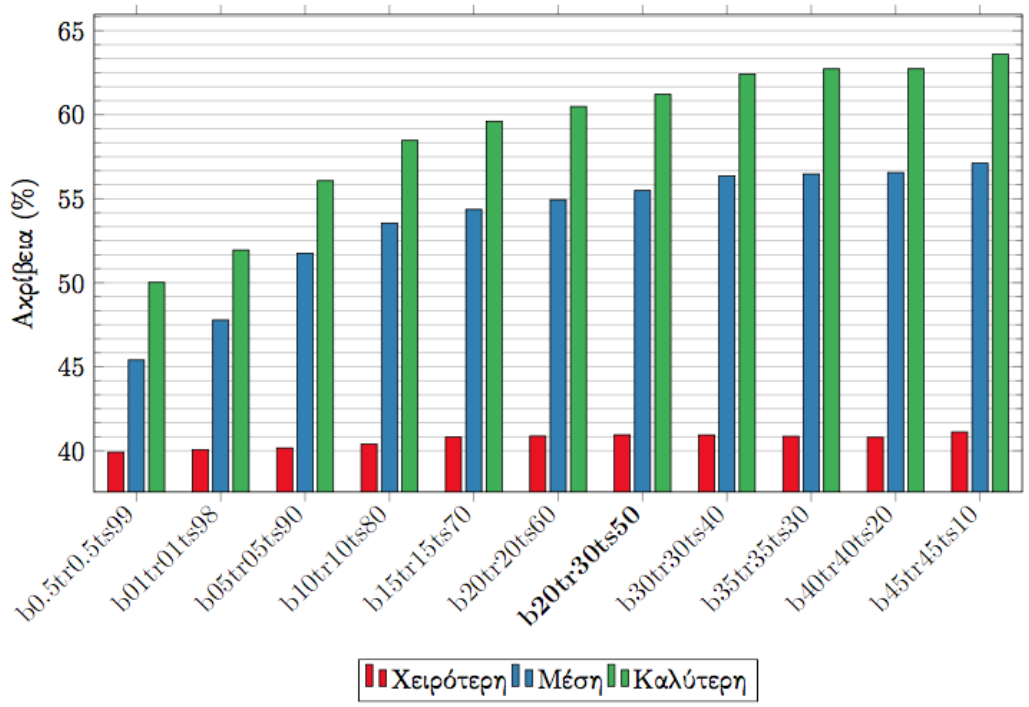


Σχήμα 4.5: Κατανομή κλάσεων πολικότητας ανά υποσύνολο δεδομένων

Μία από τις κυριότερες αποφάσεις κατά τη διαδικασία αξιολόγησης ενός μοντέλου Επιβλεπόμενης Μηχανικής Μάθησης είναι η επιλογή της κατάλληλης αναλογίας μεταξύ των συνόλων εκπαίδευσης και εξέτασης. Για δεδομένο πλήθος προτύπων, με τη χρήση ενός μεγαλύτερου συνόλου εξέτασης η εκτιμώμενη ακρίβεια εμφανίζει μικρότερη διακύμανση και προκύπτει μία πιο αξιόπιστη εικόνα της συμπεριφοράς του μοντέλου. Εντούτοις, με τη χρήση ενός μεγαλύτερου συνόλου εκπαίδευσης επιτυγχάνεται πιο αντιπροσωπευτική μάθηση. Στην περίπτωσή μας, το σύνολο εκπαίδευσης συντίθεται από τα δεδομένα κατασκευής και μάθησης.

Για τον προσδιορισμό της βέλτιστης αναλογίας ανάμεσα στα δεδομένα κατασκευής (builset), μάθησης (training set) και εξέτασης (test set) διερευνήθηκαν διάφοροι συνδυασμοί των ποσοστών διάσπασης.

Στο ακόλουθο σχήμα απεικονίζεται ο μέσος όρος της ακρίβειας των ταξινομητών στην καλύτερη, μέση και χειρότερη περίπτωση, όπως αυτές προκύπτουν χρησιμοποιώντας μέγεθος ν-γράμματος από 1 έως και 7 και για τις διάφορες ποσοστώσεις διάσπασης του συνόλου των δεδομένων.



Σχήμα 4.6: Ακριβεία κατηγοριοποίησης ανά συνδυασμό ποσοστών διάσπασης

Παρατηρούμε ότι το εξεταζόμενο μοντέλο Ανάλυσης Συναισθήματος εμφανίζει αρκετά ικανοποιητική συμπεριφορά όταν χρησιμοποιείται πολύ μικρό σύνολο εκπαίδευσης, για παράδειγμα στην περίπτωση του συνδυασμού b0.5tr0.5ts99, με αναλογία 1:99 σε πληθυσμό 121401 tweets, επιτυγχάνει μέση ακρίβεια κατηγοριοποίησης 45.9%, αυξημένη κατά 12% σε σχέση με αυτή ενός τυχαίου ταξινομητή (33,33%). Αυτή η συμπεριφορά συνήθως δεν παρατηρείται σε άλλες μεθόδους επιβλεπόμενης μάθησης.

Ωστόσο, διαπιστώνουμε πως εκπαιδεύοντας το μοντέλο με περισσότερα πρότυπα, βελτιώνεται η μέγιστη ακρίβεια - η οποία είναι και το βασικό αντικείμενο του συστήματος - αλλά οδηγούμαστε γρήγορα σε κορεσμό : εξετάζοντας τις περιπτώσεις των συνδυασμών b30tr30tr40 έως και b45tr45ts10 παρατηρούμε πως η μέση τιμή της ακρίβειας των ταξινομητών και στις τρεις περιπτώσεις - καλύτερη, μέση, χειρότερη - κυμαίνεται σε παρόμοια επίπεδα, καθώς το πλήθος των δεδομένων εκπαίδευσης αυξάνεται και αντίστοιχα το πλήθος των δεδομένων εξέτασης μειώνεται, σημειώνοντας βέλτιστη επίδοση 68% στην περίπτωση b45tr45ts10. Συμπεραίνουμε πως για το συγκεκριμένο μέγεθος και

εσωτερική διάρθρωση του συνόλου, η μέθοδος Ανάλυσης Συναισθήματος με γράφους ν-γραμμάτων δεν είναι ισχυρά εξαρτώμενη από το πλήθος των δεδομένων εκπαίδευσης.

Στο υπόλοιπο μέρος της ενότητας εξετάζουμε την περίπτωση του συνδυασμού b20tr30ts50 όπου επιτυγχάνεται επαρκής αντιστάθμιση ανάμεσα στην ακρίβεια κατηγοριοποίησης και στο μέγεθος του συνόλου εξέτασης.

Έχοντας απαντήσει το ερώτημα του κριτηρίου επιλογής των ποσοστών κατασκευής p_b, εκπαίδευσης p_{tr} και ελέγχου p_{ts}, θα προσδιορίσουμε το μέγεθος ν-γράμματος που δίνει τα βέλτιστα αποτελέσματα.

Όπως αναφέρεται στο [3], ένας γράφος ν-γραμμάτων χαρακτηρίζεται από τρεις παραμέτρους :

I. τον ελάχιστο βαθμό ν-γράμματος L_{min}
II. το μέγιστο βαθμό ν-γράμματος L_{max} και
III. τη μέγιστη απόσταση γειτνίασης ή μήκος κυλιόμενου παράθυρου D_{win}.

Ακολουθούμε την προσέγγιση των Aisopos et al. και εξετάζουμε αποκλειστικά την περίπτωση όπου $L_{min} = L_{max} = D_{win}$ η οποία πειραματικά έχει αποδειχθεί ότι οδηγεί σε επίδοση συγκρίσιμη της βέλτιστης, όπως προέκυψε από ενδελεχή αναζήτηση βέλτιστων παραμέτρων (fine-tuning) [12].

Στον πίνακα που ακολουθεί παρουσιάζονται συγκεντρωτικά τα αποτελέσματα εκτέλεσης των αλγορίθμων ταξινόμησης για μεγέθη ν-γράμμματος n = {1, 2, · · · , 7}.

NGram	MLP	SVM	Logistic	kNN	NB	MNB	C4.5
1	43.08	39.7	44.92	40.25	38.56	36.97	43.52
2	57.78	56.27	59.19	48.48	44.75	36.97	55.17
3	64.22	64.47	64.61	55.19	51.71	36.98	62.2
4	65.48	65.62	65.61	57.25	55.9	40.15	63.88
5	65.15	65.08	65.31	56.84	57.51	42.82	63.58
6	64.00	63.75	64.27	55.9	55.35	44.94	62.82
7	62.5	62.15	62.91	54.62	51.23	46.31	62.2

