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

Ανάλυση Συναισθήματος σε κείμενο από tweets με Μεθόδους Μη Επιβλεπόμενης Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Θεοφίλη Σ. Ασπρογέρακα

Επιβλέπων : Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2015

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

Ανάλυση Συναισθήματος σε κείμενο από tweets με Μεθόδους Μη Επιβλεπόμενης Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Θεοφίλη Σ. Ασπρογέρακα

Επιβλέπων : Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 27η Μαρτίου 2015.

.....	.....	.....
Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.	Ανδρέας- Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2015

.....
Θεοφίλη Ασπρογέρακα

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

Περίληψη

Αντικείμενο της παρούσας διπλωματικής είναι η Ανάλυση Συναισθήματος σε κείμενο μικρού μήκους που προέρχεται από δικτυακές πηγές, με χρήση Μηχανικής Μάθησης. Το σύνολο δεδομένων που χρησιμοποιείται αποτελείται από δημοσιεύσεις του ιστοχώρου κοινωνικής δικτύωσης www.twitter.com, οι οποίες έχουν κοινό θέμα τον τουρισμό στην Ελλάδα. Στόχος μας είναι η κατάταξη των δημοσιεύσεων (tweets) στις εξής τέσσερις κλάσεις συναισθήματος: θετική (positive) κλάση, αρνητική (negative) κλάση, ουδέτερη (neutral) κλάση, απροσδιόριστη (undefined) κλάση με τη βοήθεια Μη Επιβλεπόμενης Μηχανικής Μάθησης.

Αρχικά, γίνεται μια εισαγωγή στην έννοια της Ανάλυσης Συναισθήματος και ακολούθως γίνεται εκτενής αναφορά στις δύο βασικές μεθόδους που εφαρμόζονται στα προβλήματα Ανάλυσης Συναισθήματος. Αυτές είναι η βασισμένη σε λεξικό μέθοδος και η βασισμένη σε Μηχανική Μάθηση μέθοδος, με το ενδιαφέρον μας να επικεντρώνεται στη δεύτερη. Στο τμήμα που αναφέρεται στην βασισμένη σε λεξικό μέθοδο γίνεται περιγραφή της μεθόδου, των δυσκολιών και των τρόπων αντιμετώπισης τους και τέλος περιγράφεται η δημιουργία διαδεδομένων Λεξικών Συναισθήματος και γίνεται αναφορά σε σημαντικές σχετικές δημοσιεύσεις. Στο τμήμα που αφορά τη μέθοδο που βασίζεται στη Μηχανική Μάθηση γίνεται περιγραφή των κατηγοριών Μηχανικής Μάθησης γενικά και ειδικά για το πρόβλημα της Ανάλυσης Συναισθήματος. Αναλύονται τα στάδια και οι παράμετροι της μεθόδου και περιγράφονται βασικοί αλγόριθμοι για κάθε μία από τις κατηγορίες Μηχανικής Μάθησης στην Ανάλυση Συναισθήματος. Ακολούθως, γίνεται αναφορά στα προβλήματα της μεθόδου, στους τρόπους αξιολόγησης ενός συστήματος Ανάλυσης Συναισθήματος και σε διάφορα διαθέσιμα datasets. Τέλος, γίνεται ειδική αναφορά στη διαχείριση των tweets και κλείνουμε με μια σύντομη περιγραφή σχετικών δημοσιεύσεων και με την παράθεση διαθέσιμων APIs Ανάλυσης Συναισθήματος.

Στη συνέχεια, έχοντας πλέον περιγράψει θεωρητικά όλα τα απαραίτητα στάδια και παραμέτρους εστιάζουμε στην εφαρμογή της μεθόδου Μηχανικής Μάθησης στα δεδομένα μας. Δοκιμάζουμε διάφορες εκδοχές προεπεξεργασίας των δεδομένων, διαφορετικούς αλγορίθμους και μεταβολή διάφορων παραμέτρων του προβλήματος. Έμφαση δόθηκε στην Μη Επιβλεπόμενη Μηχανική Μάθηση (συσταδοποίηση), ωστόσο για βαθύτερη κατανόηση και μελέτη των αποτελεσμάτων έγιναν δοκιμές και με τη μέθοδο Επιβλεπόμενης Μάθησης.

Μελετώντας με προσοχή τα αποτελέσματα που προκύπτουν από τις διάφορες δοκιμές, συμπεραίνουμε ότι η συσταδοποίηση δεν μπορεί να δώσει ικανοποιητικά αποτελέσματα στο πρόβλημα της Ανάλυσης Συναισθήματος σε tweets λόγω του ότι η μέθοδος προσπαθεί να εντοπίσει ομοιότητες μεταξύ των tweets αγνοώντας εντελώς τις διαφορές που ενδεχομένως να είναι πιο σημαντικές. Ωστόσο, μέσα από τη μελέτη των αποτελεσμάτων καταλήγουμε σε σημαντικά ευρήματα που θα μπορούσαν να φανούν χρήσιμα σε μελλοντικές έρευνες.

Λέξεις κλειδιά

Μηχανική Μάθηση, Ανάλυση Συναισθήματος, Συσταδοποίηση, k-means, Expectation-Maximization, Tweets, κλάσεις συναισθήματος

Abstract

The object of this dissertation is Sentiment Analysis of short texts from web sources, using Machine Learning. The dataset that is used consists of posts from the social networking site www.twitter.com and their common subject is tourism in Greece. Our goal is to classify the posts (tweets) in four classes of sentiment: positive class, negative class, neutral class, undefined class using Unsupervised Machine Learning.

In the beginning, we make an introduction to the concept of Sentiment Analysis and then we make an extensive reference to the two basic methods that are used in the Sentiment Analysis problems. These are the lexicon - based method and the Machine-learning based method, while our interest focuses on the latter. In the part that refers to the lexicon-based method we describe the method, the difficulties and the ways to deal with them and finally, we describe the creation of some common Sentiment Lexicons. In the part that refers to the Machine Learning - based method we make a description of Machine Learning categories, in general as well as specifically for the problem of Sentiment Analysis. We analyze method's steps and parameters and we describe basic algorithms that are used in each Machine Learning category for Sentiment Analysis. Subsequently, we make reference to the method's problems, to the methods that are used for the evaluation of a system that performs Sentiments Analysis and to some available datasets. Finally, we make special reference to the management of tweets and we close this part giving brief descriptions of relevant publications and mentioning available APIs for Sentiment Analysis.

Consequently, since we have described theoretically all the steps and parameters we focus on the application of Machine Learning on our data. We test several versions of data preprocessing, several algorithms and variation of problem parameters. We paid attention to Unsupervised Machine Learning (clustering). However we also applied the method of Supervised Machine Learning for better understanding and study of the results.

Studying carefully the results we obtained from the several tests, we conclude that clustering cannot give satisfying results to the problem of Sentiment Analysis of tweets because the method tries to find similarities between the tweets, ignoring differences that may be more important. However, through the study of the results we arrive at findings that may be useful in future surveys.

Keywords

Machine Learning, Sentiment Analysis, Clustering, k-means, Expectation-Maximization, tweets, classes of sentiment

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον Καθηγητή Ε.Μ.Π. κύριο Στέφανο Κόλλια για την εμπιστοσύνη που μου έδειξε δίνοντας μου την ευκαιρία να ασχοληθώ με ένα τόσο ενδιαφέρον αντικείμενο και για τις πολύτιμες συμβουλές του.

Επίσης, θα ήθελα να ευχαριστήσω το Διευθυντή Ερευνών Ε.Π.Ι.Σ.Ε.Υ. – Ε.Μ.Π. κύριο Κώστα Καρπούζη για την βοήθεια και την καθοδήγηση του κατά την εκπόνηση της διπλωματικής εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου, που είναι πάντα δίπλα μου και με στηρίζει σε κάθε μου βήμα.

Περιεχόμενα

1. Εισαγωγή

1.1 Εισαγωγή στην έννοια της Ανάλυσης Συναισθήματος

Η αλματώδης εξέλιξη του Web 2.0 την τελευταία δεκαπενταετία έχει αλλάξει εντυπωσιακά τον τρόπο επικοινωνίας των ανθρώπων μέσα από την χρήση υπηρεσιών και εφαρμογών. Το Web 2.0, τοποθετώντας το χρήστη του Διαδικτύου στο επίκεντρο, του δίνει τη δυνατότητα να μοιράζεται πληροφορίες, να αλληλεπιδρά με άλλους χρήστες, να σχολιάζει και να φιλτράρει κάθε είδους πληροφορία με χρήση εργαλείων όπως το search (αναζήτηση), το tag (ετικέτα), η παράθεση links ή το authoring (συγγραφή περιεχομένου). Τα παραπάνω εργαλεία και δυνατότητες εφαρμόζονται καθημερινά από δισεκατομμύρια χρήστες παγκοσμίως, με αποτέλεσμα τη δημιουργία ενός τεράστιου όγκου πληροφορίας. Η εξερεύνηση του περιεχομένου και η αξιοποίηση αυτής της πληροφορίας αποτελεί αντικείμενο έρευνας τα τελευταία χρόνια και έχει οδηγήσει μεταξύ των άλλων στην ανάπτυξη της έννοιας της Ανάλυσης Συναισθήματος.

Με τον όρο Ανάλυση Συναισθήματος (Sentiment Analysis) ή αλλιώς Εξόρυξη Γνώμης (Opinion Mining) εννοούμε την διαδικασία αυτόματου προσδιορισμού του συναισθήματος που εκφράζεται από ένα άτομο ως προς κάποιο αντικείμενο μέσω ενός κειμένου γραμμένου από τον ίδιο, με τη βοήθεια μεθόδων Επεξεργασίας Φυσικής Γλώσσας (NLP), Στατιστικής και Μηχανικής Μάθησης. Στόχος του προβλήματος της Ανάλυσης Συναισθήματος σε κείμενο είναι η αυτόματη κατάταξη του προς ανάλυση κειμένου σε μία από τις προβλεπόμενες από το εκάστοτε πρόβλημα κλάσεις συναισθήματος.

Αν και οι πρώτες προσεγγίσεις της Ανάλυσης Συναισθήματος έγιναν γύρω στο 1997 [59] [62], το πρόβλημα έγινε ευρύτερα δημοφιλές στις αρχές του 21^{ου} αιώνα, με τους Bo Pang, Lilian Lee [18] και Peter Turney [69] να δημοσιεύουν έρευνες που αποτελούν μέχρι και σήμερα σημείο αναφοράς στην μελέτη του προβλήματος της Ανάλυσης Συναισθήματος. Έκτοτε η έρευνα πάνω στο αντικείμενο συνεχίζεται με όλο και αυξανόμενους ρυθμούς, καθώς το διαδίκτυο ως μέσο έκφρασης της ανθρώπινης άποψης καθιερώνεται όλο και περισσότερο.

1.2 Χρησιμότητα και εφαρμογές της Ανάλυσης Συναισθήματος

Πέρα από το ότι αποτελεί ένα πολύ ενδιαφέρον αντικείμενο μελέτης και έρευνας, η αυτόματη Ανάλυση Συναισθήματος σε κείμενο μπορεί να βρει πολλές χρήσιμες εφαρμογές. Το καινούριο που έχει να προσφέρει στις ήδη υπάρχουσες πληροφορίες που βρίσκονται στο διαδίκτυο είναι ότι ουσιαστικά μπορεί να παράγει μια περίληψη του συναισθήματος που εκφράζεται από τα δεδομένα που αναλύονται, κάνοντας πολύ απλό και γρήγορο στον εκάστοτε ενδιαφερόμενο να διαμορφώσει μια εικόνα χωρίς να χρειαστεί να μελετήσει ο ίδιος αναλυτικά την κάθε άποψη που εκφράζεται σε κάποιο δικτυακό τόπο, γύρω από

οποιοδήποτε αντικείμενο. Τα τελευταία χρόνια πολλές είναι μάλιστα οι εταιρείες που αναλαμβάνουν το έργο της Ανάλυσης Συναισθήματος στο διαδίκτυο, σχετικά με κάποιο αντικείμενο που τους ανατίθεται από τους πελάτες τους, ενώ αμέτρητα είναι και τα εργαλεία αυτόματης Ανάλυσης Συναισθήματος που έχουν αναπτυχθεί και είναι στη διάθεση οποιουδήποτε απλού χρήστη του διαδικτύου. Η Ανάλυση Συναισθήματος σε κείμενο διαδικτυακών πηγών, χρησιμοποιείται από πολλές εταιρείες που επιθυμούν να ερευνήσουν την αποδοχή των προϊόντων τους από το ευρύ κοινό, να σχεδιάσουν τις μελλοντικές κινήσεις τους και να βελτιώσουν τυχόν ατέλειες τους. Επίσης, δημοφιλή πρόσωπα αξιοποιούν την Ανάλυση Συναισθήματος διαδικτυακής πληροφορίας για να μάθουν τη γνώμη του κοινού προς το πρόσωπο τους, ενώ απλοί χρήστες του διαδικτύου θα μπορούσαν με τη βοήθεια της να διαμορφώσουν μια γενική εικόνα για μια ταινία, κάποιο προϊόν, κάποιο πρόσωπο ή κάποιον προορισμό που τους ενδιαφέρει και τυχαίνει να έχει αξιολογηθεί από πολλούς άλλους χρήστες. Έχουν γίνει προσπάθειες ακόμα και για πρόβλεψη αποτελεσμάτων πολιτικών εκλογών με χρήση Ανάλυσης Συναισθήματος.

1.3 Επίπεδα Ανάλυσης Συναισθήματος

Η Ανάλυση Συναισθήματος γίνεται συνήθως σε κάποιο από τα εξής επίπεδα:

- *Επίπεδο κειμένου (document- level)*: Σε αυτό το επίπεδο θεωρούμε ότι το κείμενο εκφράζει μία ενιαία άποψη ενός ατόμου για ένα αντικείμενο, την οποία προσπαθούμε να προσδιορίσουμε. Εφαρμόζεται σε περιπτώσεις κειμένων από τα οποία μας ενδιαφέρει να διαμορφώσουμε μια σφαιρική γνώμη για κάποιο αντικείμενο (π.χ. κριτικές ταινιών), ενώ δεν ενδείκνυται για περιπτώσεις κειμένων μέσα στα οποία σχολιάζονται διαφορετικές πτυχές ενός αντικειμένου ή διαφορετικά αντικείμενα.

- *Επίπεδο πρότασης (sentence- level)*: Σε αυτό το επίπεδο, θεωρείται ότι η κάθε πρόταση ενός κειμένου εκφράζει ένα διαφορετικό συναίσθημα.

- *Επίπεδο χαρακτηριστικού (feature-level)*: Τόσο η ανάλυση σε επίπεδο κειμένου όσο και η ανάλυση σε επίπεδο πρότασης δεν ανακαλύπτουν το ακριβές αντικείμενο του κάθε συναισθήματος. Με την ανάλυση σε επίπεδο χαρακτηριστικού επιδιώκεται ο προσδιορισμός τόσο του συναισθήματος όσο και του συγκεκριμένου αντικειμένου - στόχου του συναισθήματος. Η ανάλυση αυτή περιλαμβάνει τα εξής στάδια: Αρχικά, προσδιορίζονται και εξάγονται τα επιμέρους χαρακτηριστικά του σχολιασμένου αντικειμένου. Στη συνέχεια προσδιορίζεται το συναίσθημα που εκφράζεται για καθένα απ' αυτά και τέλος, αναζητούνται και συνδυάζονται συνώνυμα χαρακτηριστικά. Τελικός στόχος είναι η δημιουργία μιας περίληψης του συναισθήματος για κάθε χαρακτηριστικό. Αυτό το επίπεδο ανάλυσης ενδείκνυται για τη μελέτη λεπτομερούς σχολιασμού αντικειμένων, η οποία συνήθως γίνεται για

σύγκριση παρόμοιων αντικειμένων (π.χ. έρευνα αγοράς κάποιας ηλεκτρονικής συσκευής).

- *Επίπεδο οντότητας (entity - level)*: Η ανάλυση σε επίπεδο οντότητας είναι ανάλογη της ανάλυσης σε επίπεδο χαρακτηριστικού, με τη διαφορά ότι στην ανάλυση σε επίπεδο οντότητας επιδιώκεται ο εντοπισμός των διαφορετικών οντοτήτων που σχολιάζονται μέσα σε ένα κείμενο και η εξαγωγή των αντίστοιχων συναισθημάτων. Χαρακτηριστικό παράδειγμα εφαρμογής είναι τα ειδησεογραφικά άρθρα μέσα στα οποία συνήθως γίνεται λόγος για πολλές διαφορετικές οντότητες.

1.4 Αντικείμενο της διπλωματικής εργασίας

Αντικείμενο της παρούσας διπλωματική εργασίας είναι η μελέτη του προβλήματος της Ανάλυσης Συναισθήματος σε κείμενο μικρού μήκους, διαδικτυακής προέλευσης με χρήση Μη Επιβλεπόμενης Μηχανικής Μάθησης. Συγκεκριμένα θα ασχοληθούμε με την Ανάλυση Συναισθήματος σε δημοσιεύσεις χρηστών του κοινωνικού δικτύου Twitter (tweets), και θα επιχειρήσουμε την αυτόματη κατάταξη τους σε τέσσερις κλάσεις συναισθήματος (θετική, αρνητική, ουδέτερη, απροσδιόριστη) με τη βοήθεια της πλατφόρμας WEKA. Ο λόγος που επιλέξαμε το Twitter ως πηγή των δεδομένων μας είναι ότι οι δημοσιεύσεις του παρουσιάζουν ιδιαίτερο ενδιαφέρον λόγω του περιορισμένου μήκους τους (140 χαρακτήρες μέγιστο μήκος) αλλά και της απλής σύνταξης και ελεύθερης έκφρασης που συνήθως χρησιμοποιείται από τους χρήστες. Στα πλαίσια της εργασίας θα μελετήσουμε τη χρήση Μηχανικής Μάθησης, εστιάζοντας στη Μη Επιβλεπόμενη Μηχανική Μάθηση (clustering). Το σύνολο δεδομένων που χρησιμοποιείται περιλαμβάνει tweets με περιεχόμενο σχετικό με τον τουρισμό στην Ελλάδα, καθώς τα δεδομένα αντλήθηκαν με κριτήριο την αναφορά στο username @VisitGreecegr.

1.5 Διάρθρωση της διπλωματικής εργασίας

Στο επόμενο κεφάλαιο (Κεφάλαιο 2) θα γίνει μια θεωρητική εισαγωγή στις δύο βασικότερες κατηγορίες μεθόδων Ανάλυσης Συναισθήματος σε κείμενο. Θα μελετηθούν τα επιμέρους στάδια, οι παράμετροι και οι δυσκολίες του προβλήματος, ενώ θα αναλυθούν και κάποιοι βασικοί αλγόριθμοι για την μέθοδο που στηρίζεται στη Μηχανική Μάθηση. Ακόμη, θα αναφερθούμε στις ιδιαιτερότητες που εμφανίζει η ανάλυση των tweets και στις μεθόδους αξιολόγησης ενός συστήματος αυτόματης Ανάλυσης Συναισθήματος. Τέλος, θα γίνει αναφορά σε σημαντικές σχετικές έρευνες αλλά και σε διαθέσιμα APIs για Ανάλυση Συναισθήματος σε κείμενο.

Έχοντας πλέον χτίσει το απαραίτητο θεωρητικό υπόβαθρο, στο Κεφάλαιο 3 θα προχωρήσουμε στην εφαρμογή της μεθόδου Ανάλυσης Συναισθήματος με χρήση

Μηχανικής Μάθησης πάνω σε πραγματικά δεδομένα που προέρχονται από το Twitter εστιάζοντας κυρίως στη Μη Επιβλεπόμενη Μηχανική Μάθηση. Αφού γίνει περιγραφή του διαθέσιμου συνόλου δεδομένων και της πλατφόρμας στην οποία θα εκτελεστούν τα πειράματα, θα γίνει αναφορά των διάφορων δοκιμών που επιχειρήσαμε, παράθεση και σχολιασμός των σχετικών αποτελεσμάτων.

Στο Κεφάλαιο 4, για λόγους πληρότητας της μελέτης των αποτελεσμάτων του προηγούμενου κεφαλαίου θα μελετήσουμε την εφαρμογή μεθόδου Επιβλεπόμενης Μηχανικής Μάθησης σε ένα από τα προηγούμενα σύνολα δεδομένων.

Τέλος, στο Κεφάλαιο 5 θα κλείσουμε με κάποια συμπεράσματα σχετικά με το πρόβλημα της Ανάλυσης Συναισθήματος που μελετήθηκε και με προτάσεις για βελτίωση των αποτελεσμάτων σε μελλοντικές προσπάθειες.

2. Μέθοδοι Ανάλυσης Συναισθήματος σε κείμενο

Στο παρόν κεφάλαιο περιγράφονται οι δύο βασικές προσεγγίσεις του προβλήματος της Ανάλυσης Συναισθήματος σε κείμενο: Η βασισμένη σε λεξικό προσέγγιση (lexicon-based approach) και η βασισμένη σε Μηχανική Μάθηση προσέγγιση (machine learning – based approach).

2.1 Ανάλυση Συναισθήματος βασισμένη σε λεξικό

2.1.1 Περιγραφή της μεθόδου

Σύμφωνα με αυτή την προσέγγιση, το κείμενο στο οποίο πρόκειται να γίνει η συναισθηματική ανάλυση αντιμετωπίζεται σαν ένα σύνολο ανεξαρτήτων μεταξύ τους λέξεων, η σειρά και οι γραμματικές ιδιότητες των οποίων αγνοούνται. Αντιμετωπίζουμε δηλαδή το κείμενο σαν ένα σάκο με λέξεις (bag of words).

Για την απόδοση συναισθηματικού περιεχομένου στο κείμενο, γίνεται χρήση Λεξικών Συναισθήματος (sentiment lexicons). Τα λεξικά αυτά περιέχουν λέξεις που εκφράζουν συναίσθημα (sentiment words) στις οποίες έχουν αποδοθεί βαθμολογίες που εκφράζουν κατά πόσο το νόημα της λέξης ταιριάζει σε συγκεκριμένες κατηγορίες συναισθήματος. Συνήθως, οι κατηγορίες ως προς τις οποίες γίνεται η βαθμολόγηση στα λεξικά είναι οι δύο βασικές: θετικό συναίσθημα (positive sentiment) και αρνητικό συναίσθημα (negative sentiment), οπότε οι βαθμολογίες έχουν και το αντίστοιχο πρόσημο. Ωστόσο υπάρχουν και λεξικά με πιο εξειδικευμένες κατηγορίες συναισθημάτων, όπως χαρά, έκπληξη, λύπη, θυμός κ.α.

Κάθε λέξη του προς ανάλυση κειμένου αναζητείται στο Λεξικό Συναισθήματος και σημειώνεται η βαθμολογία της. Σε περίπτωση που το λεξικό που χρησιμοποιείται περιλαμβάνει πολλές κατηγορίες συναισθημάτων, μπορούμε να τις αντιστοιχήσουμε στις επιθυμητές για την ανάλυσης μας κατηγορίες (π.χ. positive, negative, high positive, high negative). Τέλος, το συνολικό συναίσθημα του κειμένου προσδιορίζεται από το άθροισμα των βαθμολογιών των επιμέρους λέξεων, και με τη βοήθεια κατωφλίων (thresholds) σε περίπτωση πολυεπίπεδης συναισθηματικής κατάταξης.

2.1.2 Αιτίες αστοχίας μεθόδου

Η παραπάνω μέθοδος Ανάλυσης Συναισθήματος, οδηγεί σε αποτελέσματα που είναι πιθανό να απέχουν αρκετά από την αντίστοιχη ανθρώπινη ερμηνεία. Αυτό συμβαίνει επειδή η μέθοδος δεν λαμβάνει υπόψη την αλληλεπίδραση των λέξεων, η οποία μπορεί να τροποποιήσει σημαντικά το νόημα μιας πρότασης. Ακολουθούν στοιχεία του κειμένου, τα

οποία αν δε ληφθούν υπόψη κατά την ανάλυση μπορεί να οδηγήσουν σε αστοχία της μεθόδου:

- ΑΡΝΗΣΗ
  Σημαντικός παράγοντας που αγνοείται από τη μέθοδο είναι η ύπαρξη λέξεων άρνησης (π.χ. όχι, δεν, ούτε) σε μια πρόταση. Η ύπαρξη κάποιας λέξης άρνησης επηρεάζει το νόημα μίας ή και περισσοτέρων λέξεων που ακολουθούν. Για παράδειγμα, μια σειρά αρνητικών λέξεων, μπορεί να έχει συνολικά θετικό νόημα, αλλά και μια λέξη με θετικό νόημα μπορεί να αντιστραφεί αν προηγείται κάποια αρνητική λέξη. Επομένως, η αντιμετώπιση λέξεων άρνησης ατομικά μόνο εισάγει σημαντικά λάθη στην ανάλυση.

  π.χ. *1. Not an inhuman monster*
  2. Not good

- ΛΕΞΕΙΣ ΕΝΤΑΣΗΣ (INTENSIFIERS)
  Οι Intensifiers είναι λέξεις που αυξάνουν ή μειώνουν την ένταση της λέξης που τις ακολουθεί (π.χ. πολύ, πραγματικά, απίστευτα, απερίγραπτα, λίγο, ελάχιστα), δρουν δηλαδή είτε ενισχυτικά είτε αποσβεστικά. Η ύπαρξη και ο συνυπολογισμός τους στην Ανάλυση Συναισθήματος είναι σημαντικά σε περίπτωση που επιδιώκουμε πολυεπίπεδη συναισθηματική κατάταξη.

  π.χ. *1. Truly terrible*
  2. Slightly ugly

- ΣΕΙΡΑ ΛΕΞΕΩΝ
  Η σειρά των λέξεων σε μία πρόταση πολλές φορές παίζει καθοριστικό ρόλο στο τελικό νόημα που αποδίδεται, αφού μπορεί ακόμα και να αντιστρέψει την συναισθηματική πολικότητα.

  π.χ. *That's true, I am not a fan of Pink Floyd.*
  That's not true, I am a fan of Pink Floyd.

- ΣΕΙΡΑ ΦΡΑΣΕΩΝ
  Σημαντικό ρόλο στον προσδιορισμό του συνολικού συναισθηματικού περιεχομένου στην πρόταση παίζει και η σειρά των επιμέρους τμημάτων της, αν αυτά εκφράζουν αντίθετα συναισθήματα, αφού κατά την ανάγνωση από άνθρωπο συνήθως είτε υπερισχύει το συναίσθημα της τελευταίας φράσης, είτε η εντύπωση που μένει εξαρτάται από την κρίση του αναγνώστη.

  π.χ. *Beautifully filmed and well-acted, but hollow in its narratives*

- ΙΔΙΩΜΑΤΙΣΜΟΙ

  Αντιμετωπίζοντας την κάθε λέξη ξεχωριστά αγνοούμε το σχηματισμό φράσεων που αποδίδουν διαφορετικό νόημα από αυτό που αποδίδουν οι λέξεις αν μελετηθούν χωριστά (ιδιωματισμοί).

  π.χ. *Castles in the air* (= Plans that are impractical and will never work out are castles in the air.)

- ΠΟΛΛΑΠΛΟΙ ΣΤΟΧΟΙ

  Ένα ακόμη στοιχείο το οποίο αγνοεί η παραπάνω προσέγγιση είναι η έννοια του ‘στόχου’ (target) του κειμένου. Με τον όρο “στόχος” εννοούμε την οντότητα (πρόσωπο, πράγμα, κατάσταση) για την οποία γίνεται λόγος στο υπό ανάλυση κείμενο. Σε ορισμένες μορφές κειμένου (π.χ. τα άρθρα ειδήσεων) ο στόχος δεν είναι τόσο εμφανής όσο σε άλλες (π.χ. κριτική προϊόντος) με αποτέλεσμα να αποδίδεται στον στόχο συναίσθημα που έχει αλλοιωθεί από λέξεις του κειμένου που δεν αναφέρονται σε αυτόν, αλλά σε άλλους δευτερεύοντες στόχους που δεν είναι αντικείμενα της ανάλυσης.

  π.χ. 1. *“Dear <hardware store>, Yesterday I visited <your competitor>. They had an excellent selection, friendly and helpful salespeople, and the lowest prices in town. I hate you. Sincerely, a customer”*

  2. *Nokia is better than Sony.*

- ΕΙΡΩΝΕΙΑ

  Ο ανθρώπινος λόγος πολλές φορές εμπεριέχει ειρωνεία, η οποία ξεγελά ως προς το συναίσθημα του συγγραφέα του κειμένου αν δε γίνει αντιληπτή, πράγμα που συμβαίνει σε περίπτωση ανάλυσης από αυτόματους μηχανισμούς, αφού διακρίνεται μόνο αν διαβαστεί ως σύνολο λέξεων.

  π.χ. *The wine was as delicious as a glass full of vinegar.*

- ΝΟΗΜΑΤΙΚΟΣ ΤΟΜΕΑΣ ΤΟΥ ΚΕΙΜΕΝΟΥ

  Πέρα από τις λέξεις που διατηρούν ένα συγκεκριμένο συναισθηματικό περιεχόμενο ανεξαρτήτως νοηματικού πλαισίου, υπάρχουν πολλές λέξεις οι οποίες δεν έχουν σταθερό συναισθηματικό περιεχόμενο. Επομένως, Λεξικά Συναισθήματος που έχουν δημιουργηθεί για Ανάλυση Συναισθήματος σε κείμενα που ανήκουν σε οποιοδήποτε τομέα σημειώνουν ως ουδέτερες λέξεις των οποίων το περιεχόμενο ίσως και να μην ήταν ουδέτερο αν γνωρίζαμε τον νοηματικό πλαίσιο του κειμένου.

π.χ. Η λέξη unpredictable (απρόβλεπτος) έχει θετικό νόημα όταν αναφέρεται στην πλοκή μιας ταινίας, ενώ έχει αρνητικό νόημα όταν περιγράφει την οδήγηση ενός αυτοκινήτου.

Σύμφωνα με όλα τα παραπάνω, καταλήγουμε στο συμπέρασμα ότι η Ανάλυση Συναισθήματος που βασίζεται σε λεξικό, αν και απλή στην κατανόηση και την υλοποίηση, μπορεί εύκολα να αστοχήσει λόγω της επιφανειακής προσέγγισης του κειμένου, κατά την οποία παραλείπεται η οποιαδήποτε προσπάθεια κατανόησης συνολικού νοήματος του.

2.1.3 Λύσεις – Βελτιώσεις

Λαμβάνοντας υπόψη τις δυσκολίες που αναφέρθηκαν παραπάνω, έχουν γίνει και συνεχίζουν να γίνονται αρκετές προσπάθειες αντιμετώπισης τους και βελτίωσης της βασισμένης σε λεξικό προσέγγισης.

Μια βασική προσπάθεια βελτίωσης των Λεξικών Συναισθήματος αφορά την επέκταση τους ώστε να περιλαμβάνουν διαδεδομένους ιδιωματισμούς με συναισθηματικό περιεχόμενο. Σημειώνουμε ως δυσκολία του εγχειρήματος ότι συνήθως οι ιδιωματισμοί δεν μεταφράζονται από γλώσσα σε γλώσσα, οπότε το λεξικό κάθε γλώσσας απαιτεί ξεχωριστή έρευνα και επεξεργασία.

Για την αξιοποίηση των λέξεων άρνησης έχουν γίνει διάφορες προσεγγίσεις. Μια από αυτές αρκείται στην αλλαγή του προσήμου της συναισθηματικής βαθμολογίας μιας λέξης η οποία έπεται μιας λέξης άρνησης (switch negation) [1]. Παρατηρώντας την αστοχία της παραπάνω προσέγγισης σε αρκετές περιπτώσεις (π.χ. προκύπτει not good > not excellent, ενώ ισχύει το αντίθετο) οι Ngoc και Yoo (2014) [2] και οι Taboada et al.(2011)[3] υπολογίζουν την τελική βαθμολογία της λέξης αφαιρώντας από την αρχική της βαθμολογία μια καθορισμένη τιμή που έχει αντιστοιχεί στην αρνητική λέξη που προηγείται (shift negation).

Σε πολλές δημοσιεύσεις οι Intensifiers χωρίζονται σε δύο κατηγορίες: τους amplifiers (ενισχυτές) και τους downtoners (εξομαλυντές). Πολλοί ερευνητές [4] [5] αξιοποιούν την ύπαρξη των Intensifiers με χρήση απλής πρόσθεσης και αφαίρεσης για τον υπολογισμό της βαθμολογίας της λέξης που συνοδεύουν. Η τεχνική αυτή όμως υστερεί καθώς η τιμή που προστίθεται ή αφαιρείται δε λαμβάνει υπόψη την λέξη στην οποία εφαρμόζεται ο εκάστοτε intensifier. Γι αυτό, οι Taboada et al.[3] πρότειναν την αντιστοίχηση κάποιου ποσοστού σε κάθε Intensifier, το οποίο προστίθεται ή αφαιρείται από το 100% και ακολούθως πολλαπλασιάζεται με την βαθμολογία της λέξης που προσδιορίζει ο Intensifier.

Η ειρωνεία είναι πολύ δύσκολο να εντοπιστεί αυτόματα, καθώς δεν έχει συγκεκριμένο ορισμό και δομή. Ο μόνος τρόπος εντοπισμού της είναι με τη βοήθεια των προτάσεων που την περιβάλλουν.

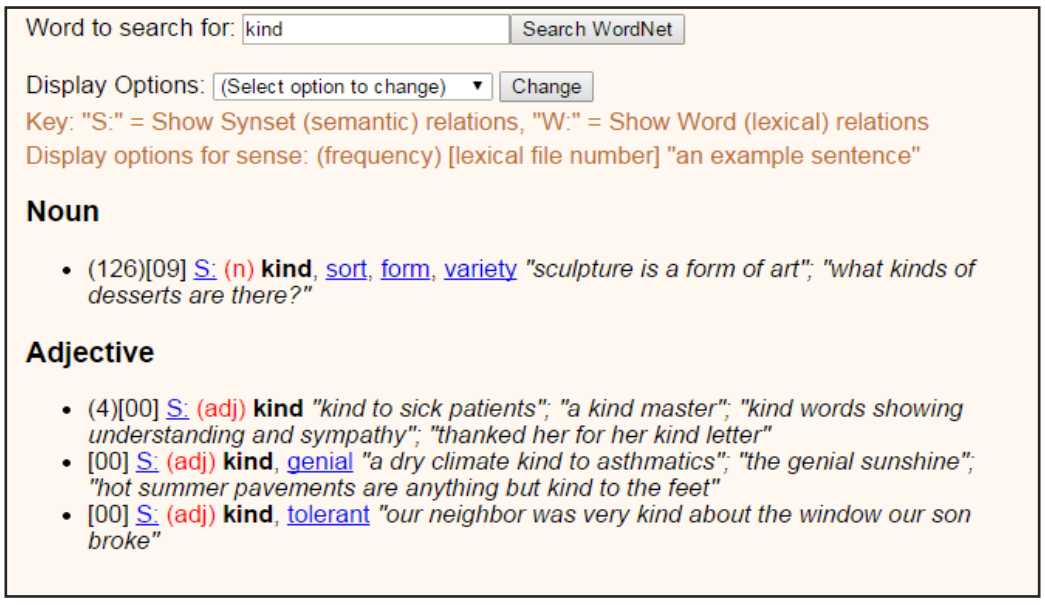
Η χρήση σημασιολογικών ρόλων (semantic roles), δηλαδή σχέσεων των οντοτήτων μιας πρότασης και του κύριου ρήματος της πρότασης, μπορεί να βοηθήσει στον εντοπισμό του "στόχου" του συναισθήματος που μας ενδιαφέρει.

Τέλος, αν και χρονοβόρο, είναι επιθυμητή η δημιουργία Λεξικού Συναισθήματος εξειδικευμένου στον νοηματικό τομέα στον οποίο ανήκει το κείμενο που επιδιώκουμε να αναλύσουμε για μεγαλύτερη ακρίβεια αποτελεσμάτων. Οι Muhammad et al. σε δημοσίευση τους [6] παρουσιάζουν μια τεχνική για δημιουργία Λεξικού Συναισθήματος που εστιάζει στα social media.

2.1.4 Τεχνικές δημιουργίας Λεξικών Συναισθήματος

Υπάρχουν διάφορες τεχνικές για την δημιουργία Λεξικών Συναισθήματος που περιγράφηκαν παραπάνω. Αναφέρουμε τις τεχνικές αυτές που έχουν εφαρμοστεί για την δημιουργία λεξικών σε διάφορες δημοσιεύσεις [27].

Σημαντικό ρόλο στη δημιουργία πολλών διαδεδομένων Λεξικών Συναισθήματος έπαιξε, όπως θα φανεί παρακάτω, το WordNet, το οποίο είναι μια λεξιλογική βάση δεδομένων για την Αγγλική γλώσσα που δημιουργήθηκε στο Πανεπιστήμιο Princeton το 1985. Ομαδοποιεί τις αγγλικές λέξεις σε σύνολα συνωνύμων (synsets), δίνει σύντομους ορισμούς τους και παραδείγματα χρήσης τους, και καταγράφει έναν αριθμό σχέσεων μεταξύ των συνόλων συνωνύμων ή των λέξεων που αυτά περιέχουν. Η πιο πρόσφατη έκδοση του (WordNet 3.1, 2012) περιλαμβάνει 155.287 λέξεις οργανωμένες σε 117.659 synsets και 206.941 σημασιολογικές σχέσεις μεταξύ synsets και μεταξύ λέξεων. Περιέχει ουσιαστικά, ρήματα, επίθετα και επιρρήματα όμως αγνοεί τις προθέσεις και τους προσδιορισμούς. Η δομή του φαίνεται στο ακόλουθο αποτέλεσμα αναζήτησης, η οποία έγινε μέσω του ιστοτόπου wordnetweb.princeton.edu/perl/webwn στον οποίο διατίθεται προς ελεύθερη χρήση το WordNet 3.1:



Εικόνα 1. Απόσπασμα Wordnet 3.1

Οι Hu and Liu [7] σε μια προσέγγιση τους για την Ανάλυση Συναισθήματος σε κριτικές πελατών για προϊόντα, ξεκινούν με ένα σύνολο από επίθετα - σπόρους (seed adjectives) (“good” και “bad”) και εφαρμόζουν σχέσεις αντωνυμίας και συνωνυμίας του WordNet για την συναισθηματική αξιολόγηση των λέξεων. Το τελικό αποτέλεσμα (Opinion Lexicon) είναι μια λίστα θετικών λέξεων και μια λίστα αρνητικών λέξεων (περίπου 6.800 λέξεις), στις οποίες σκόπιμα συμπεριλαμβάνονται και λέξεις με συνηθισμένα ορθογραφικά λάθη, ενώ δε δίνεται άλλη πληροφορία για την κάθε λέξη.

