ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάλυση Συναισθημάτων σε Κείμενα Μικρού Μήκους

Κωνσταντίνος Γ. Μαραγκός

Επιβλέπων: Παναγιώτης Σταματόπουλος, Επίκουρος Καθηγητής

ΑΘΗΝΑ

ΟΚΤΩΒΡΙΟΣ, 2019

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάλυση Συναισθημάτων σε Κείμενα Μικρού Μήκους

Κωνσταντίνος Γ. Μαραγκός
Α.Μ.: 1115201400095

ΕΠΙΒΛΕΠΟΝΤΕΣ: Παναγιώτης Σταματόπουλος, Επίκουρος Καθηγητής

ΠΕΡΙΛΗΨΗ

Το θέμα αυτής της πτυχιακής εργασίας είναι η ανάλυση συναισθήματος σε κείμενα μικρού μήκους με την χρήση ακολουθιακών κανόνων κλάσεων, γνωστοί και ως Class Sequential Rules(CSR). Τα κείμενα μικρού μήκους έχουν κάποια ιδιαίτερα χαρακτηριστικά τα οποία εμποδίζουν παλαιότερες τεχνικές μηχανικής μάθησης να αποδώσουν αξιοπρεπώς. Για τον συγκεκριμένο λόγο παρουσιάζεται αυτή η τεχνική για την ανάλυση συναισθημάτων. Στόχος είναι η επεξεργασία των κειμένων, η εξαγωγή χαρακτηριστικών από αυτά και ταξινόμησή τους σε μία από τις κατηγορίες συναισθήματος. Τα συναισθήματα αυτά είναι η χαρά, η έκπληξη, η θλίψη, ο θυμός ή το κενό(κανένα συναίσθημα). Αρχικά έχουμε δύο σετ δεδομένων με πραγματικά παραδείγματα κείμενων μικρού μήκους - ένα κύριο και ένα βοηθητικό, τα οποία επεξεργαζόμαστε και βρίσκουμε τις προτάσεις κάθε κειμένου. Με βάση ένα λεξικό βρίσκονται τα δύο κυρίαρχα συναισθήματα για κάθε μια πρόταση κειμένου. Το λεξικό περιέχει χρήσιμες λέξεις μαζί με μια ετικέτα για κάθε κατηγορία συναισθήματος. Με βάση τα παραπάνω συναισθήματα, οι προτάσεις μετατρέπονται σε ακολουθίες συναισθημάτων. Έπειτα, με την χρήση του CSR παράγονται οι ακολουθιακοί κανόνες, οι οποίοι χρησιμοποιούνται για την δημιουργία των χαρακτηριστικών εκπαίδευσης ενός ταξινομητή. Αυτός εξάγει το τελικό συναίσθημα κάθε κειμένου. Τέλος υπολογίζεται το ποσοστό επιτυχίας του ταξινομητή.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Ανάλυση συναισθήματος

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: μηχανική μάθηση, κείμενα μικρού μήκους, διαδοχικοί κανόνες κλάσης, SVM

ABSTRACT

The present thesis deals with the sentiment analysis in microblog texts with the use of class sequential rules, known as CSR. Microblog texts have some special features that prevent older machine learning techniques from performing properly. For this reason, this technique of emotion classification is presented. The purpose is to edit texts, extract features from them and classify them into one of the categories of emotion. These feelings are joy, surprise, sadness, anger or emptiness (no emotion). Initially we have two datasets with real world examples of microblog texts – a training and a test set, which we process and find the sentences of each text. With the use of a lexicon, two dominant emotions are extracted for each text sentence. This lexicon contains words along with a label for each emotion category. Based on the above feelings, the sentences are converted into sequences of emotion labels. Then, using CSR produces the sequential rules, which are used to create a classifier's training features. This extracts the emotion of each text. Finally, the success rate of the classifier is calculated.

SUBJECT AREA: Sentiment analysis

KEYWORDS: machine learning, microblog texts, class sequential rules, SVM

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

1. ΕΙΣΑΓΩΓΗ

1.1 Ανάλυση Συναισθήματος

Σε αυτό το σημείο θα γίνει μια εισαγωγή των βασικών εννοιών της ανάλυσης συναισθήματος με σκοπό να γίνουν κατανοητές οι ενέργειες και τα βήματα που ακολουθούν οι προσεγγίσεις των επόμενων κεφαλαίων. Ακόμα αναφέρονται τα προβλήματα των προηγούμενων τεχνολογιών, που έγιναν λόγοι χρήσης των ακολουθιακών κανόνων κλάσεων(CSR) σε αυτόν τον τομέα.
Η ανάλυση συναισθημάτων, γνωστή και ως εξόρυξη γνώμης, αναφέρεται στη χρήση της επεξεργασίας φυσικής γλώσσας, της ανάλυσης κειμένου και της βιομετρίας για τον συστηματικό εντοπισμό, εξαγωγή και μελέτη των συναισθηματικών καταστάσεων. Είναι μια από τις πιο ενεργές και συνεχώς αναπτυσσόμενες περιοχές μελέτης της επεξεργασίας φυσικής γλώσσας (NLP). Παρόλο που η περιοχή έχει μεγάλη διάρκεια ιστορίας, λίγες έρευνες είχαν γίνει σχετικά με τις απόψεις και τα συναισθήματα των ανθρώπων πριν από το 2000. Υπάρχουν διάφοροι λόγοι για αυτό. Πρώτον, υπάρχει μεγάλο εύρος εφαρμογών σχεδόν σε κάθε τομέα και λόγω της τεράστιας χρήσης της από εμπορικές εφαρμογές, η βιομηχανία που την περιβάλλει ευδοκιμεί. Δεύτερον προσφέρει πολλά ερευνητικά προβλήματα, τα οποία δεν είχαν μελετηθεί ποτέ πριν. Τέλος, με την εξέλιξη του διαδικτύου και των μέσων κοινωνικής δικτύωσης, όπως το Twitter και το Facebook, έχουμε για πρώτη φορά έναν τεράστιο όγκο δεδομένων.

1.2 Επίπεδα Ανάλυσης

Η ανάλυση συναισθήματος έχει ερευνηθεί κυρίως σε τρία επίπεδα. Το πρώτο ονομάζεται επίπεδο εγγράφων. Ο ρόλος αυτού του επιπέδου είναι να αποφανθεί κατά πόσο μια συνολική γνώμη του εγγράφου εκφράζει θετικό ή αρνητικό συναίσθημα. Για παράδειγμα, έχοντας μία κριτική ενός προϊόντος, το σύστημα θα αποφασίσει εάν η κριτική εκφράζει συνολικά μια θετική ή αρνητική γνώμη για το προϊόν. Αυτό μας δείχνει πως το συγκεκριμένο επίπεδο θεωρεί ότι κάθε έγγραφο εκφράζει μια γνώμη για μια μοναδική οντότητα, πράγμα που το καθιστά μη αποδεκτό για έγγραφα που κρίνουν πολλές οντότητες.
Το δεύτερο επίπεδο ανάλυσης συναισθήματος ονομάζεται προτασιακό. Αυτή τη φορά, το επίπεδο κοιτάει εάν κάποια πρόταση εκφράζει θετικό, αρνητικό ή ουδέτερο συναίσθημα. Ουδέτερο συναίσθημα έχουμε όταν η πρόταση δεν εκφράζει κάποια γνώμη. Το επίπεδο αυτό διακρίνει τις προτάσεις που εκφράζουν πραγματικές πληροφορίες, από προτάσεις που εκφράζουν υποκειμενικές απόψεις. Ωστόσο, πρέπει να σημειώσουμε ότι η υποκειμενικότητα δεν είναι ισοδύναμη με το συναίσθημα καθώς πολλές αντικειμενικές προτάσεις μπορούν να υποδηλώσουν κάποια γνώμη, παραδείγματος χάριν «Αγοράσαμε το αμάξι τον προηγούμενο μήνα και χάλασαν οι υαλοκαθαριστήρες». Τέλος, παρά τις προσπάθειες που έχουν γίνει στην ανάλυση σύνθετων φράσεων, το αποτέλεσμα δεν είναι ικανοποιητικό, για παράδειγμα «Η Apple είναι πολύ καλά σε αυτήν την κακή οικονομία».
Το επίπεδο οντότητας και άποψης είναι το τρίτο και τελευταίο επίπεδο ανάλυσης συναισθήματος. Τα δύο προηγούμενα επίπεδα δεν ανακαλύπτουν ακριβώς τι αρέσει και τι όχι στους ανθρώπους. Το επίπεδο αυτό εκτελεί μια ανάλυση υψηλής λεπτομέρειας. Αντί να λαμβάνει υπόψιν την δομή της γλώσσας, όπως έγγραφα, παραγράφους, προτάσεις, φράσεις ή λέξεις, το επίπεδο οντότητας κοιτάει απευθείας στην γνώμη. Βασίζεται δηλαδή στην ιδέα ότι η ανθρώπινη γνώμη αποτελείται από ένα συναίσθημα και ένα στόχο. Μια γνώμη χωρίς στόχο δεν έχει νόημα. Η κατανόηση αυτής της ιδέας μας βοηθάει να καταλάβουμε καλύτερα το πρόβλημα της ανάλυσης συναισθήματος. Για παράδειγμα, η πρόταση «Παρά την κακή εξυπηρέτηση του προσωπικού, το εστιατόριο είναι ένα από τα