Πίνακας 4.4: Ποσοστά Ακρίβειας Ταξινομητών για ν-γράμματα 1 έως 7

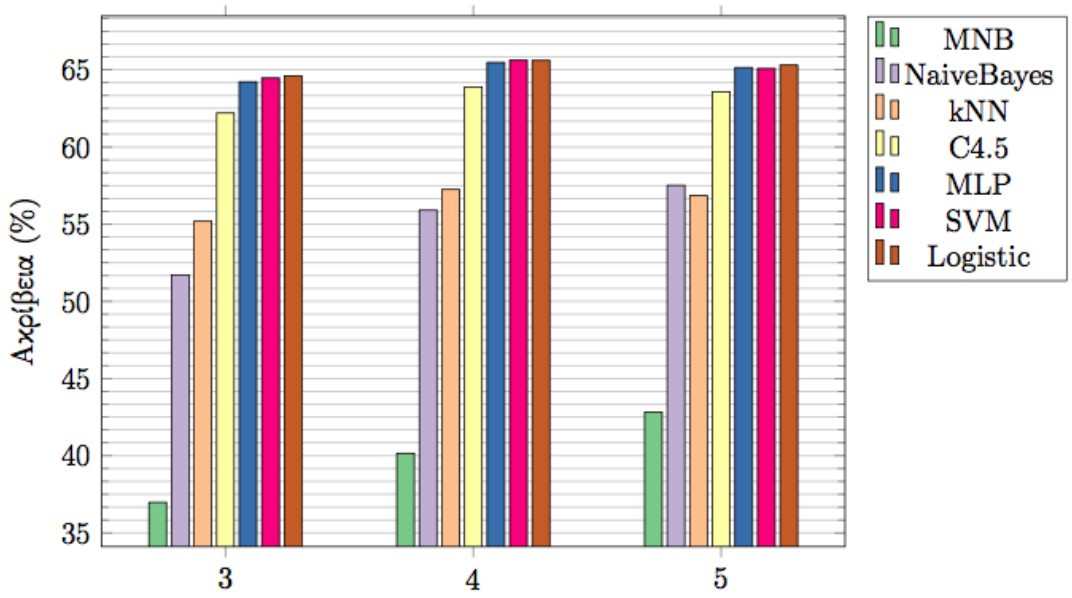
Όσο αφορά το μέγεθος ν-γράμματος, παρατηρούμε ότι, με εξαίρεση τον Πολυωνυμικό Naive Bayes (MNB), οι υπόλοιποι αλγόριθμοι ταξινόμησης εμφανίζουν παρόμοια συμπεριφορά. Καθώς αυξάνεται η τιμή του n, βελτιώνονται η ακρίβεια μέχρι την τιμή n = 4 (στην περίπτωση του NaiveBayes n = 5) όπου επιτυγχάνεται η βέλτιστη επίδοση η οποία στη συνέχεια φθίνει όταν χρησιμοποιούνται γράφοι μεγαλύτερου βαθμού.

Συμπεραίνουμε πως η τιμή n = 4 αποτελεί την καλύτερη επιλογή στην πλειονότητα των ταξινομητών όπου αξιοποιούνται στο μέγιστο τα πλεονεκτήματα της μεθόδου, δηλαδή ανεξαρτησία γλώσσας και υψηλή ανοχή στο θόρυβο. Γράφοι ν-γραμμάτων μικρού βαθμού (μονογράμματα, διγράμματα) εξαιτίας του μικρού μεγέθους τους αδυνατούν να συλλάβουν τα πιο χαρακτηριστικά εκφραστικά μοτίβα του πολυγλωσσικού περιεχομένου και οδηγούν σε χαμηλά ποσοστά ακρίβειας. Στους γράφους ν-γραμμάτων μεγάλου βαθμού αυξάνεται εκθετικά το πλήθος των κορυφών και ακμών με αποτέλεσμα η επιπλέον πληροφορία που αποτυπώνεται στους γράφους να μην αντιστοιχεί στην πραγματική εξάρτηση που υπάρχει ανάμεσα στις λέξεις-συμβοακολουθίες και να αποτελεί ουσιαστικά θόρυβο. Αυτό εξηγεί το λόγο για τον οποίο τα ποσοστά ακρίβειας για n = 6, 7 υπολείπονται μεν της βέλτιστης επίδοσης, κυμαίνονται δε σε ικανοποιητικά επίπεδα, αρκετά υψηλότερα των αντίστοιχων ποσοστών για n = 1, 2.

Ο Πολυωνυμικός Naive Bayes ακολουθεί τη συμπεριφορά των υπόλοιπων ταξινομητών, βελτιώνοντας τη ακρίβεια καθώς αυξάνεται το μέγεθος ν-γράμματος αλλά με πολύ πιο αργό ρυθμό οπότε και παρουσιάζει το καλύτερο αποτέλεσμα στην περίπτωση n = 7. Διαπιστώνουμε πως ο συγκεκριμένος ταξινομητής δεν εμφανίζει την ίδια ευαισθησία ως προς το μέγεθος ν-γράμματος με τους υπόλοιπους. Το φαινόμενο αυτό εξηγεί τα πολύ χαμηλά ποσοστά ακρίβειας στο εύρος n □ [1, 7] και οφείλεται πιθανότατα στην εσφαλμένη υπόθεση ότι οι δείκτες ομοιότητας ακολουθούν πολυωνυμική κατανομή.

Έχοντας απαντήσει τα ερωτήματα του κριτηρίου επιλογής των ποσοστών κατασκευής p_b, εκπαίδευσης p_{tr} και ελέγχου p_{ts}, καθώς επίσης του βέλτιστου μεγέθους ν-γράμματος, θα εντοπίσουμε το βέλτιστο ταξινομητή.

Προς τούτο, διεξάγουμε μία σειρά μετρήσεων για μεγέθη ν-γράμματος 3, 4 και 5. Παραθέτουμε τα αποτελέσματα στο ακόλουθο διάγραμμα.



Σχήμα 4.7: Ακρίβεια κατηγοριοποίησης για μεγέθη ν-γράμματος 3, 4 και 5

Η βέλτιστη ακρίβεια κατηγοριοποίησης επιτυγχάνεται στην περίπτωση n = 4 από τους ταξινομητές SVM (65.62%), Λογιστικής Παλινδρόμησης (65.61%) και Πολυεπίπεδου Perceptron (65.48%). Ακολουθούν σε φθίνουσα σειρά επίδοσης οι C4.5 (63.88%), Naive Bayes (57.51%, n = 5), k-NN (57.25%) και Πολυωνυμικός Naive Bayes (46.31%, n = 7).

Συγκρίνοντας τους τρεις καλύτερους αλγορίθμους ως προς τη γενικότερη επίδοση, παρατηρούμε πως αν και ανήκουν σε διαφορετικές κατηγορίες Μηχανικής Μάθησης, εμφανίζουν παρόμοια αποτελέσματα σε όλες τις τιμές ν-γράμματος που εξετάσαμε. Επομένως, χρειάζεται να μελετήσουμε πιο αναλυτικά τον τρόπο με τον οποίο επιτεύχθηκε το συγκεκριμένο ποσοστό ακρίβειας εξετάζοντας την ικανότητα των ταξινομητών στην κατηγοριοποίηση ανά κατηγορία πολικότητας με τη βοήθεια του πίνακα σφάλματος - σύγχυσης (confusion matrix). Στον ακόλουθο πίνακα παρουσιάζονται οι τιμές της ακρίβειας (precision-PR), ανάκλησης (recall-RC) και του συνδυασμού τους F1 για κάθε κατηγορία όπως επίσης και ο σταθμισμένος μέσος όρος τους για τους τρεις ταξινομητές για n = 4.

	MLP			SVM			Logistic		
Κλάση	PR	RC	F1	PR	RC	F1	PR	RC	F1
Θετικά	0.759	0.587	0.662	0.729	0.620	0.670	0.689	0.672	0.680
Αρνητικά	0.646	0.657	0.651	0.661	0.620	0.640	0.660	0.620	0.639
Ουδέτερα	0.593	0.745	0.660	0.592	0.740	0.658	0.617	0.676	0.645

Πίνακας 4.5: Precision, Recall και F_1-score των MLP, SVM και Logistic

Διαπιστώνουμε πως οι ταξινομητές εμφανίζουν παρεμφερή χαρακτηριστικά και κατά την ανάλυση σε επίπεδο κλάσεων πολικότητας. Σύμφωνα με τους δείκτες ακρίβειας και ανάκλησης προκύπτει ότι και οι τρεις αλγόριθμοι είναι περισσότερο ακριβείς στην ταξινόμηση δεδομένων στη θετική κατηγορία και αναγνωρίζουν επιτυχώς μεγαλύτερο ποσοστό των ουδέτερων tweets χωρίς σημαντικές επιπτώσεις στα αντίστοιχα μεγέθη των υπόλοιπων κατηγοριών. Η ισορροπία που επικρατεί μεταξύ στις δύο αντιστρόφως ανάλογες έννοιες αντικατοπτρίζεται στην αρκετά υψηλή τιμή του δείκτη F1 στις επιμέρους κατηγορίες καθώς και στην περίπτωση του σταθμισμένου μέσου όρου. Επομένως, χρησιμοποιώντας οποιονδήποτε από τους ταξινομητές SVM, Λογιστική Παλινδρόμηση και Πολυεπίπεδο Perceptron, το σύστημα Ανάλυσης Συναισθήματος ανταποκρίνεται επαρκώς και στις τρεις κατηγορίες των tweets χωρίς αισθητή προκατάληψη. Αυτό το φαινόμενο που υποδηλώνει αντιπροσωπευτική εκπαίδευση και οδηγεί σε ικανοποιητικά ποσοστά ακρίβειας κατηγοριοποίησης.

4.3 Αρχιτεκτονική και αξιολόγηση της υπηρεσίας

Η διαδικτυακή υπηρεσία που έχουμε σχεδιάσει πραγματοποιεί ανάλυση συναισθήματος κατά απαίτηση (on demand) και διαθέτει τη δυνατότητα σύνδεσης σε ευρύτερες πλατφόρμες και λειτουργίας εντός αυτών (plug and play).