Μια παρόμοια προσέγγιση έγινε κατά τη δημιουργία του WordNet Affect από τους Strapparava και Valitutti [8] ξεκινώντας τη δημιουργία του από ένα μεγαλύτερο σύνολο συναισθηματικών λέξεων, ταξινομημένων σύμφωνα με τις έξι βασικές κατηγορίες συναισθήματος (χαρά, λύπη, φόβος, έκπληξη και αηδία) και επεκτείνοντας το με χρήση μονοπατιών από το σημασιολογικό γράφο του λεξικό WordNet. Ουσιαστικά, αποτελείται από το υποσύνολο των synsets του Wordnet (2.874 synsets, 4.787 λέξεις) τα οποία έχουν συναισθηματικό περιεχόμενο και στα οποία αποδίδεται επιπλέον τουλάχιστον μία επιγραφή (a-label) που προσδιορίζει το είδος της συναισθηματικής έννοιας την οποία περιγράφουν οι λέξεις. Ακολουθεί πίνακας με το σύνολο των διαθέσιμων επιγραφών (a-labels) και αντίστοιχα χαρακτηριστικά παραδείγματα.

A-Labels	Examples
EMOTION	noun anger#1, verb fear#1
MOOD	noun animosisy#1, adjective amiable#1
TRAIT	noun aggressiveness#1, adjective competitive#1
COGNITIVE STATE	noun confusion#2, adjective dazed#2
PHYSICAL STATE	noun illness#1, adjective all_in#1
EDONIC SIGNAL	noun hurt#3, noun suffering#4
EMOTION-ELICITING SITUATION	noun awkwardness#3, adjective out_of_danger#1
EMOTIONAL RESPONSE	noun cold_sweat#1, verb tremble#2
BEHAVIOUR	noun offense#1, adjective inhibited#1
ATTITUDE	noun intolerance#1, noun defensive#1
SENSATION	noun coldness#1, verb feel#3

Εικόνα 2. Απόσπασμα του Wordnet Affect

Μια ακόμη παρόμοια μέθοδος εφαρμόστηκε και για τη δημιουργία του SentiWordNet (Esuli and Sebastiani, 2006) [9]. Η ιδέα πίσω από τη δημιουργία του είναι ότι “όροι με παρόμοιο χαρακτηρισμό στο WordNet τείνουν να έχουν παρόμοια συναισθηματική πολικότητα”. Επομένως, το SentiWordNet δημιουργήθηκε με χρήση λέξεων - σπόρων και επεκτάθηκε αξιοποιώντας την ομοιότητα χαρακτηρισμών με τη βοήθεια του WordNet. Σε κάθε του λέξη έχει αποδοθεί βαθμολογία ως προς τη θετικότητα (PosScore) και την αρνητικότητα (NegScore) της, ενώ η βαθμολογία για την αντικειμενικότητα(ObjScore) της μπορεί να υπολογιστεί από τον τύπο: ObjScore = 1 - (PosScore + NegScore). Πέρα από τις βαθμολογίες κάθε λέξης δίνεται και η ερμηνεία της καθώς και προσδιορισμός του synset στο οποίο ανήκει. Ακολουθεί απόσπασμα του λεξικού:

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	00001740	0.125	0	able#1	having the necessary means or skill ....
a	00002098	0	0.75	unable#1	not having the necessary means or .....
a	00002312	0	0	dorsal#2 abaxial#1	facing away from the axis of an organ...

Εικόνα 3. Απόσπασμα του SentiWordNet

Η δημιουργία του Λεξικού Συναισθήματος MicroWNOp από τους Cerini et al. [10], στηρίχθηκε σε ένα σύνολο όρων (100 όροι για κάθε μία από τις θετικές, αρνητικές και ουδέτερες κατηγορίες συναισθημάτων) που προήλθαν από το General Inquirer Lexicon (Harvard) και στη συνέχεια επεκτάθηκε με την προσθήκη όλων των synsets του WordNet που περιείχαν αυτούς τους όρους (1.105 synsets). Το λεξικό χωρίζεται σε τρία τμήματα (Common, Group1, Group2). Κάθε γραμμή ενός τμήματος αντιστοιχεί σε ένα synset και περιλαμβάνει βαθμολογίες για τη θετικότητα του synset (Positive - Score) και την αρνητικότητα του (Negative - Score), σε πλήθος που διαφέρει ανάλογα με το τμήμα του λεξικού, καθώς και το σύνολο των λέξεων που ανήκουν στο συγκεκριμένο synset.Κάθε λέξη συνοδεύεται από προσδιορισμό του μέρους του λόγου και της συγκεκριμένης ερμηνεία της, με αναφορά το WordNet. Ακολουθεί απόσπασμα του λεξικού, που ανήκει στο τμήμα Common:

# begin Common		
# Positive-Score	Negative-Score	Synset
1	0	true#a#2 real#a#4
1	0	illustrious#a#1 famous#a#1 far-famed#a#1 .....
0.5	0	real#a#6 tangible#a#2
0.25	0	existent#a#2 real#a#1
0.125	0.125	real#a#2
0	0	real#a#7
0	0	real#a#11

Εικόνα 4. Απόσπασμα του MicroWNOp

2.1.5 Σχετικές Δημοσιεύσεις

Ακολουθεί συνοπτική περιγραφή δημοσιεύσεων ερευνών στις οποίες έχει γίνει Ανάλυση Συναισθήματος σε κείμενο με χρήση Λεξικού Συναισθήματος.

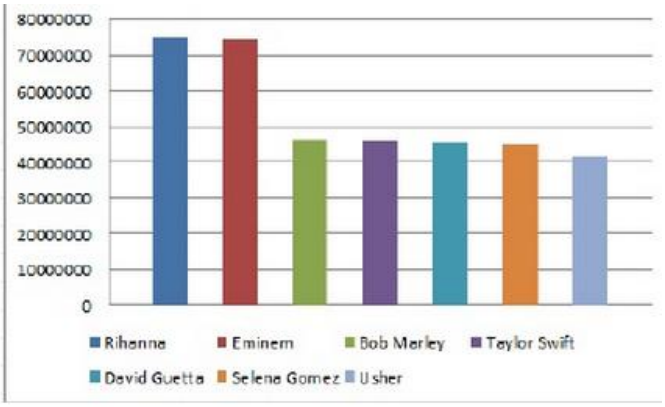
- Οι Balahur et al. [11] εφάρμοσαν την προσέγγιση του Λεξικού Συναισθήματος για την εξαγωγή συναισθήματος - άποψης από εκφράσεις σε εισαγωγικά (quotations) που έχουν αντληθεί από άρθρα ειδήσεων. Η κατάταξη έγινε σε τέσσερις κατηγορίες (positive, negative, high positive, high negative) και χρησιμοποιήθηκαν τέσσερα διαφορετικά λεξικά ξεχωριστά (JRC, WordNet Affect, SentiWordNet, MicroWNOp) αλλά και συνδυασμός τους. Τα αποτελέσματα έδειξαν ότι η εγκυρότητα των αποτελεσμάτων εξαρτάται από την ποιότητα του κάθε λεξικού, και ότι ο συνδυασμός λεξικών αποδίδει τα καλύτερα δυνατά αποτελέσματα για τη δεδομένη μέθοδο. Πειραματίστηκαν επίσης με το φιλτράρισμα των δεδομένων προς ανάλυση ως προς την υποκειμενικότητα τους πριν γίνει η Ανάλυση Συναισθήματος, οπότε

παρατήρησαν βελτίωση των αποτελεσμάτων. Ακολουθεί συγκεντρωτικός πίνακας της αξιολόγησης των αποτελεσμάτων, στον οποίο παρουσιάζονται τα Precision και Recall για κάθε κατηγορία, με παρουσία ή απουσία φιλτραρίσματος υποκειμενικότητας (S/O):

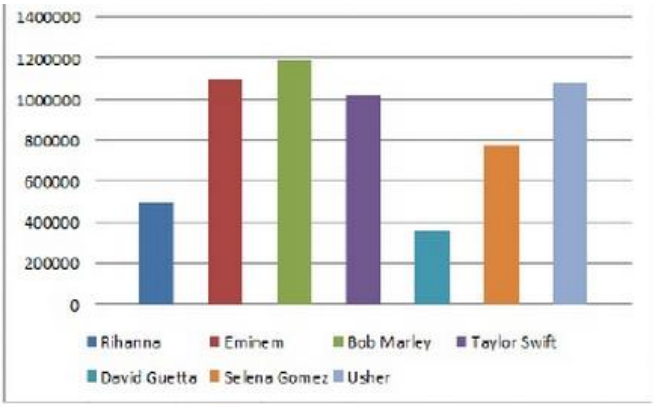
Resource	-S/O	+ S/O	P_{pos}	P_{neg}	R_{pos}	R_{neg}
JRCLists	X		0.77	0.3	0.54	0.55
		X	0.81	0.35	0.6	0.625
SentiWN	X		1	0	0.51	0
		X	1	0	0.54	0
WNAffect	X		0	1	0	0.51
		X	0	1	0	0.54
MicroWN	X		0.62	0.36	0.52	0.48
		X	0.73	0.35	0.57	0.53
SentiWN+ WNAffect	X		0.22	0.66	0.42	0.45
		X	0.24	0.67	0.47	0.41
All	X		0.68	0.64	0.7	0.62
		X	0.73	0.71	0.75	0.69

Πίνακας 1. Συγκεντρωτικός πίνακας της αξιολόγησης των αποτελεσμάτων

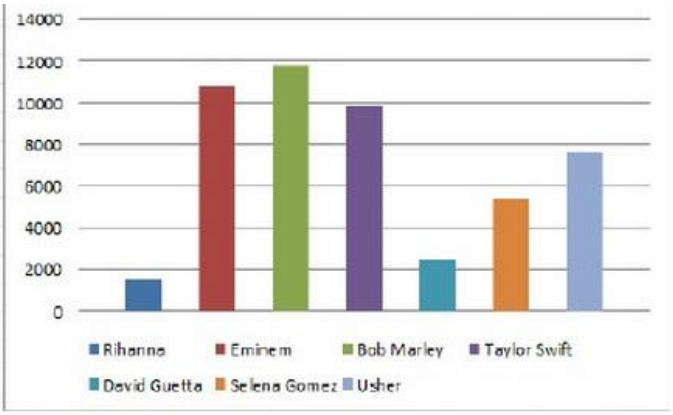
- Οι Ngoc και Yoo [2] χρησιμοποίησαν την μέθοδο ανάλυσης με χρήση Λεξικού Συναισθήματος για να αξιολογήσουν fan pages στο Facebook, στην προσπάθεια τους να ενισχύσουν την διαδεδομένη μέθοδο που στηρίζεται στην απλή καταμέτρηση των fans, των "likes", των posts και των σχολίων στην fan page. Υπολογίζονται δύο βαθμολογίες για τη σελίδα. Η πρώτη βαθμολογία (power score) προκύπτει από την καταμέτρηση των "likes" και η δεύτερη βαθμολογία (normal sentiment score) από το συναισθηματικό περιεχόμενο των σχολίων των χρηστών. Το τελικό σκορ της fan page (satisfaction score) προκύπτει από το συνδυασμό των δύο παραπάνω scores. Το normal sentiment score υπολογίζεται με τη βοήθεια του Λεξικού Συναισθήματος αγγλικών λέξεων AFINN. Το AFINN βαθμολογεί τις λέξεις με τιμές από -5 (negative) μέχρι +5 (positive). Οι λέξεις κάθε πρότασης αναζητούνται στο AFFIN και σημειώνονται οι αντίστοιχες βαθμολογίες, ενώ σε περίπτωση που μια λέξη δε βρεθεί, βαθμολογείται με 0. Η συνολική βαθμολογία της πρότασης δίνει το τελικό συναίσθημα. Κατά των υπολογισμό της βαθμολογίας μιας λέξης λαμβάνονται υπόψη η παρουσία λέξεων έντασης και λέξεων άρνησης, με χρήση ποσοστού και shift negation, αντίστοιχα. Ακολουθούν τα συγκριτικά αποτελέσματα για τη δημοτικότητα διάφορων fan pages, τα οποία προέκυψαν με τρία διαφορετικά κριτήρια: τον αριθμό των fans, τον αριθμό των χρηστών που εμφανίζουν κάποια αλληλεπίδραση με τη fan page (like, comment, share κ.α.) και το satisfaction score.



Εικόνα 5. Σύγκριση fan pages με βάση τον αριθμό των fans



Εικόνα 6. Σύγκριση fan pages με βάση την αλληλεπίδραση των χρηστών με την κάθε σελίδα



Εικόνα 7. Σύγκριση fan pages με βάση το Satisfaction Score

fan Page	Parameter		
	Number of fans	People talking about (PTA)	Satisfaction score
Rihanna	74.817.131	496.482	1.501,846
Eminem	74.446.071	1.099.341	10.804,71
Bob Marley	46.257.715	1.194.784	11.790,76
Taylor Swift	46.003.794	1.021.593	9.898,569
David Guetta	45.514.576	355.208	2.441,18
Selena Gomez	44.936.493	775.489	5.400,327
Usher	41.596.583	1.081.726	7.607,518

Πίνακας 2. Συγκεντρωτικά αποτελέσματα των εικόνων 5,6,7

2.2 Ανάλυση Συναισθήματος βασισμένη σε Μηχανική Μάθηση

2.2.1 Ορισμός και περιγραφή Μηχανικής Μάθησης

Το 1959, σχεδιαστής παιχνιδιών Arthur Samuel όρισε ως Μηχανική Μάθηση "Το πεδίο μελέτης όπου δίνει στους υπολογιστές την δυνατότητα να μαθαίνουν χωρίς να έχουν προγραμματιστεί". Το 1997, ακολούθησε ο εξής επίσημος ορισμός της Μηχανικής Μάθησης από τον καθηγητή Tom M. Mitchell: "Ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από την εμπειρία E σε σχέση με μια κατηγορία εργασιών T και μια μετρική απόδοσης P, αν η απόδοση του σε εργασίες της T, όπως μετριούνται από την P, βελτιώνονται με την εμπειρία E". Πιο απλά, Μηχανική Μάθηση είναι μια περιοχή της τεχνητής νοημοσύνης η οποία περιλαμβάνει την εκπαίδευση ενός υπολογιστικού συστήματος με τη βοήθεια ενός συνόλου δεδομένων εκπαίδευσης (training set) και ενός αλγορίθμου, ώστε να είναι σε θέση να εκτελεί τις απαιτούμενες σε διαφορετικές περιπτώσεις λειτουργίες, χωρίς να απαιτείται άμεσος προγραμματισμός του για την κάθε ειδική περίπτωση. Τη φάση της εκπαίδευσης του συστήματος ακολουθεί η φάση της αξιολόγησης η οποία γίνεται με τη χρήση ενός συνόλου δεδομένων δοκιμής (test set) τα οποία το σύστημα βλέπει πρώτη φορά και για τα οποία καλείται να παραγάγει τις αντίστοιχες εξόδους ή να τα τοποθετήσει στις κατάλληλες ομάδες. Ο εκπαιδευτής του συστήματος πρέπει με χρήση άλλης μεθόδου (συνήθως ανθρώπινη ερμηνεία) να γνωρίζει εκ των προτέρων τα σωστά αποτελέσματα για το δεδομένο test set ώστε να μπορεί να εκτιμήσει τις ιδιότητες του μοντέλου που έχει δημιουργηθεί. Τα δύο παραπάνω σύνολα δεδομένων πρέπει να είναι αντιπροσωπευτικά δείγματα του τύπου των δεδομένων στα οποία πρόκειται να εφαρμοστεί το μοντέλο γνώσης που δημιουργείται στο σύστημα και συνήθως προκύπτουν από ένα ενιαίο αρχικά σύνολο δεδομένων. Η ικανότητα ενός συστήματος Μηχανικής Μάθησης να προβλέπει αποτελέσματα για άγνωστες εισόδους με βάση τα δεδομένα εκπαίδευσης ονομάζεται ικανότητα γενίκευσης.

2.2.2 Κατηγορίες Μηχανικής Μάθησης

- *Επιβλεπόμενη Μηχανική Μάθηση (Supervised Machine Learning) ή Μάθηση με παραδείγματα (Learning from examples)*

  Σ' αυτή την περίπτωση το σύστημα εκπαιδεύεται με τη βοήθεια δεδομένων τα οποία αποτελούνται από εισόδους (inputs) και τις αντίστοιχες εξόδους (outputs). Μέσω της αξιοποίησης αυτών, με τη βοήθεια κάποιου αλγορίθμου, το σύστημα επαγωγικά δημιουργεί γενικούς κανόνες για την παραγωγή εξόδου από οποιαδήποτε είσοδο δεδομένων αντίστοιχου τύπου με αυτόν των δεδομένων εκπαίδευσης.

- *Μη Επιβλεπόμενη Μηχανική Μάθηση (Unsupervised learning) ή Μάθηση από παρατήρηση (Learning from observation)*

  Στην περίπτωση της Μη Επιβλεπόμενης Μάθησης, το σύστημα τροφοδοτείται μόνο με εισόδους και καλείται να ανακαλύψει πιθανές κρυμμένες δομές ανάμεσα τους, ώστε να τις ταξινομήσει σε ομάδες δεδομένων που παρουσιάζουν κάποια ομοιότητα. Η ύπαρξη, ο αριθμός και οι ιδιότητες των ομάδων είναι άγνωστα στο σύστημα αρχικά.

- *Μηχανική Μάθηση με Μερική Επίβλεψη (Semi- Supervised Machine Learning)*

  Η Μηχανική Μάθηση με Μερική Επίβλεψη αποτελεί συνδυασμό των δύο προηγουμένων κατηγοριών μάθησης, καθώς σ' αυτή το σύστημα τροφοδοτείται με λίγα ζεύγη εισόδου - εξόδου και αρκετές εισόδους χωρίς τις αντίστοιχες εξόδους.

- *Ενισχυτική Μηχανική Μάθηση (Reinforcement Learning)*

  Στην περίπτωση Ενισχυτικής Μηχανικής Μάθησης το σύστημα βελτιώνεται διαρκώς μέσω αλληλεπίδρασης με το περιβάλλον και παρατήρησης των αποτελεσμάτων αυτής της αλληλεπίδρασης. Η ενισχυτική μάθηση δεν χρησιμοποιείται στην ταξινόμηση κειμένου, επομένως στη συνέχεια θα ασχοληθούμε μόνο με τις τρεις πρώτες κατηγορίες Μηχανικής Μάθησης.

2.2.3 Μηχανική Μάθηση και Ανάλυση Συναισθήματος

Στον τομέα της Ανάλυσης Συναισθήματος, η Μηχανική Μάθηση χρησιμοποιείται ώστε να είναι σε θέση το σύστημα να εκτιμήσει αυτόματα το συναισθηματικό περιεχόμενο ενός οποιουδήποτε κειμένου του δοθεί προς ανάλυση, αφού πρώτα εκπαιδευτεί με βάση κάποια από τις παραπάνω μεθόδους Μηχανικής Μάθησης.

2.2.3.1 Σύνολο δεδομένων

Το σύνολο δεδομένων (dataset) περιλαμβάνει τα δεδομένα που θα χρησιμοποιηθούν κατά την υλοποίηση της Μηχανικής Μάθησης**.** Πρέπει να αποτελείται από κείμενα που ανήκουν στον ίδιο τύπο και στον ίδιο θεματικό τομέα με αυτά που θέλουμε να ταξινομεί το σύστημα μετά την εκπαίδευση του (π.χ. reviews, tweets, news articles, facebook posts, blog posts κ.α.), και θα πρέπει να περιέχει ισοδύναμα αντιπροσωπευτικά δείγματα από όλες τις επιθυμητές κλάσεις συναισθήματος στις οποίες θα γίνει η ταξινόμηση. Η ανάγκη για χρήση συγκεκριμένου τύπου δεδομένων προκύπτει από το γεγονός ότι ανάλογα με το είδος και τη θεματολογία συνήθως αλλάζει η χρήση της γλώσσας, ως προς τη σύνταξη αλλά και το νόημα του λεξιλογίου [12]. Όπως αναφέρθηκε προηγουμένως, τα σύνολα δεδομένων που χρησιμοποιούνται στις διάφορες φάσης της Μηχανικής Μάθησης προκύπτουν από τη διαίρεση του αρχικά ενιαίου συνόλου δεδομένων σε σύνολο εκπαίδευσης (training set) και σύνολο ελέγχου (test set). Αυτή η τεχνική ονομάζεται Hold out. Ο συνήθης διαχωρισμός του αρχικού συνόλου δεδομένων γίνεται με βάση κάποια αναλογία που κρίνεται λογική για την εκάστοτε χρήση (π.χ. 75% - 25%). Σε περιπτώσεις περιορισμένου αριθμού δεδομένων χρησιμοποιείται η μέθοδος n-fold validation, κατά την οποία τα δεδομένα χωρίζονται σε n ίσα υποσύνολα και στη συνέχεια με βάση αυτά δημιουργούνται n μοντέλα γνώσης, σε καθένα από τα οποία εξαιρείται από το σύνολο εκπαίδευσης κάποιο από τα n υποσύνολα, το οποίο χρησιμοποιείται ως το σύνολο ελέγχου.

2.2.3.2 Προεπεξεργασία δεδομένων

Η επιτυχία της εφαρμογής της Μηχανικής Μάθησης σε ένα υπολογιστικό σύστημα εξαρτάται σε πολύ μεγάλο βαθμό από την ποιότητα των δεδομένων στα οποία βασίζεται. Αν υπάρχει πλεονάζουσα και άσχετη πληροφορία ή τα δεδομένα είναι θορυβώδη και μη αξιόπιστα τότε η ανακάλυψη της γνώσης κατά τη φάση της εκπαίδευσης του συστήματος καθίσταται αρκετά δύσκολη [13]. Επομένως, είναι απαραίτητο να προηγηθεί της Μηχανικής Μάθησης η προεπεξεργασία δεδομένων (data preprocessing) ώστε να βελτιωθεί η απόδοση του ταξινομητή και να μειωθεί ο απαιτούμενος χρόνος ταξινόμησης [14]. Οι όροι του κειμένου που προκύπτουν από την προεπεξεργασία δεδομένων ονομάζονται χαρακτηριστικά (features).

Στην προεπεξεργασία δεδομένων μπορεί να περιλαμβάνονται τα εξής διαδομένα στάδια [15] [16] [17] [14]:

- *Μετατροπή των κεφαλαίων γραμμάτων σε μικρά:*
  Συνηθίζεται, κατά την προεπεξεργασία, η μετατροπή όλων των γραμμάτων σε μικρά, ώστε να ταυτίζονται οι διαφορετικές εμφανίσεις κάποιας λέξης. Ωστόσο, προτείνεται από ερευνητές [16] να προηγηθεί εντοπισμός των λέξεων που είναι ολόκληρες γραμμένες με κεφαλαία γράμματα και να τοποθετηθεί μπροστά τους η

φράση - κλειδί ALL_CAPS, γιατί συνήθως λέξεις σε κεφαλαία υποδηλώνουν ένταση συναισθήματος.

- *Αφαίρεση αριθμών:*
  Στις περισσότερες περιπτώσεις, οι αριθμοί σε ένα κείμενο δε σχετίζονται με το συναίσθημα που εκφράζεται και επομένως πολλοί ερευνητές θεωρούν την ανάλυση τους περιττή. Αξίζει ωστόσο να σημειωθεί ότι, ειδικά σε περιπτώσεις κειμένου από δικτυακή πηγή, ενδέχεται κάποιος αριθμός να χρησιμοποιηθεί ως συντομογραφία λέξης. Για παράδειγμα ο αριθμός 2 (two) μπορεί να χρησιμοποιηθεί αντί της αγγλικής λέξης 'too' ή 'to', ενώ ο αριθμός 8 μπορεί να χρησιμοποιηθεί στην υβριδική λέξη gr8 για αντικατάσταση της λέξης great.

- *Αφαίρεση σημείων στίξης:*
  Σε πολλές έρευνες [16] κατά την προεπεξεργασία συνηθίζεται η αφαίρεση των σημείων στίξης. Ωστόσο, πρέπει να αναφέρουμε ότι πολλές φορές η ύπαρξη σημείων στίξης υποδηλώνει την ύπαρξη κάποιου συναισθήματος. Για παράδειγμα, το θαυμαστικό μπορεί να μαρτυρά έντονα θετικό ή αρνητικό συναίσθημα. Αντίστοιχα, η χρήση ερωτηματικού μπορεί να δηλώνει προβληματισμό ή σύγχυση [17]. Επίσης, σε περίπτωση ανάλυσης κειμένου από δικτυακή πηγή ίσως πρέπει να λάβουμε υπόψη το σχηματισμό emoticons από σημεία στίξης.

- *Επέκταση συντομεύσεων:*
  Ειδικά σε κείμενα δικτυακής προέλευσης συνηθίζεται η χρήση συντομεύσεων συνηθισμένων λέξεων. Με τη βοήθεια λίστας είναι δυνατή η αντικατάσταση συντομεύσεων από τις λέξεις τις οποίες αντιπροσωπεύουν.

- *Αφαίρεση stop words:*
  Ως stop words ορίζονται κάποιες συνηθισμένες λέξεις που δε φέρουν ιδιαίτερη πληροφορία, οπότε η ανάλυση τους αφενός περιττεύει και αφετέρου μπορεί να οδηγήσει τον ταξινομητή σε λανθασμένα συμπεράσματα. Δεν υπάρχει κάποιο καθορισμένο σύνολο stop words, αλλά εξαρτάται από τις ανάγκες της εκάστοτε υλοποίησης. Συνήθως ως stop words θεωρούνται κάποιες μικρές λειτουργικές λέξεις όπως άρθρα, αντωνυμίες, προθέσεις και επιρρήματα (π.χ *the, is, at, which, on*).

- *Εντοπισμός n-grams:*
  Τα n-grams είναι ακολουθίες n στοιχείων κειμένου (χαρακτήρων, γραμμάτων συλλαβών ή λέξεων, ανάλογα με την εκάστοτε εφαρμογή) που προκύπτουν από ένα δεδομένο κείμενο και η χρήση τους βοηθάει στην αναγνώριση φράσεων που μπορεί να περιέχουν κάποιο νόημα το οποίο δεν μπορεί να εντοπιστεί σε περίπτωση ατομικής μελέτης των στοιχείων. Εισάγουν δηλαδή στην ανάλυση την έννοια της εξάρτησης των λέξεων. Η διαδικασία εξαγωγής των n-grams ενός κειμένου μπορεί να παρομοιαστεί με ένα παράθυρο n θέσεων που κινείται κατά μήκος των στοιχείων του κειμένου δημιουργώντας σε κάθε φάση ένα n-gram. Το μήκος των n-grams εξαρτάται από την εκάστοτε εφαρμογή. Μικρά n-grams ίσως δεν αρκούν ώστε να

καλύψουν μεγάλες φράσεις, ενώ μεγάλα n-grams δημιουργούν σπάνιες (ή και μοναδικές) ακολουθίες που δεν προσφέρουν, επομένως, κάτι στην ανάλυση. Συνήθως χρησιμοποιούνται unigrams, bigrams και trigrams [18][19].

- *Διαχείριση άρνησης:*
  Η ανίχνευση άρνησης κατά τη συναισθηματική ανάλυση ενός κειμένου διαφοροποιεί σημαντικά τα αποτελέσματα της. Η ανάλυση του κειμένου σε επίπεδο λέξεων ή σε επίπεδο ακατάλληλου μήκους n-grams αγνοεί την αντιστροφή του νοήματος που οφείλεται σε λέξεις άρνησης. Μια απλοϊκή προσέγγισης της ανίχνευσης της άρνησης περιλαμβάνει την σημείωση κάθε λέξης μετά την λέξη άρνησης και μέχρι το πρώτο σημείο στίξης με κάποιο συγκεκριμένη σήμανση (π.χ. με το "NOT_"). Για παράδειγμα, η φράση "I don't like this movie." γίνεται " I don't NOT_like NOT_this NOT_movie.", σύμφωνα με τον Boiy [20]. Μια πιο προηγμένη προσέγγιση δίνεται από τους Morante και Daelemans [21], οι οποίοι επιχειρούν με τη βοήθεια δύο ταξινομητών Μηχανικής Μάθησης να διαχειριστούν την ύπαρξη άρνησης σπάζοντας το έργο τους σε δύο μέρη: Αρχικά προσδιορίζουν ποιες λέξεις είναι λέξεις άρνησης και ακολούθως για κάθε λέξη διαπιστώνουν αν ανήκει στο πεδίο δράσης κάποιας αρνητικής λέξης ή όχι.

- *Εντοπισμός θέματος λέξεων (stemming):*
  Κατά το stemming αφαιρούνται από τις λέξεις οι καταλήξεις ώστε να εντοπιστεί η ρίζα της καθεμίας, με στόχο τη μείωση της πολυπλοκότητας της ανάλυσης χωρίς απώλεια σημαντικής πληροφορίας.

- *Λημματοποίηση (lemmatization) λέξεων:*
  Κατά τη λημματοποίηση, οι διάφορες μορφές μιας λέξης (κλίση, παράγωγα) αντιστοιχούνται στο ίδιο λήμμα (π.χ. το κοινό λήμμα των "τρέχοντας" και "έτρεξα" είναι το "τρέχω"). Με τον τρόπο αυτό οι λέξεις γενικεύονται και η ταξινόμηση τους γίνεται πιο εύκολα.

- *Αναγνώριση μέρους του λόγου (Part of Speech Tagging ή POS tagging):*
  Στο στάδιο αυτό γίνεται αναγνώριση και σημείωση του μέρους του λόγου (ρήμα, επίθετο, ουσιαστικό, μετοχή κ.τ.λ.) για κάθε λέξη του κειμένου με στόχο την αποκάλυψη της γραμματικής και επομένως τη βαθύτερη ανάλυση του [18] [22][19].

2.2.3.3 Επιλογή χαρακτηριστικών

Από τα χαρακτηριστικά που προέκυψαν από την προεπεξεργασία μόνο μερικά εκφράζουν έντονα κάποιο συναίσθημα, δηλαδή έχουν μεγαλύτερη επίδραση στην εκτίμηση του συνολικού συναισθηματικού προσανατολισμού του κειμένου. Η επιλογή του υποσυνόλου των χαρακτηριστικών, που θα ληφθούν υπόψη κατά τη συναισθηματική ανάλυση, ονομάζεται επιλογή χαρακτηριστικών (feature selection). Στην περίπτωση της Επιβλεπόμενης

Μάθησης, στόχος της επιλογής χαρακτηριστικών είναι η επίτευξη καλύτερης ακρίβειας του ταξινομητή ενώ στην περίπτωση της Μη Επιβλεπόμενης Μάθησης στόχος είναι ο σχηματισμός συστάδων υψηλής ποιότητας, δεδομένου του πλήθους συστάδων [23].

Οι μέθοδοι που μπορούν να εφαρμοστούν για την επιλογή χαρακτηριστικών χωρίζονται σε τέσσερις βασικές κατηγορίες: μέθοδοι Wrapper, μέθοδοι Filter και Embedded μέθοδοι, υβριδικές μέθοδοι [24].

Στις ***Ενσωματωμένες (Embedded)*** μεθόδους η επιλογή χαρακτηριστικών είναι μέρος της αντικειμενικής συνάρτησης του επιλεγμένου αλγορίθμου Μηχανικής Μάθησης. Χαρακτηριστικά παραδείγματα ενσωματωμένων μεθόδων είναι τα εξής: decision tree, LASSO, LARS, 1-norm support vector κ.α.

Οι μέθοδοι ***Wrapper*** εκτελούν την επιλογή χαρακτηριστικών θεωρώντας τον αλγόριθμο Μηχανικής Μάθησης ως ένα μαύρο κουτί. Δοκιμάζονται στον αλγόριθμο εξονυχιστικά όλα τα πιθανά σύνολα επιλεγμένων χαρακτηριστικών και το καταλληλότερο σύνολο επιλέγεται με βάση κάποιο κριτήριο, όπως η ακρίβεια ταξινόμησης. Το βασικό μειονεκτήματα των μεθόδων αυτών είναι η χρονική τους πολυπλοκότητα. Ωστόσο ο χώρος αναζήτησης μπορεί να περιοριστεί με χρήση ευρετικών (heuristic) και άπληστων (greedy) τεχνικών. Προφανώς, το βέλτιστο σύνολο χαρακτηριστικών διαφέρει από αλγόριθμο σε αλγόριθμο. Οι διάφορες ενσωματωμένες μέθοδοι διαφέρουν ως προς την τεχνικές που χρησιμοποιούν για την αναζήτηση στο χώρο τον χαρακτηριστικών, η οποία μπορεί να είναι είτε πλήρης (Complete Search), είτε σειριακή (Sequential Search), είτε τυχαία (Randomized Search).

Οι μέθοδοι ***Filter*** αποτελούν την πιο γενική προσέγγιση επιλογής χαρακτηριστικών και λειτουργούν άσχετα από τον αλγόριθμό Μηχανικής Μάθησης στον οποίο θα εισαχθούν τα επιλεγμένα χαρακτηριστικά. Χρησιμοποιούν μετρικές όπως η συσχέτιση (correlation), η εντροπία (entropy), η αμοιβαία πληροφορία (mutual information), το χ^2 (chi square) κ.α. οι οποίες αναλύουν τα γενικά χαρακτηριστικά των δεδομένων και επιλέγουν το βέλτιστο σύνολο χαρακτηριστικών. Οι παραπάνω μετρικές είναι univariate, δηλαδή βαθμολογούν κάθε χαρακτηριστικό ξεχωριστά. Υπάρχουν όμως και multivariate μετρικές, όπως οι Correlation Feature Selection (CFS), Minimum - redundancy-maximum-relevance (mRMR) feature selection. Οι μέθοδοι Filter είναι πολύ πιο απλές και γρήγορες από τις μεθόδους των δύο προηγουμένων κατηγοριών και γι’ αυτό προτιμώνται τόσο στην έρευνα όσο και στην εφαρμογή.

Τέλος, οι ***υβριδικές*** μέθοδοι συνδυάζουν την προσεγγίσεις των κατηγοριών Filter και Wrapper. Ο χώρος των δεδομένων περιορίζεται με χρήση των ιδιοτήτων της κατανομής δεδομένων (Filter) και στη συνέχεια με τη βοήθεια των τεχνικών αναζήτησης Wrapper εντοπίζεται το κατάλληλο υποσύνολο χαρακτηριστικών.

Ακολούθως, παρουσιάζονται μερικές από τις πιο διαδεδομένες μεθόδους που εφαρμόζονται για την επιλογή χαρακτηριστικών. Κοινό χαρακτηριστικό τους είναι ο υπολογισμός ενός score για κάθε ένα από τα χαρακτηριστικά (features) και στη συνέχεια η επιλογή των χαρακτηριστικών που εμφανίζουν τα υψηλότερα scores [63] [64] [65].

- Document Frequency - DF (Συχνότητα Εγγράφου)

  Η Συχνότητα Εγγράφου μετρά τον αριθμό των εγγράφων του συνόλου δεδομένων στα οποία εμφανίζεται το χαρακτηριστικό. Με χρήση αυτής της μεθόδου, απομακρύνονται τα χαρακτηριστικά των οποίων η συχνότητα εγγράφου είναι μικρότερη ενός προκαθορισμένου κατωφλίου συχνότητας. Επιλέγοντας χαρακτηριστικά μεγάλης συχνότητας αυξάνεται η πιθανότητα να εμφανιστούν αυτά τα χαρακτηριστικά σε μελλοντικά test sets. Η βασική υπόθεση είναι ότι τα πιο σπάνια χαρακτηριστικά περιέχουν τη λιγότερη πληροφορία για την ταξινόμηση σε κατηγορία. Η υπόθεση αυτή έρχεται σε αντίθεση με την αρχή της Ανάκτησης Πληροφοριών (Information Retrieval) σύμφωνα με την οποία οι όροι με τη μικρότερη συχνότητα εγγράφου περιέχουν την μεγαλύτερη πληροφορία [25]. Οι έρευνες δείχνουν ότι η συγκεκριμένη μέθοδος είναι η απλούστερη όλων, κλιμακώνεται και είναι αποτελεσματική για ταξινόμηση κειμένου [26].

- Information Gain - IG (Κέρδος Πληροφορίας)

  Το Κέρδος Πληροφορίας μετρά την ποσότητα (σε bits) της πληροφορίας που αποκτούμε σχετικά με την πρόβλεψη της κλάσης ταξινόμησης σε περίπτωση που η μόνη διαθέσιμη πληροφορία είναι η παρουσία ή η απουσία ενός χαρακτηριστικού σε ένα έγγραφο. Το Information Gain ενός χαρακτηριστικού υπολογίζεται από τη συμβολή του στη μείωση της συνολικής εντροπίας, δηλαδή της ποσότητας της πληροφορίας που απαιτείται για την κατάταξη ενός δείγματος του συνόλου δεδομένων σε μία κλάση. Έστω $\{c_1, c_2,\ldots c_m\}$ το σύνολο των διαθέσιμων κλάσεων ταξινόμησης. Το Κέρδος Πληροφορίας ενός χαρακτηριστικού f ορίζεται ως εξής:

  $$GI(f) = -\sum_{i=1}^{m} P_r(c_i)\log P_r(c_i) + P_r(f) + P_r(\bar{f}) \sum_{i=1}^{m} P_r(c_i|\bar{f})\log P_r(c_i|\bar{f}) \qquad (2.1)$$

  , όπου $P_r(c_i)$ η εκ των προτέρων πιθανότητα της κλάσης c_i, $P_r(f)$ η πιθανότητα του χαρακτηριστικού f σε ένα δοσμένο σύνολο δεδομένων, $P_r(c_i|f)$ η πιθανότητα της κλάσης c_i δεδομένου του χαρακτηριστικού f, $P_r(\bar{f})$ το συμπλήρωμα της $P_r(f)$ και $P_r(c_i|\bar{f})$ το συμπλήρωμα της $P_r(c_i|f)$.

  Από το σύνολο των χαρακτηριστικών ενός κειμένου απομακρύνονται τα χαρακτηριστικά των οποίων το Information Gain είναι κάτω από ένα προκαθορισμένο κατώφλι.

- CHI square statistic (χ - τετράγωνο)

  Το CHI είναι ένας στατιστικός δείκτης που χρησιμοποιείται για να ελέγξει την ανεξαρτησία δύο γεγονότων. Στην περίπτωση μας, τα δύο γεγονότα είναι η εμφάνισης ενός χαρακτηριστικού και η εμφάνιση μιας κλάσης συναισθήματος.

Αντιπροσωπεύει την απόκλιση από την αναμενόμενη κατανομή αν υποθέσουμε ότι η ύπαρξη του χαρακτηριστικού είναι ανεξάρτητη από την κλάση που προκύπτει:

$$\mathrm{CHI}\,(f, C_i) = \frac{N \times (AD - BE)^2}{(A+E) \times (B+D) \times (A+B) \times (E+D)} \qquad (2.2)$$

$$\mathrm{CHI}_{max}\,(f) = \max(\mathrm{CHI}\,(f, C_i)) \qquad (2.3)$$

, όπου A η συχνότητα συνύπαρξης f και C_i, B οι εμφανίσεις του f χωρίς την παράλληλη εμφάνιση της C_i, E οι εμφανίσεις της C_i χωρίς την παράλληλη εμφάνιση του f, D η συχνότητα ταυτόχρονης απουσίας f και C_i, N το πλήθος των εγγράφων στο σύνολο δεδομένων..

Μηδενικό CHI σημαίνει ανεξαρτησία χαρακτηριστικού - κλάσης. Μεγάλες τιμές CHI υποδεικνύουν την αστοχία της υπόθεσης περί ανεξαρτησίας. Επομένως, χαρακτηριστικά με μεγάλο CHI σχετίζονται με την επιλογή της εκάστοτε κατηγορίας και επομένως επιλέγονται κατά την επιλογή χαρακτηριστικών.