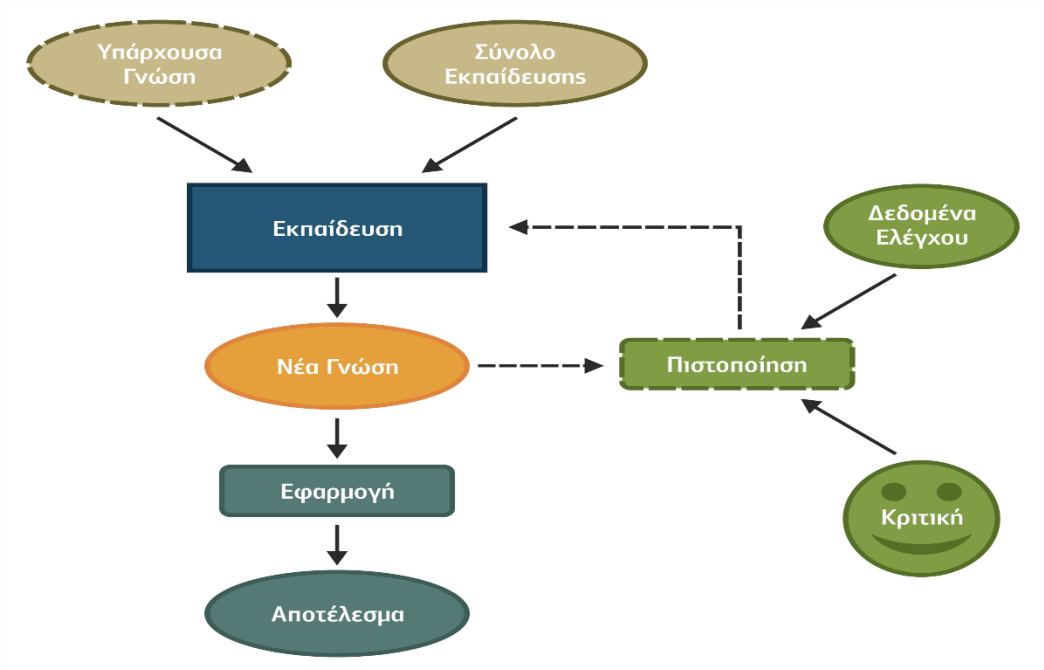
αγαπημένα μου» εκφράζει ένα θετικό ύφος, όμως η γενική γνώμη δεν είναι απόλυτα θετική. Ενώ τα προηγούμενα επίπεδα θα ήταν ικανοποιημένα με ένα θετικό συναίσθημα, το επίπεδο οντότητας ανακαλύπτει συναισθήματα με βάση τις ομάδες ή οντότητες του κειμένου. Πρακτικά, στην παραπάνω πρόταση, αναλύονται δύο οντότητες (εστιατόριο, προσωπικό), με το συναίσθημα του εστιατορίου να είναι θετικό, ενώ του προσωπικού αρνητικό. Το εστιατόριο και το προσωπικό είναι και οι στόχοι της γνώμης. Με βάση αυτή την ανάλυση καταλήγουμε σε δομές από γνώμες οντοτήτων, οι οποίες μετατρέπουν τα αδόμητα κείμενα σε δομημένα δεδομένα τα οποία μπορούν να χρησιμοποιηθούν από διάφορους αλγόριθμους ποσοτικής και ποιοτικής ανάλυσης.
Βλέπουμε πως και τα τρία επίπεδα είναι αρκετά περίπλοκα, καθώς αποτελούνται από πολλά υπο-προβλήματα. Η προσέγγιση της πτυχιακής αυτής βασίζεται πάνω στο πρώτο επίπεδο.

2. ΣΥΝΑΦΕΙΣ ΕΡΕΥΝΕΣ

Σε αυτό το κεφάλαιο αναλύονται τα αρχικά στάδια της συναισθηματικής ανάλυσης, δηλαδή ποιες έρευνες έγιναν και ποιους τομείς κάλυπταν. Ο στόχος είναι η παρουσίαση των προϋπαρχουσών τεχνολογιών που χρησιμοποιήθηκαν και εξελίχθηκαν από την προτεινόμενη προσέγγιση, καθώς και ο λόγος για τον οποίο αναπτύχθηκε αυτή.

2.1 Προσέγγιση Μηχανικής Μάθησης

Από το 2000 και μετά, η ανάλυση συναισθήματος είναι ένας τομέας συνεχούς εξέλιξης. Προηγούμενες μελέτες ερευνούσαν κυρίως αξιολογήσεις προϊόντων, παροχών υπηρεσιών και άλλων οντοτήτων, με συζητήσεις από ιστοσελίδες - φόρουμ καθώς και με τα κείμενα από ιστολόγια (blogs).
Υπάρχουν διάφοροι τρόποι προσέγγισης για την εύρεση συναισθήματος στην φυσική γλώσσα. Ένας πρώτος τρόπος που χρησιμοποιείται μέχρι σήμερα, είναι αυτός της μηχανικής μάθησης.



Εικόνα 2.1: Διάγραμμα ροής της μηχανικής μάθησης

Μια πρώτη μελέτη βασισμένη σε τεχνικές μηχανικής μάθησης είναι αυτή των **Bo Pang και Lillian Lee, το 2002[1]**. Αυτή η έρευνα μελετά το πρόβλημα της ταξινόμησης εγγράφων, όχι με βάση το θέμα, αλλά με το γενικό συναίσθημα, καθορίζοντας μια κριτική αν είναι θετική ή αρνητική. Μέχρι τότε προσπάθειες στόχευαν στην κατηγοριοποίηση με βάση το θέμα της κριτικής, παραδείγματος χάριν αθλητικά, πολιτική, οικονομία. Όμως κάποια θέματα είχαν ανάγκη την αξιολόγηση με βάση το συναίσθημα, δηλαδή αν μια κριτική για κάποια ταινία είναι θετική ή αρνητική. Η επισήμανση αυτών των άρθρων με το συναίσθημα θα μπορεί να παρέχει συνοπτικές περιλήψεις στους αναγνώστες.

2.1.1 Σετ Δεδομένων

Το πρώτο στάδιο για κάθε μέθοδο ανάλυσης συναισθήματος είναι η συλλογή ενός σετ δεδομένων (dataset). Αυτά περιέχουν πραγματικά παραδείγματα ανάλογα του τομέα μελέτης, όπως προτάσεις, έγγραφα κειμένου, κριτικές προϊόντων ή ακόμα και εικόνες ή βίντεο, και χρησιμοποιούνται ως δεδομένα για την εφαρμογή και τον έλεγχο των τεχνολογιών. Το μέγεθος τους είναι αρκετά μεγάλο, από μερικές χιλιάδες έως εκατομμύρια παραδείγματα, καθώς έχει παρατηρηθεί ότι η απόδοση της μεθόδου είναι ανάλογη του όγκου των δεδομένων. Ακόμα, ένα σετ δεδομένων περιέχει για κάθε ένα παράδειγμα του και μια ετικέτα (label), η οποία έχει ανατεθεί συνήθως από κάποιον άνθρωπο (μπορούν να παραχθούν και αυτόματα) και αντιπροσωπεύει το παράδειγμα. Στην ανάλυση συναισθήματος, αυτές οι ετικέτες είναι το γενικό συναίσθημα του παραδείγματος. Η διαδικασία όπου ένας «γνώστης» παρέχει στον υπολογιστή κάποια παραδείγματα ως είσοδο και κάποιες επιθυμητές εξόδους, και έπητα παράγετε ένας γενικός κανόνας που χαρτογραφεί εισόδους σε εξόδους, ονομάζεται εποπτευόμενη μάθηση. Όταν δεν δίνονται ετικέτες, αφήνοντας τον αλγόριθμο να βρει μια δομή στα παραδείγματα, τότε η διαδικασία ονομάζεται μη εποπτευόμενη μάθηση.
Στην μέθοδο που αναφέραμε προηγουμένως **[1],** το σετ δεδομένων αποτελείται από κριτικές ταινιών. Αυτός ο τομέας είναι αρκετά βολικός επειδή υπάρχουν μεγάλες συλλογές δεδομένων στο διαδίκτυο. Ταυτόχρονα, οι διαφορετικοί χρήστες αξιολογούν τις ταινίες, με αποτέλεσμα οι ετικέτες να υπάρχουν ήδη. Σε διαφορετική περίπτωση, η προγραμματιστές θα έπρεπε να επισημάνουν τις ετικέτες χειροκίνητα για κάθε ένα παράδειγμα, πράγμα που απαιτεί πολύ χρόνο.