Στα πλαίσια μίας πλατφόρμας που εξορύσσει δεδομένα από κάποιο κοινωνικό δίκτυο σε πραγματικό χρόνο, η υπηρεσία δύναται να προσφέρει ανάλυση συναισθήματος σε πραγματικό χρόνο (real time sentiment analysis) υπό την προϋπόθεση ότι τροφοδοτείται με το ρεύμα δεδομένων - μηνυμάτων (data stream) του κοινωνικού δικτύου.

Υπό το πρίσμα της απλουστευμένης εκδοχής ενός πελάτη και ενός εξυπηρετητή που περιέχει την υπηρεσία, ο πελάτης υποβάλει στον εξυπηρετητή μία αίτηση HTTP στο μονοπάτι sentiment_ws/rest/sentiment/service. Το περιεχόμενο της αίτησης είναι το κείμενο ενός tweet και

στέλνεται με τη μέθοδο GET ως παράμετρος με το όνομα tweet. Επομένως, το URL που ζητείται είναι το εξής

{Server Name or IP}:{Port}/emotion/rest/service/analyze/?tweet=Hello+World

για την περίπτωση που θέλουμε να εντοπίσουμε το συναίσθημα της φράσης “Hello World”.

Αφού λάβει την αίτηση, ο εξυπηρετητής εξάγει το κείμενο του tweet και δημιουργεί την αναπαράσταση του κειμένου με ένα γράφο τριγράμματος χαρακτήρα. Κατόπιν, συγκρίνει το γράφο με τους γράφους που αντιστοιχούν στην αναπαράσταση του θετικού, του αρνητικού και του ουδετέρου συναισθήματος και εξάγει ένα διάνυσμα χαρακτηριστικών για κάθε σύγκριση. Τέλος, δίνει τα διανύσματα χαρακτηριστικών ως εισόδους στον ταξινομητή και λαμβάνει την εκτίμηση του ταξινομητή για το συναίσθημα του κειμένου, δηλαδή αν είναι θετικό, αρνητικό ή ουδέτερο.

Τόσο οι γράφοι αναπαράστασης του αρνητικού, του θετικού και του ουδετέρου συναισθήματος όσο και ο ταξινομητής είναι πόροι που βρίσκονται αποθηκευμένοι στον εξυπηρετητή και φορτώνονται κατά την αρχικοποίηση της υπηρεσίας στον εξυπηρετητή. Το μέγεθος του ν-γράμματος που χρησιμοποιήσαμε, καθώς επίσης οι ταξινομητές προκύπτουν από την ανάλυση της υποενότητας που προηγήθηκε. Ωστόσο, αξίζει να σημειώσουμε ότι τόσο το μέγεθος του ν-γράμματος όσο και οι ταξινομητές που διαθέτει η υπηρεσία μπορούν εύκολα να αντικατασταθούν από το διαχειριστή. Με αυτό τον τρόπο, παράμετροι όπως η ταχύτητα και η ακρίβεια της υπηρεσίας μπορούν να βελτιστοποιηθούν εφόσον ένας νέος ταξινομητής γίνει διαθέσιμος.

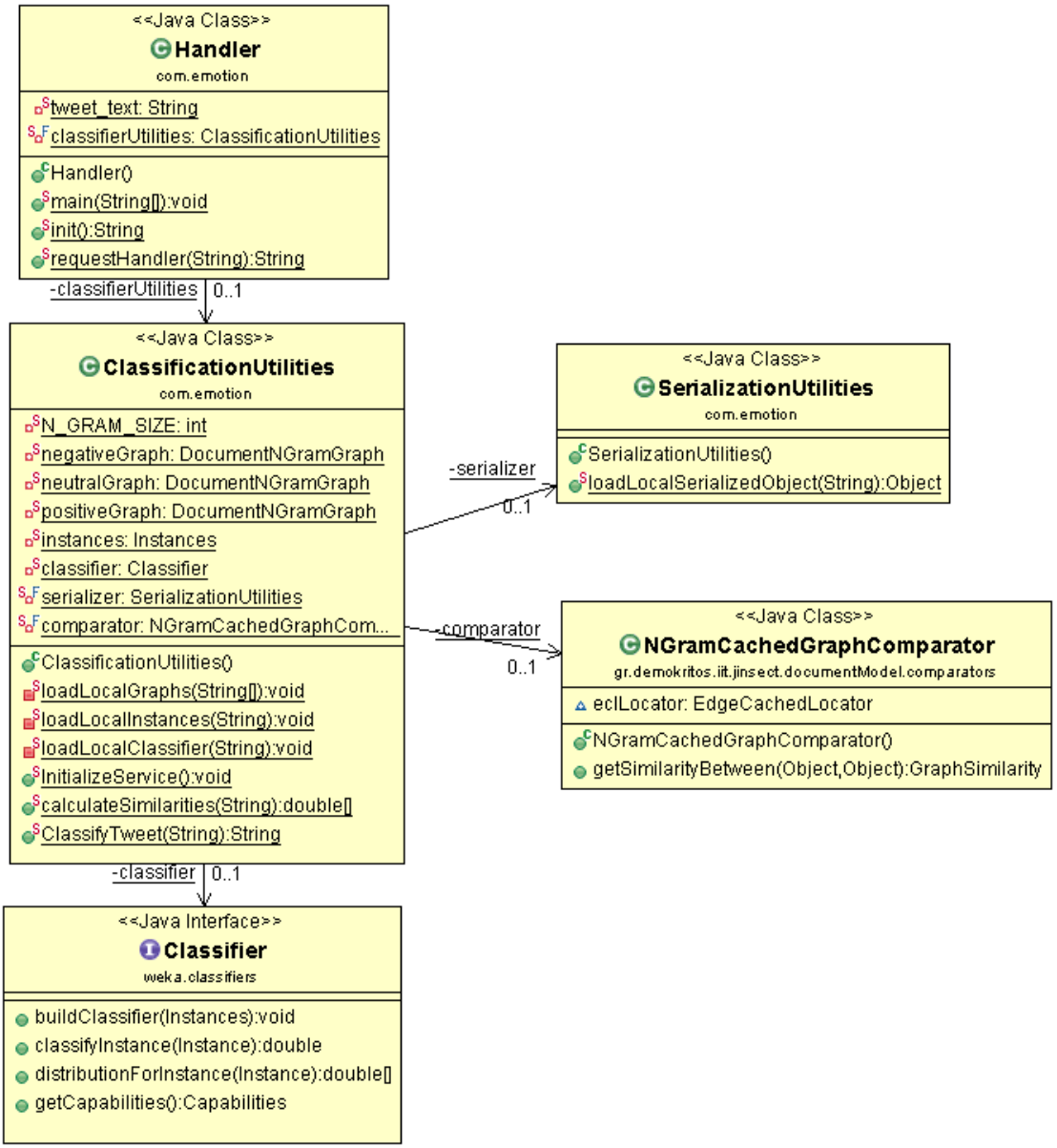
Στην περίπτωση που η υπηρεσία συνδεθεί σε κάποια που εξορύσσει δεδομένα από κάποιο κοινωνικό δίκτυο σε πραγματικό χρόνο και δέχεται ρεύματα δεδομένων ως είσοδο, κάθε μήνυμα του ρεύματος δεδομένων εντάσσεται σε μία αίτηση HTTP που υποβάλεται στην υπηρεσία με τον τρόπο που περιγράψαμε.

Στο παράρτημα παραθέτουμε παραδείγματα χρήσης της υπηρεσίας.

Στη συνέχεια της παρούσας ενότητας, περιγράφουμε διεξοδικά τη δομή και τον τρόπο λειτουργίας της υπηρεσίας, καθώς επίσης μελετάμε την επιδοσή της σε συγκεκριμένες περιπτώσεις χρήσης.

4.5.1 Διάγραμμα Κλάσεων

Παραθέτουμε το διάγραμμα κλάσεων που περιγράφει τη δομή της διαδικτυακής υπηρεσίας και επεξηγούμε τη χρησιμότητα κάθε κλάσης.



Σχήμα 4.8: Διάγραμμα κλάσεων διαδικτυακής υπηρεσίας

Κλάση Handler

Η κλάση Handler χειρίζεται αφενός μεν την αρχικοποίηση της υπηρεσίας με τη μέθοδο init αφετέρου δε τη λήψη και επεξεργασία του αιτήματος ανάλυσης ενός μηνύματος με τη μέθοδο requestHanlder.

Κλάση Classification Utilities

Η κλάση Classification Utilities προσφέρει τις μεθόδους για την αρχικοποιήση της υπηρεσίας και την ταξινόμηση του συναισθήματος ενός μηνύματος.

Για την αρχικοποίηση της υπηρεσίας διαθέτει τις μεθόδους loadLocalGraphs, loadLocalInstance και loadLocalClassifier που φορτώνουν τους αποθηκευμένους στον εξυπηρετητή γράφους, ένα πρότυπο instance και τον ταξινομητή. Τόσο οι γράφοι όσο και ο ταξινομητής είναι αποθηκευμένοι σε σειριοποιημένη μορφή (serialized objects). Για την αποσειριοποίηση τους καλείται κάθε φορά η μέθοδος loadLocalSerializedObject της κλάσης SerializationUtilities. Η μέθοδος InitializeService ορίζει της σειρά κλήσης των μεθόδων φόρτωσης και περιέχει τα ορίσματα κάθε μεθόδου.