- Mutual Information – MI (Αμοιβαία Πληροφορία)

  Το Mutual Information μετρά πόση πληροφορία προσφέρει η παρουσία ή η απουσία ενός συγκεκριμένου χαρακτηριστικού για την επιλογή της σωστής κλάσης ταξινόμησης. Το Mutual Information ενός χαρακτηριστικού f υπολογίζεται από τον ακόλουθο τύπο:

$$\mathrm{MI}\,(f, c) = \sum_{e_f \in \{0,1\}} \sum_{e_c \in \{0,1\}} P(U_f = e_f, C_c = e_c) \log_2 \frac{P(U_f = e_f, C_c = e_c)}{P(U_f = e_f)\, P(C_c = e_c)} \qquad (2.4)$$

, όπου U_f είναι μια τυχαία μεταβλητή που παίρνει τιμή $e_f = 1$ όταν το έγγραφο περιλαμβάνει το χαρακτηριστικό f και $e_f = 0$ όταν το έγγραφο δεν περιλαμβάνει το χαρακτηριστικό f, ενώ C_c είναι μια τυχαία μεταβλητή που παίρνει τιμή $e_c = 1$ όταν το έγγραφο ανήκει στην κλάση c και $e_c = 0$ όταν το έγγραφο δεν ανήκει στην κλάση c.

Το MI ενός χαρακτηριστικού f για μια κλάση c μεγιστοποιείται στην περίπτωση που το χαρακτηριστικό f εμφανίζεται αν και μόνο αν το έγγραφο στο οποίο εμφανίζεται, ανήκει στην κλάση c. Κατά την επιλογή χαρακτηριστικών επιλέγονται τα χαρακτηριστικά μεγαλύτερο MI, εκείνα δηλαδή τα οποία παίζουν καθοριστική σημασία για την κατάταξη ενός εγγράφου σε κάποια κλάση.

2.2.3.4 Το πρόβλημα της υπερπροσαρμογής στα δεδομένα εκπαίδευσης

Πολλοί αλγόριθμοι Μηχανικής Μάθησης αντιμετωπίζουν το πρόβλημα της υπερπροσαρμογής στα δεδομένα της εκπαίδευσης (Overfitting). Κατά την εμφάνιση του overfitting, ο αλγόριθμος εσφαλμένα εστιάζει σε πολύ συγκεκριμένα χαρακτηριστικά των δεδομένων εκπαίδευσης, με αποτέλεσμα να παρουσιάζει μικρό σφάλμα εκπαίδευσης, δηλαδή λίγα σφάλματα στην αναγνώριση εισόδων που προέρχονται από το σύνολο εκπαίδευσης, αλλά μεγάλο σφάλμα γενίκευσης, δηλαδή πολλά σφάλματα στην αναγνώριση αγνώστων εισόδων. Αυτό συμβαίνει διότι το σύστημα δεν έχει καταφέρει να εξάγει ορθούς κανόνες γενίκευσης κατά την εκπαίδευση του, οπότε στηρίζει τις προβλέψεις του στην απομνημόνευση των δεδομένων εκπαίδευσης. Αιτίες εμφάνισης του φαινομένου είναι οι εξής :

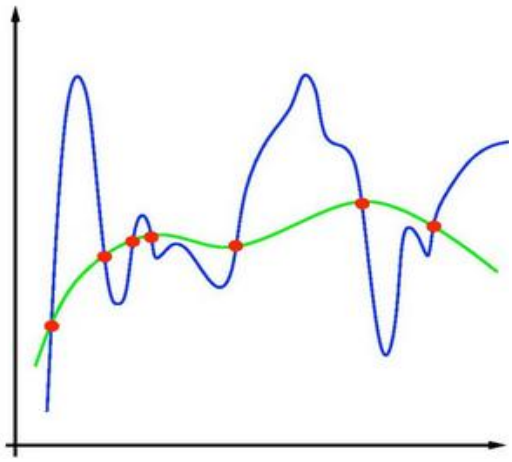
- Η χρήση μικρού συνόλου δεδομένων εκπαίδευσης, η οποία δεν επιτρέπει την εξαγωγή ικανοποιητικών κανόνων γενίκευσης.

- Η ύπαρξη τυχαίων παραπλανητικών κανονικοτήτων και θορύβου, συνήθως σε μικρά training sets, που οδηγούν τον αλγόριθμο σε εσφαλμένες γενικεύσεις.

- Η επιλογή συνόλου εκπαίδευσης που δεν περιλαμβάνει εξίσου δεδομένα από όλες τις διαθέσιμες κλάσεις ταξινόμησης, με αποτέλεσμα το σύστημα να έχει στη συνέχεια, την τάση να προβλέπει πιο συχνά τις κλάσεις που υπερίσχυαν στα δεδομένα εκπαίδευσης.

- Η μη κατάλληλη επιλογή χαρακτηριστικών.

- Ο μεγάλος αριθμός παραμέτρων του μοντέλου, ή πιο γενικά η ικανότητα του αλγορίθμου μάθησης να κατασκευάζει ιδιαίτερα πολύπλοκα μοντέλα.



Εικόνα 8. Overfitting

Η διπλανή εικόνα παρουσιάζει το πρόβλημα του overfitting. Οι κόκκινες κουκίδες αποτελούν τα δεδομένα εκπαίδευσης, η πράσινη γραμμή είναι η πραγματική συναρτησιακή σχέση τους και η μπλε γραμμή είναι η συνάρτηση που προέκυψε από την εκπαίδευση και υποφέρει από το overfitting. (Από www.wikipedia.org)

Για τον περιορισμό του φαινομένου έχουν υιοθετηθεί διάφορες μέθοδοι, οι οποίες στηρίζονται στην "αρχή του ξυραφιού" του Occam (Occam's Razor). Σύμφωνα με την αρχή αυτή "Κανείς δεν θα πρέπει να προβαίνει σε περισσότερες εικασίες από όσες είναι απαραίτητες" που σε ελεύθερη απόδοση σημαίνει ότι "Όταν δύο θεωρίες παρέχουν εξίσου ακριβείς προβλέψεις, πάντα επιλέγουμε την απλούστερη". Επίσης, στην πρόληψη του φαινομένου συμβάλει η επιλογή ενός dataset ικανοποιητικού μεγέθους, το οποίο περιλαμβάνει ισάριθμα αντιπροσωπευτικά δείγματα από όλες τις κλάσεις ταξινόμησης. Η ισορροπία των κλάσεων πρέπει να διατηρείται και κατά τον χωρισμό του αρχικού dataset σε training set και test set, ενώ πρέπει πριν την εκπαίδευση να προηγηθεί μια εύστοχη μέθοδος επιλογής χαρακτηριστικών, ανάλογα με το είδος του προς ανάλυση κειμένου.

2.2.3.5 Κατηγορίες Μηχανικής Μάθησης στην Ανάλυση Συναισθήματος

Στο σημείο αυτό θα αναφερθούμε στις διάφορες κατηγορίες Μηχανικής Μάθησης όπως αυτές εφαρμόζονται στον τομέα της Ανάλυσης Συναισθήματος και θα περιγράψουμε τους πιο διαδεδομένους αλγόριθμους που χρησιμοποιούνται στην κάθε κατηγορία.

Αναπαράσταση κειμένων

Σημειώνουμε ότι στις περιγραφές αλγορίθμων που θα ακολουθήσουν, θα γίνει αναπαράσταση των κειμένων στο διανυσματικό χώρο με χρήση της προσέγγισης του σάκου χαρακτηριστικών (bag of features), σύμφωνα με την οποία τα χαρακτηριστικά, που έχουν προκύψει από την διαδικασία της προεπεξεργασίας και της επιλογής χαρακτηριστικών, είναι μεταξύ τους ανεξάρτητα. Έστω $\{f_1, f_2, \ldots, f_m\}$ το σύνολο των m χαρακτηριστικών που μπορεί να εμφανίζονται σε ένα κείμενο d που ανήκει σε κλάση c και $n_i(d)$ ο αριθμός εμφανίσεων του χαρακτηριστικού f_i στο κείμενο. Κάθε κείμενο d αναπαρίσταται από το διάνυσμα $\vec{d} = (n_1(d), n_2(d) \ldots n_i(d))$. Στην περίπτωση μας, με τον όρο κλάση εννοούμε την κατηγορία συναισθήματος.

α) Επιβλεπόμενη Μηχανική Μάθηση

Για την Ανάλυση Συναισθήματος με χρήση Επιβλεπόμενης Μηχανικής Μάθησης το υπό εκπαίδευση σύστημα τροφοδοτείται με δεδομένα κειμένου τα οποία είναι ήδη χαρακτηρισμένα ως προς το συναισθηματικό τους περιεχόμενο, με στόχο την εκπαίδευση του. Τα δεδομένα κειμένου μπορεί είτε να έχουν σχολιαστεί με χρήση Λεξικού Συναισθήματος, είτε να έχουν σχολιαστεί χειρονακτικά (hand-annotated data) από άνθρωπο, είτε να εμπεριέχουν το σχολιασμό τους (self-annotated data) εξαρχής (π.χ. ένα review συνήθως συνοδεύεται από κάποιο rating). Τα δεδομένα αυτά αξιοποιούνται από κάποιον αλγόριθμο ώστε να προκύψει η γνώση που θα επιτρέψει μελλοντικά στο σύστημα να προβλέψει αυτόματα την απόκριση σε οποιαδήποτε είσοδο κειμένου προς συναισθηματική ανάλυση.

Μερικοί από τους πιο διαδεδομένους αλγόριθμους που χρησιμοποιούνται στην Επιβλεπόμενη Μηχανική Μάθηση είναι οι εξής: Support Vector Machines, Maximum entropy, Naive Bayes, Decision tree, κ.α..

Ακολουθεί σύντομη περιγραφή της εφαρμογής των παραπάνω αλγορίθμων για Ανάλυση Συναισθήματος σε κείμενο:

- **Αλγόριθμος Naive Bayes**

  Ο αλγόριθμος αυτός είναι ένας απλός πιθανοτικός αλγόριθμος που βασίζεται στο θεώρημα Bayes σε συνδυασμό με την υπόθεση ότι οι πιθανότητες των στοιχείων που εμπλέκονται είναι μεταξύ τους ανεξάρτητες, στην οποία οφείλει και το χαρακτηρισμό του ως naive (απλός).

  Στο πρόβλημα της Ανάλυσης Συναισθήματος η ταξινόμηση με χρήση του αλγόριθμου Naive Bayes ερμηνεύεται ως εξής:
  Η πιθανότητα ενός χαρακτηριστικού να ανήκει σε μια συγκεκριμένη κλάση δεν σχετίζεται με την πιθανότητα των υπόλοιπων χαρακτηριστικών να ανήκουν στην ίδια κλάση. Ο υπολογισμός της συνολικής πιθανότητας του κειμένου προκύπτει από τον πολλαπλασιασμό των πιθανοτήτων των επιμέρους χαρακτηριστικών του κειμένου.

  Με μαθηματικούς όρους τα παραπάνω εκφράζονται ως εξής:

  Σύμφωνα με το θεώρημα Bayes, η πιθανότητα ενός δεδομένου κειμένου d να ανήκει στην κλάση c δίνεται από τον τύπο:

$$P(c|d) = \frac{P(c)*P(d|c)}{P(d)} \qquad (2.5)$$

  , όπου P(c) η πιθανότητα της κλάσης c, P(d) η πιθανότητα του κειμένου d και P(d|c) η πιθανότητα του κειμένου d δεδομένης της κλάσης c.
  Έστω N το πλήθος των κειμένων του συνόλου εκπαίδευσης, N_j ο αριθμός των εμφανίσεων ενός κειμένου d μέσα στο σύνολο, K το πλήθος των διαφορετικών κλάσεων που εμφανίζονται στο σύνολο και K_j οι εμφανίσεις μιας κλάσης c. Τότε $P(c) = \frac{K_j}{K}$ και $P(d) = \frac{N_j}{N}$. Η πιθανότητα $P(d|c)$ υπολογίζεται με βάση την υπόθεση ανεξαρτησίας των χαρακτηριστικών δεδομένης της κλάσης του κειμένου. Σύμφωνα με την αυτή, η πιθανότητα ενός κειμένου d δεδομένης της κλάσης c ισούται με το γινόμενο των πιθανοτήτων όλων των χαρακτηριστικών από τα οποία αποτελείται το κείμενο d, δεδομένης της c. Τελικά προκύπτει ο ακόλουθος τύπος για την πιθανότητα μιας κλάσης c, δεδομένου ενός κειμένου προς ανάλυση:

$$P_{NB}(c|d) = \frac{P(c)*\prod_{i=1}^{m} P(f_i|c)^{n_i(d)}}{P(d)} \qquad (2.6)$$

Ο αλγόριθμος επιστρέφει ως έξοδο την κλάση c με τη μέγιστη πιθανότητα P_{NB} δηλαδή την $c^* = \arg\max_c P(c|d)$.

Χρήση και επιδόσεις Naive Bayes

Παρά την απλότητα του αλγορίθμου και το γεγονός ότι η υπόθεση περί ανεξαρτησίας των χαρακτηριστικών ενός κειμένου δεν ευσταθεί στον πραγματικό κόσμο, η Ανάλυση Συναισθήματος με χρήση του ταξινομητή Naive Bayes είναι εντυπωσιακά αποτελεσματική [29]. Δηλαδή, παρόλο που η εκτίμηση των πιθανοτήτων είναι χαμηλής ποιότητας, με αποτέλεσμα να απέχουν σημαντικά από τις πραγματικές τους τιμές, η τελική ταξινόμηση είναι πολύ ακριβής αφού σύμφωνα με τους υπολογισμούς του Naive Bayes η "νικήτρια" κλάση προκύπτει να έχει πολύ μεγαλύτερη πιθανότητα από τις υπόλοιπες κλάσεις [29][30]. Ωστόσο, πολλοί άλλοι ταξινομητές παρουσιάζουν καλύτερες επιδόσεις από το Naive Bayes στις περισσότερες εφαρμογές. Ο ταξινομητής Naive Bayes προτιμάται σε περιπτώσεις περιορισμένων διαθέσιμων υπολογιστικών πόρων (CPU και μνήμη) καθώς και σε περιπτώσεις στις οποίες επιθυμούμε τη γρήγορη εκπαίδευση του συστήματος λόγω της απλότητας και της μικρής υπολογιστικής πολυπλοκότητας του. Τέλος, χρησιμοποιείται σε πολλές έρευνες ως ταξινομητής αναφοράς (baseline), ώστε να γίνει αξιολόγηση της επίδοσης άλλων ταξινομητών που μελετούνται [29].

- **Αλγόριθμος Maximum Entropy**

Ο αλγόριθμος Μέγιστης Εντροπίας είναι ένας ακόμα πιθανοτικός αλγόριθμός που χρησιμοποιείται στην Μηχανική Μάθηση. Η λειτουργία του βασίζεται στην αρχή της μέγιστης εντροπίας, σύμφωνα με την οποία η κατανομή πιθανότητας που αναπαριστά καλύτερα την υπάρχουσα γνώση, λαμβάνοντας υπόψη τα δεδομένα εκπαίδευσης που έχουμε μελετήσει μέχρι στιγμής, είναι αυτή με τη μεγαλύτερη εντροπία, δηλαδή αυτή για την οποία οι μοναδικές υποθέσεις που έχουμε κάνει είναι αυτές που επιβάλλονται από τα τρέχοντα δεδομένα εκπαίδευσης ή αλλιώς η πιο κοντινή στην ομοιόμορφη κατανομή.

Η πιθανότητα ενός δεδομένου κειμένου d να ανήκει στην κλάση συναισθήματος c υπολογίζεται από τον ακόλουθο τύπο:

$$P_{ME}(c|d) := \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c}\, F_{i,c}(d,c)\right) \qquad (2.7)$$

, όπου $Z(d)$ είναι μια συνάρτηση κανονικοποίησης για μετατροπή των εκθετικών τιμών σε τιμές πραγματικής πιθανότητας. Η συνάρτηση $F_{i,c}(d,c)$ είναι μια

συνάρτηση που περιγράφει το χαρακτηριστικό f_i και την κλάση c και προσδιορίζεται ως εξής:

$$F_{i,c}(d, c') = \begin{cases} 1 & , n_i(d) > 0 \text{ and } c' = c \\ 0 & , \text{otherwise} \end{cases} \qquad (2.8)$$

Ο παραπάνω τύπος εκφράζει ότι για ένα κείμενο d και μια κλάση c', η $F_{i,c}(d, c')$ είναι ίση με τη μονάδα μόνο αν το f_i εμφανίζεται στο κείμενο d τουλάχιστον μία φορά και c=c'.

Η παράμετρος $\lambda_{i,c}$ (Πολλαπλασιαστής Lagrange) είναι το βάρος του χαρακτηριστικού f_i ως προς την κλάση συναισθήματος c. Μελετώντας τον ορισμό της P_{ME} παρατηρούμε ότι μεγάλο βάρος $\lambda_{i,c}$ σημαίνει ότι το χαρακτηριστικό f_i είναι ισχυρή ένδειξη ότι το κείμενο ανήκει στην κλάση c. Οι τιμές της παραμέτρου $\lambda_{i,c}$ επιλέγονται έτσι ώστε να μεγιστοποιείται η δεσμευμένη πιθανότητα P_{ME}. Η πιο διαδεδομένη προσέγγιση για τον υπολογισμό τον παραμέτρων είναι η χρήση αλγορίθμων Επαναληπτικής Κλιμάκωσης (Iterative Scaling algorithms), οι οποίοι έχουν ως κοινό χαρακτηριστικό την επίλυση ενός υποπροβλήματος μιας μεταβλητής σε κάθε φάση. Τέτοιοι αλγόριθμοι είναι οι εξής: Generalized Iterative Scaling algorithm (GIS) [31], Improved Iterative Scaling algorithm (IIS) [32], Sequential Conditional Generalized Iterative Scaling algorithm (SCGIS) [33].

<u>Χρήση και επιδόσεις Maximum Entropy</u>

Σε αντίθεση με τον αλγόριθμο Naive Bayes, ο αλγόριθμος Μέγιστης Εντροπίας δεν κάνει κάποια υπόθεση για την σχέση των χαρακτηριστικών μεταξύ τους, οπότε θα μπορούσε να αποδώσει καλύτερα από τον NB σε περιπτώσεις που η ανεξαρτησία των λέξεων δεν ισχύει. Αυτό βέβαια μεταφράζεται σε μεγαλύτερο υπολογιστικό κόστος και χρόνο εκπαίδευσης, λόγω του υπολογισμού των παραμέτρων. Ωστόσο, μετά τον υπολογισμό τους, ο ταξινομητής παρέχει αξιόπιστα αποτελέσματα και είναι ανταγωνιστικός σε όρους CPU και κατανάλωσης μνήμης.

- **Αλγόριθμος Support Vector Machines (SVM)**

Σε αντίθεση με τους αλγορίθμους Naive Bayes και Maximum Entropy, ο αλγόριθμος SVM παρέχει ταξινόμηση μεγάλου περιθωρίου (large margin) αντί πιθανοτικής. Σε περίπτωση που επιδιώκεται συναισθηματική κατάταξη δύο κλάσεων (negative και positive), η βασική ιδέα πίσω από τη μηχανική εκπαίδευση είναι η κατασκευή ενός υπερεπιπέδου που αναπαρίσταται από το διάνυσμα $\vec{w}$. Το υπερεπίπεδο αυτό διαχωρίζει τα διανύσματα των κειμένων που ανήκουν στην κατηγορία positive από τα διανύσματα εκείνων που ανήκουν στην κατηγορία negative και ο διαχωρισμός αυτός γίνεται με τη μεγαλύτερη δυνατή απόσταση

μεταξύ του υπερεπιπέδου και των κοντινότερων διανυσμάτων της κάθε πλευράς που ορίζει. Όσο μεγαλύτερη είναι αυτή η απόσταση τόσο μικρότερο είναι το σφάλμα γενίκευσης (generalization error), το οποίο εκφράζει πόσο καλά προσαρμόζεται (γενικεύει) το εκπαιδευμένο σύστημα σε άγνωστα δεδομένα προς ανάλυση. Επομένως, η εύρεση του επιπέδου αποτελεί πρόβλημα βελτιστοποίησης υπό περιορισμούς. Έστω ότι c_j είναι η κατηγορία στην οποία ανήκει το κείμενο d_j του συνόλου κειμένων εκπαίδευσης. Το c_j μπορεί να πάρει τις τιμές 1 και -1 για την κατηγορία positive και negative αντίστοιχα. Το ζητούμενο επίπεδο δίνεται από τον ακόλουθο τύπο.

$$\vec{w} := \sum_j a_j c_j \vec{d_j} \quad , a_j \geq 0 \qquad (2.9)$$

, όπου το a_j προκύπτει από την επίλυση ενός προβλήματος διπλής βελτιστοποίησης (dual optimization problem).

Τα διανύσματα $\vec{d_j}$ για τα οποία $a_j \geq 0$ ονομάζονται support vectors, αφού είναι τα μοναδικά που συνεισφέρουν στο $\vec{w}$.
Μετά την παραπάνω εκπαίδευση του συστήματος, η λύση στο πρόβλημα της Ανάλυσης Συναισθήματος ενός οποιουδήποτε κειμένου προκύπτει προσδιορίζοντας σε ποια πλευρά από τις δύο που ορίζει το υπερεπιπέδο προσπίπτει το διάνυσμα του εκάστοτε κειμένου.

Χρήση και επιδόσεις SVM

Χάρις στην προστασία από το overfitting που προσφέρει, η οποία δεν εξαρτάται από το πλήθος των χαρακτηριστικών του προς ανάλυση κειμένου, ο αλγόριθμος SVM είναι κατάλληλος για διαχείριση κειμένων που αναπαρίστανται από διανύσματα μεγάλης διάστασης, δηλαδή με πολλά χαρακτηριστικά. Με την ικανότητα του να γενικεύει παίρνοντας ως είσοδο διανύσματα μεγάλης διάστασης, περιορίζει την ανάγκη της επιλογής χαρακτηριστικών (feature selection).

β) Μη Επιβλεπόμενη Μηχανική Μάθηση

Για την Ανάλυση Συναισθήματος με χρήση Μη Επιβλεπόμενης Μηχανικής Μάθησης, το υπό εκπαίδευση σύστημα τροφοδοτείται με δεδομένα κειμένου τα οποία δεν είναι χαρακτηρισμένα ως προς το συναισθηματικό τους περιεχόμενο. Το σύστημα, μην έχοντας πληροφορία για το τι αποτελεί σωστή δράση ή επιθυμητή κατάσταση, δε μπορεί να εξάγει κανόνες γενίκευσης στους οποίους θα στηρίξει τις προβλέψεις του. Επομένως, προσπαθεί να ανακαλύψει ανάμεσα στα δεδομένα εκπαίδευσης κάποιες κρυμμένες δομές, ώστε να τα ταξινομήσει σε αγνώστων ιδιοτήτων ομάδες, οι οποίες θα αποτελέσουν τις κλάσεις συναισθήματος. Μια από τις πιο διαδεδομένες μεθόδους προσέγγισης του προβλήματος

είναι η συσταδοποίηση (clustering), της οποίας το όνομα συχνά ταυτίζεται με τη Μη Επιβλεπόμενη Μηχανική Μάθηση. Προφανώς, αφού η συσταδοποίηση είναι Μη Επιβλεπόμενη μορφή Μάθησης, δεν είναι γνωστό σε ποιά κλάση συναισθήματος αντιστοιχεί η κάθε μια από τις συστάδες που προκύπτουν. Επομένως, μετά τη συσταδοποίηση πρέπει να ακολουθήσει κάποια διαδικασία ανάθεσης κλάσεως συναισθήματος σε κάθε μια από τις συστάδες. Η πιο συνηθισμένη μέθοδος είναι αυτή του υπολογισμού του βάρους των χαρακτηριστικών ενός κειμένου. Σύμφωνα με αυτή, για κάθε όρο υπολογίζεται η ομοιότητα του με λέξεις αναφοράς (π.χ. "good", "bad") με τη βοήθεια κάποια λεξικού συνωνύμων και ανάλογα με την τιμή αυτής της ομοιότητας υπολογίζεται ένα βάρος για κάθε όρο, με αποτέλεσμα να εντάσσεται το κείμενο συνολικά σε κάποια κλάση συναισθήματος.

- **Αλγόριθμος k-means**

  Πρόκειται για έναν από τους πιο απλούς αλγορίθμους Μη Επιβλεπόμενης Μηχανικής Μάθησης που επιλύουν το πρόβλημα της συσταδοποίησης (clustering). Ο αλγόριθμος k-means χρησιμοποιεί συσταδοποίηση βασισμένη σε κέντρα (centroids), κατά την οποία μια συστάδα αναπαρίσταται από το centroid της (c_j), που είναι ο μέσος όρος των σημείων που την αποτελούν. Επομένως, ένα σημείο ανήκει σε μία συστάδα αν η απόσταση του από το κέντρο της είναι η μικρότερη από τις αντίστοιχες αποστάσεις από τα κέντρα των υπόλοιπων συστάδων του προβλήματος. Δεδομένου ενός συνόλου n δεδομένων (dataset) D ={d_1, d_2,d_3,…d_n}, όπου $\vec{d} = (n_1(d), n_2(d) \dots n_i(d))$ όπως έχει ήδη αναφερθεί, και ενός αριθμού k που αντιστοιχεί στο πλήθος των συστάδων, ο αλγόριθμος k-means στοχεύει στην ελαχιστοποίηση μιας αντικειμενικής συνάρτησης απόστασης, έστω της συνάρτησης τετραγωνικού σφάλματος (squared error function).

  Έχουμε, επομένως:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| d_i^{(j)} - c_j \right\|^2 \qquad (2.10)$$

  , όπου $\left\| d_i^{(j)} - c_j \right\|^2$ είναι η επιλεγμένη συνάρτηση που μετρά την απόσταση μεταξύ ενός σημείου – δεδομένου $d_i^{(j)}$ που ανήκει στη συστάδα j και του κέντρου c_j της συστάδας του, ενώ το J είναι άθροισμά των αποστάσεων κάθε σημείου –δεδομένου από το κέντρο της αντίστοιχης του συστάδας.

  Η ελαχιστοποίηση αυτής της συνάρτησης επιτυγχάνεται μέσω μιας επαναληπτικής διαδικασίας, σε κάθε επανάληψη της οποίας επαναπροσδιορίζονται τα centroids.

Ο αλγόριθμος k-means συνοψίζεται στα εξής βήματα:

1. Τυχαία τοποθέτηση k κέντρων στο χώρο των δεδομένων.
2. Ταξινόμηση των δεδομένων στις συστάδες που ορίζονται από τα παραπάνω centroids, ανάλογα με τις αποστάσεις τους από τα centroids.
3. Επαναπροσδιορισμός των centroids με χρήση μέσου όρου των ταξινομημένων δεδομένων.
4. Επανάληψη των βημάτων 2 και 3 μέχρι να μη σημειωθεί μετακίνηση κάποιου centroid.

Χρήση και επιδόσεις k - means

Ο αλγόριθμος k - means είναι ένας απλός στο σκεπτικό και την υλοποίηση αλγόριθμος, που μπορεί να εφαρμοστεί μόνο σε προβλήματα με εξαρχής γνωστό αριθμό συστάδων. Παρόλο που μπορεί να αποδειχθεί ότι πάντα τερματίζει, ο αλγόριθμος k-means δε βρίσκει απαραίτητα τη βέλτιστη λύση του προβλήματος, που αντιστοιχεί στο ολικό ελάχιστο της συνάρτησης. Τέλος, επειδή ο αλγόριθμος επηρεάζεται σημαντικά από την τυχαία αρχική επιλογή των centroids μπορεί να χρειαστεί να γίνουν πολλαπλές εκτελέσεις του μέχρι την επίτευξη της επιθυμητής ποιότητας των αποτελεσμάτων.

γ) Μηχανική Μάθηση με μερική επίβλεψη

Για την Ανάλυση Συναισθήματος με χρήση Μηχανικής Μάθησης με Μερική Επίβλεψη, το υπό εκπαίδευση σύστημα τροφοδοτείται με λίγα δεδομένα κειμένου τα οποία είναι χαρακτηρισμένα ως προς το συναισθηματικό τους περιεχόμενο και με αρκετά δεδομένα τα οποία δεν είναι χαρακτηρισμένα. Είναι, δηλαδή, μια μέθοδος που είναι σαφώς πιο ακριβής από την Μη Επιβλεπόμενη Μάθηση, ενώ δεν απαιτεί το δύσκολο και χρονοβόρο έργο του σχολιασμού όλων των δεδομένων που χαρακτηρίζει την Επιβλεπόμενη Μάθηση. Ωστόσο, για να είναι η μέθοδος τελικά πιο αποδοτική και από την Επιβλεπόμενη Μάθηση θα πρέπει η κατανομή των παραδειγμάτων που θα προκύψει από τα μη χαρακτηρισμένα δεδομένα να είναι αντιπροσωπευτική της ταξινόμησης στην οποία στοχεύουμε, δηλαδή τα μη σχολιασμένα δεδομένα να φέρουν χρήσιμη πληροφορία και να μην είναι παραπλανητικά [34]. Ακολουθεί περιγραφή διαδεδομένων αλγορίθμων για το πρόβλημα της Μηχανικής Μάθησης.

- **Αλγόριθμος Expectation – Maximization σε συνδυασμό με ταξινομητή Naive Bayes**

  Ο αλγόριθμος Expectation - Maximization (E-M) είναι μια επαναληπτική μέθοδος για την ανεύρεση βέλτιστων εκτιμήσεων των παραμέτρων στατιστικών μοντέλων με χρήση της μέγιστης πιθανοφάνειας, σε περιπτώσεις που τα δεδομένα είναι ελλιπή. Οι ζητούμενες παράμετροι επαναϋπολογίζονται μέχρι να επιτευχθεί η επιθυμητή σύγκλιση. Η διαδικασία ξεκινά με μια αρχική υπόθεση των τιμών των αγνώστων

παραμέτρων. Κατά τη διάρκεια της διαδικασίας γίνεται εναλλαγή μεταξύ των εξής δύο βημάτων:

Βήμα Ε (expectation): Στο βήμα αυτό γίνεται πρόβλεψη των δεδομένων που λείπουν με βάση τα υπάρχοντα δεδομένα και την παρούσα εκτίμηση των παραμέτρων.

Βήμα Μ (maximization): Στο βήμα αυτό γίνεται επανεκτίμηση των παραμέτρων του μοντέλου με βάση τις προβλέψεις του σταδίου Ε που προηγήθηκε. Δηλαδή, θεωρώντας πλέον γνωστά τα δεδομένα που λείπουν, γίνεται μεγιστοποίηση της συνάρτησης πιθανοφάνειας ώστε να υπολογιστούν εκ νέου οι παράμετροι του μοντέλου.

Για τις ανάγκες της Μηχανικής Μάθησης με μερική επίβλεψη, ο αλγόριθμος ΕΜ συχνά συνδυάζεται με τον ταξινομητή Naive Bayes που μελετήθηκε στην Επιβλεπόμενη Μηχανική Μάθηση. Συγκεκριμένα, ο αλγόριθμος που προκύπτει από τον παραπάνω συνδυασμό περιλαμβάνει τα εξής στάδια [28]:

Στάδιο 1: Χρησιμοποιώντας τα σχολιασμένα δεδομένα εκπαιδεύεται ένας ταξινομητής Naive Bayes (Επιβλεπόμενη Μάθηση).

Στάδιο 2 (Βήμα Ε): Με χρήση του ταξινομητή που προέκυψε στο Στάδιο 1 γίνεται σχολιασμός των μη σχολιασμένων δεδομένων.

Στάδιο 3 (Βήμα Μ): Θεωρώντας σωστές τις εκτιμήσεις για το σχολιασμό των δεδομένων που προέκυψαν στο Στάδιο 2, γίνεται επανεκτίμηση των παραμέτρων του ταξινομητή Naive Bayes (Επιβλεπόμενη Μάθηση).

Στάδιο 4: Επαναλαμβάνονται τα στάδια 2 και 3 μέχρι να επιτευχθεί η επιθυμητή σύγκλιση, δηλαδή μέχρι να μην παρατηρείται αλλαγή στις τιμές των παραμέτρων του ταξινομητή.

2.2.3.6 Προβλήματα της προσέγγισης με Μηχανική Μάθηση

Ένα από τα σημαντικότερα εμπόδια που συναντώνται κατά τη χρήση Μηχανικής Μάθησης στην Ανάλυση Συναισθήματος είναι η εξάρτηση του εκπαιδευμένου υπολογιστικού συστήματος από τον συγκεκριμένο θεματικό τομέα και τον τύπο των δεδομένων κειμένου με βάση τα οποία έγινε η εκπαίδευση του (genre and domain dependence), όπως έχει ήδη αναφερθεί. Αυτό σημαίνει ότι θα δούμε την απόδοση του συστήματος να πέφτει σημαντικά σε περίπτωση εισαγωγής προς ανάλυση κειμένου διαφορετικής δομής και νοηματικού τομέα [56]. Ωστόσο, υπάρχουν πολλοί τομείς για τους οποίους είναι δύσκολο να βρεθεί αρκετό υλικό εκπαίδευσης (training data). Επίσης, πολλές σύγχρονες εφαρμογές, ειδικά όσες τροφοδοτούνται από τον παγκόσμιο ιστό, απαιτούν συναισθηματική ανάλυση σταθερής ποιότητας, κειμένων με μεγάλη ποικιλία στη δομή και το περιεχόμενο. Επομένως,

έχουν γίνει και συνεχίζουν να γίνονται προσπάθειες [56] [57] με στόχο την επίτευξη μεταφερσιμότητας του εκπαιδευόμενου συστήματος (system portability) ως προς το θεματικό τομέα και τον τύπο των κειμένων που είναι σε θέση να αναλύσει αποδοτικά. Ένα άλλο πρόβλημα είναι η δυσκολία ανεύρεσης ήδη σχολιασμένων συνόλων δεδομένων σε οποιοδήποτε θεματικό τομέα. Επιπλέον, η δημιουργία νέων συνόλων δεδομένων είναι πολύ απαιτητικό έργο ως προς τη συλλογή των κειμένων αλλά κυρίως ως προς τον σχολιασμό τους, αφού είναι χρονοβόρα και κουραστική διαδικασία που προφανώς ενέχει τον κίνδυνο της υποκειμενικότητας του σχολιαστή. Τέλος, σε περίπτωση χρήσης λεξικού για το σχολιασμό (annotation) των δεδομένων, εισάγονται όλοι οι κίνδυνοι που αναλύθηκαν στην βασισμένη σε λεξικό προσέγγιση Ανάλυσης Συναισθήματος.

2.2.3.7 Αξιολόγηση συστήματος Ανάλυσης Συναισθήματος

Η απόδοση ενός συστήματος αυτόματης Ανάλυσης Συναισθήματος κρίνεται από τη συμφωνία που παρουσιάζει η ταξινόμηση που παράγει ως έξοδο, με την αντίστοιχη ανθρώπινη κρίση.

Από τη στιγμή που η ανθρώπινη κρίση χρησιμοποιείται ως σημείο αναφοράς για την αξιολόγηση του συστήματος θα πρέπει, αρχικά, να λάβουμε υπόψη το γεγονός ότι αυτή συνήθως διαφέρει από άνθρωπο σε άνθρωπο. Αυτό μπορεί να συμβαίνει λόγω διαφορετικού γνωστικού υποβάθρου, λόγω ασάφειας της ανθρώπινης γλώσσας και γενικότερα λόγω υποκειμενικής αντίληψης διαφόρων πραγμάτων. Επομένως, είναι απαραίτητο να εκτιμήσουμε σε τι βαθμό συγκλίνουν οι γνώμες των ανθρώπων - κριτών. Στη στατιστική, η ***συμφωνία των κριτών*** (inter-rater agreement) εκφράζει την ομοιογένεια μεταξύ των κριτικών τους και προφανώς ορίζει το πάνω όριο στο οποίο μπορεί να φτάσει η απόδοση του συστήματος αυτόματης Ανάλυσης Συναισθήματος. Για τον υπολογισμό της μπορούν να χρησιμοποιηθούν οι εξής στατιστικοί δείκτες: joint-probability of agreement, Cohen's kappa, Fleiss' kappa, inter-rater correlation, concordance correlation coefficient και intra-class correlation. Ακολούθως περιγράφονται οι πιο συχνά χρησιμοποιούμενοι στις έρευνες της Ανάλυσης Συναισθήματος στατιστικοί δείκτες

- *Cohen's kappa coefficient (Συντελεστής Κάπα του Cohen)*

  Ο συντελεστής Κάπα (κ) του Cohen αποτελεί έναν από τους πιο αξιόπιστους δείκτες συμφωνίας μεταξύ δύο κριτών που ταξινομούν N αντικείμενα σε C αμοιβαία αποκλειόμενες κλάσεις, καθώς λαμβάνει υπόψη την περίπτωσης συμφωνίας κατά τύχη. Υπολογίζεται από τον εξής τύπο:

  $$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \qquad (2.11)$$

  , όπου Pr(a) είναι η σχετική παρατηρούμενη συμφωνία μεταξύ των δύο κριτών και Pr(e) είναι η υποθετική πιθανότητα τυχαίας συμφωνίας, χρησιμοποιώντας τα

δεδομένα που παρατηρήθηκαν για τον υπολογισμό των πιθανοτήτων του κάθε κριτή για κάθε πιθανή κατηγορία ταξινόμησης.

Εάν οι κριτές είναι σε πλήρη συμφωνία, τότε κ = 1. Αν δεν υπάρχει συμφωνία μεταξύ των κριτών, εκτός από αυτή που θα αναμενόταν από την τύχη (όπως ορίζεται από το Pr(e)), τότε κ = 0. Γενικά, το κ μπορεί να πάρει τιμή μικρότερη ή ίση του 1, με τις αρνητικές τιμές να υποδηλώνουν συμφωνία μικρότερη και από αυτή που οφείλεται στην τύχη. Λόγω του ότι ο συντελεστής του Cohen αναφέρεται σε δύο μόνο κριτές, σε περιπτώσεις ύπαρξης πολλών κριτών υπολογίζονται οι συντελεστές κ για όλα τα πιθανά ζεύγη κριτών και ως τελικός συντελεστής κ του συγκεκριμένου προβλήματος ορίζεται ο μικρότερος από όσους έχουν υπολογιστεί. Εναλλακτικά, μπορεί να χρησιμοποιηθεί ο συντελεστής Κάπα του Fleiss, ο οποίος περιγράφεται στη συνέχεια.

- *Fleiss kappa coefficient (Συντελεστής Κάπα του Fleiss)*

  Σε αντίθεση με το συντελεστή κάπα του Cohen, ο συντελεστής κάπα του Fleiss, ο οποίος προέκυψε από γενίκευση του συντελεστή pi του Scott, εκτιμά τη συμφωνία μεταξύ ενός πλήθους K κριτών που ταξινομούν ένα πλήθος N αντικειμένων σε C κλάσεις, επίσης λαμβάνοντας υπόψη τον παράγοντα της τύχης. Σημειώνουμε ότι δεν είναι απαραίτητο κάθε αντικείμενο να έχει αξιολογηθεί από τους ίδιους κριτές. Έστω n το πλήθος των αξιολογήσεων για κάθε αντικείμενο, i αντικείμενο, j κλάση και n_{ij} το πλήθος των κριτών που ταξινόμησαν το αντικείμενο i στην κλάση j.