2.1.2 Προεπεξεργασία Κειμένου και Εξαγωγή Χαρακτηριστικών

Το δεύτερο κομμάτι στην ανάλυση συναισθήματος είναι η προεπεξεργασία των δεδομένων, οι εργασίες δηλαδή προετοιμασίας των δεδομένων. Πολλές φορές τα δεδομένα έχουν πολλά προβλήματα. Κάποια από αυτά είναι η ύπαρξη αλληλοσυγκρουόμενων πληροφοριών ή διπλότυπων, η ύπαρξη ασυνεπειών ως προς την κωδικοποίηση, την ονοματοδοσία πεδίων και τις μονάδες μέτρησης, καθώς και η ύπαρξη χαμένων τιμών και θορύβου, τυχαία δηλαδή κυμαινόμενων δεδομένων χωρίς ουσιαστικό περιεχόμενο. Η προεπεξεργασία των δεδομένων περιλαμβάνει τον καθαρισμό τους, αλλά δεν περιορίζεται σε αυτόν. Ειδικές απαιτήσεις των μεθόδων επεξεργασίας συχνά επιβάλλουν τον μετασχηματισμό των δεδομένων. Δύο συνήθεις εργασίες μετασχηματισμού είναι η διακριτοποίηση και η κανονικοποίηση. Ο όρος διακριτοποίης αναφέρεται στον μετασχηματισμό αριθμητικών τιμών σε ονομαστικές τιμές. Η κανονικοποίηση είναι η μετατροπή τιμών σε άλλες, πιο «κατάλληλες». Η τελευταία είναι αυτή που χρησιμοποιείται στην ανάλυση συναισθήματος. Ένα επιπλέον θετικό στοιχείο της προεπεξεργασίας των δεδομένων είναι η μείωση του όγκου τους. Συγκεκριμένα, επιλέγονται οι μεταβλητές που είναι απαραίτητες για την εξόρυξη του συναισθήματος. Η έρευνα που προαναφέρθηκε ακολουθεί την εξής λογική για την επιλογή των κατάλληλων μεταβλητών. Η διάκριση θετικών από αρνητικών κριτικών είναι σχετικά εύκολη για τον άνθρωπο, ειδικά για το πρόβλημα της κατηγοριοποίησης κριτικών, καθώς τα θέματα μπορεί να σχετίζονται στενά. Ακόμα υπάρχουν ορισμένες λέξεις που χρησιμοποιούν οι άνθρωποι για να εκφράσουν έντονα συναισθήματα. Με βάση αυτούς τους λόγους, ζητήθηκε από δύο ανθρώπους να επιλέξουν καλές λέξεις-δείκτες για θετικά και αρνητικά συναισθήματα σε κριτικές ταινιών, με σκοπό την δημιουργία μιας λίστας με τις πιο κατάλληλες λέξεις η οποία θα χρησιμοποιηθεί για την ταξινόμηση των κειμένων.
Έπειτα από την προεπεξεργασία του κειμένου ακολουθεί η εξαγωγή των χαρακτηριστικών (feature extraction). Καταρχάς, ένα χαρακτηριστικό είναι μια μεμονωμένη μετρήσιμη ιδιότητα

και συνήθως είναι αριθμητικές τιμές. Σε ένα σετ δεδομένων υπάρχουν πολλά παραδείγματα κειμένου με διαφορετικές λέξεις και διαφορετικά μήκη. Μέσω διαφόρων αλγορίθμων επιτυγχάνεται η εξαγωγή χαρακτηριστικών. Η ανάπτυξη συστημάτων για να γίνει κάτι τέτοιο είναι γνωστή ως τεχνολογία χαρακτηριστικών. Απαιτεί τον πειραματισμό πολλαπλών δυνατοτήτων και τον συνδυασμό αυτοματοποιημένων τεχνικών με τη διαίσθηση και τη γνώση του ειδικού τομέα . Η αυτοματοποίηση αυτής της διαδικασίας είναι η εκμάθηση χαρακτηριστικών , όπου μια μηχανή όχι μόνο χρησιμοποιεί δυνατότητες για μάθηση, αλλά μαθαίνει τα ίδια τα χαρακτηριστικά. Τα εξαγόμενα features τροφοδοτούν τους αλγόριθμους της μηχανικής μάθησης για την εκμάθηση μοτίβων που θα μπορούν να εφαρμοστούν σε καινούργια δεδομένα για την απόκτηση αποτελέσματος. Οι αλγόριθμοι αυτοί, περιμένουν features με την μορφή αριθμητικών διανυσμάτων επειδή κάθε αλγόριθμος είναι, κατά κύριο λόγο, μια μαθηματική λειτουργία βελτιστοποίησης και ελαχιστοποίησης απώλειας και λάθους όταν κάνει την προσπάθεια εκμάθησης μοτίβων από σημεία δεδομένων και παρατηρήσεων. Για αυτό τον λόγο, στα δεδομένα κειμένων υπάρχει μια επιπλέον δυσκολία στον μετασχηματισμό των δεδομένων και στην εξαγωγή αριθμητικών χαρακτηριστικών από αυτό. Κάποιες τεχνικές εξαγωγής δεδομένων είναι ονομαστικά το μοντέλο Vector Space, TF-IDF, NLP, Bag of Words. Το τελευταίο χρησιμοποιείται και από την έρευνα που προαναφέραμε (Pang and Lee, 2002). Στην κύρια μέθοδο που θα αναλύσουμε σε επόμενο κεφάλαιο, ο αλγόριθμος CSR είναι αυτός που εξάγει τα δεδομένα.

Πολλές φορές υπάρχει η πεποίθηση πως όσο περισσότερα τα χαρακτηριστικά τόσο καλύτερο και το αποτέλεσμα. Αυτό όμως είναι λάθος. Ο αριθμός των χαρακτηριστικών μεταβάλλεται ανάλογα το πρόβλημα. Γενικά δεν πρέπει να είναι ούτε πολλά ούτε και λίγα. Σε αντίθετες περιπτώσεις εμφανίζονται δύο φαινόμενα, η υπερπροσαρμογή και η υποπροσαρμογη (Overfitting - Underfitting).

Η υπερπροσαρμογή λαμβάνει χώρα όταν ένα στατιστικό μοντέλο ή ένας αλγόριθμος εκμάθησης μηχανής καταγράφει το θόρυβο των δεδομένων. Διαισθητικά, η υπερπροσαρμογή συμβαίνει όταν το μοντέλο ή ο αλγόριθμος ταιριάζει πολύ καλά στα δεδομένα. Συγκεκριμένα, το μοντέλο ή ο αλγόριθμος εμφανίζει χαμηλή μεροληψία (low bias) αλλά μεγάλη διακύμανση (high variance). Η υπερπροσαρμογή είναι συχνά αποτέλεσμα υπερβολικά περίπλοκου μοντέλου και μπορεί να προληφθεί με την τοποθέτηση πολλαπλών μοντέλων και με τη χρήση επικύρωσης ή διασταυρούμενης επικύρωσης για να συγκριθούν η προβλεπτική ακρίβειά τους στα δεδομένα δοκιμών.

Το αντίθετο είναι το φαινόμενο της υποπροσαρμογής. Η υποπροσαρμογή συμβαίνει όταν ένα στατιστικό μοντέλο ή ο αλγόριθμος εκμάθησης δεν μπορεί να καταγράψει την υποκείμενη τάση των δεδομένων. Γενικά συμβαίνει όταν το μοντέλο ή ο αλγόριθμος δεν ταιριάζει αρκετά καλά στα δεδομένα. Συγκεκριμένα, συμβαίνει εάν το μοντέλο ή ο αλγόριθμος δείχνει χαμηλή διακύμανση αλλά μεγάλη μεροληψία. Η υποπροσαρμογή είναι συχνά αποτέλεσμα υπερβολικά απλού μοντέλου.

Και τα δύο παραπάνω φαινόμενα οδηγούν σε κακές προβλέψεις σε νέα σύνολα δεδομένων, αν και το φαινόμενο της υποπροσαρμογής δεν εμφανίζεται τόσο συχνά στην πράξη.

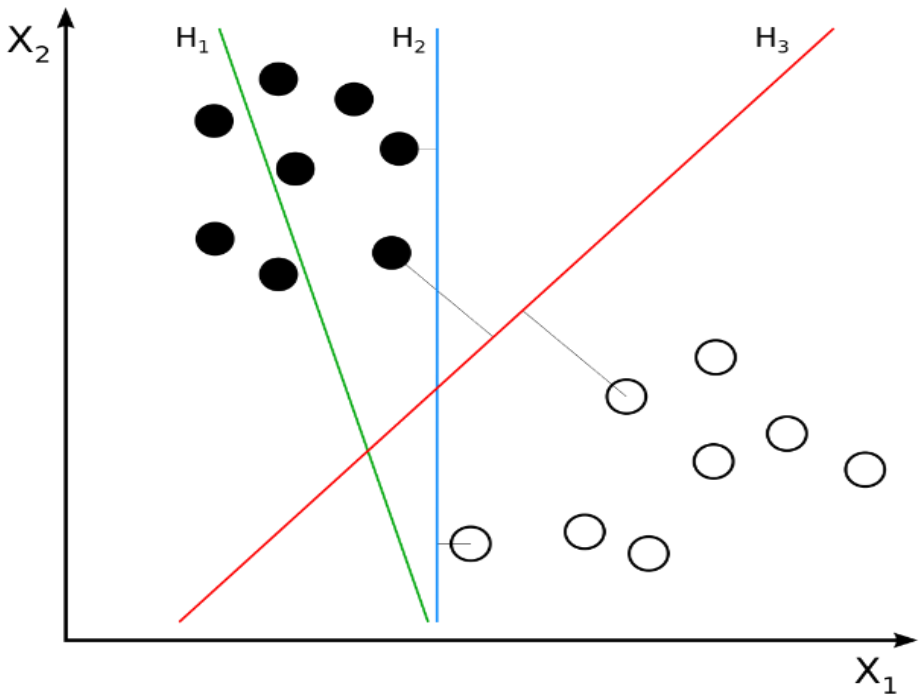
2.1.3 Αλγόριθμοι Μηχανικής Μάθησης

Από την στιγμή που θα παραχθούν αυτά τα features, επιλέγεται ένας αλγόριθμος Μηχανικής Μάθησης για την εκπαίδευση του μοντέλου. Η εκπαίδευση του μοντέλου περιλαμβάνει την τροφοδότηση των διανυσμάτων με τα χαρακτηριστικά-features των εγγράφων και τις αντίστοιχες ετικέτες-labels, ώστε ο αλγόριθμος να μπορέσει να «μάθει» διάφορα μοτίβα για κάθε μια κλάση και να μπορέσει ξαναχρησιμοποιήσει αυτήν την γνώση για να προβλέψει τις κλάσεις νέων δεδομένων. Με τον όρο κλάσεις εννοούμε τα διαφορετικά αποτελέσματα-στόχους, για παράδειγμα στην ανάλυση συναισθήματος οι διαφορετικές κλάσεις είναι τα διαφορετικά συναισθήματα που θέλουμε να εξάγουμε. Η έρευνα που προαναφέρθηκε **[1]** είναι ένα πρόβλημα δύο κλάσεων, καθώς προβλέπει μόνο δύο συναισθήματα, θετικό ή