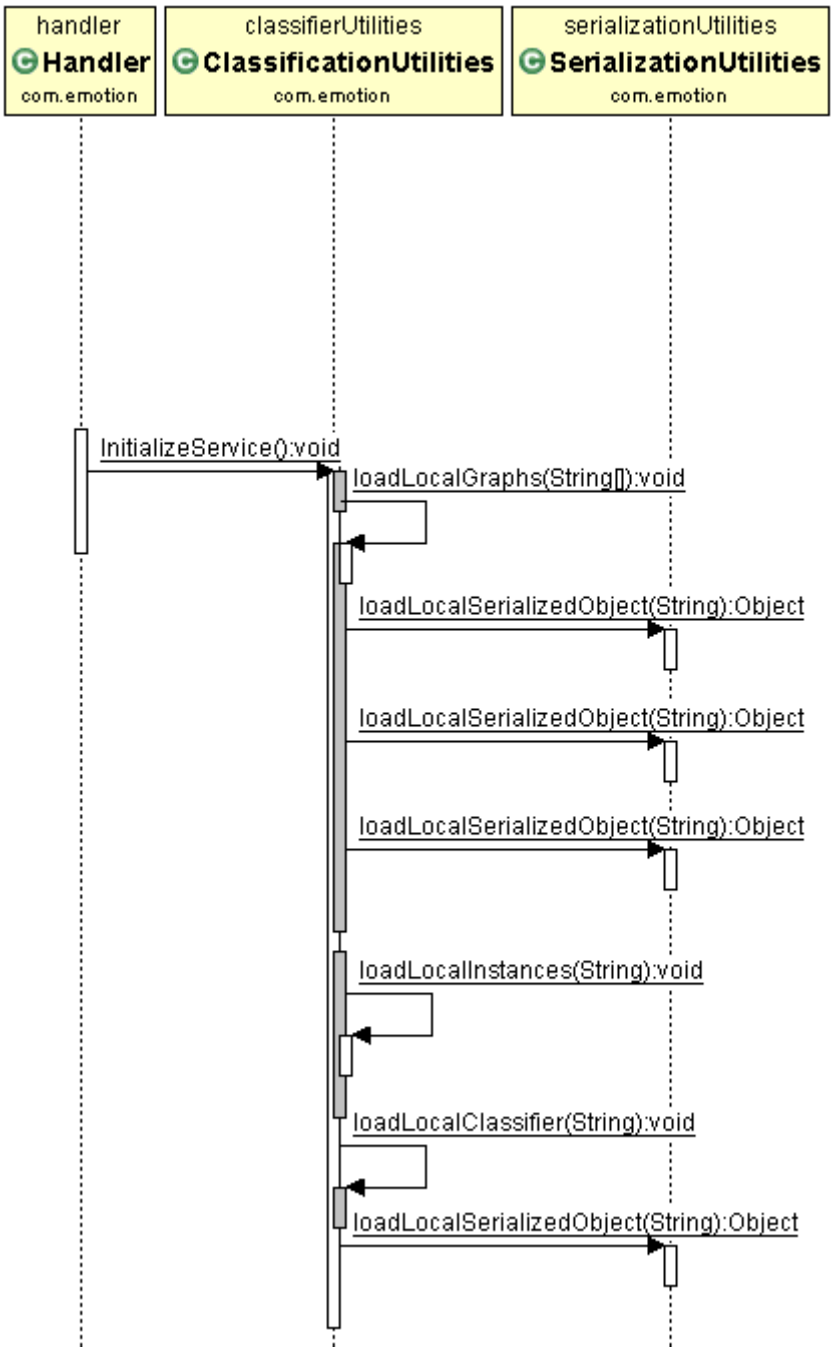
Για την ταξινόμηση του συναισθήματος ενός μηνύματος διαθέτει τις μεθόδους calculateSimilarities και ClassifyTweet. Η μέθοδος calculateSimilarities συνθέτει το γράφο του μηνύματος, υπολογίζει με τη βοήθεια της μεθόδου getSimilarityBetween της κλάσης NGramCachedGraphComparator τους δείκτες ομοιότητας του γράφου μηνύματος με τους γράφους αναπαράστασης του θετικού, του αρνητικού και του ουδετέρου συναισθήματος ξεχωριστά και συνθέτει το διάνυσμα χαρακτηριστικών από τους επιμέρους δείκτες ομοιότητας που έχουν προέκυψαν από τη σύγκριση των γράφων. Η μέθοδος ClassifyTweet υπολογίζει το διάνυσμα χαρακτηριστικών με τη μέθοδο calculateSimilarities, δημιουργεί ένα instance για το διάνυσμα χαρακτηριστικών, εντάσσει το instance στο πρότυπο - στην πραγματικότητα κενό - σύνολο instances για τη σωστή λειτουργία της μεθόδου ταξινόμησης classifyInstance και τελικά καλεί τη μέθοδο classifyInstance της κλάσης Classifier ώστε να ολοκληρώσει την πρόβλεψη συναισθήματος του μηνύματος.

Κλάση Serialization Utilities

Η κλάση Serialization Utilities διαθέτει τη μέθοδο loadLocalSerializedObject η οποία λαμβάνει το τοπικό ως προς την εφαρμογή μονοπάτι του σειριοποιημένου αντικειμένου που πρόκειται να φορτώσει και κατόπιν ανοίγει, διαβάζει και αποσειριοποιεί το αντικείμενο.

4.5.1 Ακολουθιακά Διαγράμματα

Αρχικά, παραθέτουμε το ακολουθιακό διάγραμμα που περιγράφει την αρχικοποίηση της υπηρεσίας.

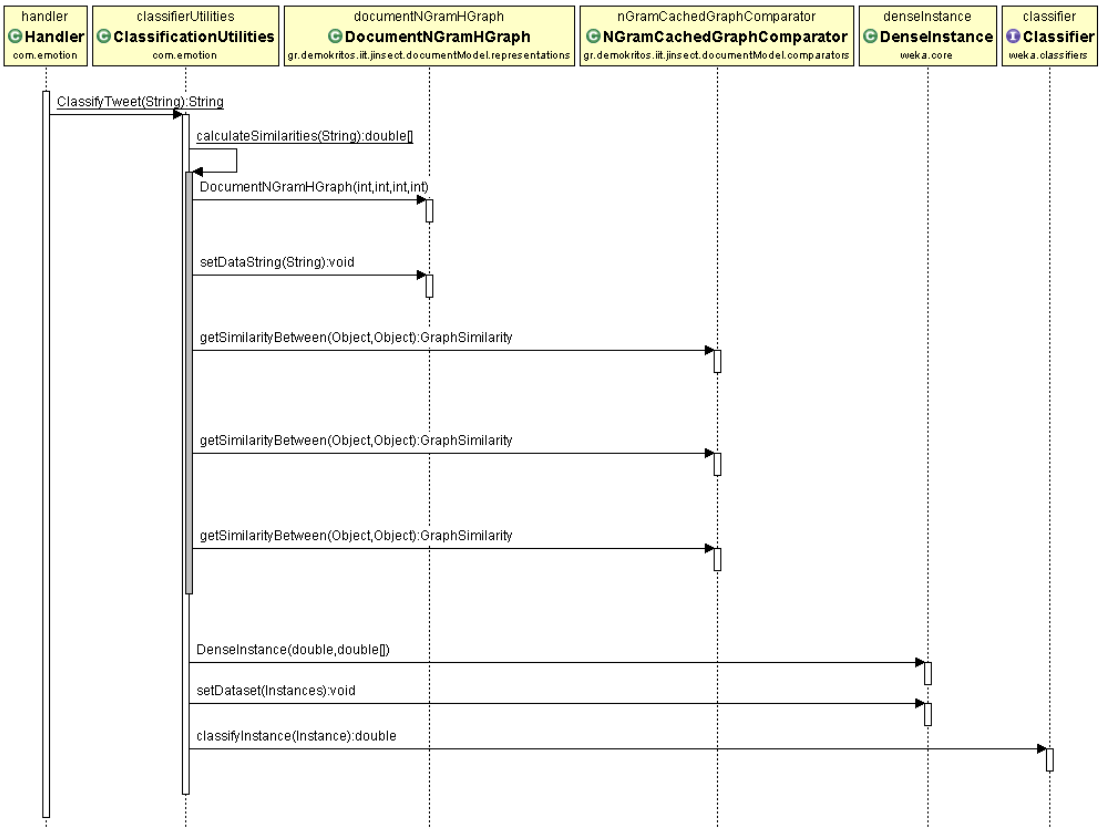


Σχήμα 4.9: Ακολουθιακό διάγραμμα αρχικοποίησης της υπηρεσίας

Η αίτηση του URL {Server Name or IP}:{Port}/emotion/rest/service/init έχει ως αποτέλεσμα την κλήση της μεθόδου init της κλάσης Handler. Η μέθοδος init καλεί τη μέθοδο InitializeService της

κλάσης ClasificationUtilities που με τη σειρά της καλεί διαδοχικά τις μεθόδους loadLocalGraphs, loadLocalInstances και loadLocalClassifier της ίδιας κλάσης. Η μέθοδος loadLocalGraphs καλεί τρεις φορές τη μέθοδο loadLocalSerializedObject για την αποσειριοποίηση του θετικού, του αρνητικού και του ουδετέρου γράφου, ενώ η μέθοδος loadLocalClassifier καλεί μία φορά τη μέθοδο loadLocalSerializedObject για την αποσειριοποίηση του ταξινομητή. Η μέθοδος loadLocalInstances καλεί κάποιες μεθόδους βοηθητικών κλάσεων. Οι συγκεκριμένες κλήσεις δεν απεικονίζονται στο διάγραμμα για την ευκολότερη κατανόηση της ουσίας της αρχικοποίησης της υπηρεσίας.