  Ο τύπος υπολογισμού του κ είναι ο ακόλουθος:

$$\kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}} \qquad (2.12)$$

  Το $\overline{P}$ προκύπτει από το μέσο όρο των πιθανοτήτων συμφωνίας των κριτών για κάθε ένα από τα N αντικείμενα δηλαδή:

$$\overline{P} = \frac{1}{N} \sum_{i=1}^{N} P_i = \frac{1}{N * n * (n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{C} n_{ij}^{\ 2} - Nn \right) \qquad (2.13)$$

  Το $\overline{P_e}$ προκύπτει από το άθροισμα των τετραγώνων των ποσοστών ταξινόμησης σε κάθε μια από τις C κλάσεις.

$$\overline{P_e} = \sum_{j=1}^{C} P_j^{\ 2} = \frac{1}{N*n} \sum_{i=1}^{N} n_{ij} \qquad (2.14)$$

Προφανώς και για το συντελεστή του Fleiss, η τέλεια συμφωνία επιτυγχάνεται για κ =1, η συμφωνία που οφείλεται μόνο σε τύχη για κ = 0, ενώ μικρότερη από την τυχαία συμφωνία για κ <0.

Ακολουθεί συγκεντρωτικός πίνακας ερμηνείας όλων των τιμών ενός οποιουδήποτε συντελεστή κ [58]:

Τιμή κ	Συμφωνία
<0,00	Less than chance agreement
0,00	Chance agreement
0,01 – 0.20	Slight agreement
0,21 – 0,40	Fair agreement
0,41 - 0,60	Moderate agreement
0,61 – 0,80	Substantial agreement
0,81 – 0,99	Almost perfect agreement
1	Perfect agreement

Πίνακας 3. Ερμηνεία τιμών κ

Δεδομένων των παραπάνω παραμέτρων μπορούμε να προχωρήσουμε πλέον στην αξιολόγηση του συστήματος αυτού καθαυτού. Για τον ορισμό των διάφορων διαδεδομένων μετρικών, θα χρησιμοποιήσουμε τη Μήτρα Σύγχυσης (Confusion Matrix). Έστω ότι επιχειρούμε Ανάλυση Συναισθήματος σε αντικείμενα (κείμενα κάποιας μορφής),τα οποία αποτελούν το σύνολο ελέγχου (test set), με στόχο την ταξινόμηση τους σε m κλάσεις συναισθήματος. Δεδομένων m κλάσεων συναισθήματος, μια Μήτρα Σύγχυσης είναι ένας πίνακας m x m διαστάσεων, όπου κάθε κελί του (i, j) φέρει έναν ακέραιο αριθμό C_{ij} που αναπαριστά το πλήθος των αντικειμένων που κανονικά ανήκουν στην κλάση i και το σύστημα τα τοποθέτησε στην κλάση j. Παρακάτω φαίνεται η γενικευμένη μορφή της Μήτρας Σύγχυσης:

		Προβλεπόμενη από τον ταξινομητή κλάση			
		Κλάση 1	**Κλάση 2**	**…..**	**Κλάση m**
Πραγματική κλάση	**Κλάση 1**	C_{11}	C_{12}	….	C_{1m}
	Κλάση 2	C_{21}	C_{22}	….	C_{2m}
	…..	….	….	….	….
	Κλάση m	C_{m1}	C_{m2}	….	C_{mm}

Πίνακας 4. Γενικευμένη μορφή της Μήτρας Σύγχυσης

Από την παραπάνω περιγραφή προκύπτει ότι το άθροισμα των στοιχείων της διαγωνίου αποτελούν τον αριθμό των σωστών ταξινομήσεων, ενώ το άθροισμα των υπολοίπων στοιχείων αποτελεί τον αριθμό των λάθος ταξινομήσεων του συστήματος.

Υιοθετούμε τον εξής συμβολισμό:
- TP_i : σωστές ταξινομήσεις στην κλάση i (true positive)
- FN_{ij} : λάθος ταξινομήσεις στην κλάση j, αντί της i (false negative)
Σύμφωνα με τα παραπάνω η Μήτρα Σύγχυσης γίνεται:

		Προβλεπόμενη από τον ταξινομητή κλάση			
		Κλάση 1	**Κλάση 2**	**…..**	**Κλάση m**
Πραγματική κλάση	**Κλάση 1**	TP_1	FN_{12}	….	FN_{1m}
	Κλάση 2	FN_{21}	TP_2	….	FN_{2m}
	…..	….	….	….	….
	Κλάση m	FN_{m1}	FN_{m2}	….	TP_{mm}

Πίνακας 5. Μήτρα Σύγχυσης με συμβολισμό TP και FN

Μπορούμε τώρα να ορίσουμε τις εξής μετρικές απόδοσης του συστήματος ταξινόμησης:

- <u>*Accuracy*</u> *(Ορθότητα):* Εκφράζει το ποσοστό των επιτυχημένων ταξινομήσεων του συστήματος και υπολογίζεται από το λόγο των σωστών ταξινομήσεων προς τις συνολικές ταξινομήσεις του συστήματος.

$$\text{Accuracy} = \frac{\text{Σωστές ταξινομήσεις}}{\text{Συνολικές Ταξινομήσεις}} = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i + \sum_{i=1}^{m} \sum_{\substack{j=1, \\ j \neq i}}^{m} FN_{ij}} \qquad (2.15)$$

- <u>*Error rate*</u> *(Λόγος σφάλματος):* Συχνά, αντί του Accuracy χρησιμοποιείται το μέτρο του error rate, το οποίο εκφράζει το ποσοστό των εσφαλμένων ταξινομήσεων του συστήματος και δίνεται από τον ακόλουθο τύπο:

$$\text{Error rate} = 1 - \text{Accuracy} \qquad (2.16)$$

- <u>*Precision*$_i$</u> *(Ακρίβεια):* Αναφέρεται σε μια συγκεκριμένη κλάση και εκφράζει πόσες από τις ταξινομήσεις που έγιναν σε αυτή την κλάση είναι σωστές.

$$\text{Precision}_i = \frac{\text{Σωστές ταξινομήσεις στην κλάση i}}{\text{Συνολικές ταξινομήσεις στην κλάση i}} = \frac{TP_i}{TP_i + \sum_{\substack{j=1 \\ j \neq i}}^{m} FN_{ji}} \qquad (2.17)$$

Για τον υπολογισμό του συνολικού Precision του συστήματος, υπάρχουν δύο μέθοδοι: η micro – averaging και η macro – averaging, η διαφορά των οποίων έγκειται στο ότι η πρώτη αποδίδει ίδιο βάρος σε κάθε έγγραφο προς ταξινόμηση (document - pivoted measure), ενώ η δεύτερη αποδίδει ίδιο βάρος σε κάθε κλάση της ταξινόμησης (class - pivoted measure) [67].

Προκύπτουν οι εξής τύποι υπολογισμού της συνολικής Precision:

$$\text{Precision}_{\text{micro}} = \frac{\sum_{i=1}^{m} \text{TP}_i}{\sum_{i=1}^{m} \text{TP}_i + \sum_{i=1}^{m} \sum_{\substack{j=1, \\ j \neq i}}^{m} \text{FN}_{ji}} \qquad (2.18)$$

$$\text{Precision}_{\text{macro}} = \frac{1}{m} * \sum_{i=1}^{m} \frac{TP_i}{TP_i + \sum_{\substack{j=1, \\ j \neq i}}^{m} \text{FN}_{ji}} \qquad (2.19)$$

- *Recall$_i$ (Ανάκληση):* Αναφέρεται σε μια συγκεκριμένη κλάση και εκφράζει το ποσοστό των αντικειμένων που ανήκουν στην πραγματικότητα στην κλάση i τα οποία κατάφερε να ταξινομήσει σωστά το σύστημα.

$$\text{Recall}_i = \frac{\text{Σωστές ταξινομήσεις στην κλάση } i}{\text{Ανήκουν στην πραγματικότητα στην κλάση } i} = \frac{\text{TP}_i}{\text{TP}_i + \sum_{\substack{j=1 \\ j \neq i}}^{m} \text{FN}_{ij}} \qquad (2.21)$$

Για το συνολικό Recall έχουμε:

$$\text{Recall}_{\text{micro}} = \frac{\sum_{i=1}^{m} \text{TP}_i}{\sum_{i=1}^{m} \text{TP}_i + \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \text{FN}_{ij}} \qquad (2.22)$$

$$\text{Recall}_{\text{macro}} = \frac{1}{m} * \sum_{i=1}^{m} \frac{\text{TP}_i}{\text{TP}_i + \sum_{\substack{j=1, \\ j \neq i}}^{m} \text{FN}_{ij}} \qquad (2.23)$$

- F – measure (Σταθμισμένος Αρμονικός Μέσος): Η μετρική αυτή συνδυάζει το Precision και το Recall υπολογίζοντας τον αρμονικό τους μέσο.

$$\text{F} - \text{measure} = \frac{2 * \text{Precis ion} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (2.24)$$

2.2.3.8 Σύνολα δεδομένων για Ανάλυση Συναισθήματος

Σημαντική διευκόλυνση στο έργο της Ανάλυσης Συναισθήματος προσφέρουν τα διάφορα σύνολα δεδομένων για Μηχανική Μάθηση που έχουν δημιουργηθεί μέχρι σήμερα και τα οποία βρίσκονται σε μεγάλη ποικιλία ως προς το θεματικό τομέα αλλά και το είδος του κειμένου. Η χρήση αυτών των συνόλων δεδομένων μειώνει σημαντικά το χρόνο προετοιμασίας του συστήματος Ανάλυσης Συναισθήματος, αφού παρακάμπτεται η χρονοβόρα και αμφιβόλου ποιότητας διαδικασία δημιουργίας συνόλου δεδομένων. Επίσης, η κοινή χρήση τους από πολλούς ερευνητές οδηγεί στην συνεχή βελτίωση τους αλλά και επιτρέπει την σύγκριση των διαφορετικών μεθόδων που εφαρμόζονται πάνω σε ίδια σύνολα δεδομένων. Ακολουθεί περιγραφή των πιο διαδεδομένων datasets.

- **Pang & Lee dataset:** Είναι μια συλλογή από 1.000 αρνητικές και 1.000 θετικές κριτικές ταινιών, η οποία δημιουργήθηκε από τους Pang & Lee. (http://www.cs.cornell.edu/people/pabo/movie-review-data/).

- **Blitzer et al Multi-domain sentiment dataset:** Συλλογή από κριτικές προϊόντων από το Amazon για διάφορες κατηγορίες προϊόντων σε πλήθος που ποικίλει ανά κατηγορία. Οι κριτικές περιλαμβάνουν βαθμονόμηση με αστέρια (από 1 μέχρι 5) που μπορούν να μετατραπούν σε δυαδικές ετικέτες αν χρειαστεί. (http://www.cs.jhu.edu/~mdredze/datasets/sentiment/)

- **MPQA opinion corpus:** Συλλογή από άρθρα ειδήσεων που προέρχονται από ποικιλία ειδησεογραφικών πηγών και έχουν σχολιαστεί από ως προς την άποψη, τις πεποιθήσεις, το συναίσθημα και τις εικασίες που εκφράζουν. (http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/)

- **ISEAR corpus:** Συλλογή από αναφορές καταστάσεων στις οποίες διάφοροι άνθρωποι βίωσαν τα 7 βασικά συναισθήματα (χαρά, φόβος, θυμός, λύπη, αηδία, ενοχή). (http://www.affective-sciences.org/researchmaterial)

- **EmotiBlog corpus:** Συλλογή από blog posts σχολιασμένα ως προς το συναίσθημα και την πολικότητα τους σε διάφορα επίπεδα (λέξεων, φράσεων, προτάσεων κτλ).

Στη συνέχεια αναφέρουμε τα σημαντικότερα datasets που έχουν δημιουργηθεί για συναισθηματική ανάλυση σε tweets [35].

- **Stanford Twitter Sentiment Corpus (STS Corpus**): Δημιουργήθηκε από τους Go et al. [36] και αποτελείται από ένα σύνολο εκπαίδευσης (training set) και ένα σύνολο ελέγχου (test set). Το σύνολο εκπαίδευσης περιέχει 1.600.000 tweets αυτόματα χαρακτηρισμένα ως αρνητικά ή θετικά (negative, positive) ανάλογα με τα emoticons που περιέχονται. Αντίθετα, το σύνολο εκπαίδευσης (STS - Test) είναι

χαρακτηρισμένο χειρονακτικά και περιέχει 177 αρνητικά, 182 θετικά και 139 ουδέτερα tweets. Τα tweets συλλέχθηκαν με χρήση του API αναζήτησης του Twitter με όρους αναζήτησης ονόματα προϊόντων, ανθρώπων και εταιρειών. Παρά το μικρό του μέγεθος, το σύνολο STS - Test έχει χρησιμοποιηθεί σε πολλές εφαρμογές για αξιολόγηση ταξινόμησης σε δύο κλάσεις [36] [37] [61] [38] [39] ή αξιολόγηση ταξινόμησης ως προς την υποκειμενικότητα [40].
(http://help.sentiment140.com/)

- **Health Care Reform (HCR):** Δημιουργήθηκε από tweets που περιέχουν το hashtag #hcr (health care reform) από τους Speriosu et al. [38]. Ένα υποσύνολο του χαρακτηρίστηκε χειρονακτικά από τους δημιουργούς του με 5 κλάσεις συναισθήματος **(**positive, negative, neutral, irrelevant, unsure(other)) και χωρίστηκε σε σύνολα εκπαίδευσης (839 tweets), ανάπτυξης (838 tweets) και ελέγχου (839 tweets). Οι δημιουργοί, επίσης, εξήγαγαν 8 διαφορετικούς στόχους συναισθήματος (Health Care Reform, Obama, Democrats, Republicans, Tea Party, Conservatives, Liberals, Stupak) από τα τρία αυτά σύνολα και τους απέδωσαν ετικέτες συναισθήματος. Ωστόσο, τόσο στα tweets όσο και στους στόχους αποδόθηκαν οι ίδιοι χαρακτηρισμοί. Το σύνολο HCR έχει χρησιμοποιηθεί για την αξιολόγηση ταξινόμησης δύο κλάσεων [38] [45] καθώς και ταξινόμησης υποκειμενικότητας αφού αναγνωρίζει και τα ουδέτερα tweets.
(https://bitbucket.org/speriosu/updown)

- **Obama - McCain Debate (OMD):** Δημιουργήθηκε από 3.238 tweets κατά το πρώτο τηλεοπτικό προεδρικό debate των Obama – McCain στις Ηνωμένες Πολιτείες το Σεπτέμβριο του 2008 από τους Shamma et al. [41]. Με χρήση του Amazon Mechanical Turk[1] αποδόθηκαν σ' αυτά τα tweets ετικέτες συναισθήματος (positive, negative, mixed, other) μετά από αξιολογήσεις τουλάχιστον τριών κριτών για το καθένα. Η συμφωνία των κριτών έχει υπολογιστεί σε 0,655, τιμή αρκετά ικανοποιητική [42]. Το σύνολο OMD έχει χρησιμοποιηθεί σε εφαρμογές τόσο Επιβλεπόμενης [43] [44] [45] όσο και Μη Επιβλεπόμενης Μηχανικής Μάθησης [46] για την αξιολόγηση συναισθηματικής ταξινόμησης σε tweets.
(https://bitbucket.org/speriosu/updown)

- **Sentiment Strength Twitter Dataset (SS-Tweet):** Το σύνολο αποτελείται από 4.242 tweets χειρονακτικά επισημασμένα με τιμές από -5 (extremely negative) μέχρι -1 (not negative) και 1(not positive) μέχρι 5 (extremely positive). Δημιουργήθηκε από τους Thelwall et al. [47] με στόχο την αξιολόγηση του Sentistrength, μιας βασισμένης σε λεξικό εφαρμογής εκτίμησης έντασης συναισθήματος.
(http://sentistrength.wlv.ac.uk/documentation/).

- **Sanders Twitter Dataset:** Αποτελείται από 5.512 tweets για τα εξής τέσσερα θέματα: Apple, Google, Microsoft, Twitter. Κάθε tweet έχει χαρακτηριστεί χειρονακτικά από κάποιον κριτή ως positive, negative, neutral ή irrelevant σε σχέση

με το κάθε θέμα. Από το χαρακτηρισμό προέκυψαν 654 negative, 2.503 neutral, 570 positive και 1.786 irrelevant tweets. Το σύνολο έχει χρησιμοποιηθεί για ταξινόμηση συναισθήματος και υποκειμενικότητας [48] [49] [50].
(http://www.sananalytics.com/lab)

- **The Dialogue Earth Twitter Corpus:** Αποτελείται από τρία υποσύνολα από tweets. Τα δύο πρώτα υποσύνολα (WA, WB) περιέχουν 4.490 και 8.850 tweets σχετικά με τον καιρό, ενώ το τρίτο υποσύνολο (GASP) περιέχει 12.770 tweets σχετικά με τις τιμές του φυσικού αερίου. Αυτά τα σύνολα δημιουργήθηκαν ως μέρος του Dialogue Earth Project[2] και επισημάνθηκαν χειρονακτικά από αρκετούς με τους χαρακτηρισμούς positive, negative, neutral, not related, can't tell (other). Τα σύνολα WAB και GASP έχουν χρησιμοποιηθεί στην αξιολόγηση απόδοσης ταξινομητών Μηχανικής Μάθησης (π.χ. Naive Bayes, SVM, KNN) σε συναισθηματική ταξινόμηση tweets [51].

- **SemEval-2013 Dataset (SemEval):** Αυτό το σύνολο δεδομένων δημιουργήθηκε για την Ανάλυση Συναισθήματος στο Twitter στα πλαίσια του Semantic Evaluati-Evaluation of System challenge (SemEval - 2013, Task 2). Αποτελείται από 20.000 tweets, τα οποία μοιράζονται σε σύνολα εκπαίδευσης, ανάπτυξης και ελέγχου. Όλα τα tweets έχουν χαρακτηριστεί χειρονακτικά από 5 κριτές του Amazon Mechanical Turk ως negative, positive και neutral. Οι κριτές, επίσης αξιολόγησαν τα tweets ως υποκειμενικά ή αντικειμενικά. Στο SemEval 2013 - Task2 το σύνολο δεδομένων χρησιμοποιήθηκε για αξιολόγηση των συστημάτων στον εντοπισμό υποκειμενι-κότητας σε επίπεδο έκφρασης [52][53] καθώς και σε επίπεδο tweet [54] [55].

Στον παρακάτω πίνακα συνοψίζονται τα χαρακτηριστικά των παραπάνω datasets:

Dataset	No.of Tweets	Negative	Neutral	Positive	Mixed	Other	Irrelevant
STS –Test	498	177	139	182	-	-	-
HCR	2,516	1.381	470	541	-	45	79
OMD	3.238	1.196	-	710	245	1.087	-
SS-Twitter	4.242	1.037	1.953	1.252	-	-	-
Sanders	5.513	654	2.503	570	-	-	1.786
GASP	12.771	5.235	6.268	1.050	-	218	-
WAB	13.340	2.580	3.707	2.915	-	420	3.718
SemEval	13.975	2.186	6.440	5.349	-	-	-

Πίνακας 6. Συνοπτικός πίνακας των datasets για Ανάλυση Συναισθήματος σε tweets

2.2.3.9 Σχετικές Δημοσιεύσεις

Ακολουθεί περιγραφή σημαντικών δημοσιευμένων εφαρμογών Ανάλυσης Συναισθήματος και συγκριτική παράθεση τους:

- Οι Pang και Lee σε δημοσίευση τους ("Thumbs up? Sentiment Classification using Machine Learning Techniques", 2002)[18] επιχειρούν Ανάλυση Συναισθήματος σε κριτικές ταινιών που προέρχονται από τη βάση ταινιών IMDb, με χρήση ***Επιβλεπόμενης Μηχανικής Μάθησης***, και μελετούν την απόδοση της στην προσπάθεια τους να αντιπαραθέσουν την εργασία Ανάλυσης Συναισθήματος με την εργασία εξαγωγής θέματος από κείμενο. Οι κριτικές - δεδομένα που χρησιμοποιηθήκαν για την εκπαίδευση συμπεριλάμβαναν και βαθμολογία εκφρασμένη με αριθμητική τιμή ή αστέρια, η οποία στη συνέχεια αντιστοιχήθηκε στις εξής κατηγορίες συναισθήματος : αρνητικό, θετικό, ουδέτερο. Σε πρώτη φάση, οι δύο ερευνητές με τη βοήθεια ανθρώπων δημιούργησαν συλλογές λέξεων που σύμφωνα με τη γνώμη τους εκφράζουν αρνητικό ή θετικό συναίσθημα σε κριτικές ταινιών. Με εφαρμογή των συλλογών αυτών πάνω σε δεδομένα, προέκυψαν ποσοστά ακρίβειας (accuracy) που κυμαίνονται από 50-69%, τα οποία στη συνέχεια της έρευνας χρησιμοποιήθηκαν ως αναφορά για αξιολόγηση αποτελεσμάτων της Μηχανικής Μάθησης. Για την Ανάλυση Συναισθήματος χρησιμοποιήθηκαν οι ταξινομητές ***Naive Bayes***, ***SVM*** και ***Maximum Entropy***. Επιλέχθηκαν τυχαία 700 θετικές και 700 αρνητικές κριτικές ταινιών και μοιράστηκαν ομοιόμορφα σε τρία υποσύνολα. Ακολούθησε προεπεξεργασία των τριών υποσυνόλων δεδομένων κατά την οποία τα σημεία στίξης αντιμετωπίστηκαν σαν στοιχεία του λεξιλογίου, ενώ δεν αφαιρέθηκαν stop words, ούτε εφαρμόστηκε stemming.Τα πειράματα έγιναν τόσο με χρήση unigrams όσο και με χρήση bigrams. Για τη διαχείριση της άρνησης, στην περίπτωση των unigrams, υιοθετήθηκε η προσέγγιση της επισήμανσης των λέξεων μεταξύ μιας λέξης άρνησης και του αμέσως επόμενου σημείου στίξης με την ετικέτα NOT_. Επίσης, χρησιμοποιήθηκε αναγνώριση μέρους του λόγου (POS – tagging) αλλά και αναγνώριση μόνο των επιθέτων (adjective – tagging).Τα πειράματα που περιείχαν μόνο unigrams πραγματοποιήθηκαν είτε λαμβάνοντας υπόψη μόνο την ύπαρξη ενός χαρακτηριστικού (presence) είτε λαμβάνοντας υπόψη τη συχνότητα εμφάνισης του χαρακτηριστικού (frequency). Τέλος, λήφθηκε υπόψη και η θέση των χαρακτηριστικών στην κριτική(position), αφού μια κριτική ταινίας συνήθως ξεκινά με τη δήλωση του γενικού συναισθήματος του κριτή για την ταινία, ακολουθεί μια ανάλυση της πλοκής και κλείνει με τη σύνοψη των απόψεων του συγγραφέα.

  Ακολουθούν τα ποσοστά ακρίβειας που προέκυψαν από τα διάφορα πειράματα που περιγράφηκαν παραπάνω:

	Features	# of features	frequency or presence?	NB	MaxEnt	SVM
(1)	unigrams	16.165	freq.	**78,7**	N/A	72,8
(2)	unigrams	16.165	pres.	81,0	80,4	**82,9**
(3)	unigrams + bigrams	32.330	pres.	80,6	80,8	**82,7**
(4)	bigrams	16.165	pres.	77,3	**77,4**	77,1
(5)	unigrams +POS	16.695	pres.	81,5	80,4	**81,9**
(6)	adjectives	2.633	pres.	77,0	**77,7**	75,1
(7)	top 2633 unigrams [3]	2.633	pres.	80,3	81,0	**81,4**
(8)	unigrams + position	22.430	pres.	81,0	80,1	**81,6**

Πίνακας 7. Accuracy πειραμάτων

Σε σύγκριση με τα ποσοστά αναφοράς που προέκυψαν, τα αποτελέσματα των πειραμάτων είναι αρκετά καλά. Ο ταξινομητής Naive Bayes έχει τη χειρότερη ακρίβεια, ενώ ο ταξινομητής SVM την καλύτερη, με μικρές διαφορές βέβαια. Οι συγγραφείς καταλήγουν στο συμπέρασμα ότι απαιτείται πιο λεπτομερής προσέγγιση των κριτικών ταινιών, η οποία θα περιλαμβάνει το διαχωρισμό λέξεων που εκφράζουν άποψη του κριτή από αυτές που περιγράφουν την ίδια την ταινία.

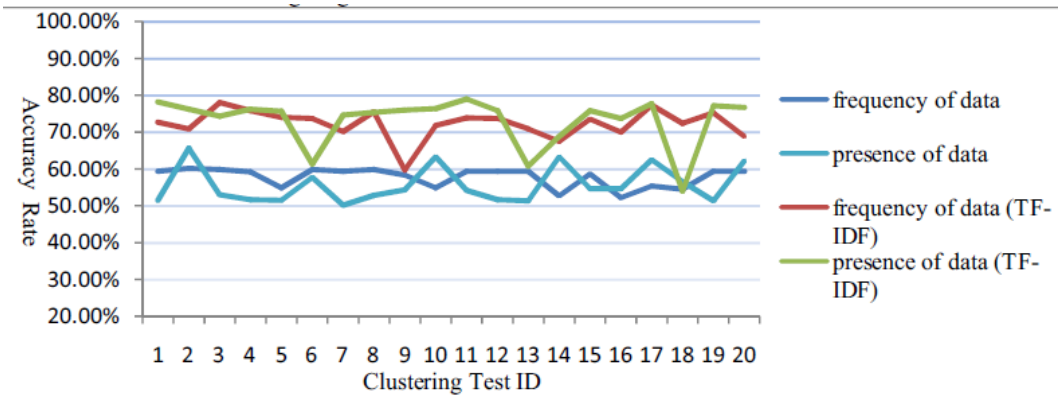
- Οι Alec Go et al (2009) σε δημοσίευση τους (Twitter Sentiment Classification using Distant Supervision)[36] επιχειρούν τη δημιουργία ενός συστήματος αυτόματης Ανάλυσης Συναισθήματος, το οποίο ταξινομεί αγγλικά tweets σε θετικά ή αρνητικά σε σχέση με τον όρο αναζήτησης (query term) που εισάγεται. Το σύστημα βασίζεται σε **Επιβλεπόμενη *Μηχανική Μάθηση με χρήση των αλγορίθμων Naive Bayes, Maximum Entropy και SVM***. Για την αποφυγή του χρονοβόρου και απαιτητικού χειρονακτικού σχολιασμού επιλέγεται η χρήση της μεθόδου distant supervision, κατά την οποία μέσω του Twitter API επιλέγονται μόνο tweets με emoticons, ώστε ο σχολιασμός των tweets του training set να προκύψει από τα emoticons που περιέχουν. Σημειώνεται, επίσης, ότι εξαιρέθηκαν τα ουδέτερα tweets. Ως χαρακτηριστικά στα διάφορα πειράματα χρησιμοποιήθηκαν είτε unigrams,είτε bigrams, είτε unigrams με bigrams, είτε unigrams με POS tags, ενώ χρησιμοποιήθηκε η ετικέτα QUERY TERM στις εμφανίσεις του όρου αναζήτησης για την αποφυγή αλλοίωσης της ταξινόμησης από το συναισθηματικό περιεχόμενο που ίσως ενέχει ο όρος αναζήτησης. Τα emoticons μετά τη χρήση τους ως ετικέτες συναισθήματος αφαιρούνται από το training set, ώστε να μην επηρεάσουν την εκπαίδευση των ταξινομητών SVM και Maxent (Ο Naive Bayes δεν επηρεάζεται). Κατά την προεπεξεργασία επισημαίνονται με ετικέτες οι αναφορές σε ονόματα χρηστών (@username) και οι σύνδεσμοι (links), ώστε να αγνοηθούν κατά την ανάλυση, ενώ οι συνεχόμενες πολλαπλές επαναλήψεις ενός γράμματος αντικαθίστανται από δύο επαναλήψεις του. Επίσης αφαιρούνται από το training set τα tweets που περιλαμβάνουν και αρνητικό και θετικό περιεχόμενο, τα Retweets, τα tweets emoticon :P (το οποίο λανθασμένα επιστρέφεται στο query ":(" από το API),

και τα επαναλαμβανόμενα tweets. Μετά από αυτή την επεξεργασία, το σύστημα εκπαιδεύεται με 800.000 θετικά και 800.000 αρνητικά tweets. Το test set αντλείται από το Twitter API (με queries προϊόντα, ονόματα και εταιρείες) και χαρακτηρίζεται χειρονακτικά ώστε τελικά να περιλαμβάνει 177 αρνητικά και 182 θετικά tweets, τα οποία δεν έχει σημασία αν περιέχουν ή όχι emoticons, αφού κατά την εκπαίδευση αγνοήθηκαν. Ως αποτελέσματα αναφοράς χρησιμοποιήθηκαν τα αποτελέσματα που προέκυψαν από το site Ανάλυσης Συναισθήματος Twitrratr το οποίο υλοποιεί ανάλυση βασισμένη σε keywords και Λεξικό Συναισθήματος. Ακολουθούν τα συγκεντρωτικά αποτελέσματα.

Features	Keyword	NB	MaxEnt	SVM
Unigram	65,2	81,3	80,5	82,2
Bigram	N/A	81,6	79,1	78,8
Unigram + Bigram	N/A	82,7	83,0	81,6
Unigram + POS	N/A	79,9	79,9	81,9

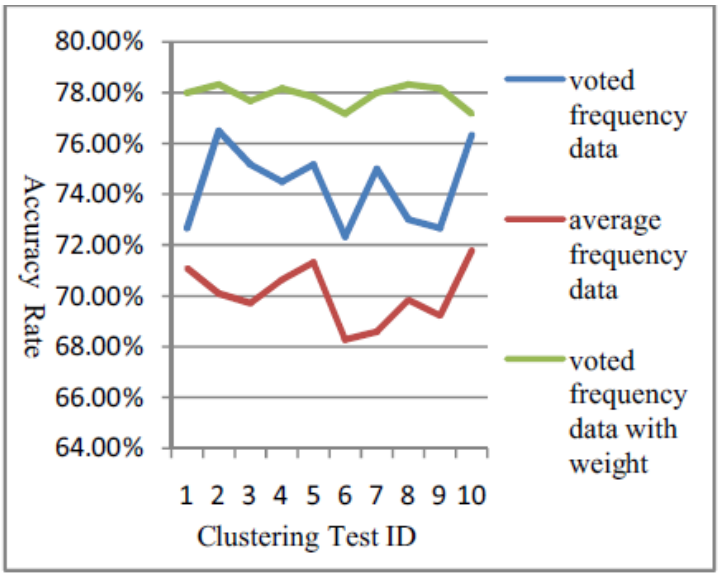
Πίνακας 8. Accuracy ταξινομητή

- Οι Li και Liu σε δημοσίευση τους [60] προτείνουν την χρήση ***συσταδοποίησης*** για τη ***Μη Επιβλεπόμενη*** συναισθηματική ανάλυση αγγλικών κριτικών ταινιών, σε επίπεδο κειμένου, σε αντιπαράθεση με τις διαδεδομένες συμβολικές μεθόδους (με χρήση λεξικού) οι οποίες παρουσιάζουν χαμηλά ποσοστά ακρίβειας και τις μεθόδους Επιβλεπόμενης Μάθησης που αν και ακριβείς είναι χρονοβόρες και απαιτούν ανθρώπινη συμμετοχή. Στόχος τους ήταν η ταξινόμηση των δεδομένων σε αρνητική και θετική κλάση συναισθήματος. Η προεπεξεργασία των 300 θετικών και 300 αρνητικών κειμένων περιλαμβάνει stemming, POS – tagging, μετατροπή των κειμένων σε διανύσματα τόσο με συνιστώσες που εκφράζουν συχνότητα των χαρακτηριστικών όσο και με συνιστώσες που εκφράζουν παρουσία ή απουσία των χαρακτηριστικών, ενώ απομακρύνθηκαν οι ετικέτες των δεδομένων. Για τη συσταδοποίηση χρησιμοποιήθηκε ο αλγόριθμος ***k-means*** με συνάρτηση απόστασης την απόσταση συνημιτόνου (cosine distance). Λόγω της αστάθειας του k-means έγιναν 20 επαναλήψεις του πειράματος για κάθε μια από τις δύο διανυσματικές αναπαραστάσεις. Ωστόσο, τα μέγιστα ποσοστά ακρίβειας που επιτεύχθηκαν ήταν 60,17% και 65,67%, αντίστοιχα. Επομένως για την βελτίωση της μεθόδου κρίθηκε απαραίτητη κατά την προεπεξεργασία η χρήση του κριτηρίου tf-ifd[4] για την επιλογή των πιο σημαντικών χαρακτηριστικών. Η αστάθεια των αποτελεσμάτων παρέμεινε, ωστόσο επιτεύχθηκε μέγιστη ακρίβεια 79%.



Εικόνα 9. Αποτελέσματα μετά την εφαρμογή του tf-idf

Για τον περιορισμό της αστάθειας των αποτελεσμάτων χρησιμοποιήθηκε μηχανισμός συμψηφισμού των πολλαπλών επαναλήψεων. Στη συνέχεια, το πείραμα εμπλουτίζεται υπολογίζοντας τιμές-βάρη για τα χαρακτηριστικά ανάλογα με την σχέση τους με τις λέξεις αναφοράς "good" και "bad", με τη βοήθεια του WordNet, και κρατώντας μόνο τα χαρακτηριστικά με υψηλά βάρη. Η προσθήκη αυτή αφενός μειώνει τις διαστάσεις των διανυσμάτων, κάνοντας τη διαδικασία της ανάλυσης πιο σύντομη και αφετέρου επιτρέπει να εντοπίσουμε σε ποια κλάση αντιστοιχεί η κάθε συστάδα υπολογίζοντας το μέσο όρο των βαρών, ενώ βελτιώνει σημαντικά την ακρίβεια στην περίπτωση των διανυσμάτων συχνότητας χαρακτηριστικών, όπου επιτυγχάνεται ακρίβεια μέχρι και 78,33%.



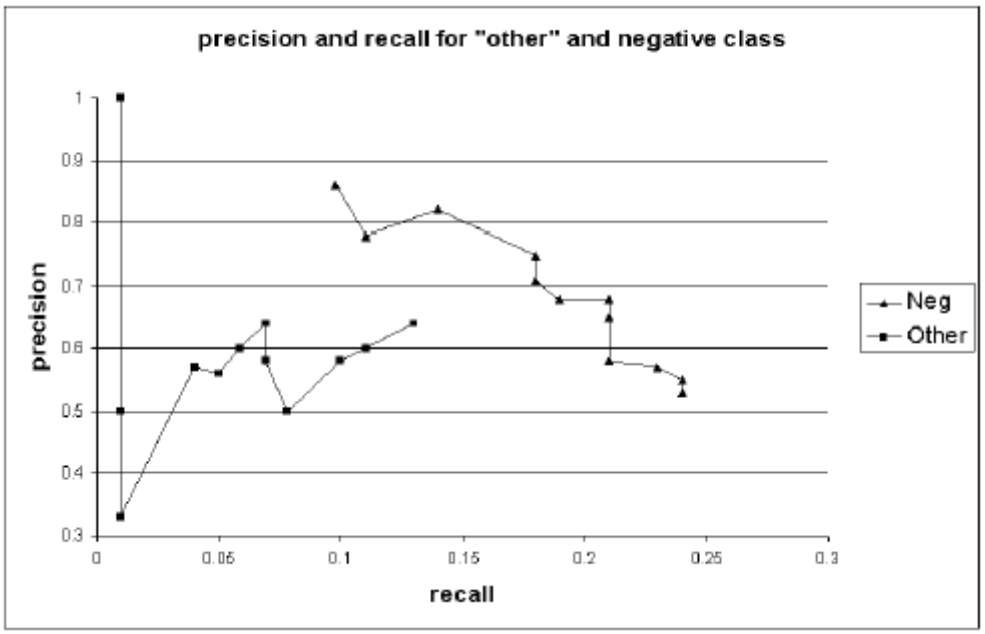
Εικόνα 10. Αποτελέσματα μετά τη χρήση βαρών

Ακολούθως παρουσιάζεται πίνακας σύγκρισης της τελικής μεθόδου συσταδοποίησης με τις συμβολικές και επιβλεπόμενες τεχνικές.

	Accuracy	Efficiency	Human participation
Symbolic Techniques	65,83 % (Turney 2002)	Very fast	Mostly No
Supervised Learning	77% - 82 % (Pang 2002)	Slow on training & fast on test	Yes
Clustering – based Approach	77,17% - 78,33%	Fast	No

Πίνακας 9. Αξιολόγηση των τριών μεθόδων

- Οι Gamon et al. [66] δημιούργησαν ένα σύστημα (Pulse) το οποίο εξάγει αυτόματα το θέμα (topic) και το συναίσθημα (sentiment) από σχόλια πελατών για αυτοκίνητα. Η ανάλυση τόσο του θέματος όσο και του συναισθήματος έγινε σε επίπεδο πρότασης (sentence- level). Ως σύνολο δεδομένων χρησιμοποιήθηκε ένα δείγμα 406.818 κριτικών αυτοκινήτων με μέγεθος από μία έως 50 προτάσεις η καθεμία, οι οποίες συνοδεύονταν από μία βαθμολογία για κάθε συνολική κριτική. Όσον αφορά την Ανάλυση Συναισθήματος, η αρχική βαθμολογία αγνοήθηκε και από το σύνολο των κριτικών επιλέχθηκαν 3.000 προτάσεις στις οποίες έγινε χειρονακτικός σχολιασμός ως positive, negative ή other, από τις οποίες οι 2.500 χρησιμοποιήθηκαν στη συνέχεια για την εκπαίδευση ενώ οι υπόλοιπες 500 για την αξιολόγηση. Το μέγιστο inter - annotator agreement υπολογίστηκε ίσο με 79,93%, γεγονός που υποδηλώνει την δυσκολία του προβλήματος. Λόγω της έλλειψης αρχικού σχολιασμού των προτάσεων επιλέχθηκε μια μέθοδος μάθησης με μερική επίβλεψη, που απαιτεί τα ελάχιστα δυνατά σχολιασμένα δεδομένα. Συγκεκριμένα χρησιμοποιήθηκε ο αλγόριθμος ***Expectation- Maximization σε συνδυασμό με Naive Bayes.*** Στο γνωστό αλγόριθμο εισήχθηκε μια παράμετρος δ, η οποία ρυθμίζει το βάρος που αποδίδεται στα μη σχολιασμένα δεδομένα. Για τη δημιουργία του μοντέλου χρησιμοποιήθηκαν συνολικά 9.000 μη σχολιασμένες προτάσεις και 3.000 χειρονακτικά σχολιασμένες. Η προεπεξεργασία των δεδομένων περιλάμβανε μετατροπή όλων των γραμμάτων σε μικρά και κάθε αριθμός μετατράπηκε σε μια απλή λέξη. Κάθε πρόταση μετατράπηκε σε διάνυσμα με δυαδικές συνιστώσες που αντιστοιχούν σε κάθε διαφορετική λέξη ή σημείο στίξης. Λόγω του μεγάλου πλήθους θετικών προτάσεων έναντι των άλλων κλάσεων και επομένως της τάσης του συστήματος προς τη θετική κλάση, η αξιολόγηση έγινε ως προς τις κλάσεις negative και other. Προέκυψε ότι τι σύστημα μπορεί να πετύχει υψηλό precision στις δύο αυτές μειονεκτούσες κλάσεις, εις βάρος του recall, πράγμα αποδεκτό σε τομείς με πολλά σχόλια πελατών. Το recall της θετικής κλάσης κυμάνθηκε από 0,95 μέχρι 0,97. Ακολουθεί το διάγραμμα precision – recall για τις κλάσεις negative και other.