αρνητικό. Αργότερα θα μελετήσουμε έρευνες που χρησιμοποιούν πολλές περισσότερες κλάσεις, με την κύρια έρευνα της πτυχιακής αυτής να προβλέπει έξι συναισθήματα.
Υπάρχουν πολλοί τύποι αλγορίθμων ταξινόμησης. Ένας από αυτούς είναι και τα νευρωνικά δίκτυα (Neural Networks), τα οποία τα τελευταία χρόνια έχουν δεχθεί μεγάλη αναγνώριση καθώς είναι πολύ χρήσιμα. Στην έρευνα (Pang and Lee, 2002), ελέγχεται η απόδοση τριών διαφορετικών ταξινομητών, του “Naive Bayes”, της μέγιστης εντροπίας (Maximum Entropy) και του “Support Vector Machines” (SVM). Στόχος σε αυτό το έργο ήταν να εξεταστεί κατά πόσο αρκεί η αντιμετώπιση της κατάταξης των συναισθημάτων απλά ως μια ειδική περίπτωση κατηγοριοποίησης βάσει θέματος ή εάν θα έπρεπε να αναπτυχθούν καινούριοι αλγόριθμοι. Και οι τρεις αλγόριθμοι ακολουθούν διαφορετικές φιλοσοφίες, όμως πολλές μελέτες του τομέα αποδεικνύουν πως ο SVM έχει την καλύτερη απόδοση.

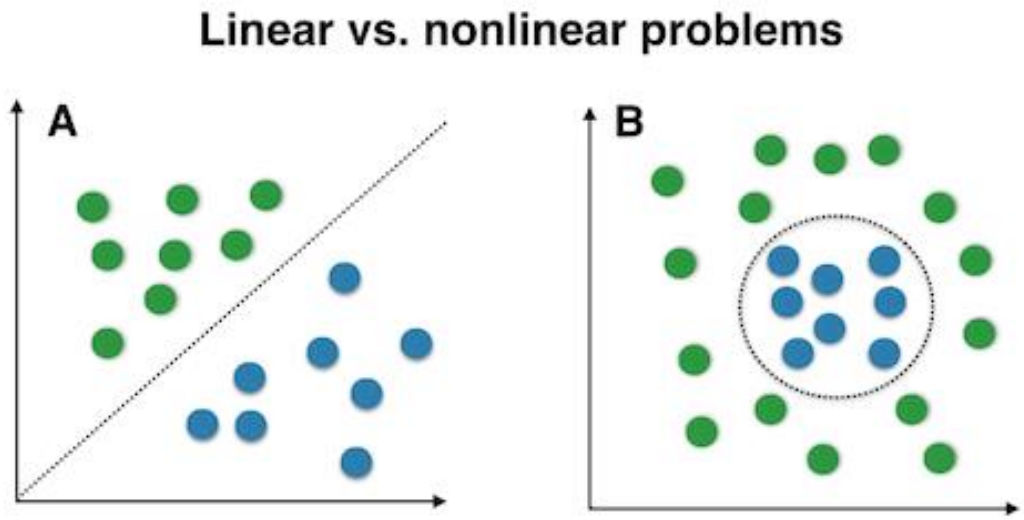
2.1.4 Ο αλγόριθμος SVM

Σε αυτό το σημείο θα αναλυθεί ο αλγόριθμος SVM, καθώς χρησιμοποιείται και από το σύστημα που παρουσιάζει η πτυχιακή αυτή.
Τα SVM, μηχανές υποστήριξης διανυσμάτων, είναι υπό εποπτεία μαθησιακά μοντέλα με συναφείς αλγορίθμους εκμάθησης που αναλύουν
δεδομένα που χρησιμοποιούνται για την ανάλυση ταξινόμησης και παλινδρόμησης. Σε ένα πρόβλημα δύο κλάσεων, υπάρχουν σημεία δεδομένων που ανήκουν σε μία από τις κλάσεις. Στόχος είναι να αποφασιστεί σε ποια κατηγορία θα ανήκει ένα νέο σημείο δεδομένων. Ένα τέτοιο σημείο θεωρείται ένα n-διάστατο διάνυσμα (λίστα με n αριθμούς). Πιο τυπικά, μια μηχανή υποστήριξης διανυσμάτων κατασκευάζει ένα υπερπλαίσιο (hyperplane) ή ένα σύνολο υπερπλαιδίων σε ένα χώρο υψηλής ή απεριόριστης διαστάστασης, ο οποίος μπορεί να χρησιμοποιηθεί για ταξινόμηση , παλινδρόμηση ή άλλα καθήκοντα όπως η ανίχνευση των ακραίων τιμών. Διαισθητικά, ένας καλός διαχωρισμός επιτυγχάνεται με το υπερεπίπεδο που έχει τη μεγαλύτερη απόσταση από το πλησιέστερο σημείο δεδομένων οποιασδήποτε κατηγορίας (λέγεται λειτουργικό περιθώριο), δεδομένου ότι σε γενικές γραμμές όσο μεγαλύτερο είναι το περιθώριο, τόσο χαμηλότερη είναι η περίπτωση λάθους του ταξινομητή.



Σχήμα 2.1: Οι H1, H2 και H3 είναι τρία πιθανά hyperplanes του προβλήματος δύο διαφορετικών κλάσεων, μαύρες – άσπρες κουκίδες

Στο σχήμα 2.1 το H1 δεν διαχωρίζει τις κλάσεις. Το H2 το επιτυγχάνει αυτό αλλά μόνο με ένα μικρό περιθώριο. Το H3 διαχωρίζει τις κλάσεις με το μέγιστο περιθώριο. Ο SVM προσπαθεί να βρεί αυτό το υπερπλαίσιο.
Ενώ το αρχικό πρόβλημα μπορεί να αναφερθεί σε μια πεπερασμένη διάσταση, συχνά συμβαίνει ότι τα σύνολα που διακρίνονται δεν είναι γραμμικά διαχωρίσιμα σε αυτό το χώρο. Για το λόγο αυτό, προτάθηκε ότι ο αρχικός χώρος πεπερασμένων διαστάσεων έχει χαρτογραφηθεί σε ένα χώρο πολύ υψηλότερων διαστάσεων, πιθανώς καθιστώντας τον διαχωρισμό ευκολότερο σε αυτόν τον χώρο. Για να διατηρηθεί εύλογο το υπολογιστικό φορτίο, οι χαρτογραφήσεις που χρησιμοποιούνται από τα σχήματα SVM σχεδιάζονται έτσι ώστε να εξασφαλίζουν ότι τα προϊόντα ζευγών δεδομένων εισόδου μπορούν να υπολογιστούν εύκολα από την άποψη των μεταβλητών στον αρχικό χώρο, καθορίζοντάς τα με κάποια λειτουργία πυρήνα (kernel), επιλεγμένη για να ταιριάζει στο πρόβλημα. Υπάρχουν «γραμμικά και μη-γραμμικά kernels».



Σχήμα 2.2: Γραμμικό και μη-γραμμικο παράδειγμα hyperplane

Κάποια μη-γραμμικά kernels είναι ονοματικά το πολυωνυμικό, η υπερβολική εφαπτομένη και η λειτουργία ακτινικής βάσης «Gaussian».
Μετά την εκμάθηση του ταξινομητή μπορεί να γίνει πρόβλεψη καινούργιων δεδομένων. Πάνω σε αυτά τα αποτελέσματα γίνονται διάφορες παρατηρήσεις σχετικά με την απόδοση του αλγορίθμου. Αυτό το στάδιο ονομάζεται έλεγχος αποτελεσμάτων και είναι πολύ σημαντικό καθώς βοηθάει στην περαιτέρω κατανόηση του προβλήματος, παρουσιάζοντας ποσοστά λάθους και απόκλισης από το επιθυμητό αποτέλεσμα. Με βάση αυτά μπορεί να γίνουν αλλαγές στα χαρακτηριστικά-features ή στον αλγόριθμο ταξινόμησης, παραδείγματος χάριν να χρησιμοποιηθεί διαφορετικός πυρήνας, και να επαναληφθεί η εκπαίδευση του ταξινομητή μέχρι να επιτευχθεί το επιθυμητό αποτέλεσμα.