Στη συνέχεια, παραθέτουμε το ακολουθιακό διάγραμμα που περιγράφει τη λειτουργία της υπηρεσίας κατά την ταξινόμηση ενός μηνύματος.



Σχήμα 4.10: Ακολουθιακό διάγραμμα λειτουργίας ταξινόμησης μηνύματος

Η αίτηση του URL με μονοπάτι emotion/rest/service/analyze/?tweet=Hello+World έχει ως αποτέλεσμα την κλήση της μεθόδου requestHandler της κλάσης Handler με παράμετρο το μήνυμα προς αξιολόγηση συναισθήματος. Η requestHandler καλεί την ClassifyTweet της κλάσης ClasificationUtilities με όρισμα το μήνυμα. Η μέθοδος ClassifyTweet αρχικά καλεί τη μέθοδο calculateSimilarities της κλάσης ClasificationUtilities με όρισμα το μήνυμα. Η μέθοδος

calculateSimilarities καλεί τον τυποκατασκευαστή της κλάσης DocumentNGramHGraph και ακολούθως τη μέθοδο setDataString που ουσιαστικά απεικονίζει το μήνυμα σε ένα γράφο. Κατόπιν, η μέθοδος calculateSimilarities καλεί τρεις φορές διαδοχικά τη μέθοδο getSimilarityBetween για τον υπολογισμό της ομοιότητας των γράφων. Η μέθοδος calculateSimilarities επιστρέφει το διάνυσμα χαρακτηριστικών στη μέθοδο ClassifyTweet και η μέθοδος ClassifyTweet καλεί με τη σειρά τις DenseInstance και setDataset για τη δημιουργία του στιγμιοτύπου (instance) του διανύσματος χαρακτηριστικών. Τέλος, η μέθοδος ClassifyTweet καλεί τη μέθοδο classifyInstance για την ταξινόμηση του συναισθήματος που αποτυπώνεται στο στιγμιότυπο και επιστρέφει το συναίσθημα στη μέθοδο requestHandler της κλάσης Handler.

4.5.3 Μελέτη Επίδοσης Διαδικτυακής Υπηρεσίας

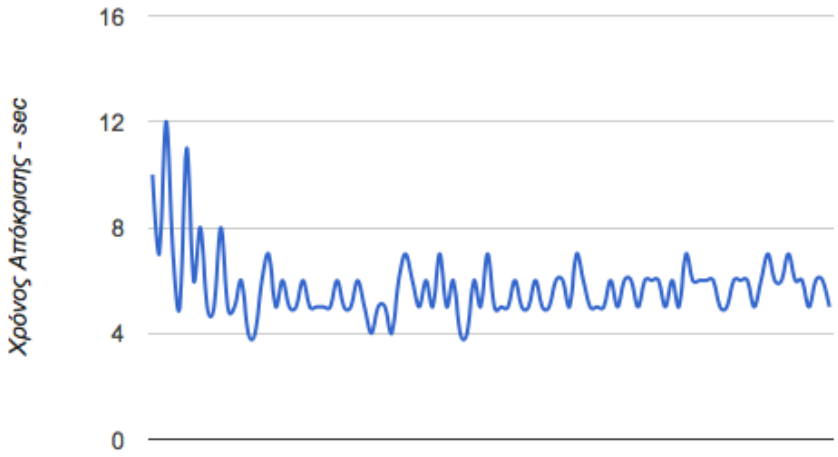
Στην παρούσα ενότητα μελετάμε την επίδοση της υπηρεσίας για περιπτώσεις χρήσεις που διαφοροποιούνται ως προς το πλήθος των χρηστών που υποβάλλουν αιτήματα στην υπηρεσία ή ως προς το πλήθος των διαδοχικών αιτήσεων που υποβάλλει ένας χρήστης.

Αρχικά, μελετάμε την απόκριση της υπηρεσίας σε περιπτώσεις όπου δέχεται ταυτόχρονα αιτήματα από πολλαπλούς χρήστες.

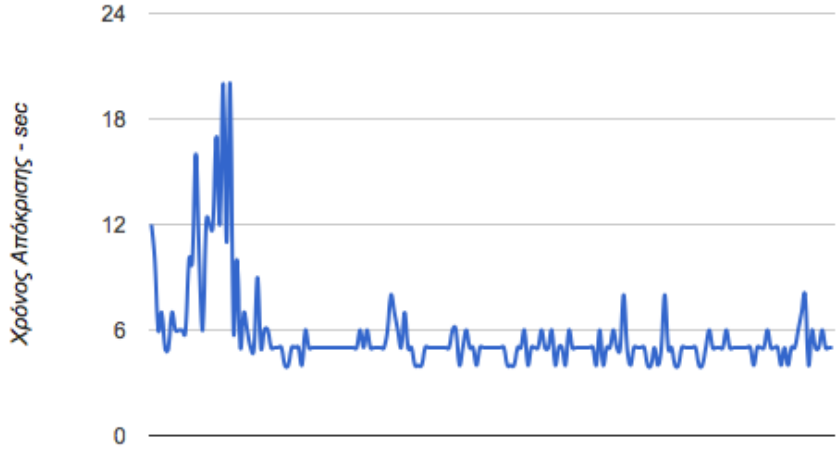
Παραθέτουμε τους μέσους χρόνους απόκρισης ανά πλήθος χρηστών στον παρακάτω πίνακα που ακολουθείται από τα διαγράμματα του χρόνου απόκρισης ανά χρήστη.

# Χρηστών	Μέσος χρόνος απόκρισης ανά χρήστη - Sec
100	5.76
200	5.75
300	6.28
400	6.06
500	7.85

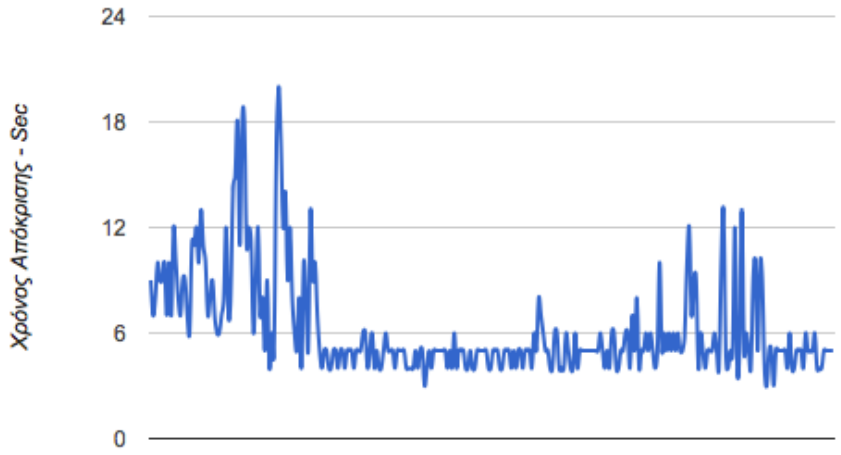
Πίνακας 4.6: Μέσος χρόνος απόκρισης ανά χρήστη



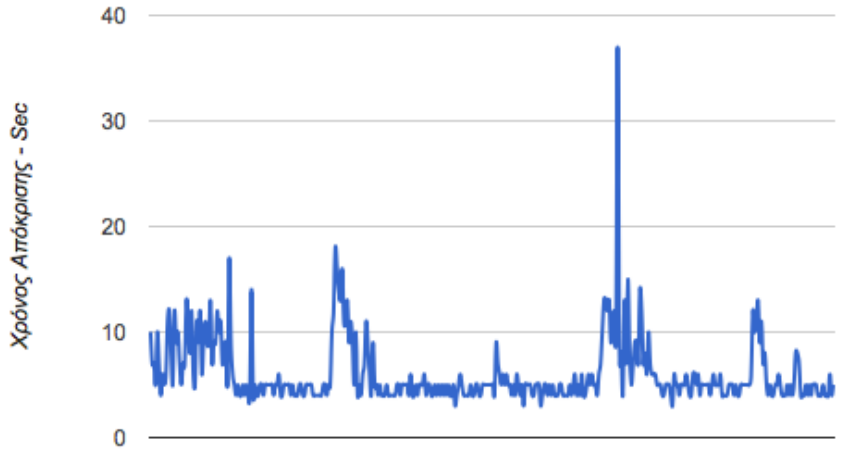
Σχήμα 4.11: Χρόνος απόκρισης για 100 χρήστες



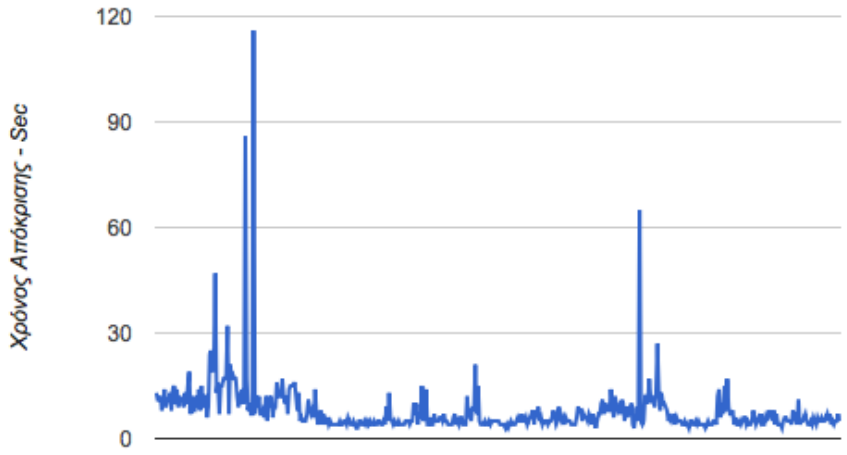
Σχήμα 4.12: Χρόνος απόκρισης για 200 χρήστες