Εικόνα 11. Διάγραμμα precision – recall για τις κλάσεις negative και other.

2.2.3.10 Διαχείριση tweets

Στην παρούσα εργασία εστιάζουμε στην αυτόματη Ανάλυση Συναισθήματος μικρών κειμένων που προέρχονται από δικτυακές πηγές (microblogging), όπως οι ιστοχώροι κοινωνικής δικτύωσης (social networking sites) ή σελίδες στις οποίες ο χρήστης καλείται να σχολιάσει σύντομα κάποιο αντικείμενο. Αντιπροσωπευτικό παράδειγμα τέτοιας μορφής κειμένου είναι τα tweets, δηλαδή οι αναρτήσεις χρηστών του ιστοχώρου κοινωνικής δικτύωσης Twitter, τα οποία και πρόκειται να μελετήσουμε στα επόμενα κεφάλαια. Χαρακτηριστική ιδιότητα των tweets είναι το μικρό τους μήκος, που περιορίζεται στους 140 χαρακτήρες. Λόγω των περιορισμένων διαθέσιμων χαρακτήρων, οι χρήστες του Twitter συνηθίζουν να χρησιμοποιούν συντομογραφίες λέξεων (π.χ. *lol, cu, bff*), emoticons[5] (:) :P :S) και εσκεμμένα ή μη εσκεμμένα τροποποιημένες ορθογραφικά λέξεις (π.χ. *luv, lovin*). Επίσης, συχνά συμπεριλαμβάνουν hastags[6], αναφορές[7] σε άλλους χρήστες του Twitter και συνδέσμους (URL) προς άλλες σελίδες. Τέλος, οι χρήστες δεν επιδιώκουν τη δημιουργία ορθών συντακτικά φράσεων, αλλά την παράθεση ατελών φράσεων ή ασύνδετων λέξεων μέσω των οποίων εκφράζονται σύντομα και περιεκτικά. Τα παραπάνω μαρτυρούν ότι η Ανάλυση Συναισθήματος σε tweets απαιτεί ειδική προσέγγιση με ιδιαίτερη προσοχή κατά την οποία πρέπει οπωσδήποτε να ληφθούν υπόψη τα διάφορα στοιχεία της γλώσσας του διαδικτύου, τα οποία μπορεί να φέρουν σημαντικότατη συναισθηματική πληροφορία. Η δυσκολία έγκειται στο γεγονός ότι τα παραπάνω στοιχεία είναι αμέτρητα, αλλάζουν διαρκώς, και συνεχώς προστίθενται νέα, ανάλογα με τις τάσεις του διαδικτύου.
Ακολουθούν παραδείγματα από tweets στα οποία διαφαίνονται οι ιδιαιτερότητες που περιγράφηκαν παραπάνω:

“I bought iPad yesterday, just lovvee it :-)”.

“@epiphanygirl we gon’ be like 15 deep coming to see u on Saturday in atlantic city. We luuuuvvvvv you!!! Lol!”

“It’s soooo funny... Guys, U should watch it! (=” -

- Στην περίπτωση Ανάλυσης Συναισθήματος σε tweets με χρήση Λεξικού Συναισθήματος, το λεξικό που χρησιμοποιείται θα πρέπει να εμπλουτίζεται με τα βασικά τουλάχιστον emoticons, καθώς και τις διαδεδομένες διαφορετικές εκδοχές γραφής ορισμένων λέξεων (π.χ. love = looove = loveee = luv, you = U = ya κ.α.). Βέβαια, μια τέτοια προσπάθεια πρέπει συνεχώς να ανανεώνεται, χωρίς και πάλι να είναι εγγυημένα επαρκής.

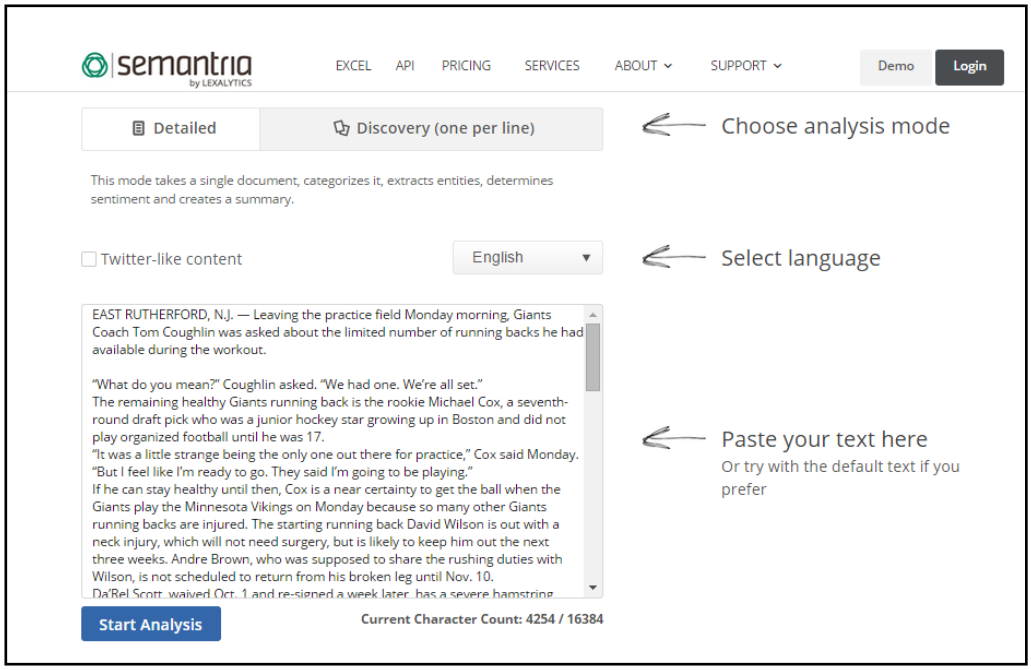
- Στην περίπτωση Ανάλυσης Συναισθήματος σε tweets με χρήση Μηχανικής Μάθησης, θα πρέπει να δοθεί ιδιαίτερη προσοχή κατά την προεπεξεργασία του συνόλου δεδομένων. Όπως έχουμε δει προηγουμένως, ένα στάδιο της προεπεξεργασίας μπορεί να είναι η αφαίρεση σημείων στίξης και αριθμών. Στην περίπτωση των tweets θα ήταν προτιμότερο να εντοπιστούν αρχικά πιθανά emoticons και στη συνέχεια να αφαιρεθούν τα άχρηστα σημεία στίξης και αριθμοί. Επίσης, απαραίτητη κρίνεται η απομάκρυνση των συνδέσμων και η επέκταση συντομεύσεων.

2.2.3.11 APIs για Ανάλυση Συναισθήματος

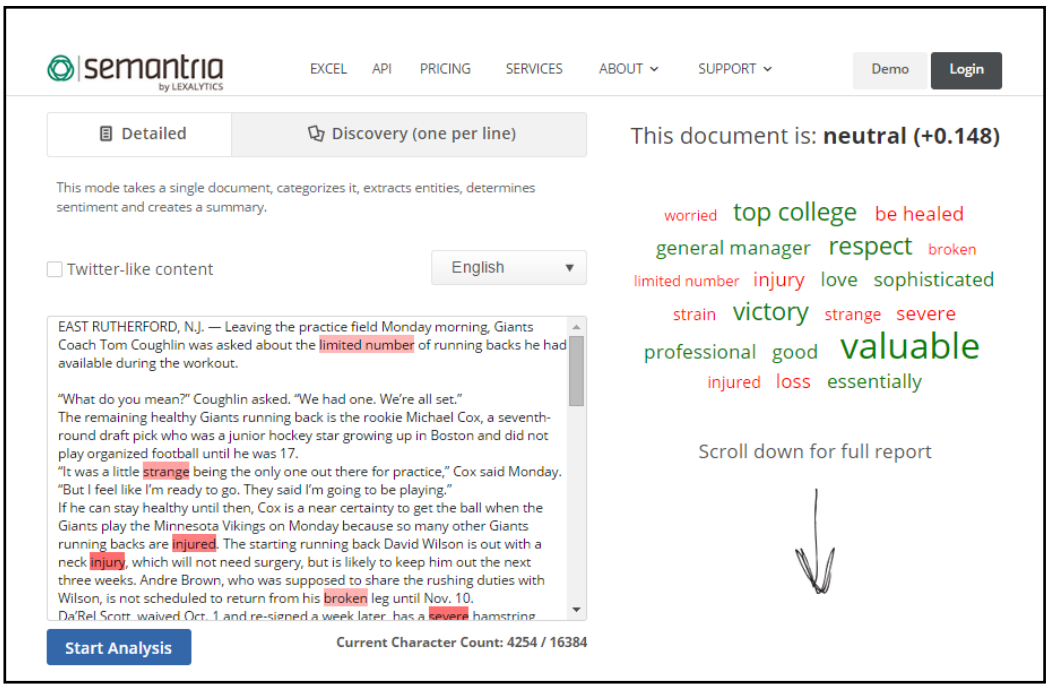
Τα τελευταία χρόνια έχουν αναπτυχθεί πολλά software για Ανάλυση Συναισθήματος τα οποία είναι διαθέσιμα στο διαδίκτυο ως APIs που μπορούν να ενσωματωθούν σε οποιαδήποτε εφαρμογή. Τα περισσότερα από αυτά πέρα από το συναίσθημα, εξάγουν και άλλα χαρακτηριστικά από το κείμενο, όπως το θέμα και οι οντότητες για τις οποίες γίνεται λόγος. Ακολουθεί η περιγραφή μερικών από αυτά.

- Το API της εταιρείας ανάλυσης κειμένου **Semantria** επιτρέπει την ανάλυση οποιουδήποτε κειμένου (μέχρι 16.384 χαρακτήρες) είναι γραμμένο σε κάποια από τις έντεκα διαθέσιμες γλώσσες. Από την ανάλυση του κειμένου προκύπτει το συνολικό του συναίσθημα που μπορεί να είναι αρνητικό (negative), θετικό (positive) ή ουδέτερο (neutral), συνοδευόμενο από μια αντίστοιχη βαθμολογία. Επίσης εξάγονται οι οντότητες που εμφανίζονται σ’ αυτό, οι κατηγορίες στις οποίες ανήκουν οι οντότητες, τα θέματα στα οποία αναφέρεται το κείμενο και η περίληψη του. Καθένα από τα επιμέρους χαρακτηριστικά που εξάγονται συνοδεύεται και από την αντίστοιχη βαθμολογία ως προς το συναισθηματικό περιεχόμενο. Η Ανάλυση Συναισθήματος υλοποιείται με συντακτική ανάλυση του κειμένου αρχικά και εντοπισμό φράσεων με συναισθηματικό περιεχόμενο στη συνέχεια. Η τελική

βαθμολογία προκύπτει από το συνδυασμό των βαθμολογιών των επιμέρους φράσεων. Εκτός από το σχετικό API διατίθεται και plug-in της εφαρμογής για το για το excel. Ακολουθούν βασικά στιγμιότυπα από το demo της εφαρμογής. (https://semantria.com/demo)

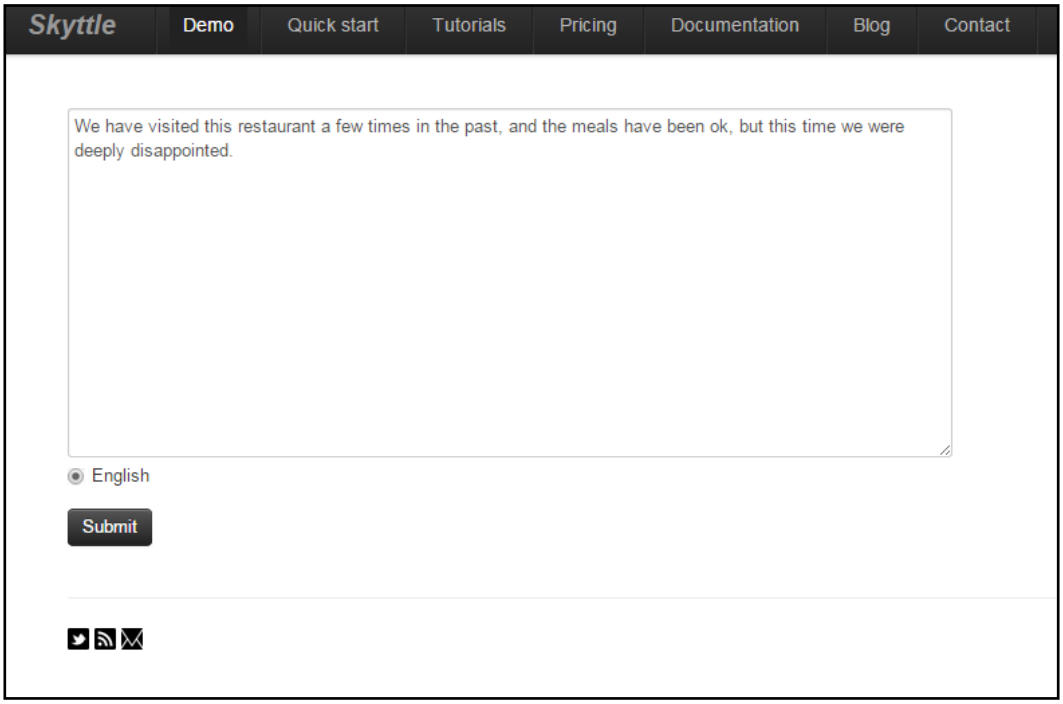


Εικόνα 12. Στιγμιότυπο Semantria 1

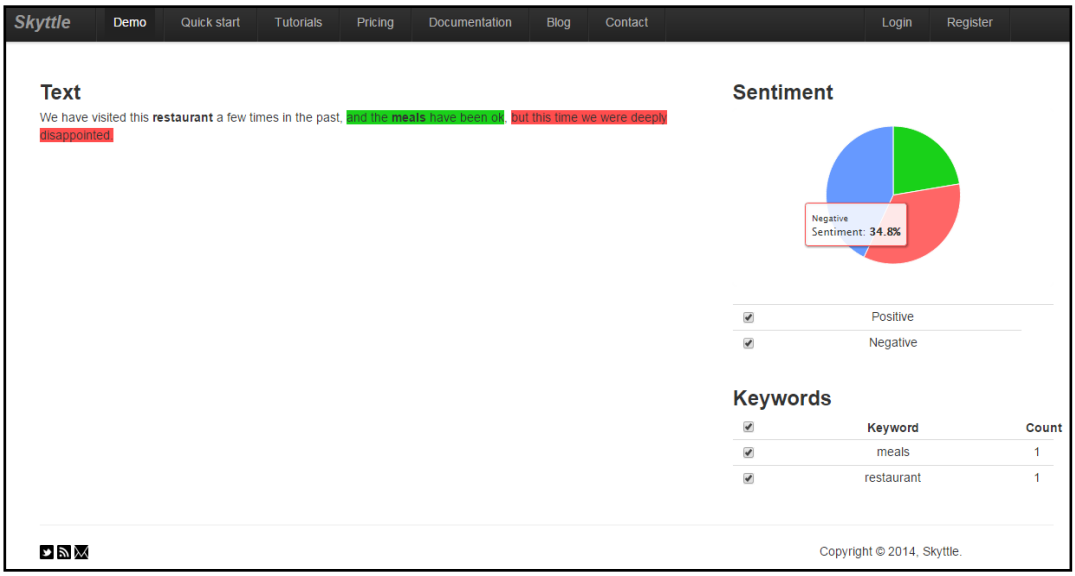


Εικόνα 13. Στιγμιότυπο Semantria 2

- Το **Skyttle** είναι ένα API που επιτρέπει την ανάλυση κειμένου σε επίπεδο φράσης, σε τέσσερις αποδεκτές γλώσσες. Η ανάλυση περιλαμβάνει την απόδοση συναισθηματικού χαρακτηρισμού (positive, negative, neutral) στις φράσεις του κειμένου, τον υπολογισμό ποσοστών για τις κατηγορίες συναισθήματος που εμφανίζονται και τον εντοπισμό των keywords του κειμένου. Ακολουθούν στιγμιότυπα από το demo της εφαρμογής (http://www.skyttle.com/demoin).

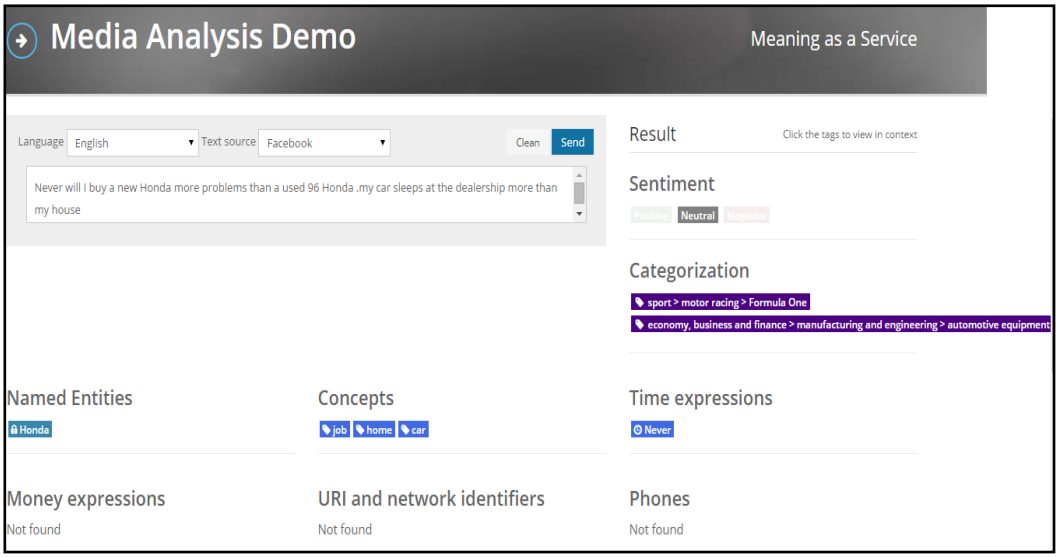


Εικόνα 14. Στιγμιότυπο Skyttle 1



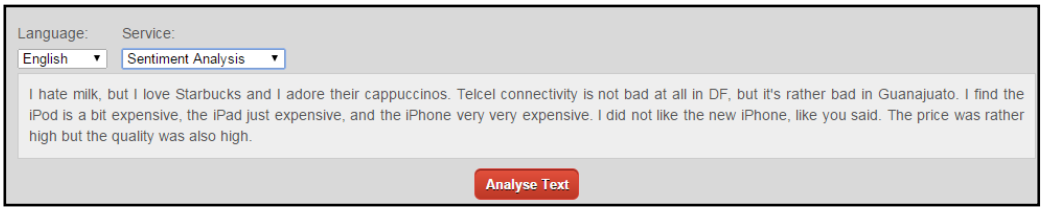
Εικόνα 15. Στιγμιότυπο Skyttle 2

- Το **Textalytics Media Analysis API** επιτρέπει την ανάλυση κειμένου που προέρχεται από κοινωνικά δίκτυα, forums, blogs αλλά και sites ειδήσεων σε Αγγλική ή Ισπανική γλώσσα. Από την ανάλυση προκύπτει η κατηγορία συναισθήματος του κειμένου (negative, positive, neutral) καθώς και κατηγορίες, οντότητες και άλλες χαρακτηριστικές πληροφορίες. Ακολουθεί στιγμιότυπο από το demo της εφαρμογής (https://textalytics.com/api-text-analysis-demo-en).

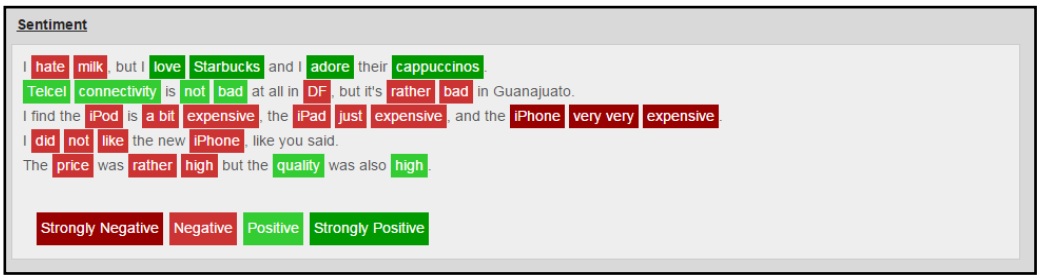


Εικόνα 16. Στιγμιότυπο Textalytics Media Analysis

- Το **Bittext API** υλοποιεί ανάλυση κειμένου σε επτά γλώσσες με τη μέθοδο Deep Linguistic Analysis. Με αυτή την ανάλυση είναι ικανό να εξάγει συναίσθημα, οντότητες, έννοιες και κατηγορίες από το κείμενο. Ακολουθούν στιγμιότυπα από το demo της εφαρμογής (http://www.bitext.com/api-demo.html).

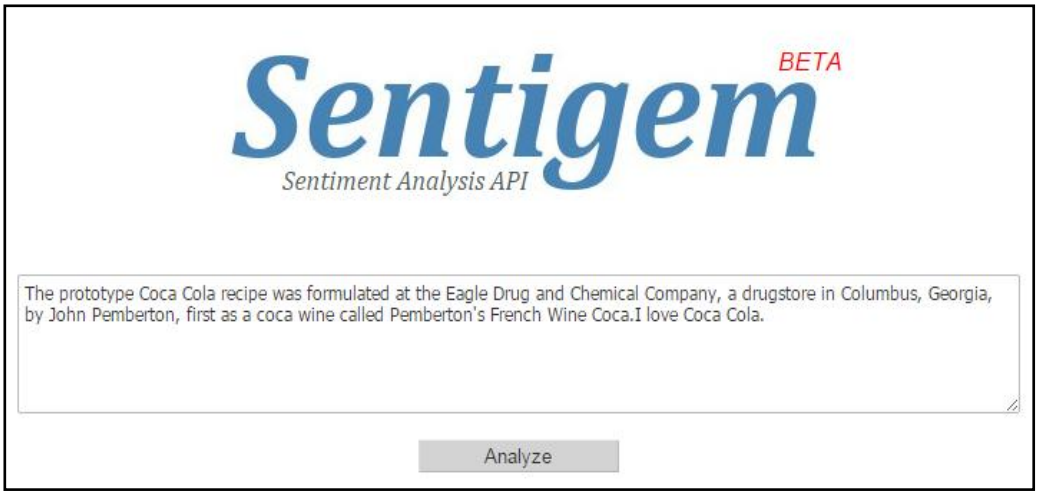


Εικόνα 17. Στιγμιότυπο Bitttext 1

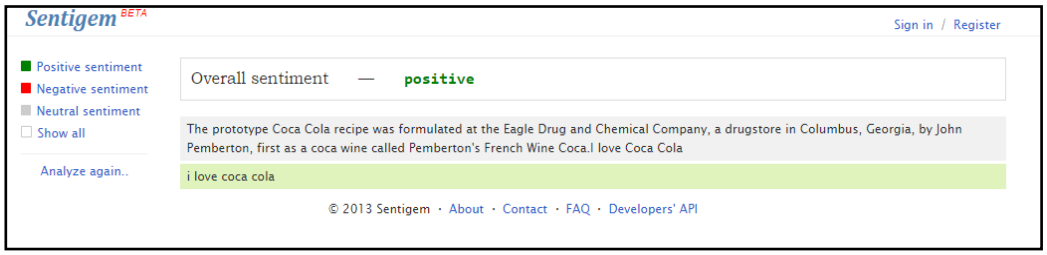


Εικόνα 18. Στιγμιότυπο Bitttext 2

- Το **Sentigem** είναι API που υλοποιεί μόνο Ανάλυση Συναισθήματος σε κείμνο στα Αγγλικά. Υπολογίζει το συνολικό συναίσθημα του κειμένου καθώς και το συναίσθημα των επιμέρους φράσεων. Ακολουθούν στιγμιότυπα από το demo της εφαρμογής(http://sentigem.com/#!).



Εικόνα 19. Στιγμιότυπο Sentigem 1



Εικόνα 20. Στιγμιότυπο Sentigem 2

3. Εφαρμογή Μη Επιβλεπόμενης Μηχανικής Μάθησης σε πραγματικά δεδομένα

3.1 Περιγραφή του πειράματος

Έχοντας μελετήσει το θεωρητικό υπόβαθρο της αυτόματης Ανάλυσης Συναισθήματος σε κείμενο στα προηγούμενα κεφάλαια, στο παρόν κεφάλαιο θα προχωρήσουμε στην εφαρμογή της βασισμένης σε Μηχανική Μάθηση μεθόδου Ανάλυσης Συναισθήματος σε πραγματικά δεδομένα. Η μέθοδος που επιλέξαμε να εφαρμόσουμε είναι αυτή της Μη Επιβλεπόμενης Μηχανικής Μάθησης (clustering, συσταδοποίηση), λόγω του ότι το μεγαλύτερο μέρος των μέχρι σήμερα ερευνών στο πεδίο της Ανάλυσης Συναισθήματος εστιάζει στην χρήση Λεξικού Συναισθήματος και στην Επιβλεπόμενη Μηχανική Μάθηση, παραμελώντας την Μη Επιβλεπόμενη Μηχανική Μάθηση. Στόχος μας είναι να μελετήσουμε κατά πόσο μπορεί να είναι αποτελεσματική μια Μη Επιβλεπόμενη μέθοδος Ανάλυσης Συναισθήματος αλλά και να εντοπίσουμε τους λόγους για τους οποίους διστάζουν να επενδύσουν σ' αυτή οι ερευνητές.

3.2 Το σύνολο δεδομένων του πειράματος

Στα πλαίσια της παρούσας διπλωματικής χρησιμοποιήθηκε ως σύνολο δεδομένων της Μηχανικής Μάθησης ένα σύνολο από 4.009 tweets (μαζί με retweets) σχετικά με τον τουρισμό στην Ελλάδα. Τα tweets μας παραχωρήθηκαν από την εταιρεία ανάλυσης δεδομένων Brandwatch (http://www.brandwatch.com/) και το κοινό χαρακτηριστικό τους, με βάση το οποίο έγινε η εξόρυξη τους είναι η παρουσία του username @VisitGreecegr**.** Τα tweets μας δόθηκαν χαρακτηρισμένα ως προς το συναισθηματικό περιεχόμενο ως positive (θετικά), negative (αρνητικά) και neutral (ουδέτερα). Ο χαρακτηρισμός αυτός έγινε με αυτόματο τρόπο, με βάση την εμφάνιση κάποιων πολύ χαρακτηριστικών συναισθηματικών λέξεων. Το σύνολο των tweets χαρακτηρίστηκε από εμάς χειρονακτικά εκ νέου, διότι παρατηρήσαμε ότι το σύστημα με βάση το οποίο είχε γίνει ο αρχικός σχολιασμός δεν ήταν ιδιαίτερα ευαίσθητο, με αποτέλεσμα να χαρακτηρίζει τα περισσότερα tweets ως ουδέτερα. Επίσης, θεωρήσαμε χρήσιμη την προσθήκη μιας νέας κλάσης συναισθήματος, στην οποία θα ταξινομούνται τα tweets των οποίων είναι αδύνατη η ταξινόμηση με αξιολόγηση μόνο της πληροφορίας του κειμένου του tweet και όχι τυχόντων διευκρινιστικών συνοδευτικών συνδέσμων, καθώς είτε περιλαμβάνουν ταυτόχρονα αρνητικό και θετικό συναίσθημα, είτε χρησιμοποιούν συναισθηματικές φράσεις οι οποίες μπορεί να εμπεριέχουν κατά περίπτωση είτε αρνητικό είτε θετικό νόημα. Στην κατηγορία αυτή, εντάξαμε επίσης τα tweets που ήταν γραμμένα σε γλώσσες εκτός Ελληνικών και Αγγλικών.

Συγκεκριμένα, ορίσαμε τις εξής κατηγορίες συναισθήματος :

Κατηγορία Συναισθήματος	**Σύμβολο**	**Περιγραφή**
Positive (Θετικό)	+	Το tweet εκφράζει θετικό συναίσθημα.
Negative (Αρνητικό)	-	Το tweet εκφράζει θετικό συναίσθημα.
Neutral (Ουδέτερο)	=	Το tweet δεν εκφράζει συναίσθημα.
Undefined (Απροσδιόριστο)	*	Το tweet φαίνεται να εκφράζει συναίσθημα το οποίο δεν μπορούμε να προσδιορίσουμε διαβάζοντας μόνο το κείμενο.

Πίνακας 10. Σύμβολα που χρησιμοποιήθηκαν για το σχολιασμό

Κατά τον χειρωνακτικό σχολιασμό μας αγνοήθηκαν οι σύνδεσμοι, λήφθηκαν υπόψη τα hashtags που περιείχαν συναισθηματικές λέξεις (π.χ. #loveGreece), οι αναφορές σε χρήστες των οποίων το όνομα αποκαλύπτει συναίσθημα (π.χ. @ilovekavala, @WeLoveKefalonia, @loveZante) καθώς και τα emoticons. Επίσης, όπως συνηθίζεται, προσαρμόσαμε την ανάλυση στον συγκεκριμένο νοηματικό τομέα στον οποίο αναφέρονται τα tweets του dataset, δηλαδή τον τουρισμό στην Ελλάδα, ώστε να είναι πιο ακριβής. Για παράδειγμα, το tweet RT @VisitGreecegr: Clear waters in #Halkidiki via @VisitHalkidiki #Greece #ttot @VeryMacedonia #travel pic.twitter.com/pEyPrbxelR", που μιλά για καθαρά νερά στη Χαλκιδική θεωρούμε ότι εκφράζει θετικό συναίσθημα στα πλαίσια του τουρισμού στην Ελλάδα.

Ακολουθούν χαρακτηριστικά παραδείγματα του χειρονακτικού σχολιασμού μας:

Κατηγορία	**Παραδείγματα από το dataset**
Positive (Θετικό)	1. Fascinating Blue Caves on #Zakynthos island #Greece #travelpics photography by Alistair Ford @alistair_ford pic.twitter.com/zy4ctJbT5r 2. RT @VisitGreecegr: #Nafplio: One of the most romantic cities in #Greece! Find out more here: ow.ly/ynHx2 ow.ly/i/60yas 3. RT @penandpalate: #Greece : hospitality and cuisine as legendary as its mythology exm.nr/TiBZV9

	@VisitHalkidiki @VisitGreecegr pic.twitter.com/IF784enegP
Negative (Αρνητικό)	1. @mstoysav @VisitGreecegr who are those monsters who use cruelty against poor animals? Please make them pay to save dogs life 2. It was a bad idea putting the @VisitGreecegr on my Google+ feed. #BagsArePacked 3. @mstoysav @VisitGreecegr Whoever did this murder has not only shamed the village of Ziros and Sitia but the whole island of Crete.
Neutral (Ουδέτερο)	1. RT @VisitGreecegr: Rockwave Festival will take place this year on July 11 and 12, 2014. #ttot #Greece ow.ly/yAjJc 2. @VisitGreecegr We shared your #travel post on geotravellers.com under #Travel #GeoTravellers rbl.ms/1jasYoi 3. In #Greece, “YAH sahs!” is the way to say hello @VisitGreecegr
Undefined (Απροσδιόριστο)	1. @mstoysav @VisitGreecegr I don't understand please explain if to me.What's happening to those dogs...it's Incredible what I'm seeing 2. RT @VisitGreecegr: Nights in #Andros by @JamesCiccone @Gemstars81 #Greece pic.twitter.com/oVVf6Bn6gG" 3. A Crete Thrill Seeker’s Guide for Adventure blog.visitgreece.gr/a-crete-thrill… via @visitgreecegr

Πίνακας 11. Χαρακτηριστικά παραδείγματα από κάθε κλάση

Μετά τον σχολιασμό μας προέκυψε dataset που αποτελείται από 2.280 θετικά, 115 αρνητικά, 954 ουδέτερα, 660 απροσδιόριστα. Παρατηρούμε ότι ο αριθμός των αρνητικών tweets είναι πολύ μικρός, οπότε αναμένεται να αντιμετωπίσουμε δυσκολία ως προς τον

προσδιορισμό της αρνητικής συστάδας. Επίσης, η ανισότητα των κλάσεων είναι ένα ακόμη χαρακτηριστικό που ενδέχεται να προκαλέσει προβλήματα.

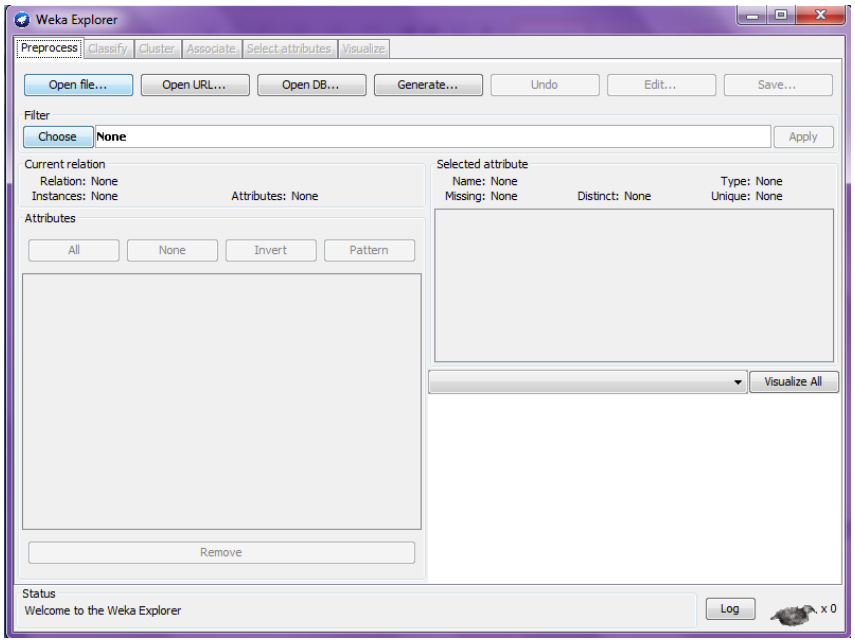
3.3 Η πλατφόρμα WEKA

Ως εργαλείο για την υλοποίηση της Ανάλυσης Συναισθήματος επιλέξαμε την πλατφόρμα WEKA (http://www.cs.waikato.ac.nz/ml/weka/). Το WEKA είναι ένα δωρεάν λογισμικό ανοιχτού κώδικα που περιλαμβάνει μια συλλογή αλγορίθμων Μηχανικής Μάθησης με αντικείμενο εργασίες σχετικές με την εξόρυξη δεδομένων, η οποία δημιουργήθηκε από στο Πανεπιστήμιο του Waikato της Νέας Ζηλανδίας. Οι αλγόριθμοι του WEKA είναι γραμμένοι σε Java και παρέχουν τις εξής δυνατότητες:

- Προεπεξεργασία δεδομένων
- Ταξινόμηση
- Συσταδοποίηση
- Εύρεση Κανόνων Συσχέτισης

Οι παραπάνω εργασίες μπορούν να γίνουν είτε μέσω του γραφικού περιβάλλοντος της εφαρμογής που έχει τρεις διαφορετικές εκδοχές, είτε μέσω γραμμής εντολών. Για τις ανάγκες τις διπλωματικής έγινε χρήση του γραφικού περιβάλλοντος, και συγκεκριμένα του Explorer.

Εικόνα 21. Οθόνη εκκίνησης Weka



Εικόνα 22. Γραφικό περιβάλλον Explorer του WEKA

Τα βασικά αρχεία τα οποία δέχεται το WEKA προς επεξεργασία είναι αρχεία τύπου arff (Attribute - Relation File Format). Ένα αρχείο τύπου arff είναι ένα αρχείο κειμένου χαρακτήρων ASCII (ASCII text file) που περιλαμβάνει μία λίστα παραδειγμάτων (instances) τα οποία περιγράφονται από χαρακτηριστικά (attributes). Αποτελείται από δύο διακριτά τμήματα: το τμήμα Header και το τμήμα Data. Το τμήμα Header περιλαμβάνει το όνομα του συνόλου δεδομένων, τα ονόματα και τους τύπους των χαρακτηριστικών (attributes), ενώ το τμήμα Data περιλαμβάνει τα παραδείγματα και τις τιμές που αντιστοιχούν στο κάθε χαρακτηριστικό τους.
Στην περίπτωση μας, χρειάστηκε να μετατρέψουμε το εκάστοτε dataset από αρχείο excel σε αρχείο csv και τελικά μέσω του WEKA σε αρχείο arff. Το κάθε arff αρχείο μας περιλαμβάνει τρία attributes: το attribute tweets, που περιγράφει τα tweets και επομένως είναι τύπου string, το attribute my_annotation που περιγράφει το δικό μας χειρονακτικό σχολιασμό και είναι τύπου nominal και το attribute first_annotation που περιγράφει τον αρχικό σχολιασμό των δεδομένων και είναι επίσης τύπου nominal. Το τρίτο attribute αγνοήθηκε κατά την εκτέλεση των πειραμάτων.

3.4 Περιγραφή σταδίων της πειραματικής διαδικασίας

Στην παράγραφο που ακολουθεί περιγράφουμε τα διάφορα πειράματα που εκτελέσαμε και τα αντίστοιχα αποτελέσματα που λάβαμε.

3.4.1 Τροποποιήσεις του αρχικού dataset πριν την εκτέλεση των πειραμάτων

Όπως γίνεται εύκολα κατανοητό, τα επαναλαμβανόμενα όμοια tweets (retweets) δεν προσφέρουν κάτι ουσιαστικό στην απόδοση του συστήματος στην περίπτωση της συσταδοποίησης. Απεναντίας, αλλοιώνουν τα πραγματικά αποτελέσματα και βελτιώνουν πλασματικά την απόδοση, αφού προφανώς πανομοιότυπα tweets ταξινομούνται στην ίδια συστάδα. Επομένως, από το αρχικό dataset των 4.009 tweets, αφαιρώντας τα πολλαπλά tweets παίρνουμε ένα dataset με 1.546 tweets (831 θετικά, 89 αρνητικά, 453 ουδέτερα, 173 απροσδιόριστα). Επίσης, για λόγους συμβατότητας με την πλατφόρμα μετατρέψαμε τα ελληνικά γράμματα στα αντίστοιχα λατινικά.