2.2 Προσέγγιση Λεξικού

Μια δεύτερη προσέγγιση ανάλυσης συναισθήματος είναι αυτή του λεξικού. Η προσέγγιση με βάση το λεξικό περιλαμβάνει τον υπολογισμό του προσανατολισμού ενός εγγράφου από τον σημασιολογικό προσανατολισμό των λέξεων ή των φράσεων στο έγγραφο. Ονομάζεται έτσι γιατί ολόκληρη η μέθοδος στηρίζεται πάνω σε λεξικά λέξεων σχολιασμένων με τον σημασιολογικό προσανατολισμό της λέξης ή την πολικότητα τους. Αυτά τα λεξικά μπορούν να δημιουργηθούν χειροκίνητα ή και αυτόματα. Η πρώτη περίπτωση προϋποθέτει πως μια ομάδα ανθρώπων θα συλλέξει ένα ικανοποιητικό πλήθος λέξεων που καλύπτουν και εκφράζουν τις ανάγκες του προβλήματος καλύτερα, καθώς και την ανάθεση της ετικέτας

κάθε λέξης. Η αυτόματη δημιουργία ενός λεξικού επιτυγχάνεται χρησιμοποιώντας λέξεις σπόρων για να επεκτείνουν τη λίστα λέξεων.
Ακόμα, μεγάλο μέρος της έρευνας που βασίζεται σε λεξικά έχει επικεντρωθεί στη χρήση επίθετων ως δείκτες του σημασιολογικού προσανατολισμού του κειμένου. Πρώτον, μια λίστα των επίθετων και των αντίστοιχων τιμών SO (Semantic Orientation) καταρτίζεται σε ένα λεξικό. Στη συνέχεια, για κάθε δεδομένο κείμενο, όλα τα επίθετα εξάγονται και σχολιάζονται με την τιμή SO τους, χρησιμοποιώντας τα λεξικά. Οι βαθμολογίες SO με τη σειρά τους συγκεντρώνονται σε ένα ενιαίο σκορ για κάθε κείμενο.
Αυτή η προσέγγιση είναι σχετικά πιο εύκολη από αυτή της μηχανικής μάθησης, αλλά είναι και λιγότερο αποδοτική. Ο συνδυασμός και των δύο επιφέρει αρκετά καλά αποτελέσματα και έχει χρησιμοποιείθει και από το σύστημα αυτής της πτυχιακής.

2.3 Πολλαπλές Τάξεις Πολυπλοκότητας

Εκτός από την ταξινόμηση ενός κειμένου απλά ως θετικό ή αρνητικό, μερικές μελέτες αποσκοπούν στον εντοπισμό του συναισθήματος ενός κειμένου, όπως ο θυμός, η ευτυχία, η λύπη και άλλα διάφορα συναισθήματα. Στην έρευνα του **(Mishne, 2005)[2]** χρησιμοποιείται ο SVM για να εκπαιδεύσει έναν ταξινομητή συναισθημάτων με 132 διαθέσεις πάνω σε αναρτήσεις ιστολογίου (blog posts).
Η ταξινόμηση της διάθεσης είναι χρήσιμη για διάφορες εφαρμογές, όπως να βελτιώσει την αλληλεπίδραση μεταξύ γιατρού και ασθενή. Στη περίπτωση, των αναρτήσεων ιστολογίου (και άλλων μεγάλων ποσών υποκειμενικών δεδομένων), μπορεί επίσης να επιτρέψει νέες προσεγγίσεις της ταξινόμησης κειμένου, παραδείγματος χάριν, το φιλτραρίσματα αποτελεσμάτων αναζήτησης με βάση τη διάθεση, τον εντοπισμό κοινοτήτων και τη συσσωμάτωση (clustering).
Συνεχίζοντας με την έρευνα του **(Mishne, 2005)[2]**, τα δεδομένα που έχουν αποκτήσει αποτελούνται από μια μεγάλη συλλογή από αναρτήσεις ιστολογίου - καταχωρήσεις ηλεκτρονικών ημερολογίων - οι οποίες περιλαμβάνουν ενδείξεις της διάθεσης του συγγραφέα. Είναι ένα ιδιαίτερα ενδιαφέρον έργο γιατί προσφέρει επίσης έναν αριθμό νέων επιστημονικών προκλήσεων. Πρώτον, οι συγγραφείς δημιουργούν μια μυριάδα διαφορετικών μορφών και ορισμών διαθέσεων. Έτσι, ο εντοπισμός χαρακτηριστικών που είναι συνεπής μεταξύ των συγγραφέων είναι μια πολύπλοκη εργασία. Επιπλέον, το μικρό μήκος των καταχωρήσεων του ιστολογίου, θέτει μια πρόκληση στις μεθόδους ταξινόμησης που βασίζονται σε στατιστικά στοιχεία από ένα μεγάλο κείμενο. Βάση των αποτελεσμάτων της έρευνας, οι συγγραφείς της συμπεραίνουν ότι η περαιτέρω αύξηση του αριθμού των διαθέσιμων δεδομένων εκπαίδευσης θα οδηγήσει σε μια πρόσθετη αύξηση της ακρίβειας. Επιπλέον, αποδεικνύουν ότι, η ακρίβεια ταξινόμησης, αν και χαμηλή, δεν είναι ουσιαστικά χειρότερη από την ανθρώπινη απόδοση για το ίδιο έργο. Η κύρια τους διαπίστωση είναι ότι η ταξινόμηση της διάθεσης είναι δύσκολη υπόθεση με βάση τις τρέχουσες μεθόδους ανάλυσης κειμένου.
Παρατηρούμε ότι όσες περισσότερες κλάσεις υπάρχουν, τόσο πιο δύσκολο θα είναι το έργο του υπολογισμού, δηλαδή η εύρεση μιας γενικής δομής που θα προβλέπει πολλές διαφορετικές περιπτώσεις.

2.4 Κείμενα Μικρού Μήκους

Με την πρόοδο των μέσων κοινωνικής δικτύωσης, ένα μεγάλο μέρος ερευνών ερευνά την ανάλυση συναισθήματος σε κείμενα μικρού μήκους. Αυτά τα κείμενα είναι ένα δύσκολο κομμάτι μελέτης, καθώς η μορφή τους διαφέρει αρκετά από αυτήν των εγγράφων, άρθρων ή κριτικών. Παραδοσιακά, το μεγαλύτερο μέρος μελέτης έχει επικεντρωθεί ταξινομώντας μεγαλύτερα κομμάτια κειμένου. Κάποιες πρώτες έρευνες έγιναν πάνω σε δεδομένα του Twitter **(Go, 2009)[3]**. Το Twitter είναι μια δημοφιλής υπηρεσία microblogging όπου οι

χρήστες δημιουργούν μηνύματα κατάστασης (που ονομάζονται "tweets"). Αυτά τα tweets εκφράζουν μερικές φορές απόψεις για διαφορετικά θέματα. Τα Tweets, και γενικά τα κείμενα μικρού μήκους (microblog text) διαφέρουν από τις αξιολογήσεις κυρίως λόγω του σκοπού τους. Ενώ οι κριτικές αντιπροσωπεύουν συνοπτικές σκέψεις συγγραφέων, τα tweets είναι πιο περιστασιακά και περιορίζονται σε λίγους χαρακτήρες κειμένου (140-200 χαρακτήρες). Γενικά, τα tweets δεν είναι γραμμένα με κάποιο σοβαρό ύφος όπως οι κριτικές, και πολλές φορές δεν φέρουν κάποιο συναίσθημα. Ωστόσο, εξακολουθούν να προσφέρουν στις εταιρείες μια πρόσθετη λεωφόρο για τη συλλογή ανατροφοδοτήσεων.
Για να γίνει πιο κατανοητή η μορφή των κειμένων μικρού μήκους, οι διαφορές τους από προηγούμενους τομείς μελέτης, καθώς και οι ευκολίες και δυσκολίες που υπάρχουν στην μελέτη τους, θα αναλυθούν κάποια από τα μοναδικά χαρακτηριστικά τους.
Το πρώτο χαρακτηριστικό είναι το σαφέστερα μικρότερο μήκος τους. Ένα μέσο μέγεθος κειμένου microblog κυμαίνεται στις δέκα με είκοσι λέξεις. Αυτό είναι πολύ διαφορετικό από τις προηγούμενες έρευνες που επικεντρώνονται στην ταξινόμηση μακρύτερων σωμάτων εγγράφων, όπως κριτικές ταινιών.
Μια άλλη διαφορά είναι το μέγεθος των διαθέσιμων δεδομένων. Για παράδειγμα, με το Twitter, είναι πολύ εύκολη η συλλογή εκατομμυρίων tweets για εκπαίδευση. Σε προηγούμενες έρευνες, τα σετ δεδομένων αποτελούνταν μόνο από χιλιάδες αντικείμενα κατάρτισης.
Η τρίτη διαφορά είναι το γλωσσικό μοντέλο που χρησιμοποιείται για την σύνταξη των κειμένων. Οι χρήστες των κοινωνικών δικτύων δημοσιεύουν μηνύματα από πολλά διαφορετικά μέσα, συμπεριλαμβανομένων των κινητών τηλεφώνων τους. Η συχνότητα εμφάνισης ορθογραφικών λαθών καθώς και αργκό, είναι πολύ υψηλότερη από ότι στους άλλους τομείς.
Ακόμα οι χρήστες κοινωνικών δικτύων δημοσιεύουν microblogs πάνω σε πολλά διαφορετικά θέματα, σε αντίθεση με άλλους ιστότοπους που είναι προσαρμοσμένοι σε ένα συγκεκριμένο θέμα.
Τέλος τα κείμενα μικρού μήκους περιέχουν ακολουθίες χαρακτήρων που δεν είναι λέξεις. Κάποιες από αυτές είναι URL's, hashtags και προσωπάκια (emoticons).