Σχήμα 4.13: Χρόνος απόκρισης για 300 χρήστες



Σχήμα 4.14: Χρόνος απόκρισης για 400 χρήστες



Σχήμα 4.15: Χρόνος απόκρισης για 500 χρήστες

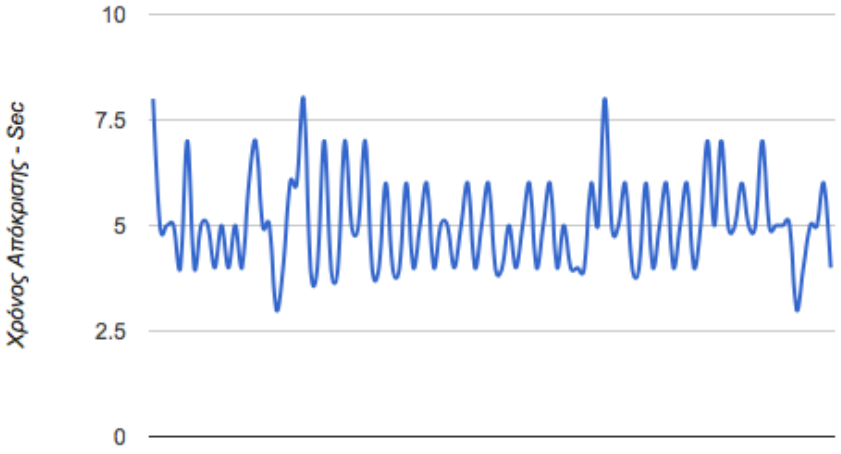
Στη συνέχεια, μελετάμε την απόκριση της υπηρεσίας σε περιπτώσεις όπου δέχεται πολλαπλά διαδοχικά αιτήματα, ένα ρεύμα αιτημάτων, από ένα χρήστη ή πηγή και χωρίς χρήση caching.

Παραθέτουμε τους μέσους χρόνους απόκρισης ανά πλήθος διαδοχικών αιτημάτων στον παρακάτω πίνακα που ακολουθείται από τα διαγράμματα του χρόνου απόκρισης ανά αίτημα.

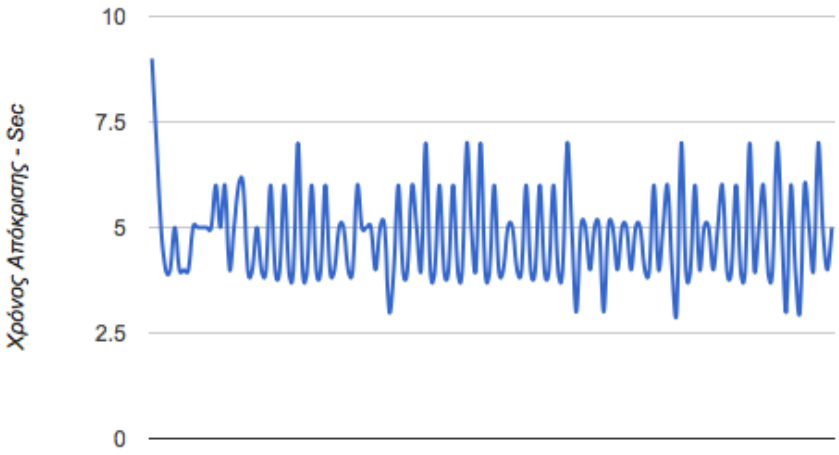
# Αιτημάτων	Μέσος Χρόνος Απόκρισης Ανά Αίτημα - Sec
100	4.80
200	5.05
300	4.80
400	4.82
500	4.96

Πίνακας 4.7: Μέσος χρόνος απόκρισης ανά αίτημα

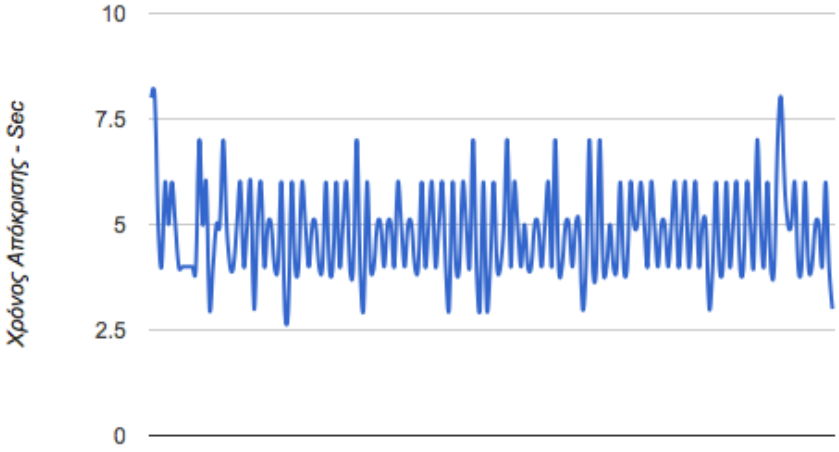
Σχήμα 4.16: Χρόνος απόκρισης για 100 αιτήματα



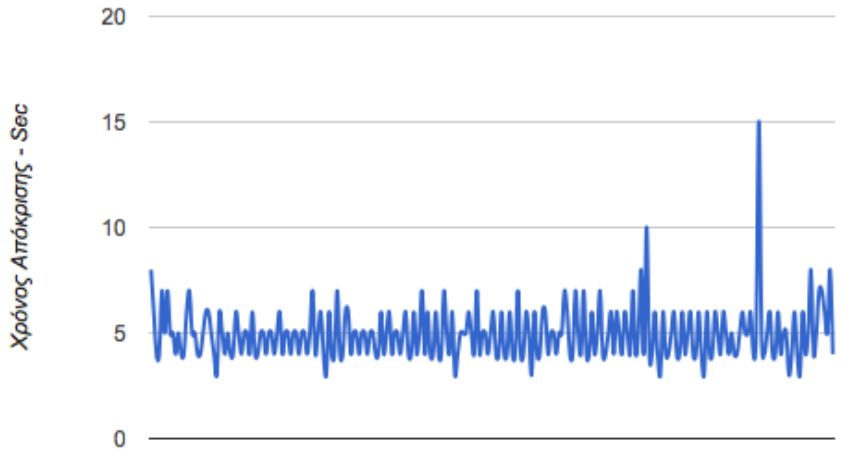
Σχήμα 4.17: Χρόνος απόκρισης για 200 αιτήματα



Σχήμα 4.18: Χρόνος απόκρισης για 300 αιτήματα



Σχήμα 4.19: Χρόνος απόκρισης για 400 αιτήματα



Σχήμα 4.20: Χρόνος απόκρισης για 500 αιτήματα

Από τα αποτελέσματα των παραπάνω μετρήσεων, συμπεραίνουμε ότι η υπηρεσία ανταποκρίνεται καλύτερα στην περίπτωση που δέχεται ένα ρεύμα αιτημάτων από μία πηγή, καθώς αφενός μεν οι μέσοι χρόνοι απόκρισης είναι μικρότεροι της περίπτωσης αποδοχής αιτημάτων από πολλαπλούς χρήστες αφετέρου δε οι μέσοι χρόνοι απόκρισης διατηρούνται πρακτικά σταθεροί ανεξαρτήτως του αριθμού των διαδοχικών αιτημάτων που δέχεται η υπηρεσία.

Κατά συνέπεια, η υπηρεσία ενδείκνυται για χρήση σε ένα περιβάλλον όπου θα δέχεται και θα αξιολογεί μία συνεχή ροή δεδομένων, καθώς η απόδοσή της παραμένει αμετάβλητη και η αξιοπιστία της υψηλή.

5

Επίλογος

5.1 Σύνοψη

Στο πλαίσιο της παρούσας εργασίας, μελετήσαμε το πρόβλημα της Ανάλυσης Συναισθήματος σε δεδομένα από το κοινωνικό δίκτυο Twitter σε πραγματικό χρόνο με το μοντέλο γράφων ν-γραμμάτων. Η προσέγγιση με το μοντέλο γράφων ν-γραμμάτων ενδείκνυται για Ανάλυση Συναισθήματος στο πολυγλωσσικό και πολυθεματικό περιβάλλον κοινωνικών δικτύων. Ο λόγος είναι η ανεξαρτησία της μεθόδου από τη γλώσσα, καθώς επίσης και η αντοχή της στους ιδιωματισμούς, τους νεολογισμούς, τις συντμήσεις, τα ιδιαίτερα συντακτικά μοτίβα και εν γένει στο θόρυβο που εμφανίζεται στα κοινωνικά δίκτυα.

Από τη βιβλιογραφική ανασκόπηση, διαπιστώσαμε ότι η μέθοδος των γράφων ν-γραμμάτων αποτελεί μία καινοτόμο προσέγγιση στην Ανάλυση Συναισθήματος. Παραθέσαμε μία συνοπτική θεωρητική θεμελίωση της μεθόδου, καθώς επίσης μία ανασκόπηση των αλγορίθμων Μηχανικής Μάθησης. Παρουσιάσαμε τη δόμηση με βέλτιστο τρόπο των συνόλων κατασκευής, εκπαίδευσης και εξέτασης, την εύρεση του βέλτιστου μεγέθους ν-γράμματος και τον εντοπισμό των ταξινομητών με τη μεγαλύτερη ακρίβεια.