3.4.2 Εισαγωγή αρχείου δεδομένων και προεπεξεργασία

Ανοίγοντας το περιβάλλον explorer του WEKA, επιλέγουμε την καρτέλα Preprocess και φορτώνουμε το κατάλληλο dataset σε μορφή arff. Για να επεξεργαστούμε τα δεδομένα με χρήση κάποιου αλγορίθμου συσταδοποίησης θα πρέπει να τα μετατρέψουμε σε μαθηματική μορφή. Αυτό γίνεται με τη χρήση του φίλτρου String-to-Word-Vector που παρέχει το WEKA(Filter: weka> filters> unsupervised> attributes> stringToWordVector). Το φίλτρο αυτό μετατρέπει κάθε ένα από τα παραδείγματα του dataset (στην περίπτωση μας κάθε tweet) σε ένα διάνυσμα του οποίου το πλήθος συνιστωσών ισούται με τον αριθμό των διαφορετικών χαρακτηριστικών (features) που υπάρχουν στο dataset αρχικά ή που επιλέγουμε να παραμείνουν. Η τιμή της κάθε συνιστώσας μπορεί να υποδηλώνει είτε τη συχνότητα εμφάνισης (απλή, tf, tf-idf), είτε την παρουσία - απουσία του αντίστοιχου χαρακτηριστικού στο εκάστοτε παράδειγμα - έγγραφο, ανάλογα με τη ρύθμιση που έχουμε κάνει στην καρτέλα ρυθμίσεων του φίλτρου πριν την εφαρμογή του. Μέσω της καρτέλας ρύθμισης μπορούμε μεταξύ των άλλων, να επιλέξουμε διάφορες από της μεθόδους προεπεξεργασίας που μελετήθηκαν στο Κεφάλαιο 2, όπως stemming, αφαίρεση stop-words, χρήση n-grams.

3.4.3 Επιλογή χαρακτηριστικών

Μετά την εφαρμογή του φίλτρου String-to-word-vector, και αφού ορίσουμε το attribute που θα αποτελέσει την κλάση συναισθήματος (στα dataset μας το attribute “my annotation”), μπορούμε να κάνουμε επιλογή χαρακτηριστικών με χρήση του φίλτρου Attribute-Selection (Filter: weka> filters> supervised> attributes> Attribute - Selection). Στην καρτέλα ρυθμίσεων του φίλτρου αυτού επιλέγουμε ως μέθοδο αναζήτησης (search) την επιλογή Ranker και ως αξιολογητή (evaluator) είτε το Info Gain είτε το Chi-square, που μελετήθηκαν στο δεύτερο κεφάλαιο. Επισημαίνουμε ότι σε ένα ρεαλιστικό πρόβλημα Μη Επιβλεπόμενης Μάθησης η επιλογή χαρακτηριστικών με τη χρήση των παραπάνω κριτηρίων δεν είναι εφικτή εφόσον δε θα διαθέταμε τις κλάσεις που απαιτούνται για τον

υπολογισμό τους. Ωστόσο στο παρόν πρόβλημα επιλέγουμε να το χρησιμοποιήσουμε ως επιπλέον στάδιο στην προεπεξεργασία των δεδομένων, αφού δεν επηρεάζει τη διαδικασία της συσταδοποίησης αυτή καθαυτή.

3.4.4 Μέθοδος αξιολόγησης συσταδοποίησης

Για την αξιολόγηση της απόδοσης τη συσταδοποίησης, το WEKA παρέχει την επιλογή εκτέλεσης "classes to clusters evaluation". Σύμφωνα με αυτή σύστημα τροφοδοτείται αρχικά με χαρακτηρισμένα ως προς το συναίσθημα δεδομένα, στη συνέχεια κάνει συσταδοποίηση αγνοώντας τον χαρακτηρισμό των δεδομένων και μετά το τέλος της συσταδοποίησης αναθέτει μία κλάση συναισθήματος το πολύ σε μία συστάδα ανάλογα με την κλάση που υπερτερεί στα ομαδοποιημένα δεδομένα και την ελαχιστοποίηση του ποσοστού αποτυχίας. Με βάση αυτή την ανάθεση κλάσεων γίνεται στην συνέχεια υπολογισμός του ποσοστού επιτυχίας της συσταδοποίησης. Σημειώνουμε ωστόσο ότι η συγκεκριμένη μέθοδος αξιολόγησης αντιβαίνει στο κυρίαρχο πλεονέκτημα της συσταδοποίησης που είναι η απουσία χαρακτηρισμού των δεδομένων. Ωστόσο την υιοθετούμε για τις ανάγκες της παρούσας μελέτης, ώστε να εστιάσουμε κυρίως στις παραμέτρους της συσταδοποίησης και όχι στις διάφορες τεχνικές ανάθεσης κλάσεων στις συστάδες που έχουν δημιουργηθεί.

Αν και δοκιμάσαμε διάφορους αλγορίθμους συσταδοποίησης, θα αναφερθούμε μόνο στην εφαρμογή των αλγορίθμων k-means και Expectation – Maximization (E-M) καθώς αφενός επιτρέπουν την προεπιλογή του πλήθους των συστάδων και αφετέρου παρουσίαζαν πάντα την καλύτερη απόδοση.

3.4.5 Πειράματα

Α) Περιγραφή των Datasets

Ακολουθεί περιγραφή των datasets στα οποία έγιναν τα πειράματα συσταδοποίησης:

1. Αρχικά χρησιμοποιήθηκε ακέραιο το dataset των 1.546 tweets που προέκυψε από τις τροποποιήσεις του αρχικού dataset, δηλαδή το "all-unique" dataset (831 θετικά, 89 αρνητικά, 453 ουδέτερα, 173 απροσδιόριστα).

2. Στην προσπάθεια μας να απομακρύνουμε οποιαδήποτε πληροφορία εισάγει πλασματική ομοιότητα μεταξύ των tweets, για το δεύτερο πείραμα αφαιρέθηκαν από το dataset "all-unique" οι αναφορές σε usernames και τα περισσότερα links. Στη συνέχεια, αφαιρέθηκαν τα επαναλαμβανόμενα tweets που προέκυψαν. Οι αναφορές σε κάποιο συγκεκριμένο username δεν μπορούν να συνδεθούν με την έκφραση όμοιου συναισθήματος. Εξαίρεση αποτελεί η αναφορά σε usernames μαρτυρούν το συναίσθημα του tweet (π.χ. @iloveKavala). Επίσης, τα links δεν προσδίδουν κάποια

πληροφορία για το συναισθηματικό περιεχόμενο. Προέκυψε το dataset “all-unique-clean” που αποτελείται από 1.435 tweets (780 θετικά, 89 αρνητικά, 428 ουδέτερα, 138 απροσδιόριστα).

3. Κατά την τρίτη προσπάθεια αφαιρέθηκαν από το dataset “all-unique-clean-reduced” μερικά θετικά και μερικά ουδέτερα tweets, αφού οι δύο κλάσεις ήταν πολύ μεγαλύτερες από τις άλλες δύο, με αποτέλεσμα να βελτιώνουν πλασματικά τα ποσοστό επιτυχίας. Προέκυψε το dataset “all-unique-clean-reduced” το οποίο αποτελείται από 567 tweets (180 θετικά, 89 αρνητικά, 160 ουδέτερα, 138 απροσδιόριστα).

Β) Περιγραφή των παραμέτρων

<u>Παράμετροι των διάφορων σταδίων προεπεξεργασίας και της συσταδοποίησης που έμειναν σταθερές</u>

Κατά τη διάρκεια των διάφορων δοκιμών κρατήθηκαν σταθερές οι εξής παράμετροι των διάφορων σταδίων:

Στις ρυθμίσεις του StringToWordVector filter:
- Μετατρέπουμε όλα τα γράμματα σε μικρά (lowerCaseTokens: True)
- Χρησιμοποιούμε λίστα αφαίρεσης stopwords (useStoplist: True)
- Επιλέγουμε απλή συχνότητα εμφάνισης λέξεων σε ένα tweet ως τιμή της αντίστοιχης συνιστώσας του διανύσματος (outputWordCounts: True)
-Στα πεδία IDFTransform, TFTransform, AttributeIndices, AttributeName - Prefix, DoNotOperateOnPerClassBasis, InvertSelection, minTermFreq, normalizeDocLength και PeriodicPruning διατηρήσαμε τις προεπιλεγμένες τιμές.

Στις ρυθμίσεις του AttribureSelection filter:
Επιλέγουμε πάντα ως μέθοδο αναζήτησης στο πεδίο search το Ranker, και στην καρτέλα ρυθμίσεων αναζήτησης αλλάζουμε μόνο το πλήθος των λέξεων που θέλουμε να παραμείνουν μετά την επιλογή (numToSelect). Η μέθοδος ranker αναζητά μεταξύ των attributes και επιλέγει αυτά που εμφανίζουν τις μεγαλύτερες τιμές ενός αξιολογητή που επιλέγουμε (π.χ Info Gain, Gain Ration, Entropy κ.α.).

Στις ρυθμίσεις του αλγορίθμου k-means:
Διατηρούμε σταθερό το seed στην τιμή 8 ύστερα από μερικές δοκιμές με διάφορες τιμές. Το seed είναι ένας τυχαίος αριθμός που απαιτείται από τον αλγόριθμο και επηρεάζει την επιλογή των κέντρων των συστάδων. Όπως παρατηρήθηκε, η επίδραση του είναι απρόβλεπτη και επομένως δεν έχει νόημα να προσπαθήσουμε να βρούμε τη βέλτιστη τιμή του.

Στις ρυθμίσεις του αλγορίθμου E-M:
Διατηρούμε το επίσης απρόβλεπτο seed στην τιμή 100. Διατηρούμε επίσης τα πεδία debug, displayModelInOldFormat, maxIterations, minStdDev, στις προεπιλεγμένες τιμές.

<u>Παράμετροι των διάφορων σταδίων προεπεξεργασίας και της συσταδοποίησης που μεταβάλλαμε μεταξύ των διάφορων εκτελέσεων</u>

Έχοντας αναφέρει τις παραμέτρους που θα διατηρούμε σταθερές σε όλες τις δοκιμές, θα προχωρήσουμε τώρα στη μελέτη των παραμέτρων τις οποίες μεταβάλλουμε με στόχο να ερευνήσουμε την επίδραση τους στο τελικό αποτέλεσμα της Ανάλυσης Συναισθήματος με την τεχνική της συσταδοποίησης.

- *Distance function (Συνάρτηση απόστασης)*: Στην περίπτωση του αλγορίθμου k-means ο χρήστης καλείται να επιλέξει τη συνάρτηση απόστασης στην οποία θα βασιστεί ο αλγόριθμος. Δοκιμάσαμε και τις δύο διαθέσιμες συναρτήσεις, Euclidean Distance και Manhattan Distance, οι οποίες δίνονται από τους ακόλουθους τύπους:

  Έστω σύνολο δεδομένων D, και δύο δεδομένα του d_1 και d_2 που περιγράφονται από m συνιστώσες ($n_1(d_1)$, $n_2(d_1)$,……,$n_m(d_1)$), ($n_1(d_2)$, $n_2(d_2)$,……,$n_m(d_2)$). Η απόσταση των δύο αυτών δεδομένων είναι η παρακάτω.

  Euclidean Distance: $dist(d_1, d_2) = \sqrt{\sum_i (n_i(d_1) - n_i(d_2))^2}$ (3.1)

  Manhattan Distance: $dist(d_1, d_2) = \sum_i |n_i(d_1) - n_i(d_2)|$ (3.2)

  Η διαφορά των δύο αυτών συναρτήσεων έγκειται στο ότι η πρώτη βασίζεται στην τετραγωνική απόκλιση, ενώ η δεύτερη στην απόλυτη απόκλιση δύο σημείων. Δίνουν περίπου ίδια αποτελέσματα, εκτός από την περίπτωση που υπάρχουν έκτροπες παρατηρήσεις (outliers)[8], οπότε η απόσταση Manhattan, λόγω μη ύψωσης στο τετράγωνο οδηγεί σε πιο ομαλά αποτελέσματα.

- *Πλήθος clusters*: Τόσο ο αλγόριθμος k-means όσο και ο αλγόριθμος E-M δέχονται ως είσοδο το πλήθος των clusters. Η δοκιμή αυτή γίνεται με στόχο να δημιουργηθούν έστω και περισσότερες από μία συστάδες για κάθε κλάση, αρκεί να είναι "καθαρές" από tweets άλλων κλάσεων. Βέβαια λόγω της λογικής του WEKA, που αναθέτει κάθε κλάση το πολύ σε μία συστάδα, μπορούμε να εκτιμήσουμε μια τέτοια βελτίωση μόνο με παρατήρηση των συστάδων, και όχι μέσω του ποσοστού αποτυχίας που υπολογίζεται αυτόματα.

- *Stemming (θεματοποίηση)*: Πειραματιστήκαμε με την χρήση ή όχι της θεματοποίησης, και συγκεκριμένα του εργαλείου θεματοποίησης lovinstemmer που παρέχει το WEKA.

- *Tokenization ('σπάσιμο του κειμένου σε features')*: Το WEKA διακρίνει τα χαρακτηριστικά ενός κειμένου με κριτήριο την εμφάνιση συγκεκριμένων χαρακτήρων - οριοθετών των που επιλέγονται από το χρήστη (delimiters). Δοκιμάσαμε τη χρήση unigrams, bigrams και trigrams. Δε μελετήθηκαν n-grams μεγαλύτερου μήκους αφού είναι σπάνια η εμφάνιση μεγαλύτερων φράσεων με ιδιαίτερο νόημα, ειδικά σε κείμενα μικρού μήκους όπως τα tweets.

- *Attribute Selection Evaluator (αξιολογητής χαρακτηριστικών)*: Το WEKA δίνει την δυνατότητα επιλογής μεταξύ διαφορετικών αξιολογητών για την επιλογή των attributes. Εμείς δοκιμάσαμε τόσο τον αξιολογητή Info Gain όσο και τον αξιολογητή Chi-square.

Γ) Αποτελέσματα των πειραμάτων

Στις δοκιμές που έγιναν για κάθε ένα από τα τρία datasets η επιλογή των τιμών των παραμέτρων έγινε με βάση την εξής λογική: Αλλάζαμε τις παραμέτρους μία-μία κρατώντας τις υπόλοιπες παραμέτρους (οι οποίες είχαν αλλάξει νωρίτερα) σταθερές στη βέλτιστη τιμή τους. Ξεκινούσαμε πάντα από την πιο απλή περίπτωση προεπεξεργασίας των δεδομένων, δηλαδή χωρίς stemming, σε επίπεδο unigrams, και με πλήθος συστάδων ίσο με 4. Στην περίπτωση του αλγορίθμου k-means, σ' αυτό το σημείο δοκιμάζαμε ως συνάρτηση απόστασης τόσο την Euclidean όσο και την Manhattan και επιλέγαμε κάθε φορά την καλύτερη. Στη συνέχεια, αυξάναμε τον αριθμό των συστάδων που δίνεται ως προεπιλογή στον αλγόριθμο συσταδοποίησης. Ακολούθως, πειραματιστήκαμε με bigrams και trigrams, με stemming, και τέλος με επιλογή χαρακτηριστικών.

Στη συνέχεια παρουσιάζονται δύο πίνακες για το κάθε dataset, στους οποίους συνοψίζονται οι διάφορες τιμές των παραμέτρων και τα αντίστοιχα αποτελέσματα τόσο για τον αλγόριθμο k-means όσο και για τον αλγόριθμο E-M. Παρουσιάζονται επίσης οι Μήτρες Σύγχυσης (Confusion Matrices) και οι αναθέσεις κλάσεων στις συστάδες, όπως προκύπτουν από το WEKA.

1) Δοκιμές στο "all-unique" dataset

Αρχικό πλήθος συνιστωσών διανύσματος: 4.066

ID	String-to-word filter		Attribute-selection filter		k-means clustering algorithm		Incorrectly clustered instances
K1	**stemmer**	-	**evaluator**	-	**clusters**	4	***42,62%***
	tokenizer	unigrams			**distance function**	Euclidean distance	
K2	**stemmer**	-	**evaluator**	-	**clusters**	4	*43,53%*
	tokenizer	unigrams			**distance function**	Manhattan distance	
K3	**stemmer**	-	**evaluator**	-	**clusters**	6	*43,07%*
	tokenizer	unigrams			**distance function**	Euclidean distance	
K4	**stemmer**	-	**evaluator**	-	**clusters**	15	*43,91%*
	tokenizer	unigrams			**distance function**	Euclidean distance	
K5	**stemmer**	-	**evaluator**	-	**clusters**	4	*46,05%*
	tokenizer	bigrams			**distance function**	Euclidean distance	
K6	**stemmer**	-	**evaluator**	-	**clusters**	4	*46,05%*
	tokenizer	trigrams			**distance function**	Euclidean distance	
K7	**stemmer**	✓	**evaluator**	-	**clusters**	4	*51,48%*
	tokenizer	unigrams			**distance function**	Euclidean distance	
K8	**stemmer**	-	**evaluator**	InfoGain (2000 attributes)	**clusters**	4	*43,53%*
	tokenizer	unigrams			**distance function**	Euclidean distance	
K9	**stemmer**	-	**evaluator**	Chi-square (2000 attributes)	**clusters**	4	*54,98%*
	tokenizer	unigrams			**distance function**	Euclidean distance	

Πίνακας 12. Πίνακας εκτέλεσης αλγορίθμου k-means για το 1ο dataset

ID	String-to-word filter		Attribute-selection filter		EM clustering algorithm		Incorrectly clustered instances
E1	**stemmer**	-	**evaluator**	**-**	**clusters**	4	*69,46%*
	tokenizer	unigrams					
E2	**stemmer**	-	**evaluator**	**-**	**clusters**	6	*66,49%*
	tokenizer	unigrams					
E3	**stemmer**	-	**evaluator**	**-**	**clusters**	15	*72,63%*
	tokenizer	unigrams					
E4	**stemmer**	-	**evaluator**	**-**	**clusters**	4	*63,51%*
	tokenizer	bigrams					
E5	**stemmer**	-	**evaluator**	**-**	**clusters**	4	***57,76%***
	tokenizer	trigrams					
E6	**stemmer**	✓	**evaluator**	**-**	**clusters**	4	*58,08%*
	tokenizer	trigrams					
E7	**stemmer**	-	**evaluator**	Info Gain (2000 attributes)	**clusters**	4	*59,31%*
	tokenizer	trigrams					
E8	**stemmer**	-	**evaluator**	Chi-square (2000 attributes)	**clusters**	4	*59,31%*
	tokenizer	trigrams					

Πίνακας 13. Πίνακας εκτέλεσης αλγορίθμου Ε-Μ για το 1^{ο} dataset

Ακολουθούν τα Confusion Matrices των παραπάνω εκτελέσεων, από το WEKA:

K1:

```
Clustered Instances
0          6 (0%)
1        121 (8%)
2       1418 (92%)
3          1 (0%)

Classes to Clusters:
  0    1    2  3 ← cluster
  0   53   36  0 | -
  1   58  113  1 | *
  1    5  825  0 | +
  4    5  444  0 | =

Cluster 0 ← =
Cluster 1 ← *
Cluster 2 ← +
Cluster 3 ← No class
```

K2:

```
Clustered Instances
0          6 (0%)
1         42 (3%)
2       1497 (97%)
3          1 (0%)

Classes to Clusters:
  0    1    2  3 ← cluster
  0    2   87  0 | -
  2   40  130  1 | *
  1    0  830  0 | +
  3    0  450  0 | =

Cluster 0 ← =
Cluster 1 ← *
Cluster 2 ← +
Cluster 3 ← No class
```

K3:

```
Clustered Instances
0          6 ( 0%)
1        121 (8%)
2       1411 (91%)
3          1 (0%)
4          1 (0%)
5          6 (0%)

Classes to Clusters:
  0    1    2  3  4  5 ← cluster
  0   53   36  0  0  0 | -
  1   58  113  1  0  0 | *
  1    5  818  0  1  6 | +
  4    5  444  0  0  0 | =

Cluster 0 ← =
Cluster 1 ← *
Cluster 2 ← +
Cluster 3 ← No class
Cluster 4 ← No class
Cluster 5 ← No class
```

K4:

Clustered Instances	
0	6 (0%)
1	121 (8%)
2	1380 (89%)
3	1 (0%)
4	1 (0%)
5	6 (0%)
6	3 (0%)
7	2 (0%)
8	2 (0%)
9	12 (1%)
10	2 (0%)
11	2 (0%)
12	2 (0%)
13	4 (0%)
14	2 (0%)

Classes to Clusters:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	← cluster
0	53	36	0	0	0	0	0	0	0	0	0	0	0	0	| -
1	58	109	1	0	0	0	1	0	3	0	0	0	0	0	| *
1	5	805	0	1	6	3	0	0	8	0	0	0	0	2	| +
4	5	430	0	0	0	0	1	2	1	2	2	2	4	0	| =

Cluster 0 ← No class
Cluster 1 ← *
Cluster 2 ← +
Cluster 3 ← No class
Cluster 4 ← No class
Cluster 5 ← No class
Cluster 6 ← No class
Cluster 7 ← No class
Cluster 8 ← No class
Cluster 9 ← No class
Cluster 10 ← No class
Cluster 11 ← No class
Cluster 12 ← No class
Cluster 13 ← =
Cluster 14 ←No class

K5:

Clustered Instances
0 1542 (100%)
1 2 (0%)
2 1 (0%)
3 1 (0%)

Classes to Clusters:

0	1	2	3	← cluster
88	0	1	0	-
170	2	0	1	*
831	0	0	0	+
453	0	0	0	=

Cluster 0 ←+
Cluster 1 ← *
Cluster 2 ← -
Cluster 3 ←No class

K6:

Clustered Instances
0 1542 (100%)
1 2 (0%)
2 1 (0%)
3 1 (0%)

Classes to Clusters:

0	1	2	3	← cluster
88	0	1	0	-
170	2	0	1	*
831	0	0	0	+
453	0	0	0	=

Cluster 0 ← +
Cluster 1 ← *
Cluster 2 ← -
Cluster 3 ← No class

K7:

Clustered Instances
0 810 (52%)
1 42 (3%)
2 692 (45%)
3 2 (0%)

Classes to Clusters:

0	1	2	3	← cluster
9	2	78	0	-
60	40	72	1	*
499	0	331	1	+
242	0	211	0	=

Cluster 0 ← +
Cluster 1 ← *
Cluster 2 ← =
Cluster 3 ← No class

K8:

Clustered Instances
0 5 (0%)
1 42 (3%)
2 1498(97%)
3 1 (0%)

Classes to Clusters:

0	1	2	3	← cluster
0	2	87	0	-
1	40	131	1	*
1	0	830	0	+
3	0	450	0	=

Cluster 0 ←=
Cluster 1 ← *
Cluster 2 ← +
Cluster 3 ← No class

K9:

Clustered Instances

0	1023 (66%)
1	42 (3%)
2	5 (0%)
3	476 (31%)

Classes to Clusters:

0	1	2	3	← cluster
73	2	5	9	-
90	40	0	43	*
520	0	0	311	+
340	0	0	113	=

Cluster 0 ← =
Cluster 1 ← *
Cluster 2 ← -
Cluster 3 ← +

E1:

Clustered Instances

0	401 (26%)
1	495 (32%)
2	298 (19%)
3	352 (23%)

Log likelihood: 8650.98994

Classes to Clusters:

0	1	2	3	← cluster
46	34	0	9	-
45	81	18	29	*
161	280	163	227	+
149	100	117	87	=

Cluster 0 ← -
Cluster 1 ← +
Cluster 2 ← =
Cluster 3 ← *

E2:

Clustered Instances

0	1 (0%)
1	292 (19%)
2	461 (30%)
3	249 (16%)
4	235 (15%)
5	308 (20%)

Log likelihood: 8818.5235

Classes to Clusters:

0	1	2	3	4	5	← cluster
0	8	17	0	28	36	-
0	20	31	17	68	37	*
1	188	306	124	87	125	+
0	76	107	108	52	110	=

Cluster 0 ←No class
Cluster 1 ← No class
Cluster 2 ← +
Cluster 3 ← =
Cluster 4 ← *
Cluster 5 ← -

E3:

Clustered Instances

0	3 (0%)
1	1 (0%)
2	58 (4%)
3	89 (6%)
4	2 (0%)
5	27 (2%)
6	63 (4%)
7	166 (11%)
8	58 (4%)
9	322 (21%)
10	3 (0%)
11	5 (0%)
12	194 (13%)
13	354 (23%)
14	201 (13%)

Log likelihood: 9410.26635

Classes to Clusters:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	← cluster
0	1	0	52	0	0	0	6	0	1	3	0	10	9	7	| -
0	0	4	28	0	1	2	10	3	1	0	0	20	58	33	| *
3	0	25	3	0	20	34	110	36	229	0	4	80	200	87	| +
0	0	29	6	2	6	27	40	19	78	0	1	84	87	74	| =

Cluster 0 ← No class
Cluster 1 ← No class
Cluster 2 ← No class
Cluster 3 ← -
Cluster 4 ← No class
Cluster 5 ← No class
Cluster 6 ← No class
Cluster 7 ← No class
Cluster 8 ← No class
Cluster 9 ← +
Cluster 10 ← No class
Cluster 11 ← No class
Cluster 12 ← =
Cluster 13 ← *
Cluster 14 ← No class

E4:

```
Clustered Instances
0       569 (37%)
1         3 (0%)
2       592 (38%)
3       382 (25%)

Log likelihood: 20935.25638

Classes to Clusters:
     0   1    2     3  ← cluster
    16   3   64     6 | -
    84   0   61    28 | *
   324   0  258   249 | +
   145   0  209    99 | =

Cluster 0 ← +
Cluster 1 ← -
Cluster 2 ← =
Cluster 3 ← *
```

E5:

```
Clustered Instances
 0       808 (52%)
1          3 (0%)
2       602 (39%)
3       133 (9%)

Log likelihood: 25132.77463

Classes to Clusters:
    0   1    2    3  ← cluster
   29   3   54    3 | -
   98   0   60   15 | *
  455   0  308   68 | +
  226   0  180   47 | =

Cluster 0 ← +
Cluster 1 ← -
Cluster 2 ← =
Cluster 3 ← *
```

E6:

```
Clustered Instances
0       770 (50%)
1         3 (0%)
2       637 (41%)
3       136 (9%)

Log likelihood: 24968.97776

Classes to Clusters:
    0   1    2    3  ← cluster
   25   3   58    3 | -
   99   0   56   18 | *
  433   0  329   69 | +
  213   0  194   46 | =

Cluster 0 ←+
Cluster 1 ← -
Cluster 2 ← =
Cluster 3 ← *
```

E7:

```
Clustered Instances
0       245 (16%)
1       559 (36%)
2         2 (0%)
3       740 (48%)


Log likelihood: 22369.75892

Classes to Clusters:
  0     1   2     3  ← cluster
 26    15   0    48 | -
 18    84   1    70 | *
134   338   1   358 | +
 67   122   0   264 | =

Cluster 0 ← -
Cluster 1 ← +
Cluster 2 ← *
Cluster 3 ← =
```

E8:

```
Clustered Instances
0       245 (16%)
1       559 (36%)
2         2 (0%)
3       740 (48%)


Log likelihood: 22369.75886

Classes to Clusters:
  0     1   2     3  ← cluster
 26    15   0    48 | -
 18    84   1    70 | *
134   338   1   358 | +
 67   122   0   264 | =

Cluster 0 ← -
Cluster 1 ← +
Cluster 2 ← *
Cluster 3 ← =
```

<u>Σχολιασμός αποτελεσμάτων για το “all-unique” dataset</u>

- Στην περίπτωση χρήσης του k-means παρατηρούμε τα εξής:
  - Μεταξύ των δύο πρώτων δοκιμών (K1, K2), οι οποίες διαφέρουν μόνο στην χρησιμοποιούμενη συνάρτηση απόστασης υπάρχει ελάχιστη διαφορά, με την δοκιμή K2 (Manhattan Distance) να εμφανίζει μεγαλύτερο ποσοστό αποτυχίας κατά 1% περίπου. Η κατανομή στις συστάδες είναι σχεδόν ίδια. Η μεγαλύτερη μετακίνηση μεταξύ των δύο εκτελέσεων παρατηρείται στα αρνητικά tweets. Ωστόσο, λόγω της ιδιαιτερότητας της μεθόδου αξιολόγησης classes – to – clusters, που έχει αναφερθεί προηγουμένως, σε καμία από τις δύο περιπτώσεις δεν γίνεται ανάθεση της αρνητικής κλάσης σε κάποιο cluster. Επομένως, σε όποια κλάση και να τοποθετηθούν τα αρνητικά tweets, προκαλούν το ίδιο ποσοστό λάθους. Συμπεραίνουμε λοιπόν, ότι η διαφορά στην επίδοση των δύο εκτελέσεων οφείλεται στην μετακίνηση κάποιων tweets των υπόλοιπων τριών κλάσεων σε διαφορετικά clusters, με των οποίων την κλάση δεν ταυτίζονται. Η μικρή μετακίνηση αυτή είναι αποτέλεσμα της διαφοράς στον τύπο υπολογισμού της απόστασης. Λόγω της

ελάχιστα καλύτερης επίδοσης της Euclidean distance, επιλέγουμε να συνεχίσουμε αυτήν στις ακόλουθες δοκιμές.

- Μεταξύ των δοκιμών K1, K3, K4 μεταβάλλουμε το πλήθος των συστάδων. Παρατηρώντας τις αντίστοιχες Μήτρες Σύγχυσης, βλέπουμε ότι αυξάνοντας τον αριθμό των συστάδων, γίνεται μετακίνηση μικρών ομάδων από tweets που ανήκουν στην ίδια κλάση στις νέες κενές συστάδες. Δημιουργούνται δηλαδή περισσότερες συστάδες, οι οποίες είναι πιο "εξειδικευμένες". Εφόσον δεν έχουμε δυνατότητα μελέτης του φαινομένου μέσω του συνολικού ποσοστού λάθους, για λόγους καλύτερης εποπτείας και προσαρμογής με το συγκεκριμένο πρόβλημα διατηρούμε τον αριθμό των συστάδων στις τέσσερις (4) για τις υπόλοιπες δοκιμές.

- Μεταξύ των δοκιμών K1, K5, K6 μεταβάλλουμε το μήκος των n-grams, από 1 μέχρι 3. Παρατηρείται αύξηση του ποσοστού αποτυχίας κατά τη μετάβαση από unigrams σε bigrams ή trigrams, ενώ τα το ποσοστό παραμένει ακριβώς το ίδιο στην περίπτωση bigrams και trigrams. Αν και κατά τη μετάβαση από unigrams σε bigrams ή trigrams παρατηρείται σύμπτυξη των tweets ίδιας κλάσης, ταυτόχρονα δεν γίνεται απομάκρυνση τους από το cluster που είναι κατειλημμένο από άλλη κλάση. Ως αποτέλεσμα, έχουμε τη συγκέντρωση της πλειονότητας των tweets σε ένα cluster και την πτώση της απόδοσης εν τέλει. Επιλέγουμε επομένως τα unigrams για τις ακόλουθες εκτελέσεις.

- Στη δοκιμή K7 εφαρμόζουμε stemming στα unigrams πριν τη συσταδοποίηση και παρατηρούμε την απόδοση να πέφτει σημαντικά (περίπου 9%) σε σύγκριση με την αντίστοιχη χωρίς stemming δοκιμή (K1). Παρατηρούμε στους πίνακες ότι συμβαίνει διάσπαση των tweets ίδιας κλάσης. Αυτό προφανώς συμβαίνει γιατί με το stemming δημιουργούνται νέες ομοιότητες μεταξύ των tweets οι οποίες υπερισχύουν, καθιστώντας κάποια tweets πιο όμοια από ότι ήταν πριν. Τυχαίνει στην περίπτωση μας λοιπόν να δημιουργούνται ομοιότητες μεταξύ tweets διαφορετικών κλάσεων, με αποτέλεσμα την αύξηση του ποσοστού αποτυχίας. Επομένως, δε θα χρησιμοποιήσουμε stemming στη συνέχεια.

- Στις δοκιμές K8 και K9 εισάγουμε την επιλογή χαρακτηριστικών και μεταβάλλουμε το κριτήριο επιλογής των 2000 από τα αρχικά 4066 unigrams. Μεταξύ των δύο αξιολογητών δεν παρατηρείται καμία διαφορά, ενώ σε σχέση με τη δοκιμή K1 (χωρίς αξιολογητή) παρατηρούνται ελάχιστες ασήμαντες μετακινήσεις tweets.

➢ Στην περίπτωση χρήσης του EM παρατηρούμε τα εξής:

- Μεταξύ των δοκιμών E1, E2, E3 μεταβάλλουμε και πάλι το πλήθος των συστάδων. Όπως και στην περίπτωση του k-means παρατηρούμε και πάλι τη δημιουργία πιο εξειδικευμένων clusters με την αύξηση του πλήθους συστάδων. Για τους ίδιους λόγους συνεχίζουμε με πλήθος συστάδων ίσο με 4.

- Μεταξύ των δοκιμών Ε1, Ε4, Ε5 μεταβάλλουμε το μήκος των n-grams, από 1 μέχρι 3. Παρατηρούμε σταδιακή βελτίωση καθώς αυξάνεται το μήκος των n-grams. Η βελτίωση παρατηρείται μόνο στις συστάδες της θετικής και ουδέτερης κλάσης. Ωστόσο οι συστάδες των υπόλοιπων δύο κλάσεων έχουν την καλύτερη ποιότητα στην περίπτωση των unigrams, γεγονός που υποδηλώνει ότι στο dataset εντοπίζονται bigrams ή trigrams μόνο στην περίπτωση των θετικών και ουδέτερων tweets, που είναι τα πολυπληθέστερα. Εφόσον καμία συστάδα δεν είναι απολύτως ικανοποιητική συνεχίζουμε με trigrams με την προσδοκία να βελτιώσουμε ακόμα περισσότερο έστω και δύο από τις τέσσερις συστάδες.

- Στη δοκιμή Ε6 εφαρμόζουμε stemming στα trigrams πριν τη συσταδοποίηση και παρατηρούμε την απόδοση να πέφτει ελάχιστα (περίπου 1%) σε σύγκριση με την αντίστοιχη χωρίς stemming δοκιμή (Ε5). Οι συστάδες των αρνητικών και απροσδιόριστων tweets προφανώς χειροτέρευσαν ενώ χειρότερη έγινε και η θετική συστάδα. Ελάχιστη βελτίωση παρατηρείται στη συστάδα των ουδετέρων. Αφαιρούμε το stemming στις επόμενες δοκιμές.

- Στις δοκιμές Ε7 και Ε8 εισάγουμε την επιλογή χαρακτηριστικών ενώ μεταβάλλουμε και το κριτήριο επιλογής των 10.000 από τα αρχικά 11.135 trigrams. Οι δοκιμές Ε7 και Ε8 δίνουν πανομοιότυπες συσταδοποίησεις που είναι βελτιωμένες κατά 10% σε σύγκριση με την αντίστοιχη χωρίς επιλογή χαρακτηριστικών δοκιμή (Ε5). Ολες οι συστάδες πλην αυτής των απροσδιόριστων tweets δείχνουν να βελτιώνονται.