Πίνακας 2.1: Πίνακας που δείχνει την μείωση του πλήθους των χαρακτηριστικών της κάθε κατηγορίας μετά από feature reduction

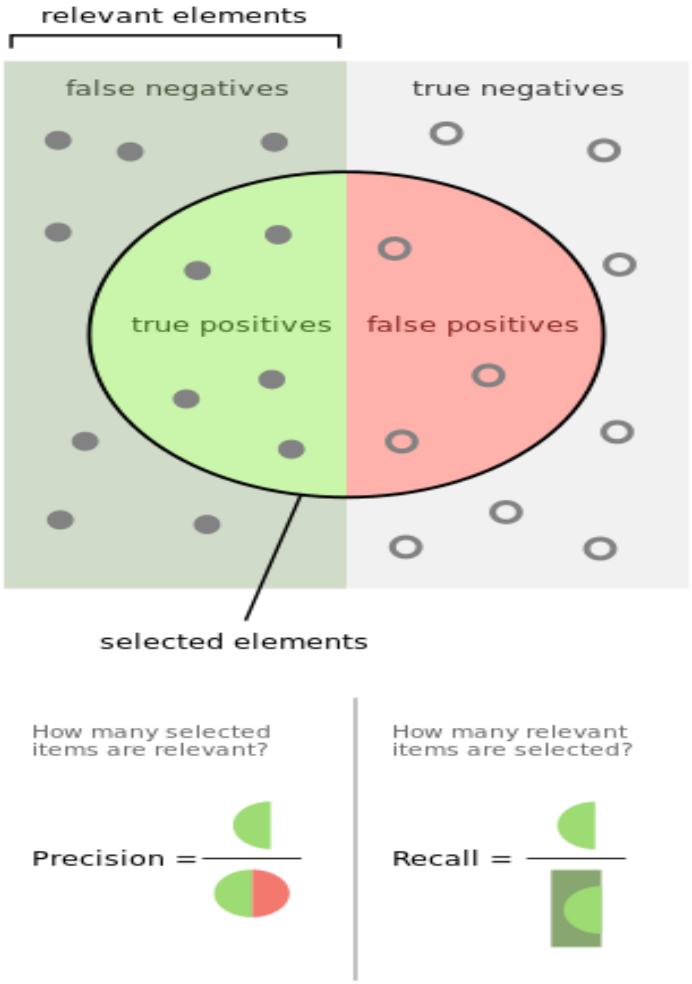
Feature Reduction	# of Features	Percent of Original
None	794876	100.00%
Username	449714	56.58%
URLs	730152	91.86%
Repeated Letters	773691	97.33%
All	364464	45.85%

Πολλά από αυτά δεν επηρεάζουν το συναίσθημα του κειμένου και έτσι θεωρούνται θόρυβος. Όμως κάποια άλλα όπως τα emoticons (😀, ☹) ή τα επαναλαμβανόμενα σημεία στίξης (!!!, ..., ???), μπορούν να βοηθήσουν σημαντικά στην εξόρυξη του συναισθήματος. Η χρήση αυτών των ιδιοτήτων ονομάζεται Feature Reduction και είναι μια πολύ σημαντική διαδικασία κατά την προετοιμασία των χαρακτηριστικών.

2.5 Εκτίμηση της Απόδοσης

Ένα πολύ σημαντικό κομμάτι της μηχανικής μάθησης είναι η ανάλυση της επίδοσης του μοντέλου που έχει εκπαιδευθεί. Συνήθως στην ανάλυση συναισθημάτων χρησιμοποιούνται η ακρίβεια-precision, η ανάκληση-recall και η F-mesure ή F1-Score. Η βαθμολογία F1 αποτελεί μέτρο της ακρίβειας της δοκιμής. Η τεχνική θεωρεί τόσο την ακρίβεια p και

την ανάκληση r της δοκιμής για να υπολογίσει την βαθμολογία: *p* είναι ο αριθμός των σωστών θετικών αποτελεσμάτων διαιρούμενο με τον αριθμό όλων των θετικών αποτελεσμάτων που επιστρέφονται από τον ταξινομητή, και *το r* είναι ο αριθμός των σωστών θετικών αποτελεσμάτων διαιρούμενο με τον αριθμό όλων των σχετικών δειγμάτων (όλα τα δείγματα που θα έπρεπε να έχουν προσδιοριστεί ως θετικά).



Εικόνα 2.2: Παράδειγμα λειτουργίας F1-Score

Η βαθμολογία F_1 είναι ο αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης , όπου η βαθμολογία F_1 φτάνει την καλύτερη τιμή της στο 1 (άριστη ακρίβεια και ανάκληση) και χειρότερη στο 0. Στην εικόνα 2.2 φαίνεται ο δειγματικός χώρος της ακρίβειας και της ανάκλησης.

Όταν υπάρχουν πολλαπλές κλάσεις χρησιμοποιείται το macro-avarage και micro-average για το F-Score. Αυτά υπολογίζονται ως εξής:

$$Macro_{Precison} = \frac{1}{n}\sum_i \frac{correct(emotion = i)}{proposed(emotion = i)}$$

$$Macro_{Recall} = \frac{1}{n}\sum_{\mathrm{i}} \frac{correct(emotion = i)}{gold(emotion = i)}$$

$$Macro_{F-measure} = \frac{2xMacro_{precision}xMacro_{recall}}{Macro_{precision} + Macro_{recall}}$$

$$Micro_{Precision} = \frac{\sum \#correct(emotion = i)}{\sum \#proposed(emotion = i)}$$

$$Micro_{Recall} = \frac{\sum \#correct(emotion = i)}{\sum \#gold(emotion = i)}$$

$$Micro_{F-measure} = \frac{2xMicro_{precision}xMicro_{recall}}{Micro_{precision} + Micro_{recall}}$$

Proposed είναι ο αριθμός των κειμένων που έχουν την ετικέτα συναισθήματος i από τον ταξινομητή, correct είναι ο αριθμός των κειμένων που έχουν την σωστή ετικέτα i από τον ταξινομητή και gold είναι ο αριθμός των κειμένων μέσα στο σετ δεδομένων που έχουν την ετικέτα i. i είναι ένα από τα n συναισθήματα.

2.6 Σύνοψη Κεφαλαίου

Όπως κάθε επιστήμη έτσι η ανάλυση συναισθήματος είναι συνεχώς εξελισσόμενη. Υπάρχουν πολλά ανεξερεύνητα πεδία μελέτης και η ταξινόμηση κειμένων μικρού μήκους είναι ένα από αυτά. Όπως αναφέρθηκε προηγουμένως, οι δυσκολίες αυτού του τομέα είναι η δομή και τα περιεχόμενα του κειμένου. Η δυσκολία αυτού του προβλήματος αυξάνεται περισσότερο εάν θελήσουμε να αναλύσουμε το συναίσθημα σε περισσότερες από δύο κλάσεις (θετικό – αρνητικό). Η μελέτη αυτής της πτυχιακής στοχεύει ακριβώς σε αυτό το θέμα και θα παρουσιαστή στο επόμενο κεφάλαιο.

3. ΥΛΟΠΟΙΗΣΗ

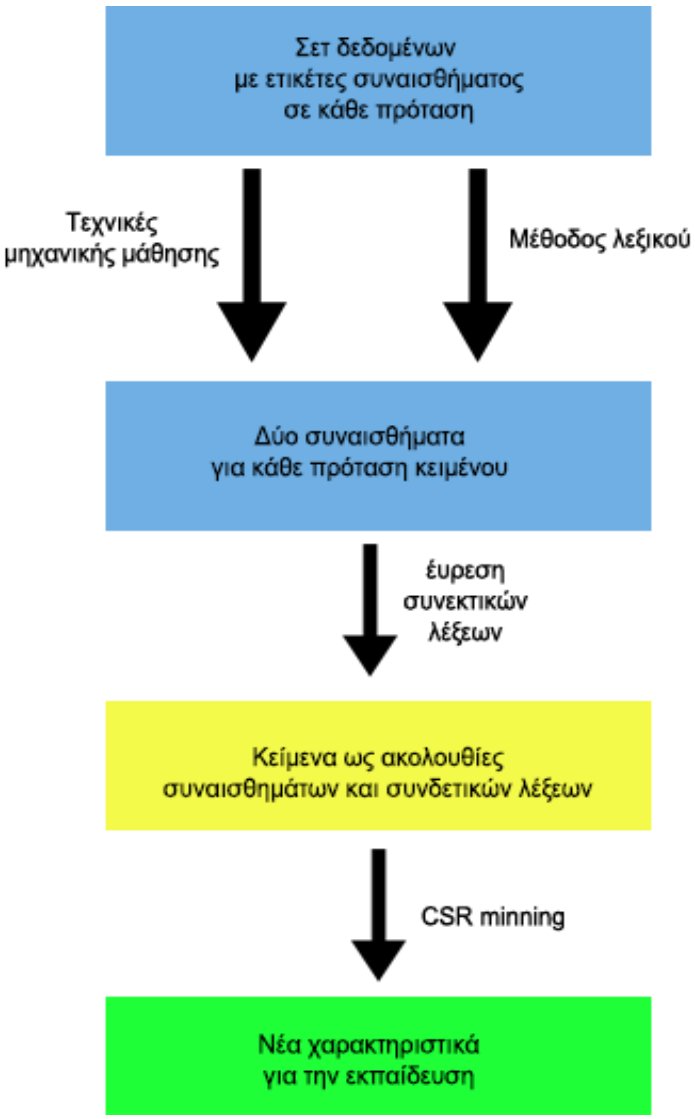
Έχοντας αναφέρει τους λόγους για τους οποίους πραγματοποιήθηκε η έρευνα που παρουσιάζεται σε αυτήν την πτυχιακή (**Shiyang Wen and Xiaojun Wan, AAAI14[4])**, σειρά έχει η ανάλυση των τεχνικών που χρησιμοποιήθηκαν σε αυτήν. Το κεφάλαιο αυτό χωρίζεται σε δύο μέρη. Το πρώτο μέρος περιλαμβάνει την παρουσίαση της υπάρχουσας έρευνας, ενώ στο δεύτερο αναλύεται η προσωπική μου υλοποίηση του θέματος. Στις αναφορές υπάρχει σύνδεσμος που οδηγεί σε σελίδα στο github, η οποία περιέχει το προγραμματιστικό μέρος της υλοποίησης**[8]**.