Παράλληλα, επισημάναμε ένα βασικό πρόβλημα που συναντάται στη στατιστική ανάλυση επιχειρησιακών δεδομένων (business analytics), δηλαδή την ανάγκη διαρκούς μεταφοράς δεδομένων από τις επιχειρησιακές βάσεις δεδομένων στις αποθήκες δεδομένων που ευθύνεται για το 80% του κόστους της στατιστικής ανάλυσης επιχειρησιακών δεδομένων.

Σε αυτό το πλαίσιο, πραγματοποιήσαμε αναφορά στη ρηξικέλευθη πλατφόρμα LeanBigData που αποσκοπεί στη μείωση του κόστους της στατιστικής ανάλυσης επιχειρησιακών δεδομένων αποτελώντας μία υπηρεσία διαχείρισης δεδομένων μεγάλου όγκου (big data) σε πραγματικό χρόνο που προσφέρει παράλληλα τις λειτουργικότητες On Line Transactional Processing (OLTP) και On Line Analytical Processing (OLAP).

Η συνεισφορά μας στην πλατφόρμα συνίσταται στη δημιουργία της υπηρεσίας που προσφέρει Ανάλυση Συναισθήματος σε πραγματικό χρόνο σε δεδομένα κοινωνικών δικτύων συνεχούς ροής. Η υπηρεσία σχεδιάστηκε έτσι, ώστε να διαθέτει τα εξής χαρακτηριστικά:

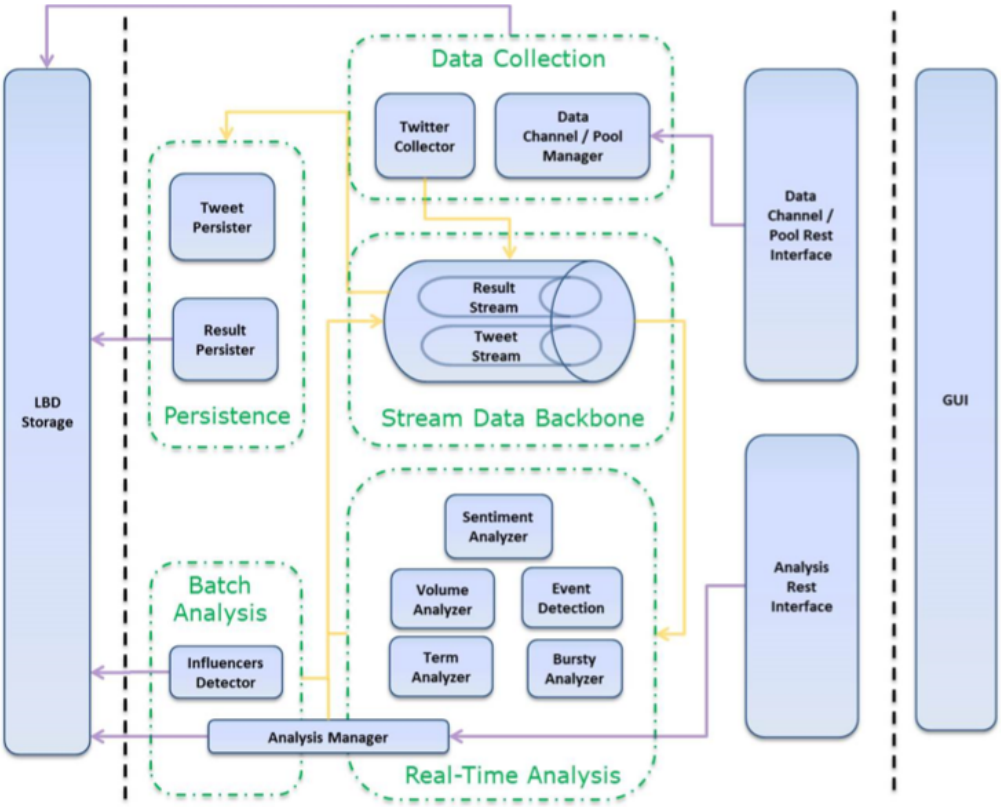
I. Δυνατότητα χρήσης ανεξάρτητα της πλατφόρμας προσαρμογής. Με άλλα λόγια, η υπηρεσία διαθέτει την ιδιότητα “Plug and play”.
II. Ευκολία αντικατάστασης των ταξινομητών. Αποτέλεσμα είναι η άμεση βελτίωση της ακρίβειας πρόβλεψης της υπηρεσίας, εφόσον χρησιμοποιηθεί ταξινομητής υψηλότερης ακρίβειας.
III. Ευκολία αντικατάστασης των γράφων ν-γραμμάτων. Σε περίπτωση που κριθεί ότι για τις ανάγκες μίας ανάλυσης εξυπηρετεί ένα διαφορετικό μέγεθος ν-γράμματος, αλλαγή μπορει να πραγματοποιηθεί άμεσα.
IV. Δυνατότητα ταξινόμησης ως προς την πολικότητα είτε μεμονωμένων μηνυμάτων ή ενός μεγάλου συνόλου μηνυμάτων. Το συγκεκριμένο χαρακτηριστικό επιτρέπει στην υπηρεσία να αξιοποιηθεί ως διαδικτυακό εργαλείο εξυπηρέτησης μεμονωμένων χρηστών.
V. Υψηλή ταχύτητα απόκρισης. Επομένως, η υπηρεσία παρουσιάζει υψηλή επίδοση ως συστατικό ενός συστήματος ανάλυσης μεγάλου όγκου δεδομένων.

5.2 Προεκτάσεις

Κατά την παρουσίαση της επίδοσης των ταξινομητών, κατέστη σαφές ότι περισσότεροι του ενός ταξινομητών εμφανίζουν συγκρίσιμη υψηλή ακρίβεια. Ως εκ τούτου, θα είχε ιδιαίτερο ενδιαφέρον ο συνδυασμός των ταξινομητών για την αξιολόγηση των μηνυμάτων με στόχο τη μείωση των ποσοστών σφάλματος κατηγοριοποίησης συγκριτικά με μεμονωμένους ταξινομητές. Ως αποτέλεσμα, το σύστημα θα εμφάνιζε ευσταθέστερη συμπεριφορά, καθώς κάποιοι ταξινομητές αντιμετωπίζουν δυσκολίες σε συγκεκριμένα συνολα δεδομένων.

Παραδείγματα συνδυασμού ταξινομητών είναι το σχήμα ψηφοφορίας ή το σχήμα συνένωσης με τριγωνικές νόρμες. Στο σχήμα ψηφοφορίας, κάθε ταξινομητής αποφασίζει μεμονωμένα και ψηφίζει την κλάσης πολικότητας που θεωρεί πιο πιθανή, ενώ το μήνυμα κατατάσσεται στην πολικότητα που συγκέντρωσε τις περισσότερες ψήφους. Στο σχήμα συνένωσης με τριγωνικές νόρμες, δύο ταξινομητές είναι θετικά συσχετισμένοι όταν κατηγοριοποιούν λανθασμένα στην ίδια κατηγορία, ενώ είναι αρνητικά συσχετισμένοι όταν κατηγοριοποιούν λανθασμένα σε διαφορετικές κατηγορίες.

Επίσης, στα πλαίσια της Ανάλυσης Συναισθήματος σε Κοινωνικά Δίκτυα με την πλατφόρμα LeanBigData που αναφέραμε στην εισαγωγή του Κεφαλαίου 4, θα είχε ιδιαίτερο ενδιαφέρον η ένταξη της διαδικτυακής υπηρεσίας που αναπτύξαμε στην εν λόγω πλατφόρμα. Η ένταξη μίας υπηρεσία ανάλυσης συναισθήματος στην πλατφόρμα LeanBigData σκιαγραφείται στο ακόλουθο σχήμα που παρατίθεται στην [59]. Η υπηρεσία ονομάζεται Sentiment Analyzer.



Σχήμα 5.1: Η Αρχιτεκτονική της πλατφόρμας LeanBigData

Η συλλογή των δεδομένων γίνεται με χρήση των Twitter Search και Stream APIs από το σύμπλεγμα Συλλογής Δεδομένων (Data Collection). Τα δεδομένα δίνονται ως είσοδος στο σύμπλεγμα της Ανάλυσης σε Πραγματικό Χρόνο (Real-Time Analysis). Η υπηρεσία Sentiment Analyzer καλείται να προσφέρει Ανάλυση Συναισθήματος των δεδομένων σε πραγματικό χρόνο.

Με σκοπό την ένταξη της υπηρεσίας που αναπτύξαμε στην πλατφόρμα LeanBigData, θα χρειαστεί να διενεργήσουμε επιπλέον πειράματα ώστε να βελτιώσουμε την ακρίβεια και την απόδοση της υπηρεσίας.

Τέλος, η μελέτη της προσέγγισης της μεθόδου ν-γραμμάτων για ανάλυση συναισθήματος επικεντρώθηκε σε δεδομένα προερχόμενα από το κοινωνικό δίκτυο Twitter. Θα είχε ενδιαφέρον η εξέταση της συμπεριφοράς του μοντέλου σε δεδομένα από διαφορετικά κοινωνικά δίκτυα. Για

παράδειγμα, θα μπορούσε να πραγματοποιηθεί μία μελέτη σε αναρτήσεις ή σχόλια στο Facebook, το Instagram ή το Pinterest, καθώς επίσης σε σχόλια στο YouTube, αλλά και σε κριτικές προϊόντων στο Amazon ή το eBay.

Βιβλιογραφία

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media, pages 30–38. Association for Computational Linguistics, 2011.