2) Δοκιμές στο "all-unique-clean" dataset

Αρχικό πλήθος συνιστωσών διανύσματος: 2.820

ID	String-to-word filter		Attribute-selection filter		k-means clustering algorithm		Incorrectly clustered instances
K1	**stemmer**	-	**evaluator**	**-**	**clusters**	4	*54,14%*
	tokenizer	unigrams			**distance function**	Euclidean distance	
K2	**stemmer**	-	**evaluator**	**-**	**clusters**	4	*54,00%*
	tokenizer	unigrams			**distance function**	Manhattan distance	
K3	**stemmer**	-	**evaluator**	**-**	**clusters**	6	*61,67%*
	tokenizer	unigrams			**distance function**	Manhattan distance	
K4	**stemmer**	-	**evaluator**	**-**	**clusters**	15	*61,81%*
	tokenizer	unigrams			**distance function**	Manhattan distance	
K5	**stemmer**	-	**evaluator**	**-**	**clusters**	4	*45,92%*
	tokenizer	bigrams			**distance function**	Manhattan distance	
K6	**stemmer**	-	**evaluator**	**-**	**clusters**	4	***45,85%***
	tokenizer	trigrams			**distance function**	Manhattan distance	
K7	**stemmer**	✓	**evaluator**	**-**	**clusters**	4	*56,37%*
	tokenizer	unigrams			**distance function**	Manhattan distance	
K8	**stemmer**	-	**evaluator**	Info Gain (2000 attributes)	**clusters**	4	*51,70%*
	tokenizer	unigrams			**distance function**	Manhattan distance	
K9	**stemmer**	-	**evaluator**	Chi-square (2000 attributes)	**clusters**	4	*51,70%*
	tokenizer	unigrams			**distance function**	Manhattan distance	

Πίνακας 14. Πίνακας εκτέλεσης αλγορίθμου k-means για το 2ο dataset

ID	String-to-word filter		Attribute-selection filter		EM clustering algorithm		Incorrectly clustered instances
E1	**stemmer**	-	**evaluator**	**-**	**Clusters**	4	*62,29%*
	tokenizer	unigrams					
E2	**stemmer**	-	**evaluator**	**-**	**Clusters**	6	*69,19%*
	tokenizer	unigrams					
E3	**stemmer**	-	**evaluator**	**-**	**Clusters**	15	*72,19%*
	tokenizer	unigrams					
E4	**stemmer**	-	**evaluator**	**-**	**Clusters**	4	*57,70%*
	tokenizer	bigrams					
E5	**stemmer**	-	**evaluator**	**-**	**Clusters**	4	***44,80%***
	tokenizer	trigrams					
E6	**stemmer**	✓	**evaluator**	**-**	**Clusters**	4	*61,95%*
	tokenizer	unigrams					
E7	**stemmer**	✓	**evaluator**	Info Gain (2000 attributes)	**Clusters**	4	*54,42%*
	tokenizer	unigrams					
E8	**stemmer**	✓	**evaluator**	Chi-square (2000 attributes)	**Clusters**	4	*54,42%*
	tokenizer	unigrams					

Πίνακας 15. Πίνακας εκτέλεσης αλγορίθμου E-M για το 2^{ο} dataset

Ακολουθούν τα Confusion Matrices των παραπάνω εκτελέσεων, από το WEKA:

K1:

```
Clustered Instances
0       71 (5%)
1      994 (69%)
2        1 (0%)
3      369 (26%)

Classes to Clusters:
   0    1   2    3  ← cluster
   0   89   0    0 | -
   6  108   0   24 | *
  38  524   1  217 | +
  27  273   0  128 | =

Cluster 0 ← *
Cluster 1 ← +
Cluster 2 ← No class
Cluster 3 ← =
```

K2:

```
Clustered Instances
0       73 (5%)
1      999 (70%)
2        1 (0%)
3      362 (25%)

Classes to Clusters:
   0    1   2    3  ← cluster
   0   89   0    0 | -
   7  109   0   22 | *
  38  527   1  214 | +
  28  274   0  126 | =

Cluster 0 ← *
Cluster 1 ← +
Cluster 2 ← No class
Cluster 3 ← =
```

K3:

```
Clustered Instances
0      276 (19%)
1      783 (55%)
2        1 (0%)
3        3 (0%)
4        1 (0%)
5      371 (26%)

Classes to Clusters:
   0    1  2  3  4    5  ← cluster
   5   84  0  0  0    0 | -
  26   88  0  0  0   24 | *
 165  395  1  0  1  218 | +
  80  216  0  3  0  129 | =

Cluster 0 ← *
Cluster 1 ← +
Cluster 2 ← No class
Cluster 3 ← No class
Cluster 4 ← No class
Cluster 5 ← =
```

K4:

Clustered Instances

0	3 (0%)
1	2 (0%)
2	1 (0%)
3	3 (0%)
4	1 (0%)
5	369 (26%)
6	2 (0%)
7	4 (0%)
8	3 (0%)
9	2 (0%)
10	255 (18%)
11	10 (1%)
12	776 (54%)
13	3 (0%)
14	1 (0%)

Classes to Clusters:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	← cluster
0	0	0	0	0	0	0	0	0	0	5	0	83	0	1	| -
0	0	0	0	0	23	0	0	1	2	23	1	88	0	0	| *
2	1	1	3	1	214	2	0	2	0	153	7	392	2	0	| +
1	1	0	0	0	132	0	4	0	0	74	2	213	1	0	| =

Cluster 0 ← No class
Cluster 1 ←No class
Cluster 2 ← No class
Cluster 3 ← No class
Cluster 4 ← No class
Cluster 5 ← =
Cluster 6 ← No class
Cluster 7 ← No class
Cluster 8 ←No class
Cluster 9 ←No class
Cluster 10 ← *
Cluster 11 ← No class
Cluster 12 ← +
Cluster 13 ←- No class
Cluster 14 ← -

K5:

```
Clustered Instances
0      1429 (100%)
1         2 (0%)
2         1 (0%)
3         3 (0%)

Classes to Clusters:
   0   1   2   3 ← cluster
  89   0   0   0 | -
 138   0   0   0 | *
 775   1   1   3 | +
 427   1   0   0 | =

Cluster 0 ← +
Cluster 1 ← =
Cluster 2 ← No class
Cluster 3 ← No class
```

K6:

```
Clustered Instances
0      1430 (100%)
1         2 (0%)
2         1 (0%)
3         2 (0%)


Classes to Clusters:
   0   1   2   3 ← cluster
  89   0   0   0 | -
 138   0   0   0 | *
 776   1   1   2 | +
 427   1   0   0 | =

Cluster 0 ← +
Cluster 1 ← =
Cluster 2 ← No class
Cluster 3 ← No class
```

K7:

```
Clustered Instances

0       901 (63%)
1       161 (11%)
2         2 (0%)
3       371 (26%)

Classes to Clusters:
    0   1   2    3 ← cluster
   89   0   0    0 | -
  100  14   0   24 | *
  482  79   2  217 | +
  230  68   0  130 | =

Cluster 0 ← +
Cluster 1 ← *
Cluster 2 ← No class
Cluster 3 ← =
```

K8:

```
Clustered Instances
0         5 (0%)
1      1055 (74%)
2         1 (0%)
3       374 (26%)

Classes to Clusters:
  0    1   2    3 ← cluster
  0   89   0    0 | -
  2  112   0   24 | *
  2  559   1  218 | +
  1  295   0  132 | =

Cluster 0 ← *
Cluster 1 ← +
Cluster 2 ← No class
Cluster 3 ← =
```

K9:

```
Clustered Instances
0          5 (0%)
1       1055 (74%)
2          1 (0%)
3        374 (26%)

Classes to Clusters:
  0    1    2    3  ← cluster
  0  89    0    0 | -
  2 112    0   24 | *
  2 559    1 218 | +
  1 295    0 132 | =

Cluster 0 ←*
Cluster 1 ← +
Cluster 2 ← No class
Cluster 3 ← =
```

E1:

```
Clustered Instances
0       553 (39%)
1       558 (39%)
2        49 (3%)
3       275 (19%)

Log likelihood: 6239.8277

Classes to Clusters:
   0    1    2        3  ← cluster
  72   16    0        1 | -
  53   58    8       19 | *
 253  347   32  148 | +
 175  137    9  107 | =

Cluster 0←=
Cluster 1 ← +
Cluster 2 ← No class
Cluster 3 ← *
```

E2:

```
Clustered Instances
0        14 (1%)
1       411 (29%)
2       431 (30%)
3       306 (21%)
4        90 (6%)
5       183 (13%)

Log likelihood: 6720.65895

Classes to Clusters:
  0    1     2    3    4     5  ← cluster
  2   24    17   45    1     0 | -
 12   38    42   24    8    14 | *
  0  253  242  156   48    81 | +
  0   96  130   81   33    88 | =

Cluster 0 ←No class
Cluster 1 ← +
Cluster 2 ← =
Cluster 3 ← -
Cluster 4 ← No class
Cluster 5 ← *
```

E3:

Clustered Instances

0	60 (4%)
1	6 (0%)
2	4 (0%)
3	128 (9%)
4	212 (15%)
5	4 (0%)
6	65 (5%)
7	502 (35%)
8	289 (20%)
9	38 (3%)
10	52 (4%)
11	63 (4%)
12	6 (0%)
13	4 (0%)
14	2 (0%)

Log likelihood: 6705.18887

Classes to Clusters:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	← cluster
1	0	0	0	9	2	0	28	40	0	9	0	0	0	0	| -
1	1	3	12	19	0	3	53	33	1	9	1	2	0	0	| *
50	2	1	65	129	2	49	283	128	15	21	29	2	4	0	| +
8	3	0	51	55	0	13	138	88	22	13	33	2	0	2	| =

Cluster 0 ← No class
Cluster 1 ← No class
Cluster 2 ← No class
Cluster 3 ← No class
Cluster 4 ← *
Cluster 5 ← No class
Cluster 6 ← No class
Cluster 7 ← +
Cluster 8 ← =
Cluster 9 ← No class
Cluster 10 ← -
Cluster 11 ← No class
Cluster 12 ← No class
Cluster 13 ← No class
Cluster 14 ← No class

E4:

```
Clustered Instances
0       4 (0%)
1       1 (0%)
2     882 (61%)
3     548 (38%)

Log likelihood: 16019.77213

Classes to Clusters:
  0   1    2    3  ← cluster
  0   0   71   18 | -
  0   0  103   35 | *
  3   1  444  332 | +
  1   0  264  163 | =

Cluster 0 ← No class
Cluster 1 ← No class
Cluster 2 ← +
Cluster 3 ← =
```

E5:

```
Clustered Instances
0       14 (1%)
1     1417 (99%)
2        2 (0%)
3        2 (0%)

Log likelihood: 16008.69974

Classes to Clusters:
  0    1  2  3  ← cluster
  2   87  0  0 | -
 12  126  0  0 | *
  0  778  2  0 | +
  0  426  0  2 | =

Cluster 0 ← *
Cluster 1 ← +
Cluster 2 ← No class
Cluster 3 ← =
```

E6:

```
Clustered Instances
0      368 (26%)
1      212 (15%)
2      672 (47%)
3      183 (13%)


Log likelihood: 5265.49811

Classes to Clusters:
   0    1    2    3  ← cluster
  39    2   46    2 | -
  38   19   69   12 | *
 178  103  407   92 | +
 113   88  150   77 | =

Cluster 0 ← -
Cluster 1 ← =
Cluster 2 ← +
Cluster 3 ← *
```

E7:

```
Clustered Instances
0       368 (26%)
1       727 (51%)
2       117 (8%)
3       223 (16%)


Log likelihood: 4078.59394

Classes to Clusters:
   0    1   2    3  ← cluster
  54   26   7    2 | -
  45   60   7   26 | *
 132  484  50  114 | +
 137  157  53   81 | =

Cluster 0 ← =
Cluster 1 ← +
Cluster 2 ← -
Cluster 3 ← *
```

E8:

```
Clustered Instances
0       368 (26%)
1       727 (51%)
2       117 (8%)
3       223 (16%)


Log likelihood: 4078.59394

Classes to Clusters:
   0    1   2    3 ←cluster
  54   26   7    2 | -
  45   60   7   26 | *
 132  484  50  114 | +
 137  157  53   81 | =

Cluster 0 ←=
Cluster 1 ← +
Cluster 2 ← -
Cluster 3 ← *
```

<u>Σχολιασμός αποτελεσμάτων για το "all-unique-clean" dataset</u>

- Στην περίπτωση χρήσης του k-means παρατηρούμε τα εξής:

  - Μεταξύ των δύο πρώτων δοκιμών (K1, K2), οι οποίες διαφέρουν μόνο στην χρησιμοποιούμενη συνάρτηση απόστασης υπάρχει αμελητέα διαφορά, με την δοκιμή K1 (Euclidean Distance) να εμφανίζει μεγαλύτερο ποσοστό αποτυχίας. Η κατανομή στις συστάδες είναι σχεδόν ίδια. Λόγω της ελάχιστα καλύτερης επίδοσης της Manhattan distance, επιλέγουμε να συνεχίσουμε αυτήν στις ακόλουθες δοκιμές.

  - Μεταξύ των δοκιμών K1, K3, K4 μεταβάλλουμε το πλήθος των συστάδων. Όπως και στην περίπτωση του προηγούμενου data set παρατηρούμε την δημιουργία πιο εξειδικευμένων clusters με την αύξηση του πλήθους των συστάδων. Για τους ίδιους λόγους συνεχίζουμε με πλήθος συστάδων ίσο με 4.

  - Μεταξύ των δοκιμών K1, K5, K6 μεταβάλλουμε το μήκος των n-grams, από 1 μέχρι 3. Παρατηρείται μείωση του ποσοστού αποτυχίας κατά τη μετάβαση από unigrams σε bigrams ή trigrams κατά περίπου 8%, ενώ μεταξύ bigrams και trigrams η διαφορά

θεωρείται αμελητέα. Σημειώνουμε ωστόσο, ότι στην περίπτωση Κ5 και Κ6 αναγνωρίζονται μόνο 2 από τις 3 κλάσεις που αναγνωρίζονται στην Κ1. Επομένως, η βελτίωση στα αποτελέσματα των Κ5 και Κ6 είναι υποτιμημένη στο τελικό ποσοστό. Η βελτίωση οφείλεται κυρίως στο ότι συμπτύσσονται τα θετικά tweets, που αποτελούν την πολυπληθέστερη ομάδα. Ωστόσο, παρατηρείται και εδώ το φαινόμενο της συγκέντρωσης της πλειονότητας των tweets σε μία συστάδα. Επομένως, παρά την φαινομενική ποσοστιαία βελτίωση επιλέγουμε τα unigrams για τις ακόλουθες εκτελέσεις.

- Στη δοκιμή Κ7 εφαρμόζουμε stemming στα unigrams πριν τη συσταδοποίηση και παρατηρούμε την απόδοση να πέφτει σε σύγκριση με την αντίστοιχη χωρίς stemming δοκιμή (Κ1). Όπως παρατηρούμε, αυτό οφείλεται σε διάσπαση κάποιων tweets της ίδια κλάσης σε μικρότερες ομάδες που ταξινομούνται σε διαφορετικά clusters, και συσπείρωση κάποιων άλλων σε clusters κατειλημμένα από άλλη κυρίαρχη κλάση. Δεν διατηρούμε το stemming στις ακόλουθες δοκιμές.

- Στις δοκιμές Κ8 και Κ9 εισάγουμε την επιλογή χαρακτηριστικών ενώ μεταβάλλουμε το κριτήριο επιλογής 2.000 από τα αρχικά 2.820 unigrams. Παρατηρούμε πολύ μικρή βελτίωση της απόδοσης (περίπου 3%) σε σχέση με τη δοκιμή Κ1. Το Info Gain και Chi-square δίνουν πανομοιότυπα αποτελέσματα.

- Στην περίπτωση χρήσης του ΕΜ παρατηρούμε τα εξής:

  - Μεταξύ των δοκιμών Ε1, Ε2, Ε3 μεταβάλλουμε και πάλι το πλήθος των συστάδων. Όπως και στην περίπτωση του k-means παρατηρούμε την δημιουργία πιο εξειδικευμένων clusters με την αύξηση του πλήθους. Για τους ίδιους λόγους συνεχίζουμε με πλήθος συστάδων ίσο με 4.

  - Μεταξύ των δοκιμών Ε1, Ε4, Ε5 μεταβάλλουμε το μήκος των n-grams, από 1 μέχρι 3. Παρατηρούμε σταδιακή βελτίωση του ποσοστού αποτυχίας καθώς αυξάνεται το μήκος των n-grams. Ωστόσο με παρατήρηση προκύπτει ότι η βελτίωση είναι πλασματική, λόγω της πολυπληθούς θετικής κλάσης. Παρατηρείται συσσώρευση των tweets στη θετική συστάδα. Επομένως, διατηρούμε τα unigrams στις ακόλουθες δοκιμές.

  - Στη δοκιμή Ε6 εφαρμόζουμε stemming στα unigrams πριν τη συσταδοποίηση και παρατηρούμε την απόδοση να βελτιώνεται ελάχιστα (περίπου 1%) σε σύγκριση με την αντίστοιχη χωρίς stemming δοκιμή (Ε1) με κάποιες ελάχιστες μετακινήσεις και αναγνώριση πλέον και της αρνητικής συστάδας.

  - Στις δοκιμές Ε7 και Ε8 εισάγουμε την επιλογή χαρακτηριστικών ενώ μεταβάλλουμε και το κριτήριο επιλογής των 2.000 από τα αρχικά 2.570 unigrams. Οι δοκιμές Ε7 και Ε8 δίνουν πανομοιότυπες συσταδοποίησεις που είναι βελτιωμένες κατά 10% σε σύγκριση με την αντίστοιχη χωρίς επιλογή χαρακτηριστικών δοκιμή (Ε1). Ωστόσο,

στις Ε7 και Ε8 παρατηρούμε ότι στη συστάδα στην οποία έχει ανατεθεί η θετική κλάση υπάρχουν περισσότερα αρνητικά tweets, γεγονός πολύ αρνητικό για την ποιότητα της συσταδοποίησης.

3) Δοκιμές στο "all – unique –clean-reduced" dataset

Αρχικό πλήθος συνιστωσών διανύσματος: 1.680

ID	String-to-word filter		Attribute-selection filter		k-means clustering algorithm		Incorrectly clustered instances
K1	**stemmer**	-	**evaluator**	**-**	**clusters**	4	*67,37 %*
	tokenizer	unigrams			**distance function**	Euclidean distance	
K2	**stemmer**	-	**evaluator**	**-**	**clusters**	4	*67,72%*
	tokenizer	unigrams			**distance function**	Manhattan distance	
K3	**stemmer**	-	**evaluator**	**-**	**clusters**	6	*67,54%*
	tokenizer	unigrams			**distance function**	Euclidean distance	
K4	**stemmer**	-	**evaluator**	**-**	**clusters**	15	*70,89%*
	tokenizer	unigrams			**distance function**	Euclidean distance	
K5	**stemmer**	-	**evaluator**	**-**	**clusters**	4	*68,60%*
	tokenizer	bigrams			**distance function**	Euclidean distance	
K6	**stemmer**	-	**evaluator**	**-**	**clusters**	4	*67,90%*
	tokenizer	trigrams			**distance function**	Euclidean distance	
K7	**stemmer**	✓	**evaluator**	**-**	**clusters**	4	*67,37%*
	tokenizer	unigrams			**distance function**	Euclidean distance	
K8	**stemmer**	-	**evaluator**	Info Gain (1500 attributes)	**clusters**	4	***66,13%***
	tokenizer	unigrams			**distance function**	Euclidean distance	
K9	**stemmer**	-	**evaluator**	Chi-square (1500 attributes)	**clusters**	4	*66,84%*
	tokenizer	unigrams			**distance function**	Euclidean distance	

Πίνακας 16. Πίνακας εκτέλεσης αλγορίθμου k-means για το 3ο dataset

ID	String-to-word filter		Attribute-selection filter		EM clustering algorithm		Incorrectly clustered instances
E1	**stemmer**	-	**evaluator**	**-**	**clusters**	4	*67,37%*
	tokenizer	unigrams					
E2	**stemmer**	-	**evaluator**	**-**	**clusters**	6	*65,25%*
	tokenizer	unigrams					
E3	**stemmer**	-	**evaluator**	**-**	**clusters**	15	*70,89%*
	tokenizer	unigrams					
E4	**stemmer**	-	**evaluator**	**-**	**clusters**	4	*69,66%*
	tokenizer	bigrams					
E5	**stemmer**	-	**evaluator**	**-**	**clusters**	4	*68,25%*
	tokenizer	trigrams					
E6	**stemmer**	✓	**evaluator**	**-**	**clusters**	4	***62,43%***
	tokenizer	unigrams					
E7	**stemmer**	✓	**evaluator**	Info Gain (1500 attributes)	**clusters**	4	*69,31%*
	tokenizer	unigrams					
E8	**stemmer**	✓	**evaluator**	Chi-square (1500 attributes)	**clusters**	4	*63,84%*
	tokenizer	unigrams					

Πίνακας 17. Πίνακας εκτέλεσης αλγορίθμου Ε-Μ για το 3ο dataset

Ακολουθούν τα Confusion Matrices των παραπάνω εκτελέσεων, από το WEKA:

K1:

```
Clustered Instances
0      429 (76%)
1       15 (3%)
2        3 (1%)
3      120 (21%)


Classes to Clusters:
     0   1   2   3  ← cluster
    89   0   0   0 | -
   119   1   0  18 | *
   118   5   0  57 | +
   103   9   3  45 | =

Cluster 0 ← *
Cluster 1 ← =
Cluster 2 ← No class
Cluster 3 ← +
```

K2:

```
Clustered Instances
0      431 (76%)
1       15 (3%)
2        3 (1%)
3      118 (21%)


Classes to Clusters:
     0   1   2   3  ← cluster
    86   0   0   3 | -
   121   1   0  16 | *
   122   5   0  53 | +
   102   9   3  46 | =

Cluster 0 ← *
Cluster 1 ← =
Cluster 2 ← No class
Cluster 3 ← +
```

K3:

```
Clustered Instances
0      422 (74%)
1       15 (3%)
2        3 (1%)
3      100 (18%)
4        9 (2%)
5       18 (3%)

Classes to Clusters:
   0   1   2   3  4   5  ← cluster
  82   0   0   0  0   7 | -
 118   1   0  16  0   3 | *
 117   5   0  50  5   3 | +
 105   9   3  34  4   5 | =

Cluster 0 ← *
Cluster 1 ← =
Cluster 2 ← No class
Cluster 3 ← +
Cluster 4 ← No class
Cluster 5 ← -
```

K4:

Clustered Instances

0	103 (18%)
1	13 (2%)
2	3 (1%)
3	52 (9%)
4	9 (2%)
5	16 (3%)
6	1 (0%)
7	23 (4%)
8	14 (2%)
9	2 (0%)
10	260(46%)
11	14 (2%)
12	1 (0%)
13	2 (0%)
14	54 (10%)

Classes to Clusters:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	← cluster
15	0	0	0	0	6	0	0	2	0	62	3	1	0	0	| -
28	1	0	8	0	2	0	5	12	0	72	1	0	0	9	| *
38	3	0	29	5	3	1	14	0	2	58	4	0	2	21	| +
22	9	3	15	4	5	0	4	0	0	68	6	0	0	24	| =

Cluster 0 ←*
Cluster 1 ← No class
Cluster 2 ← No class
Cluster 3 ← +
Cluster 4 ← No class
Cluster 5 ← No class
Cluster 6 ← No class
Cluster 7 ← No class
Cluster 8 ← No class
Cluster 9 ← No class
Cluster 10 ← -
Cluster 11 ← No class
Cluster 12 ←- No class
Cluster 13 ← No class
Cluster 14 ← =

K5:

```
Clustered Instances
0      548 (97%)
1        2 (0%)
2        3 (1%)
3       14 (2%)

Classes to Clusters:
   0   1   2   3  ← cluster
  89   0   0   0 | -
 135   0   0   3 | *
 172   1   0   7 | +
 152   1   3   4 | =

Cluster 0 ← +
Cluster 1 ← No class
Cluster 2 ← =
Cluster 3 ← *
```

K6:

```
Clustered Instances
0      560 (99%)
1        2 (0%)
2        3 (1%)
3        2 (0%)

Classes to Clusters:
   0   1   2   3  ← cluster
  89   0   0   0 | -
 138   0   0   0 | *
 179   1   0   0 | +
 154   1   3   2 | =

Cluster 0 ←+
Cluster 1 ← No class
Cluster 2 ← =
Cluster 3 ← No class
```

K7:

```
Clustered Instances
0      436 (77%)
1       15 (3%)
2        3 (1%)
3      113 (20%)

Classes to Clusters:
   0   1   2   3  ← cluster
  89   0   0   0 | -
 121   1   0  16 | *
 120   5   0  55 | +
 106   9   3  42 | =

Cluster 0 ←*
Cluster 1 ← =
Cluster 2 ← No class
Cluster 3 ← +
```

K8:

```
Clustered Instances
0      349 (62%)
1       11 (2%)
2        3 (1%)
3      204 (36%)

Classes to Clusters:
  0  1   2   3  ← cluster
 84  0   0   5 | -
 98  1   0  39 | *
 86  5   0  89 | +
 81  5   3  71 | =

Cluster 0 ← *
Cluster 1 ← =
Cluster 2 ← No class
Cluster 3 ← +
```

K9:

```
Clustered Instances

0      348 (61%)
1       34 (6%)
2        3 (1%)
3      182 (32%)

Classes to Clusters:
  0   1  2    3  ←cluster
 84   0  0    5 | -
 98   5  0   35 | *
 85  17  0   78 | +
 81  12  3   64 | =

Cluster 0 ← *
Cluster 1 ← =
Cluster 2 ←No class
Cluster 3 ← +
```

E1:

```
Clustered Instances
0        1 (0%)
1      205 (36%)
2      102 (18%)
3      259 (46%)

Log likelihood: 2385.99503

Classes to Clusters:
  0   1   2    3  ← cluster
  1  55   5   28 | -
  0  51  20   67 | *
  0  45  41   94 | +
  0  54  36   70 | =

Cluster 0 ←No class
Cluster 1 ← -
Cluster 2 ← =
Cluster 3 ← +
```

E2:

```
Clustered Instances
0      136 (24%)
1      152 (27%)
2      178 (31%)
3        1 (0%)
4        2 (0%)
5       98 (17%)

Log likelihood: 2525.48339

Classes to Clusters:
  0   1    2    3   4    5  ←  cluster
  7  27   54    0   0    1 | -
 30  35   57    0   0   16 | *
 57  56   28    1   0   38 | +
 42  34   39    0   2   43 | =

Cluster 0 ← +
Cluster 1 ← *
Cluster 2 ← -
Cluster 3 ← No class
Cluster 4 ← No class
Cluster 5 ← =
```

E3:

Clustered Instances

0	1 (0%)
1	1 (0%)
2	1 (0%)
3	4 (1%)
4	51 (9%)
5	31 (5%)
6	95 (17%)
7	2 (0%)
8	191(34%)
9	111(20%)
10	1 (0%)
11	71 (13%)
12	2 (0%)
13	4 (1%)
14	1 (0%)

Log likelihood: 2584.57548

Classes to Clusters:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	← cluster
1	0	1	2	8	5	0	0	45	23	0	4	0	0	0	| -
0	0	0	0	19	5	15	0	50	32	0	15	2	0	0	| *
0	1	0	0	17	13	32	2	42	27	1	40	0	4	1	| +
0	0	0	2	7	8	48	0	54	29	0	12	0	0	0	| =

Cluster 0 ← No class
Cluster 1 ← No class
Cluster 2 ← No class
Cluster 3 ← No class
Cluster 4 ← No class
Cluster 5 ← No class
Cluster 6 ← =
Cluster 7 ← No class
Cluster 8 ← -
Cluster 9 ← *
Cluster 10 ← No class
Cluster 11 ← +
Cluster 12 ← No class
Cluster 13 ← No class
Cluster 14 ← No class

E4:

```
Clustered Instances
0       19 (3%)
1      241 (43%)
2      306 (54%)
3        1 (0%)

Log likelihood: 6216.91389

Classes to Clusters:
  0   1   2  3  ← cluster
  3 37 49  0 | -
  4 67 67  0 | *
  7 75 97  1 | +
  5 62 93  0 | =

Cluster 0 ←*
Cluster 1 ← +
Cluster 2 ← =
Cluster 3 ← No class
```

E5:

```
Clustered Instances
0      502 (89%)
1       56 (10%)
2        8 (1%)
3        1 (0%)


Log likelihood: 5410.79863

Classes to Clusters:
    0    1   2   3  ← cluster
   79    8   2   0 | -
  122  10   6   0 | *
  153  26   0   1 | +
  148  12   0   0 | =

Cluster 0 ← =
Cluster 1 ← +
Cluster 2 ← *
Cluster 3 ← No class
```

E6:

```
Clustered Instances
0      154 (27%)
1      116 (20%)
2       14 (2%)
3      283 (50%)

Log likelihood: 2337.70823

Classes to Clusters:
   0    1    2    3  ← cluster
  45    5    2   37 | -
  40  22  12   64 | *
  33  40    0  107 | +
  36  49    0   75 | =

Cluster 0 ← -
Cluster 1 ← =
Cluster 2 ← *
Cluster 3 ← +
```

E7:

```
Clustered Instances
0      197 (35%)
1      163 (29%)
2       14 (2%)
3      193 (34%)

Log likelihood: 2133.18397

Classes to Clusters:
  0  1   2   3  ← cluster
 53  5   2  29 | -
 53 30  12  43 | *
 42 63   0  75 | +
 49 65   0  46 | =

Cluster 0 ← -
Cluster 1 ← =
Cluster 2 ← *
Cluster 3 ← +
```

E8:

```
Clustered Instances
0      197 (35%)
1      163 (29%)
2       14 (2%)
3      193 (34%)

Log likelihood: 2133.18397

Classes to Clusters:
 0     1   2   3  ← to cluster
 53    5   2  29 | -
 53   30  12  43 | *
 42   63   0  75 | +
 49   65   0  46 | =

Cluster 0 ← -
Cluster 1 ← =
Cluster 2 ← *
Cluster 3 ← +
```

<u>Σχολιασμός αποτελεσμάτων για το "all-unique-clean -reduced" dataset</u>

- Στην περίπτωση χρήσης του k-means παρατηρούμε τα εξής:

  - Μεταξύ των δύο πρώτων δοκιμών (K1, K2), οι οποίες διαφέρουν μόνο στην χρησιμοποιούμενη συνάρτηση απόστασης υπάρχει αμελητέα διαφορά, με την δοκιμή K2 (Manhattan Distance) να εμφανίζει μεγαλύτερο ποσοστό αποτυχίας. Η κατανομή στις συστάδες είναι σχεδόν ίδια. Λόγω της ελάχιστα καλύτερης επίδοσης της Euclidean distance, επιλέγουμε να συνεχίσουμε αυτήν στις ακόλουθες δοκιμές.

  - Μεταξύ των δοκιμών K1, K3, K4 μεταβάλλουμε το πλήθος των συστάδων. Όπως και στην περίπτωση του προηγούμενου data set παρατηρούμε την δημιουργία πιο εξειδικευμένων clusters με την αύξηση του πλήθους των συστάδων. Για τους ίδιους λόγους συνεχίζουμε με πλήθος συστάδων ίσο με 4.

  - Μεταξύ των δοκιμών K1, K5, K6 μεταβάλλουμε το μήκος των n-grams, από 1 μέχρι 3. Παρατηρείται μικρή αύξηση του ποσοστού αποτυχίας κατά τη μετάβαση από unigrams σε bigrams ή trigrams, καθώς παρατηρούμε συγκέντρωση των περισσότερων tweets στο πρώτο cluster.

- Στη δοκιμή Κ7 εφαρμόζουμε stemming στα unigrams πριν τη συσταδοποίηση και παρατηρούμε την απόδοση να μένει σταθερή σε σύγκριση με την αντίστοιχη χωρίς stemming δοκιμή (Κ1), αφού παρατηρούνται ελάχιστες μετακινήσεις tweets. Εφόσον δεν προσφέρει κάτι, δεν θα εφαρμόσουμε stemming στις ακόλουθες δοκιμές.

- Στις δοκιμές Κ8 και Κ9 εισάγουμε την επιλογή χαρακτηριστικών ενώ μεταβάλλουμε το κριτήριο επιλογής 1.500 από τα αρχικά 1.682 unigrams. Σε σύγκριση με την δοκιμή Κ1 οι δοκιμές Κ8 και Κ9 παρουσιάζουν παρόμοιο ποσοστό αποτυχίας, με ελάχιστες μετακινήσεις στις Κ8 και Κ9 που βελτιώνουν την απόδοση. Οι συσταδοποιήσεις των Κ8 και Κ9 είναι πανομοιότυπες.

➢ Στην περίπτωση χρήσης του k-means παρατηρούμε τα εξής:

- Μεταξύ των δοκιμών Ε1, Ε2, Ε3 μεταβάλλουμε το πλήθος των συστάδων. Όπως και στην περίπτωση του k - means παρατηρούμε την δημιουργία πιο εξειδικευμένων clusters με την αύξηση του πλήθους των συστάδων. Για τους ίδιους λόγους συνεχίζουμε με πλήθος συστάδων ίσο με 4.

- Μεταξύ των δοκιμών Ε1, Ε4, Ε5 μεταβάλλουμε το μήκος των n-grams, από 1 μέχρι 3. Παρατηρείται μικρή αύξηση του ποσοστού αποτυχίας κατά τη μετάβαση από unigrams σε bigrams ή trigrams, καθώς παρατηρούμε συγκέντρωση των περισσότερων tweets σε ένα cluster. Επομένως, κρίνεται σκόπιμο να συνεχίσουμε με unigrams στις ακόλουθες δοκιμές.

- Στη δοκιμή Ε6 εφαρμόζουμε stemming στα unigrams πριν τη συσταδοποίηση και παρατηρούμε την απόδοση βελτιώνεται κατά 5% περίπου σε σύγκριση με την αντίστοιχη χωρίς stemming δοκιμή (Ε1). Επιπλέον, στην δοκιμή Ε6 αποδίδεται και η απροσδιόριστη κλάση σε μία από τις τέσσερις συστάδες. Μελετώντας αναλυτικά τον πίνακα βλέπουμε ότι το stemming ευνοεί μόνο την θετική συστάδα.

- Στις δοκιμές Ε7 και Ε8 εισάγουμε την επιλογή χαρακτηριστικών με stemming ενώ μεταβάλλουμε το κριτήριο επιλογής 1.300 από τα αρχικά 1.388 unigrams. Σε σύγκριση με την δοκιμή Ε1 οι δοκιμές Ε7 και Ε8 παρουσιάζουν παρόμοιο ποσοστό αποτυχίας, με ελάχιστες μετακινήσεις στις Ε7 και Ε8 που βελτιώνουν λίγο την απόδοση αναγνωρίζοντας πλέον όλες τις κλάσεις.

Δ) Συνολικός Σχολιασμός αποτελεσμάτων

Επίδραση μεταβαλλόμενων παραμέτρων

- Παρατηρήθηκε ελάχιστη βελτίωση με αύξηση του πλήθους των συστάδων (μέχρι έναν αριθμό). Κατά την αύξηση των συστάδων έχουμε δημιουργία πολλών μικρών πολύ εξειδικευμένων και καθαρών συστάδων, γεγονός που αποτελεί απόρροια του overfitting. Για τους παραπάνω λόγους, αλλά και για να έχουμε ακριβή περιγραφή του συγκεκριμένου προβλήματος που περιλαμβάνει 4 κλάσεις συναισθήματος, προτιμήσαμε να διατηρήσουμε το πλήθος συστάδων στις τέσσερις (4) κατά την μεταβολή των υπόλοιπων παραμέτρων.

- Η διαφορά επίδοσης μεταξύ των συναρτήσεων απόστασης του αλγορίθμου k-means είναι πολύ μικρή με την Euclidean distance να υπερτερεί στο dataset 1 και 3. Όπως έχει ήδη αναφερθεί, οι δύο συναρτήσεις παρουσιάζουν πολύ μικρή διαφορά στον υπολογισμό της απόστασης, με αποτέλεσμα να δίνουν κοντινά αποτελέσματα στην πλειονότητα των περιπτώσεων.

- Η χρήση Stemming δημιουργεί νέες πιθανές ομοιότητες μεταξύ των tweets, που ανάλογα με το dataset μπορεί να βελτιώσει ή να χειροτερέψει τα αποτελέσματα. Για παράδειγμα, έστω ότι έχουμε τα εξής tweets *"I love Greece"*, *"I won't visit Greece again. I hate it."*, *"I visited Mykonos and hated it"*. Πριν το stemming, θεωρούνται ως περισσότερο όμοια τα δύο πρώτα tweets λόγω της λέξης Greece. Μετά το stemming, οι λέξεις visited – visit και hated – hate ταυτίζονται, με αποτέλεσμα να θεωρούνται πιο όμοια τα δύο τελευταία tweets.

- Προφανώς, όσο μεγαλύτερο είναι το μήκος των n-grams τόσο λιγότερα είναι τα κοινά n-grams που εντοπίζονται μεταξύ των tweets. Επομένως, παρόλο που τα n-grams με μήκος n = 2 και πάνω βοηθούν στον εντοπισμό ομάδων λέξεων που μπορεί να έχουν κάποιο ιδιαίτερο νόημα και συναισθηματικό περιεχόμενο, φαίνεται ότι σε περιπτώσεις μικρών datasets που αποτελούνται από μικρού μήκους κείμενα με απλή σύνταξη, όπως τα δικά μας, δεν βοηθούν στην συσταδοποίηση των κειμένων. Σε πολλές από τις περιπτώσεις που μελετήσαμε παρατηρούμε ότι κατά τη χρήση των bigrams και trigrams η σύγκλιση επιτυγχάνεται σε λιγότερες επαναλήψεις από ότι στην περίπτωση των unigrams. Αυτό σημαίνει ότι η συσταδοποίηση βασίζεται κυρίως στην πρώτη τοποθέτηση των tweets που γίνεται με βάση τα τυχαία κέντρα, η οποία αν και βέλτιστη ενδεχομένως δεν είναι καθόλου αντιπροσωπευτική της ορθής λύσης του προβλήματος, και στη συνέχεια λόγω της σπανιότητας του κάθε n-gram δεν εντοπίζονται ομοιότητες που θα οδηγούσαν σε μετακινήσεις.

- Οι αξιολογητές Info Gain και Chi-square εμφάνισαν σε όλα τα dataset σχεδόν ίδια μεταξύ τους συνολική επίδραση, με μικρές διαφορές στην κατάταξη των

χαρακτηριστικών, ενώ γενικά δεν προσέφεραν ουσιαστική βελτίωση στη συσταδοποίηση.

Πριν προχωρήσουμε σε οποιοδήποτε συμπέρασμα πρέπει να επισημάνουμε το εξής: Όπως έχει γίνει αντιληπτό από τον επιμέρους σχολιασμό των αποτελεσμάτων που προηγήθηκε, το ποσοστό αποτυχίας που προκύπτει από κάθε εκτέλεση σ δεν είναι αρκετά αντιπροσωπευτικός δείκτης της ποιότητας του. Ωστόσο τον παραθέτουμε και τον μελετάμε με σκοπό να παρατηρήσουμε τις διαφορές στην απόδοση που συταδοποίησης οφείλονται στις μεταβολές των παραμέτρων και να έχουμε μια γενική εικόνα του αποτελέσματος. Για να κατανοήσουμε σε βάθος τα αποτελέσματα της συσταδοποίησης, πρέπει να μελετήσουμε την αντίστοιχη Μήτρα Σύγχυσης. Για να είναι συσταδοποίηση πετυχημένη πρέπει οι συστάδες να είναι όσο το δυνατόν πιο "ξεκάθαρες" ως προς το είδος των tweets που περιέχουν, αλλά και να απέχουν μεταξύ τους όσο το δυνατόν περισσότερο. Ιδανικά σε κάθε συστάδα θα έπρεπε να υπάρχουν όλα τα tweets που ανήκουν μόνο σε μία κλάση συναισθήματος. Δηλαδή, η Μήτρα Σύγχυσης θα έπρεπε να έχει μόνο μια τιμή διάφορη του μηδενός σε κάθε στήλη. Για ακόμα καλύτερη εποπτεία του αποτελέσματος έχουμε τη δυνατότητα ανάγνωσης των tweets που αντιστοιχήθηκαν στην κάθε κλάση.
Μελετώντας τα παραπάνω αποτελέσματα, παρατηρούμε ότι ακόμα και αυτά με το μικρότερο ποσοστό αποτυχίας δεν είναι καθόλου ικανοποιητικά. Αναφέρουμε και πάλι ότι λόγω του τρόπου λειτουργίας της μεθόδου αξιολόγησης "classes – to – clusters", η οποία αναθέτει κάθε κλάση το πολύ σε μία συστάδα ανάλογα με την πλειονότητα των tweets της και με κριτήριο την ελαχιστοποίηση του ποσοστού λάθους, υπάρχουν περιπτώσεις που δεν είναι δυνατόν να ανατεθεί κλάση σε κάποια συστάδα. Αυτό που παρατηρούμε γενικά είναι η αδυναμία δημιουργίας "καθαρών" συστάδων, αφού όπως φαίνεται από τις Μήτρες Σύγχυσης, τα tweets κάθε κλάσης "σπάνε" σε μικρότερες ομάδες και μοιράζονται στην πλειονότητα των συστάδων, με αποτέλεσμα να δημιουργούνται είτε πίνακες σχεδόν γεμάτοι με στοιχεία διάφορα του μηδενός είτε πίνακες με συσσωρευμένα tweets σε ένα cluster. Όπως έχει ήδη αναφερθεί έχουμε αντιστοιχήσει το κάθε tweet σε ένα διάνυσμα, με συνιστώσες το σύνολο των διαφορετικών χαρακτηριστικών που προέκυψαν μετά την προεπεξεργασία των δεδομένων, και τιμή της κάθε συνιστώσας τη συχνότητα εμφάνισης του αντίστοιχου χαρακτηριστικού στο συγκεκριμένο tweet.