3.1 Παρουσίαση

Στόχος αυτής της έρευνας **(Shiyang Wen and Xiaojun Wan[4])** είναι ταξινόμηση ή ανάλυση του συναισθήματος κειμένου μικρού μήκους στην κινεζική γλώσσα. Προσπαθούν να κατηγοριοποιήσουν τα κείμενα σε μία από τις εννέα κατηγορίες - κλάσεις συναισθήματος, όπως θυμός, αηδία, φόβος, ευτυχία, αρεστό, θλίψη, έκπληξη, ουδέτερο και κενό σε περιπτώσεις που δεν εκφράζει κάποια άποψη - γνώμη το κείμενο. Παρατηρούμε πως το πρόβλημα αυτό περιέχει και τις δύο δυσκολίες που αναφέραμε στο **Κεφάλαιο 2**, καθώς προσπαθούν να βασιστούν πάνω στα ιδιαίτερα χαρακτηριστικά των κειμένων microblog για να παραχθεί μια δομή πρόβλεψης πολλών διαφορετικών συναισθημάτων. Σημειώστε ότι η εργασία που αντιμετωπίζουν βασίζεται στο επίπεδο εγγράφου η οποία είναι συνήθως πιο δύσκολη από τις εργασίες του επιπέδου προτάσεων (Κεφάλαιο 1). Δυστυχώς, υπάρχουσες μέθοδοι που βασίζονται σε λεξικά ή τεχνικές μηχανικής μάθησης δεν μπορούν να επιτύχουν ικανοποιητικά αποτελέσματα απόδοσης γιατί συνήθως αντιμετωπίζουν τα κείμενα αυτά με τεχνικές εξαγωγής χαρακτηριστικών όπως το bag of words, ή bag of sentences (Κεφάλαιο 2.1.2). Αυτές οι τεχνικές δεν λαμβάνουν υπόψιν πληροφορίες σχετικά με την σειρά των λέξεων και της διάρθρωσης του λόγου σε ένα κείμενο. Προτείνεται λοιπόν η ενσωμάτωση των κανόνων διαδοχικών κλάσεων για την συλλογή νέων πιο αποτελεσματικών χαρακτηριστικών, στο πρόβλημα της εποπτευόμενης ανάλυσης συναισθημάτων.
Περιληπτικά τα βήματα της έρευνας είναι τα εξής:

1. Έχοντας ένα σετ δεδομένων εκπαίδευσης, με κάθε κείμενο χωρισμένο σε προτάσεις, χρησιμοποιούνται παραδοσιακές προσεγγίσεις μηχανικής μάθησης και λεξικού για την απόκτηση δύο συναισθημάτων για κάθε πρόταση των κειμένων.
2. Βάση μιας λίστας συνδετικών λέξεων βρίσκονται και αποθηκεύονται για κάθε κείμενο οι συνδετικές λέξεις που περιέχονται, καθώς η θέση τους μέσα σε αυτό.
3. Με αυτά τα στοιχεία μετατρέπεται κάθε κείμενο σε μία ακολουθία συναισθημάτων και συνδετικών λέξεων. Αυτές τις επεξεργάζεται ο αλγόριθμος CSR για να παραχθούν κατατοπιστικοί κανόνες για όλο το σετ δεδομένων.
4. Με αυτούς τους κανόνες παράγονται νέα χαρακτηριστικά εκπαίδευσης τα οποία θα χρησιμοποιηθούν από έναν αλγόριθμο μηχανικής μάθησης.

Υπάρχουν τουλάχιστον δύο θετικά σε αυτή τη προσέγγιση. Το πρώτο είναι πως αυτή η μέθοδος μπορεί να αναγνωρίσει την σειρά των προτάσεων και τις σχέσεις του λόγου μεταξύ αυτών. Το δεύτερο είναι ότι χρησιμοποιούνται δύο διαφορετικές μέθοδοι για την εύρεση των αρχικών συναισθημάτων. Για αυτό το λόγο, παρόλο που η απόκτηση συναισθημάτων με τους παραπάνω τρόπους βασίζεται στο προτασιακό επίπεδο, χρησιμοποιώντας δύο διαφορετικές μεθόδους για την απόκτηση συναισθήματος για κάθε πρόταση, η προσέγγιση είναι ανεκτική σε λάθη που μπορεί να υπάρξουν από μία μόνο μέθοδο.
Όπως αναφέρεται, τα αποτελέσματα δείχνουν ότι αυτή η προσέγγιση είναι αρκετά ανταγωνιστική και μπορεί να υπερβεί φανερά ορισμένες σχετικές βάσεις.



Σχήμα 3.1: Διάγραμμα βημάτων της έρευνας

3.2 Βασικές Προσεγγίσεις

Σε αυτό το σημείο θα αναλυθούν οι δύο βασικές τεχνικές που χρησιμοποιούνται στο πρώτο βήμα της προσέγγισης. Αυτές είναι οι τεχνικές με βάση το λεξικό και την μηχανική μάθηση και βασίζονται σε δύο επίπεδα, το προτασιακό και του εγγράφου. Η βασική λογική τους έχει παρουσιαστεί στο δεύτερο κεφάλαιο, αλλά τώρα θα γίνει κατανοητό πώς χρησιμοποιούνται στην πράξη.

Προτού αναφερθούμε σε αυτές πρέπει να γίνει κατανοητή η μορφή του σετ δεδομένων εκπαίδευσης. Σε αυτή την έρευνα χρησιμοποιείται ένα σετ δεδομένων αναφοράς από την κινεζική εκτίμηση συναισθηματικής ανάλυσης κειμένων μικρού μήκους του 2013 **(Microblog Sentiment Analysis Evaluation – CMSAE)[5]**. Ένα παράδειγμα του κειμένου και τον προτάσεων του, μαζί με τις ετικέτες συναισθήματος φαίνεται στον **πίνακα 3.1**. Αυτό περιέχει εφτά διαφορετικές κατηγορίες συναισθήματος: Θυμός - Anger, Αηδία- Disgust, Φόβος - Fear, Χαρά - Happiness, Αρέσκεια - Like, Θλίψη - Sadness και Έκπληξη - Surprise. Εάν το κείμενο δεν περιέχει συναίσθημα τότε του αναθέτεται το Κενό - None. Το σετ εκπαίδευσης περιέχει 4000 κείμενα microblog και 13252 προτάσεις. Κάθε κείμενο χωρίζεται από προτάσεις που το αποτελούν. Υπάρχει μία ετικέτα συναισθήματος για ολόκληρο το κείμενο, καθώς μία κύρια και μία δευτερεύουσα για κάθε πρόταση, η οποίες ετικέτες έχουν ανατεθεί χειροκίνητα. Το σετ δεδομένων ελέγχου περιέχει 10.000 κείμενα microblog και 32185

προτάσεις. Σε αυτό έχουν ανατεθεί ετικέτες, μόνο των κύριων συναισθημάτων ολόκληρου του κειμένου. Ο **πίνακας 3.2** δείχνει την κατανομή των δεδομένων του σετ. Αυτό μας βοηθάει να κατανοήσουμε σημαντικές πληροφορίες για τα δεδομένα που θα φανούν χρήσιμες στον αλγόριθμο CSR. Είναι γενικά επιθυμητό τα σετ δεδομένων να είναι ισοζυγισμένα.