[2] Fotis Aisopos, George Papadakis, Konstantinos Tserpes, and Theodora Varvarigou. Content vs. context for sentiment analysis: a comparative analysis over microblogs. In Proceedings of the 23rd ACM conference on Hypertext and social media, pages 187–196. ACM, 2012.

[3] Fotis Aisopos, George Papadakis, Konstantinos Tserpes, and Theodora A. Varvarigou. Textual and contextual patterns for sentiment analysis over microblogs. In Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume), pages 453–454, 2012.

[4] Nathan Aston, Timothy Munson, Jacob Liddle, Garrett Hartshaw, Dane Livingston, and Wei Hu. Sentiment analysis on the social networks using stream algorithms. Journal of Data Analysis and Information Processing, 2(02):60, 2014.

[5] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44. Association for Computational Linguistics, 2010.

[6] Adam Bermingham and Alan F Smeaton. Classifying sentiment in microblogs: is brevity an advantage? In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 1833–1836. ACM, 2010.

[7] Albert Bifet and Eibe Frank. Sentiment knowledge discovery in twitter streaming data. In Discovery Science, pages 1–15. Springer, 2010.

[8] Piero Bonissone, Kai Goebel, and Weizhong Yan. Classifier fusion using triangular norms. In Multiple Classifier Systems, pages 154–163. Springer, 2004.

[9] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 241–249. Association for Computational Linguistics, 2010.

[10] Christiane Fellbaum. WordNet. Wiley Online Library, 1998.

[11] Daniel Gayo-Avello, Panagiotis Takis Metaxas, and Eni Mustafaraj. Limits of electoral predictions using twitter. In ICWSM, 2011.

[12] George Giannakopoulos, Vangelis Karkaletsis, George A. Vouros, and Panagiotis Stamatopoulos. Summarization system evaluation revisited: N-gram graphs. TSLP, 5(3), 2008.

[13] George Giannakopoulos and Themis Palpanas. Content and type as orthogonal modeling features: a study on user interest awareness in entity subscription services. International Journal On Advances in Networks and Services, 3(1 and 2):296–309, 2010.

[14] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, pages 1–12, 2009.

[15] Tobias Günther and Lenz Furrer. Gu-mlt-lt: Sentiment analysis of short messages using linguistic features and stochastic gradient descent. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 328–332, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[16] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In
Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’04, pages 168–177, New York, NY, USA, 2004. ACM.

[17] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Unsupervised sentiment analysis with emotional signals. In Proceedings of the 22nd international conference on World Wide Web, pages 607–618. International World Wide Web Conferences Steering Committee, 2013.

[18] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target- dependent twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 151–160. Association for Computational Linguistics, 2011.

[19] Soo-Min Kim and Eduard Hovy. Identifying and analyzing judgment opinions. In Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL '06, pages 200–207, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[20] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! ICWSM, 11:538–541, 2011.

[21] Akshi Kumar and Teeja Mary Sebastian. Sentiment analysis on twitter. IJCSI International Journal of Computer Science Issues, 9(3):372–378, 2012.

[22] Bing Liu. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[23] Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. Emoticon smoothed language models for twitter sentiment analysis. In AAAI, 2012.

[24] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. CoRR, abs/1308.6242, 2013.

[25] Francisco Moreno-Seco, José M Inesta, Pedro J Ponce De León, and Luisa Micó. Comparison of classifier fusion methods for classification in pattern recognition tasks. In Structural, Syntactic, and Statistical Pattern Recognition, pages 705– 713. Springer, 2006.

[26] Alessandro Moschitti. Efficient convolution kernels for dependency and constituent syntactic trees. In Machine Learning: ECML 2006, pages 318–329. Springer, 2006.

[27] Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. 2013.

[28] Sascha Narr, Michael Hülfenhaus, and Sahin Albayrak. Language-independent twitter sentiment analysis. In KDML workshop on knowledge discovery, data mining and machine learning, 2012.

[29] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. ICWSM, 11:122–129, 2010.

[30] Reynier Ortega, Adrian Fonseca, and Andrés Montoyo. Ssa-uo: unsupervised twitter sentiment analysis. In Second Joint Conference on Lexical and Computational Semantics (* SEM), volume 2, pages 501–507, 2013.

[31] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In LREC, 2010.

[32] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics, page 271. Association for Computational Linguistics, 2004.

[33] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2):1–135, 2008.

[34] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL- 02 conference on Empirical methods in natural language processing-Volume 10, pages 79–86. Association for Computational Linguistics, 2002.

[35] Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 181–189. Association for Computational Linguistics, 2010.

[36] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada, 2005.

[37] Eshrag Refaee and Verena Rieser. Subjectivity and sentiment analysis of arabic twitter feeds with limited resources. In Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme, page 16, 2014.

[38] Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. Semeval-2014 task 9: Sentiment analysis in twitter. Proc. SemEval, 2014.

[39] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. Senticircles for contextual and conceptual semantic sentiment analysis of twitter. In 11th Extended Semantic Web Conference ESWC2014, 2014.

[40] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. Senticircles for contextual and conceptual semantic sentiment analysis of twitter. In The Semantic Web: Trends and Challenges, pages 83–98. Springer, 2014.

[41] Hassan Saif, Yulan He, and Harith Alani. Semantic smoothing for twitter sentiment analysis. In Proceedings of the 10th International Semantic Web Conference (ISWC), 2011.

[42] Hassan Saif, Yulan He, and Harith Alani. Alleviating data sparsity for twitter sentiment analysis. Making Sense of Microposts (# MSM2012), 2012.

[43] Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. The Semantic Web–ISWC 2012, pages 508–524, 2012.

[44] Fabrizio Sebastiani. Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1):1–47, 2002.

[45] David A Shamma, Lyndon Kennedy, and Elizabeth F Churchill. Tweet the debates: understanding community annotation of uncollected sources. In Proceedings of the first SIGMM workshop on Social media, pages 3–10. ACM, 2009.

[46] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In Proceedings of the First workshop on Unsupervised Learning in NLP, pages 53–63. Association for Computational Linguistics, 2011.

[47] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, 63(1):163–173, 2012.

[48] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, 63(1):163–173, 2012.

[49] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12):2544–2558, 2010.

[50] Erik Tromp. Multilingual sentiment analysis on social media. Master's thesis, Eindhoven University of Technology, Eindhoven, 7 2011.

[51] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 417–424. Association for Computational Linguistics, 2002.

[52] Julio Villena Román, Sara Lana Serrano, Eugenio Martínez Cámara, and José Carlos González Cristóbal. Tass-workshop on sentiment analysis at sepln. 2013.

[53] Cynthia Whissell, Michael Fournier, René Pelland, Deborah Weir, and Katherine Makarec. A dictionary of affect in language: Iv. reliability, validity, and applications. Perceptual and Motor Skills, 62(3):875–888, 1986.

[54] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. Language resources and evaluation, 39(2- 3):165–210, 2005.

[55] Ian H. Witten, Eibe Frank, and Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[56] Ley Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. HP Laboratories, Technical Report HPL-2011, 89, 2011.

[57] Ιωάννης Βλαχάβας, Πέτρος Κεφαλάς, Νικόλαος Βασιλειάδης, Φώτης Κόκκορας, and Ηλίας Σακελλαρίου. Τεχνητή Νοημοσύνη. Εκδόσεις Πανεπιστημίου Μακεδονίας, Θεσσαλονίκη, 3rd edition, 2006.

[58] Δημήτριος Μ. Τζαννέτος. Ανάλυση Συναισθήματος σε Δεδομένα Κοινωνικών Δικτύων με χρήση Γράφων ν-γραμμάτων. Εθνικό Μετσόβιο Πολυτεχνείο. Αθήνα, Οκτώβριος 2014.

[59] Ricardo Jimenez, Marta Patino, Valerio Vianello, Ivan Brondino, Ricardo Vilaca, Jorge Teixeira, Miguel Biscaia, Giannis Drossis, Damien Michel, Chryssi Birliraki, George Margetis, Antonis

Argyros, Constantine Stephanidis, Luigi Sgaglione, Gaetano Papale, Giavanni Mazzeo, Ferdinando Campanile, Marc Sole, Victor Munt´es-Mulero, David Solans, Alberto Huelamo, Pavlos Kranas, Dora Varvarigou, Vrettos Moulos and Fotis Aisopos. Scalable and Efficient Big Data Analytics: The LeanBigData Approach.LeanXcale, Universidad Politecnica de Madrid, Altice Labs, Institute of Computer Science, Foundation for Research and Technology Hellas & Computer Science Department, University of Crete, University of Naples ”Parthenope”, Sync Lab srlSync Lab srl, CA Technologies, National Technical University Of Athens & ICCS, 2017.

Παράρτημα

Αρχικοποίηση υπηρεσίας

Service initialization has been succesful.

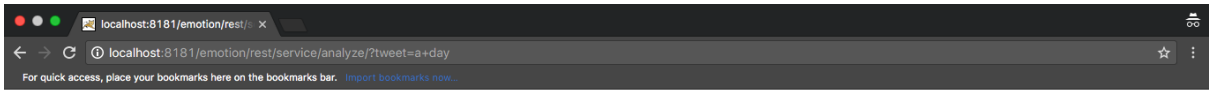
Παράδειγμα υποβολής θετικού μηνύματος: what a lovely day

Tweet sentiment is positive.

Παράδειγμα υποβολής αρνητικού μηνύματος: what a bad day

Tweet sentiment is negative.

Παράδειγμα υποβολής ουδέτερου μηνύματος: a day



Tweet sentiment is neutral.