- Στην περίπτωση του αλγορίθμου k-means, αρχικά επιλέγονται τυχαία 4 από τα tweets ως κέντρα των συστάδων. Στο επόμενο βήμα τα υπόλοιπα tweets ταξινομούνται στις 4 συστάδες με κριτήριο την απόσταση τους από τα κέντρο των συστάδων. Για κάθε tweet επιλέγεται η κοντινότερη συστάδα. Στη συνέχεια τα κέντρα των συστάδων επαναϋπολογίζονται και τα tweets ανακατανέμονται στις κοντινότερες συστάδες με βάση τα νέα κέντρα. Στην ουσία, η προσπάθεια ένταξης του κάθε tweet στη συστάδα από την οποία απέχει λιγότερο, ερμηνεύεται ως προσπάθεια ομαδοποίησης των tweets που έχουν τις περισσότερες κοινές λέξεις. Η διαδικασία επαναλαμβάνεται μέχρι να σταματήσουν να γίνονται ανακατανομές των tweets. Το πρώτο προβληματικό σημείο

της διαδικασίας που εντοπίζουμε είναι ότι η τυχαία αρχική επιλογή των κέντρων είναι πολύ κρίσιμη για τον καθορισμό του αποτελέσματος, όπως φαίνεται από την αστάθεια που προκαλεί η μεταβολή του seed (και άρα των αρχικών κέντρων). Ναι μεν, διατηρώντας το ίδιο seed σε όλα τα πειράματα μπορούμε να παρατηρήσουμε τις μεταξύ τους μεταβολές, αλλά σε καμία περίπτωση δεν μπορούμε να ισχυριστούμε ότι πετυχαίνουμε τη βέλτιστη λύση του προβλήματος. Το βασικότερο όμως μειονέκτημα της μεθόδου, όσον αφορά τη χρήση της στο πρόβλημα της Ανάλυσης Συναισθήματος, είναι ότι αντιμετωπίζει τις λέξεις σαν απλούς αριθμούς χωρίς καμία σημασιολογία, και επομένως δεν αναγνωρίζει τη συναισθηματική βαρύτητα τους. Αυτό έχει σαν αποτέλεσμα να καταλήγουν στην ίδια συστάδα tweets που μπορεί να έχουν κάποιον ικανό αριθμό κοινών λέξεων, διαφέρουν όμως σε κάποιες άλλες λέξεις, ακόμα και σε μία μόνο λέξη, που είναι οι πιο κρίσιμες για την ταξινόμηση σε κάποια κλάση συναισθήματος. Αναφέρουμε το εξής απλουστευμένο χαρακτηριστικό παράδειγμα που προέκυψε από παρατήρηση των αποτελεσμάτων μας:
Έστω ότι το θετικό tweet *'#Nafplio: One of the most romantic cities in **#Greece**! Find out more here:'* έχει τοποθετηθεί στο cluster0 και το αρνητικό tweet *'the poor poor dog and all others roaming the streets. Heartbreaking.'* στο cluster1. Στην απλούστερη εκδοχή, κατά την οποία δεν έχουν τοποθετηθεί άλλα tweets στα clusters, το αρνητικό tweet *'Come to #LiveYourMyth in **#Greece** where #Tourism in #Kerkyra #Airport Fulll of Garbage '* θα τοποθετηθεί εσφαλμένα στο cluster0. Προφανώς η όλη διαδικασία της συσταδοποίησης δεν είναι τόσο απλή, αφού υπάρχουν πολλά tweets με τα οποία γίνεται ταυτόχρονα σύγκριση, αλλά και πολλές επαναλήψεις που βελτιώνουν τα αποτελέσματα. Ωστόσο, με το παράδειγμα αυτό θέλουμε να εστιάσουμε στο γεγονός ότι αρκεί μια μικρή ομοιότητα στο λεξιλόγιο των tweets για να αγνοηθεί εντελώς το συνολικό συναισθηματικό περιεχόμενο.

- Στην περίπτωση του αλγορίθμου E-M, θεωρείται ως άγνωστη παράμετρος το cluster στο οποίο ανήκει το κάθε tweet, και επιδιώκεται η πρόβλεψη της βάσει ενός πιθανοτικού μοντέλου. Ο αλγόριθμος υποθέτει ότι τα δεδομένα προέκυψαν από κανονικές κατανομές, το πλήθος τον οποίων ταυτίζεται με το πλήθος των συστάδων και οι παράμετροι των οποίων αναζητούνται. Στόχος του αλγορίθμου είναι να κάνει υποθέσεις για τον τρόπο δημιουργίας των δεδομένων ώστε τελικά να προσδιορίσει τις κατανομές αυτές. Αρχικά, με κάποιο τρόπο που εξαρτάται από την υλοποίηση, τα δεδομένα ταξινομούνται στις συστάδες. Στην περίπτωση του WEKA η αρχική συσταδοποίηση των tweets γίνεται με χρήση του αλγορίθμου k-means. Με βάση την αρχική αυτή ανάθεση υπολογίζεται ένα πρώτο μοντέλο. Στη συνέχεια, κατά το βήμα E, για κάθε tweet υπολογίζεται πιθανότητα του να ανήκει σε κάθε μία από τις συστάδες, με βάση την τρέχουσα εκτίμηση του μοντέλου. Το κάθε tweet ταξινομείται υποθετικά στην συστάδα για την οποία εμφανίζει τη μεγαλύτερη πιθανότητα. Ακολούθως, στο βήμα M γίνεται επανεκτίμηση του μοντέλου με βάση τις υποθετικές ταξινομήσεις που έγιναν στο στάδιο E. Στόχος της επανεκτίμησης είναι να μεγιστοποιηθεί η συνάρτηση

πιθανοφάνειας που προκύπτει. Η εναλλαγή μεταξύ των βημάτων Ε και Μ συνεχίζεται μέχρι να μην υπάρχει πλέον βελτίωση της πιθανοφάνειας των δεδομένων. Και στην περίπτωση του αλγορίθμου ΕΜ παρατηρείται η αστάθεια που οφείλεται στην επιλογή του seed. Επιπλέον, η χρήση του k-means για την αρχικοποίηση των συστάδων εισάγει τα προβλήματα που αναφέρθηκαν στον αλγόριθμο k-means. Και σε αυτή την περίπτωση αλγορίθμου δίνεται έμφαση στον εντοπισμό ομοιότητας μεταξύ των tweets, με αποτέλεσμα να αγνοείται η ύπαρξη ενδεχομένως ουσιαστικών διαφορών.

Όσον αφορά το συγκεκριμένο dataset που χρησιμοποιήθηκε διαπιστώθηκαν τα εξής χαρακτηριστικά στοιχεία στα οποία ενδεχομένως πρέπει να αποδώσουμε ένα μέρος της ευθύνης για την κακή ποιότητα των αποτελεσμάτων:

Το dataset που χρησιμοποιήθηκε ήταν σχετικά μικρό και δεδομένης της τεράστιας ποικιλίας τρόπου έκφρασης που χρησιμοποιείται από τους χρήστες του Twitter ο εντοπισμός κοινών λέξεων και πολύ περισσότερο κοινών φράσεων ήταν σπάνιος. Αυτό μπορούμε να το αντιληφθούμε από το μεγάλο πλήθος των διαστάσεων των διανυσμάτων που δημιουργήθηκαν. Επίσης, πρέπει να επισημάνουμε ότι αν και στόχος μας ήταν να ασχοληθούμε με το θεματικό τομέα του τουρισμού στην Ελλάδα, ο τρόπο άντλησης των δεδομένων δεν μπορεί να εγγυηθεί κάτι τέτοιο. Συγκεκριμένα, κατά το χειρονακτικό σχολιασμό συναντήθηκαν tweets που ναι μεν ταξινομήθηκαν σε κάποια από τις τέσσερις κλάσεις αλλά διαπιστώσαμε ότι το περιεχόμενο τους δεν ήταν σχετικό με τον τουρισμό. Αυτό καθιστά ακόμα πιο απίθανο τον εντοπισμό ομοιοτήτων μέσα στα tweets. Ωστόσο τα συγκεκριμένα tweets δεν εξαιρέθηκαν για να προσομοιωθεί όσο το δυνατόν πιο ρεαλιστικά το πρόβλημα. Άλλωστε, η συσταδοποίηση δεν παρουσιάζει κάποια εξάρτηση από το νοηματικό τομέα των κειμένων που επεξεργάζεται. Τέλος, επισημαίνουμε ότι γεγονός ότι το πρώτο dataset, το οποίο περιέχει τα πιο θορυβώδη δεδομένα, με links, αναφορές σε usernames, και ανισότητα μεταξύ των διάφορων κλάσεων, εμφανίζει τα μικρότερα ποσοστά αποτυχίας οφείλεται πιθανότατα στο overfitting.

4. Εφαρμογή Επιβλεπόμενης Μηχανικής Μάθησης σε πραγματικά δεδομένα

Ακολούθως, επιχειρούμε την εφαρμογή μεθόδου Επιβλεπόμενης Μηχανικής Μάθησης στα δεδομένα, για να συγκρίνουμε τα αποτελέσματα που προκύπτουν με αυτά που προέκυψαν από την εφαρμογή συσταδοποίησης. Πέρα από τη μελέτη της επίδραση των αντίστοιχων παραμέτρων στην Επιβλεπόμενη μέθοδο θα μπορέσουμε να εκτιμήσουμε κατά πόσο το συγκεκριμένο dataset ευθύνεται για την κακή ποιότητα των αποτελεσμάτων.

Dataset

Επιλέγουμε να χρησιμοποιήσουμε το dataset "all – unique – clean – reduced", λόγω της αυξημένης ευαισθησίας της Επιβλεπόμενης Μάθησης στο φαινόμενο του overfitting σε περίπτωση άνισων κλάσεων.

Αλγόριθμος

Ο αλγόριθμος Επιβλεπόμενης Μηχανικής Μάθησης που επιλέξαμε είναι ο αλγόριθμος SMO (Sequential Minimum Optimization) του WEKA, που στηρίζεται στη χρήση διανυσμάτων υποστήριξης (SVM). Ουσιαστικά, ο αλγόριθμος αυτός χρησιμοποιείται για την εκπαίδευση των διανυσμάτων υποστήριξης, επιλύοντας το πρόβλημα βελτιστοποίησης που προκύπτει κατά τη δημιουργία τους. Ο λόγος που επιλέξαμε τον αλγόριθμο SVM είναι η ικανότητα που παρουσιάζει στο χειρισμό διανυσμάτων πολλών διαστάσεων.

Πειράματα

Οι διάφορες δοκιμές έγιναν με τη μέθοδο 10 – fold, η οποία προτιμάται σε περιπτώσεις μικρού συνόλου δεδομένων. Όπως έχει αναφερθεί και στο κεφάλαιο 2, κατά τη μέθοδο αυτή το dataset χωρίζεται σε 10 υποσύνολα. Από αυτά τα 10 υποσύνολα τα 9 χρησιμοποιούνται για την εκπαίδευση (training) του μοντέλου και το 1 υποσύνολο που απομένει χρησιμοποιείται για τον έλεγχο (test). Η διαδικασία επαναλαμβάνεται συνολικά 10 φορές, ώστε κάθε ένα από τα υποσύνολα να χρησιμοποιηθεί ακριβώς μία φορά ως υποσύνολο ελέγχου. Τα 10 διαφορετικά μοντέλα που προκύπτουν συνδυάζονται ώστε να προκύψει ένα τελικό μοντέλο.

Στις παραμέτρους του αλγορίθμου SMO διατηρούμε τις προεπιλεγμένες τιμές. Ακολουθεί συγκεντρωτικός πίνακας των δοκιμών που έγιναν.

Αρχικό πλήθος attributes: 1.680

ID	String- to -word filter		Attribute- selection filter		Incorrectly clustered instances
S1	**stemmer**	-			
	tokenizer	unigrams	**evaluator**	-	*35,62%*
S2	**stemmer**	-			
	tokenizer	bigrams	**evaluator**	-	*40,91%*
S3	**stemmer**	-			
	tokenizer	trigrams	**evaluator**	-	*44,79%*
S4	**stemmer**	✓			
	tokenizer	unigrams	**evaluator**	-	*35,09%*
S5	**stemmer**	✓		InfoGain	
	tokenizer	unigrams	**evaluator**	(1500 attributes)	***34,92%***
S6	**stemmer**	✓		Chi -squared	
	tokenizer	unigrams	**evaluator**	(1500 attributes)	*36,15%*

Πίνακας 18. Αποτελέσματα χρήσης αλγορίθμου SMO

Ακολουθούν τα αναλυτικά αποτελέσματα από το WEKA για τις παραπάνω έξι εκτελέσεις. Από το σύνολο των διαθέσιμων αποτελεσμάτων τα ενδεικτικά της επιτυχίας της ταξινόμησης που προβλέπει το εκάστοτε μοντέλο είναι το ποσοστό σωστά ταξινομημένων tweets (correctly classified instances), δηλαδή η ακρίβεια (accuracy), η περιοχή κάτω από την καμπύλη ROC (ROC area) και η γνωστή Μήτρα Σύγχυσης (Confusion Matrix). Το kappa statistic μετρά τη συμφωνία της πρόβλεψης με την πραγματική εξαιρώντας τον παράγοντα της τύχης, ενώ τα διάφορα στατιστικά λάθη (Mean Absolute error, Root mean squared error, Relative absolute error, Root relative squared error) δεν έχουν ιδιαίτερη σημασία και ερμηνεία στο πρόβλημα της ταξινόμησης που μελετάμε. Το μέγεθος ROC Area προκύπτει από την καμπύλη ROC η οποία αναφέρεται σε μία κλάση και αναπαριστά τις TP ταξινομήσεις σε συνάρτηση με τις FP, τιμές που προκύπτουν μεταβάλλοντας το κατώφλι της ταξινόμησης. Το ROC Area κυμαίνεται από 0,5 (στη χειρότερη περίπτωση) μέχρι 1 (στον ιδανικό ταξινομητή) και μαρτυρά το πόσο καλά μπορεί να διακρίνει ο ταξινομητής τα δείγματα που ανήκουν στην συγκεκριμένη κλάση από αυτά που δεν ανήκουν. Τέλος, τα TP Rate, FP Rate, Precision, Recall, F-measure είναι οι γνωστές από το 2^{ο} κεφάλαιο μετρικές απόδοσης του συστήματος.

S1:

Correctly Classified Instances	365	64.3739 %
Incorrectly Classified Instances	202	35.6261 %
Kappa statistic	0.5148	
Mean absolute error	0.2988	
Root mean squared error	0.3816	
Relative absolute error	81.2128 %	
Root relative squared error	88.9867 %	
Total Number of Instances	567	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.64	0.04	0.75	0.64	0.691	0.885	-
	0.674	0.168	0.564	0.674	0.614	0.771	*
	0.711	0.15	0.688	0.711	0.699	0.806	+
	0.544	0.13	0.621	0.544	0.58	0.729	=
Weighted Avg.	0.644	0.131	0.649	0.644	0.644	0.788	

=== Confusion Matrix ===

a	b	c	d	← classified as
57	16	9	7	a = -
7	93	15	23	b = *
5	24	128	23	c = +
7	32	34	87	d = =

S2:

Correctly Classified Instances	335	59.0829 %
Incorrectly Classified Instances	232	40.9171 %
Kappa statistic	0.4432	
Mean absolute error	0.3092	
Root mean squared error	0.3955	
Relative absolute error	84.049 %	
Root relative squared error	92.2277 %	
Total Number of Instances	567	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.371	0.008	0.892	0.371	0.524	0.749	-
	0.797	0.368	0.41	0.797	0.542	0.713	*
	0.628	0.101	0.743	0.628	0.681	0.811	+
	0.494	0.076	0.718	0.494	0.585	0.743	=
Weighted Avg.	0.591	0.144	0.679	0.591	0.595	0.758	

=== Confusion Matrix ===

a	b	c	d	←classified as
33	49	3	4	a = -
2	110	13	13	b = *
0	53	113	14	c = +
2	56	23	79	d = =

S3:

Correctly Classified Instances	313	55.2028 %
Incorrectly Classified Instances	254	44.7972 %
Kappa statistic	0.3929	
Mean absolute error	0.3161	
Root mean squared error	0.4042	
Relative absolute error	85.9265 %	
Root relative squared error	94.2417 %	
Total Number of Instances	567	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.315	0.004	0.933	0.315	0.471	0.672	-
	0.833	0.462	0.367	0.833	0.51	0.689	*
	0.556	0.072	0.781	0.556	0.649	0.795	+
	0.438	0.064	0.729	0.438	0.547	0.753	=
Weighted Avg.	0.552	0.154	0.69	0.552	0.558	0.738	

=== Confusion Matrix ===

a	b	c	d	← classified as
28	58	2	1	a = -
1	115	11	11	b = *
0	66	100	14	c = +
1	74	15	70	d = =

S4:

Correctly Classified Instances	368	64.903 %
Incorrectly Classified Instances	199	35.097 %
Kappa statistic	0.5226	
Mean absolute error	0.3001	
Root mean squared error	0.3831	
Relative absolute error	81.5723 %	
Root relative squared error	89.3228 %	
Total Number of Instances	567	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.674	0.046	0.732	0.674	0.702	0.874	-
0.659	0.161	0.569	0.659	0.611	0.763	*

	0.717	0.15	0.69	0.717	0.703	0.805	+
	0.55	0.123	0.638	0.55	0.591	0.723	=
Weighted Avg.	0.649	0.129	0.652	0.649	0.649	0.782	

=== Confusion Matrix ===

a	b	c	d	← classified as
60	15	8	6	a = -
8	91	16	23	b = *
6	24	129	21	c = +
8	30	34	88	d = =

S5:

Correctly Classified Instances	369	65.0794 %
Incorrectly Classified Instances	198	34.9206 %
Kappa statistic	0.5251	
Mean absolute error	0.3001	
Root mean squared error	0.3831	
Relative absolute error	81.5723 %	
Root relative squared error	89.3303 %	
Total Number of Instances	567	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.674	0.048	0.723	0.674	0.698	0.874	-
	0.652	0.159	0.57	0.652	0.608	0.761	*
	0.717	0.147	0.694	0.717	0.705	0.805	+
	0.563	0.123	0.643	0.563	0.6	0.726	=
Weighted Avg.	0.651	0.128	0.654	0.651	0.651	0.783	

=== Confusion Matrix ===

a	b	c	d	← classified as
60	15	8	6	a = -
8	90	16	24	b = *
6	25	129	20	c = +
9	28	33	90	d = =

S6:

Correctly Classified Instances	362	63.8448 %
Incorrectly Classified Instances	205	36.1552 %
Kappa statistic	0.5071	
Mean absolute error	0.3009	
Root mean squared error	0.3845	
Relative absolute error	81.772 %	
Root relative squared error	89.6502 %	
Total Number of Instances	567	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.551	0.033	0.754	0.551	0.636	0.878	-
	0.688	0.217	0.505	0.688	0.583	0.745	*
	0.733	0.137	0.714	0.733	0.723	0.816	+
	0.538	0.106	0.667	0.538	0.595	0.732	=
Weighted Avg.	0.638	0.131	0.656	0.638	0.639	0.785	

=== Confusion Matrix ===

a	b	c	d	← classified as
49	29	5	6	a = -
7	95	18	18	b = *
1	28	132	19	c = +
8	36	30	86	d = =

<u>Σχολιασμός</u>

Όπως ήταν αναμενόμενο τα αποτελέσματα είναι σαφώς καλύτερα από τα αντίστοιχα που προέκυψαν με συσταδοποίηση για το ίδιο dataset (all - unique - clean - reduced). Δεδομένου του ότι η ταξινόμηση γίνεται σε τέσσερις κλάσεις, η τυχαία ταξινόμηση ενός tweet έχει πιθανότητα επιτυχίας 25%. Επομένως, το ποσοστό επιτυχίας της Επιβλεπόμενης εκπαίδευσης του συστήματος που κυμαίνεται γύρω στο 65% είναι αρκετά ικανοποιητικό και αποδεκτό δεδομένης της δυσκολίας του προβλήματος. Όπως και στην περίπτωση της συσταδοποίησης του ίδιου dataset, την καλύτερη απόδοση παρουσιάζουν τα unigrams. Το stemming φαίνεται να επιδρά θετικά, όπως είχε γίνει και στην αντίστοιχη περίπτωση συσταδοποίησης (κυρίως κατά τη χρήση του E-M αλγορίθμου). Η επιλογή χαρακτηριστικών οδηγεί σε βελτίωση της απόδοσης, η οποία παραμένει σταθερή μεταξύ των δύο διαφορετικών αξιολογητών, ενώ στην περίπτωση της συσταδοποίησης η επιλογή χαρακτηριστικών είχε αμυδρά επιδράσει θετικά μόνο στην περίπτωση του αλγορίθμου k-means. Όσον αφορά την απόδοση σε κάθε κλάση ξεχωριστά, παρατηρούμε ότι στην πλειονότητα των δοκιμών, χειρότερη απόδοση εμφανίζεται στην ουδέτερη (=) και στην απροσδιόριστη (*) κλάση, πράγμα αναμενόμενο αφού οι κλάσεις αυτές ουσιαστικά χαρακτηρίζονται είτε από την απουσία συναισθηματικών λέξεων στην περίπτωση της ουδέτερης κλάσης, είτε από τη συνύπαρξη ανάμεικτων συναισθημάτων στην περίπτωση της απροσδιόριστης κλάσης. Δηλαδή, δεν υπάρχουν αντικειμενικά αντιπροσωπευτικές τους λέξεις και επομένως δύσκολα μπορούν να εντοπιστούν με αυτόματο τρόπο.

5. Σύνοψη και Συμπεράσματα

Βασικός στόχος των πειραμάτων που εκτελέστηκαν ήταν η μελέτη της απόδοσης της Μη Επιβλεπόμενης Μηχανικής Μάθησης στο πρόβλημα της Ανάλυσης Συναισθήματος. Όπως προέκυψε, η μέθοδος της συσταδοποίησης δεν αποφέρει ικανοποιητικά αποτελέσματα στο είδος των κειμένων που μελετήθηκαν και στο συγκεκριμένο νοηματικό τομέα. Όπως φάνηκε και κατά την ανάλυση των αποτελεσμάτων το πρόβλημα της συσταδοποίησης δεν έχει μία και μοναδική σωστή λύση. Μεταβάλλοντας τις παραμέτρους, παρατηρήσαμε κάποια χαρακτηριστικά να βελτιώνονται εις βάρος κάποιων άλλων. Σε τέτοιες περιπτώσεις το επιδιωκόμενο αποτέλεσμα προκύπτει συνήθως μετά από "συμβιβασμούς" και εξαρτάται από τις ανάγκες του εκάστοτε προβλήματος.

Το είδος των κειμένων που μελετήθηκε έχει πολλές ιδιαιτερότητες που καθιστούν την ανάλυση του δύσκολη. Αυτό φάνηκε τόσο στη συσταδοποίηση όσο και στην Επιβλεπόμενη μέθοδο. Το μικρό μήκος των tweets αναγκάζει τους χρήστες να εκφράζονται με ανεπίσημο λεξιλόγιο και πολλές χωρίς ορθή σύνταξη. Χρησιμοποιούνται συνήθως συντομεύσεις λέξεων και τροποποιημένες ορθογραφικά λέξεις που διαφέρουν από χρήστη σε χρήστη αλλά και σύμβολα ή λέξεις που καθιερώνονται στο διαδίκτυο και μεταβάλλονται διαρκώς. Όσο πιο πολύ απομακρυνόμαστε από το αντικειμενικά ορθό λεξιλόγιο και συντακτικό τόσο πιο απίθανο γίνεται να εντοπιστούν ομοιότητες μεταξύ των tweets, οι οποίες θα μπορούσαν να οδηγήσουν σε σωστή συσταδοποίηση. Τέλος, πολύ συνηθισμένη τακτική είναι η παράθεση εικόνας ή συνδέσμου που συμπληρώνει το tweet και μεταδίδει καλύτερα το νόημα του, στα οποία η Ανάλυση Συναισθήματος προφανώς δεν έχει πρόσβαση.

Στην προεπεξεργασία που εφαρμόσαμε συμπεριλήφθησαν σε διάφορους συνδυασμούς τα εξής: απομάκρυνση αναφορών και συνδέσμων, μετατροπή όλων των γραμμάτων σε μικρά, stemming και χρήση n-grams. Διαπιστώθηκε ότι η χρησιμότητα του stemming και των n-grams είναι κάτι που εξαρτάται καθαρά από το εκάστοτε dataset στο οποίο χρησιμοποιούνται. Ως κριτήρια επιλογής χαρακτηριστικών χρησιμοποιήθηκαν το Information Gain και το Chi - square με τις αποδόσεις τους να ταυτίζονται στις περισσότερες περιπτώσεις, χωρίς να προσφέρουν ουσιαστική βελτίωση. Όπως έχει αναφερθεί, η χρήση τέτοιων κριτηρίων επιλογής δεν είναι δυνατή σε περιπτώσεις Μη Επιβλεπόμενης Μάθησης, αλλά στην περίπτωση μας χρησιμοποιήθηκαν ως προεπεξεργασία των δεδομένων. Για την αναπαράσταση των tweets χρησιμοποιήθηκαν διανύσματα με συνιστώσες τις συχνότητες χαρακτηριστικών. Ελάχιστα tweets παρουσίασαν κάποια λέξη πάνω από μία φορά, οπότε ουσιαστικά οι συνιστώσες δήλωναν παρουσία ή απουσία ενός χαρακτηριστικού.

Όσον αφορά τους δύο αλγορίθμους συσταδοποίησης που μελετήθηκαν, παρατηρήθηκε ελάχιστα καλύτερη απόδοση στον αλγόριθμο Expectation - Maximization, η οποία όμως δε δικαιολογεί το πολύ μεγαλύτερο υπολογιστικό του κόστος.

Τα πλεονεκτήματα της χρήσης συσταδοποίησης στο πρόβλημα της Ανάλυσης Συναισθήματος είναι ότι δεν απαιτούνται σχολιασμένα δεδομένα και επομένως ανθρώπινη ανάμειξη. Επιπλέον δεν απαιτείται αρχικά χρόνος εκπαίδευσης του συστήματος και το σύστημα δεν συνδέεται με κάποιο μοναδικό νοηματικό τομέα. Ωστόσο, αυτή η πλήρης αγνόηση του οποιουδήποτε νοήματος οδηγεί σε αστοχία στην περίπτωση μας. Οι μέθοδοι της συσταδοποίησης αναζητούν ομοιότητες μεταξύ των tweets και αγνοούν τη βαρύτητα

των διαφορών. Ενδεχομένως η μέθοδος θα ήταν πιο αποτελεσματική σε κείμενα περιορισμένου λεξιλογίου, στα οποία το συναίσθημα εκφράζεται με αντικειμενικά συναισθηματικές λέξεις και όχι με λέξεις που αποκτούν συναίσθημα μόνο εφόσον κάποιος γνωρίζει τον νοηματικό τομέα του κειμένου.

Για την μελλοντική έρευνα πάνω στο συγκεκριμένο πρόβλημα προτείνεται η εισαγωγή βαρύτητας στις λέξεις των κειμένων, κατά την προεπεξεργασία, ώστε να γίνει απομάκρυνση των χαρακτηριστικών που δε φέρουν έντονο συναισθηματικό περιεχόμενο. Η βαρύτητα μπορεί να υπολογιστεί με τη βοήθεια ειδικών λεξικών συνωνύμων (π.χ. WordNet), ώστε να βρεθεί η σχετικότητα κάθε λέξης με κάποιες πολύ χαρακτηριστικές λέξεις συναισθήματος (π.χ. “good”, “bad”). Θα ήταν πολύ χρήσιμο να γίνει ο αντίστοιχος υπολογισμός βαρύτητας με αναφορά στο λεξιλόγιο του ειδικού νοηματικού τομέα στον οποίο ανήκουν τα κείμενα. Επίσης, προτείνεται είτε η επέκταση του εργαλείου WEKA, είτε η χρήση άλλου εργαλείου με περισσότερες δυνατότητες προεπεξεργασίας δεδομένων, όπως η διαχείριση άρνησης και η αναγνώριση μέρους του λόγου (POS-tagging).

Υποσημειώσεις

1. Amazon Mechanical Turk: που επιτρέπει σε ιδιώτες και επιχειρήσεις να συντονίζουν τη χρήση της ανθρώπινης νοημοσύνης με στόχο την εκτέλεση καθηκόντων που ο υπολογιστής δεν είναι σε θέση να κάνει.

2. Το Dialogue Earth, είναι ένα πρόγραμμα του Ινστιτούτου Περιβάλλοντος στο Πανεπιστήμιο της Μινεσότα. (www.dialogueearth.org)

3. Τα top 2.633 unigrams είναι τα 2.633 πιο συχνά unigrams.

4. Το tf – idf είναι ένα στατιστικό μέτρο που εκφράζει το πόσο σημαντική είναι μια λέξη σε ένα κείμενο, όταν έχουμε μια συλλογή κειμένων. Η τιμή του αυξάνεται ανάλογα με τον αριθμό εμφάνισης της λέξης στο κείμενο αλλά αντισταθμίζεται από τη συχνότητα της λέξης σε ολόκληρη τη συλλογή των κειμένων, ώστε να λαμβάνεται υπόψη το κατά πόσο η λέξη συνηθίζεται να έχει μεγάλη συχνότητα εμφάνισης στα κείμενα.

5. Emoticons είναι ανθρώπινες εκφράσεις προσώπου που σχηματίζονται με χρήση σημείων στίξης, αριθμών και γραμμάτων και μαρτυρούν το αντίστοιχο συναίσθημα

6. Hashtag ονομάζεται μια λέξη που ακολουθεί το σύμβολο # και χρησιμεύει στην σύνδεση του tweet με την κατηγορία που δηλώνει το hashtag .

7. Η αναφορά σε κάποιον χρήστη γίνεται με το σύμβολο @ ακολουθούμενο από το όνομα χρήστη.

8. Έκτροπες ονομάζονται οι παρατηρήσεις που απέχουν σημαντικά από τις υπόλοιπες παρατηρήσεις του συνόλου.

Βιβλιογραφία

[1] Sauri, R.(2008). A Factuality Profiler for Eventualities in Text. Brandeis University.

[2] Phan Trong Ngoc, Myungsik Yoo (2014). The lexicon-based sentiment analysis for fan page ranking in Facebook. ICOIN 2014: 444-448

[3] Taboada, M., J. Brooke, Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon – Based Methods for Sentiment Analysis. Computational Linguistics 37 (2): 267-307.

[4] Kennedy, A., and Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence, 22(2), 110–125.

[5] Polanyi, L., and Zaenen, A. (2006). Contextual Valence Shifters. Computing Attitude and Affect in Text: Theory and Applications (pp.1–10). Dordrecht, The Netherlands: Springer

[6] Muhammad, A., Wiratunga, N., Lothian, R., Glassey, R. (2013). Domain- Based Lexicon Enhancement for Sentiment Analysis. SMA@BCS SGAI 2013: 7-18

[7] Hu, Minqing and Bing Liu. (2004). Mining and summarizing customer reviews. in Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004).

[8] Strapparava, C., and Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pp. 1083-1086.

[9] Esuli, A. and Sebastiani, F. (2006).SentiWordNet: a high-coverage lexical resource for opinion mining. Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR)

[10] Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., and Gandini, G. (2007). Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining

[11] Balahur, A., Steinberger, R., Goot, E., Pouliquen, B.; and Kabadjov, M. (2009). Opinion mining on newspaper quotations. In WI-IAT'09, volume 3, 523–526.

[12] Remus, R., und Ziegelmayer, D. (2014). Learning from Domain Complexity. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)

[13] Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data preprocessing for supervised learning. International Journal of Computer Science, 1, 111–117

[14] Haddi, E., Liu, X., Shi, Y. (2013). The Role of Text Pre-Processing Sentiment Analysis, Information Technology and Quantitative Management, Procedia Computer Science 26 -32, Elsevier 2013.

[15] Meyer, D., and Hornik, K., and Feinerer, I. (2008) Text Mining Infrastructure in R. Journal of Statistical Software, 25 (5). pp. 1-54. ISSN 1548-7660

[16] Prasad S.(2010). Micro-blogging Sentiment Analysis Using Bayesian Classification Methods.

[17] Altrabsheh, N., Cocea, M. and Fallahkhair, S. (2014). Learning sentiment from students' feedback for real-time interventions in classrooms. In: Bouchachia, Abdelhamid, ed. Adaptive and intelligent systems: Third International Conference, ICAIS 2014, Bournemouth, UK, September 8-10, 2014. proceedings. Lecture Notes in Computer Science. Springer, Heidleberg, pp. 40-49. ISBN 9783319112978

[18] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In EMNLP, pages 79–86.

[19] Gamon, M. (2004).Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis, in Proceeding of COLING-04, the 20th International Conference on Computational Linguistics, International Conference on Computational Linguistics, Geneva, CH, August 2004

[20] Boiy, E., Hens, P., Deschacht, K., and Moens, M.F.(2007). Automatic Sentiment Analysis in On-line Text. ELPUB, page 349-360.

[21] Morante, R., Daelemans, W.(2009). A metalearning approach to processing the scope of negation. In. Proc of the 13th Conference on Computational Natural Language Learning, pp 21-29

[22] Yi, J., Nasukawa, T., Bunescu, R., Niblack, W. (2003). Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. icdm, Third IEEE International Conference on Data Mining (ICDM'03). Florida, pp.427

[23] Kim, Y. et al.,(2003). Feature Selection in Data Mining. Data Mining: Opportunities and Challenges, Idea Group Publishing , pp. 80-105.

[24] Khalilian, A., Sahamijoo, G., Avatefipour, O., Piltan, F., Nasrabad, M.R.S. (2014). Design High Efficiency-Minimum Rule Base PID Like Fuzzy Computed Torque Controller. Research and Development Department, Institute of Advance Science and Technology-SSP, Shiraz, Iran

[25] Salton, G., Buckley, C.(1988). Term-weighting approaches in automatic text retrieval. Inform.Process. Man 24(5), 513–523

[26] Tan, S. and Zhang, J. (2008).An empirical study of sentiment analysis for chinese documents. Expert Systems with Applications, 34, 2622-2629.

[27] Balahur, A. (2011) Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types. Department of Software and Computing Systems. Alacant, Univeristy of Alacant.

[28] Nigam, K., McCallum, A., Thrun, S., and Mitchell T. (2000). Text classification from labeled and unlabeled documents using EM. Machine Learning, 39(2/3):103–134.

[29] Manning, C.,D., Raghavan, P., Schütze, H. (2009).An Introduction to Information Retrieval. Cambridge University Press Cambridge, England

[30] David D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval Proceedings of ECML-98, 10th European Conference on Machine Learning, 1398, page 4--15.Chemnitz, DE, Springer Verlag, Heidelberg, DE, (1998)

[31] Darroch, J.N., and Ratcliff, D. (1972). Generalized Iterative Scaling for log-linear models. The annals of mathematical statistics.

[32] Berger, A. (1997). The Improved Iterative Scaling Algorithm: A gentle introduction, School of Compute Science, Carnegie Mellon University

[33] Goodman, J. (2002). Sequential conditional generalized iterative scaling. In ACL ’02.

[34] Chapelle, O.,Scholkopf, B., Zien, A. (2006). Semi- Supervised Learning, Adaptive computation and machine learning. MIT Press, Cambridge, Mass., USA

[35] Saif, H., Fernandez, M., He, Y. and Alani, H. (2013). Evaluation Datasets for Twitter Sentiment Analysis. In Proceedings of ESSEM

[36] Go, A., R. Bhayani, and L. Huang. (2009). Twitter sentiment classification using distant supervision. Technical report, Stanford Digital Library Technologies Project.

[37] Saif, H., He, Y., Alani, H. (2011). Semantic Smoothing for Twitter Sentiment Analysis. In:Proceeding of the 10th International Semantic Web Conference (ISWC)

[38] Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J. (2011). Twitter polarity classificationwith label propagation over lexical links and the follower graph. In:

Proceedings of the EMNLP First workshop on Unsupervised Learning in NLP. Edinburgh,Scotland

[39] Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., Varma, V. (2012). Miningsentiments from tweets. Proceedings of the WASSA 12

[40] Bravo-Marquez, F., Mendoza, M., Poblete, B. (2013). Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining.ACM

[41] Shamma, D. A., Kennedy, L., & Churchill, E. F. (2009). Tweet the debates: Understanding community annotation of uncollected sources. Proceedings of ACM Multimedia.

[42] Diakopoulos, N. and Shamma, D. A. (2010). Characterizing Debate Performance via Aggregated Twitter Sentiment. CHI'10, 1195-1198.

[43] Hu, X., Tang, L.,Tang, J., Liu, H.(2013). Exploiting social relations for sentiment analysis in microblogging. In: WSDM, pp 537-546

[44] Speriosu, M., Sudan, N., Upadhyay, S., and Baldridge, J. (2011). Twitter Polarity Classification with Label Propagation Over Lexical Links and the Follower Graph. Proceedings of the 1st Workshop on Unsupervised Learning in NLP, Edinburgh, 30 July 2011, pp. 53-63.

[45] Saif, H., Y. He, et al. (2012).Semantic sentiment analysis of twitter.Proceedings of the 11th international conference on The Semantic Web - Volume Part I. Boston, MA, Springer-Verlag: 508-524.

[46] Hu, X.; Tang, J.; Gao, H.; and Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. In Proceedings of the 22nd international conference on World Wide Web, WWW'13. ACM.

[47] Thelwall : Thelwall, M., Buckley, K., Paltoglou, G. (2012). Sentiment strength detection for the socialweb. Journal of the American Society for Information Science and Technology 63(1),163–173

[48] Bravo-Marquez, F., Mendoza, M., Poblete, B. (2013). Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining.ACM

[49] Liu, K.L., Li, W.J., Guo, M. (2012). Emoticon smoothed language models for twitter sentiment analysis. In: AAAI

[50] Deitrick, W., Hu, W. (2013). Mutually enhancing community detection and sentiment analysis on twitter networks. Journal of Data Analysis and Information Processing 1, 19{29

[51] Asiaee T, A., Tepper, M., Banerjee, A., Sapiro, G. (2012). If you are happy and you know it... tweet. In: Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 1602{1606. ACM

[52] Mohammad, S.M., Kiritchenko, S., Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In: In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), Atlanta, Georgia, USA, June 2013

[53] Chalothorn, T., Ellman, J. (2013). Tjp: Using twitter to analyze the polarity of contexts.In: In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), Atlanta, Georgia, USA.

[54] Martinez-Camara, E., Montejo-Raez, A., Martin-Valdivia, M., Urena-Lopez, L. (2013). Sinai: Machine learning and emotion of the crowd for sentiment analysis in microblogs.

[55] D-Remus, R. (2013). Asvunioeipzig.:Sentiment analysis in twitter using data-driven machine learning techniques.

[56] Andreevskaia, A. and Bergler, S. (2008). When specialists and generalists work together: overcoming domain dependence in sentiment tagging. In Proceedings of ACL-08: HLT

[57] Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classi- fication: A deep learning approach. In ICML, 513–520.

[58] Pyry Takala, Pekka Malo, Ankur Sinha, Oskar Ahlgren (2014). Gold-standard for Topic-specific Sentiment Analysis of Economic Texts. LREC 2014: 2152-2157

[59] Hatzivassiloglou, V., and McKeown, K. (1997). Predicting the semantic orientation of adjectives. In Proc. of the 35th ACL/8th EACL, pages 174–181

[60] Li, G., Liu, F. (2010) .A Clustering-based Approach on Sentiment Analysis. 978-1-4244-6793-8/10 ©2010 IEEE

[61] Saif, H., He, Y., Alani, H. (2012). Alleviating data sparsity for twitter sentiment analysis. In: Proceedings, 2nd Workshop on Making Sense of Microposts (#MSM2012) in conjunction with WWW 2012. Layon, France

[62] Wilks, Y., Stevenson, M. (1998). Word sense disambiguation using optimised combinations of knowledge sources, in: Proceedings of the 17th International

Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics. pp. 1398–1402.

[63] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. The Journal of machine learning research, 3, 1289-1305.

[64] Dey, S. (2013). Performance Investigation of Feature Selection Methods. arXiv preprint arXiv:1309.3949.

[65] Yang, Y., and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In ICML (Vol. 97, pp. 412-420).

[66] Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text. In Advances in Intelligent Data Analysis VI(pp. 121-132). Springer Berlin Heidelberg.

[67] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management,45(4), 427-437.

[68] Carstens, L. (2011).Sentiment Analysis- a multimodal approach. Imperial College London ,Department of Computing.

[69] Turney, P. D. (2002, July). 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 (pp. 417-424). Association for Computational Linguistics.