Πίνακας 3.1: Παράδειγμα κειμένου microblog με τρία συναισθήματα

	Sentence	Emotion
1	今天下雨。 (Today is rainy.)	*none*
2	我有点郁闷 [流泪]! (I am a little depressed [tears]!)	*sadness*
3	但是在家里看书也不错 [嘻嘻]。 (But staying at home to read some books is also not so bad [hee hee].)	*happiness*

Πίνακας 3.2: Αριθμός παραδειγμάτων κάθε κλάσης συναισθήματος σε επίπεδο εγγράφου στα δύο σετ δεδομένων

emotion type	train dataset	test dataset
anger	235	436
disgust	425	935
fear	49	102
happiness	371	1116
like	597	1558
sadness	388	744
surprise	112	236
none	1823	4873
total	4000	10000

3.3 Lexicon-based Approach

Όπως αναφέρθηκε, το πρώτο βήμα είναι η εξαγωγή δύο συναισθημάτων μέσω δύο προσεγγίσεων, αυτής με βάση το λεξικό και αυτής της μηχανικής μάθησης.
Η προσέγγιση με βάση το λεξικό μπορεί να ενταχθεί στην κατηγορία της μη-εποπτευόμενης μάθησης, καθώς εξάγεται το συναίσθημα χωρίς να χρειάζονται οι ετικέτες των παραδειγμάτων. Όμως, όπως αναφέρθηκε σε προηγούμενο κεφάλαιο, η προσέγγιση αυτή εξαρτάται απόλυτα από το λεξικό. Όσο πιο αντιπροσωπευτικό είναι τόσο καλύτερη θα είναι η απόδοση του συστήματος. Στην έρευνα που μελετάμε δημιουργήθηκε ένα κινέζικο λεξικό συναισθημάτων από τρεις διαφορετικές πηγές. Το βασικό λεξικό αποτελείται από τους επτά τύπους συναισθημάτων που χρησιμοποιούνται στην μελέτη αυτή. Από το λεξικό αφαιρέθηκαν ορισμένες λέξεις συναισθημάτων που δεν είναι κατάλληλες για το

συγκεκριμένο θέμα. Στην συνέχεια, συλλέγεται μία λίστα με λέξεις “αργκό” και προστίθενται στο λεξικό. Τέτοιου είδους λέξεις είναι απαραίτητο να υπάρχουν στο λεξικό καθώς εμφανίζονται συχνά στα κείμενα μικρού μήκους. Σε άλλες περιπτώσεις ανάλυσης συναισθήματος, όπως σε επιστημονικά άρθρα, αυτό το βήμα δεν θα ήταν αναγκαίο. Τέλος, για τον ίδιο λόγο συλλέγονται και χρησιμοποιούνται διάφορα emoticons. Στον **πίνακα 3.3** φαίνονται οι αριθμοί των λέξεων κάθε κατηγορίας μέσα στο λεξικό.

Πίνακας 3.3: Αριθμός συναισθηματικών λέξεων στο λεξικό της έρευνας

Emotion Type	*anger*	*disgust*	*fear*	*happiness*	*like*	*sadness*	*surprise*
Number	431	9624	1096	1859	10237	2227	218

Έπειτα από την συλλογή του απαραίτητου λεξικού ακολουθεί το στάδιο της προεπεξεργασίας των κειμένων. Στην έρευνα χρησιμοποιείται το εργαλείο κατάτμησης κειμένου, το οποίο διασπά τις προτάσεις σε μεμονωμένες λέξεις. Στην συνέχεια καταμετράται ο αριθμός των λέξεων κάθε κατηγορίας, που περιέχει η κάθε πρόταση, με βάση το λεξικό. Τέλος, το συναίσθημα της πρότασης προκύπτει, πολύ απλά, από την κατηγορία με τις περισσότερες λέξεις. Εάν δεν έχει βρεθεί κάποια συναισθηματική λέξη από το λεξικό στην πρόταση, τότε καμία κατηγορία συναισθήματος δεν θα εκφράζει την πρόταση και έτσι της αναθέτεται η συναισθηματική ετικέτα “Κενό - None”.
Αυτή η τεχνική μπορεί να χρησιμοποιηθεί σε οποιαδήποτε πρόταση για να αποκτηθεί το συναίσθημα σε προτασιακό επίπεδο.

3.4 Προσέγγιση με SVM

Προηγούμενες μελέτες έχουν αποδείξει πως ο SVM προσφέρει καλύτερες επιδόσεις από άλλους αλγόριθμους μηχανικής μάθησης. Λόγω της ανωτερότητας του χρησιμοποιείται και σε αυτήν την προσέγγιση.
Ο αλγόριθμος αυτός υποστηρίζεται από πολλές γλώσσες προγραμματισμού, καθώς υπάρχουν πολλές βιβλιοθήκες με την υλοποίηση του και με παραδοχές του. Στην συνέχεια αναφέρονται τα χαρακτηριστικά που χρησιμοποιούνται για την εκπαίδευση του SVM, και για τα δύο επίπεδα (προτασιακό, εγγράφου).
Πρώτο χαρακτηριστικό είναι οι λέξεις του κειμένου. Κάθε ξεχωριστή κινεζική λέξη που εμφανίζεται σε πρόταση του σετ δεδομένων αποθηκεύεται. Αφού ελεγχθεί ολόκληρο το σετ, θα δημιουργηθεί μια λίστα με μοναδικές λέξεις. Έπειτα για κάθε πρόταση θα σημειωθεί με 1 αν η κάθε λέξη της λίστας εμφανίζεται μέσα στην πρόταση, και με 0 αν όχι. Υποθέτοντας ότι βρέθηκαν 5.000 διαφορετικές λέξεις και ότι όλες οι προτάσεις του σετ είναι 13.252, τότε θα καταλήξουμε με έναν δισδιάστατο μαθηματικό πίνακα μεγέθους 13.252 x 5.000, με εγγραφές 0 και 1. Αυτός ο πίνακας θα χρησιμοποιηθεί μαζί με τις ετικέτες συναισθήματος για κάθε πρόταση του σετ εκπαίδευσης ως χαρακτηριστικά εκπαίδευσης του SVM.
Δεύτερο χαρακτηριστικό είναι τα σημεία στίξης. Ορισμένες ακολουθίες στίξης μπορεί να αντικατοπτρίζουν ειδικά είδη συναισθημάτων. Μια λίστα τέτοιων ακολουθιών στίξης προστίθενται στα χαρακτηριστικά. Για παράδειγμα, Το "???" μπορεί να αντικατοπτρίζει ένα συναίσθημα "θυμού" και το "!!!" μπορεί να αντικατοπτρίζει το "έκπληκτο".
Τα τελευταία χαρακτηριστικά εκπαίδευσης προέρχονται από την προσέγγιση με βάση το λεξικό. Πιο συγκεκριμένα, ακολουθούνται τα βήματα της προσέγγισης που παρουσιάστηκε στην ενότητα 3.3, και χρησιμοποιείται ο πίνακας όπου έχουν καταμετρηθεί οι συναισθηματικές λέξεις κάθε κατηγορίας. Για παράδειγμα, υπάρχουν μόνο τρεις λέξεις που εντάσσονται στην κατηγορία “happy” του κατασκευασμένου λεξικού και μία λέξη που εμφανίζεται στην κατηγορία "like", σε ένα κείμενο του σετ, και συνεπώς οι αντίστοιχες χαρακτηριστικές τιμές είναι: 0(anger), 0(disgust), 0(fear), 3(happiness),1(like), 0(sadness), 0(surprise).

Τα χαρακτηριστικά και από τις τρεις κατηγορίες έχουν αριθμητικές τιμές η οποίες πρέπει να μην έχουν μεγάλη διαφορά εύρους. Δηλαδή να μην μετριέται το ένα σε χιλιάδες και το άλλο σε δεκάδες. Αν συμβαίνει κάτι τέτοιο, τότε χρησιμοποιούνται τεχνικές κανονικοποίησης για την εξισορρόπηση των δεδομένων. Γενικά τα χαρακτηριστικά που μπορούν να χρησιμοποιηθούν είναι πολλά και η επιλογή τους εξαρτάται από την κατανόηση του προβλήματος από την μεριά του σχεδιαστή.

3.5 Προτεινόμενη Προσέγγιση

Σε αυτό το σημείο παρουσιάζεται η προτεινόμενη προσέγγιση των **Shiyang Wen** και **Xiaojun Wan[4]** για την ταξινόμηση κειμένου μικρού μήκους σε πολλαπλές τάξεις συναισθημάτων. Στόχος είναι να συνδυαστούν οι ετικέτες συναισθήματος προτασιακού επιπέδου, που έχουν εξαχθεί από τα σετ δεδομένων μαζί με τον αλγόριθμο διαδοχικών κανόνων κλάσης, για να παραχθούν χαρακτηριστικά επιπέδου εγγράφου για την τελική εκπαίδευση του συστήματος μέσω ενός SVM. Καταρχάς θα παρουσιαστεί η τεχνολογία των CSR. Στην συνέχεια θα αναφερθούν τα βήματα της προσέγγισης.

3.5.1 CSR

3.5.1.1 Κανόνες Διασύνδεσης και Διαδοχικά Μοτίβα

Οι ακολουθιακοί κανόνες κλάσης (Class Sequential Rules – CSR) δεν έχουν χρησιμοποιηθεί στον τομέα της ανάλυσης συναισθήματος πριν από αυτή την έρευνα. Αυτοί είναι ένα σημαντικό μοντέλο εξόρυξης δεδομένων το οποίο χρησιμοποιείται σε πολλά έργα του τομέα “Web Mining”. Υπάρχουν δύο βασικές τεχνικές του τομέα αυτού. Η πρώτη είναι η εξόρυξη κανόνων διασύνδεσης (Association Rules Mining), η οποία βρίσκει σετ δεδομένων από αντικείμενα που εμφανίζονται συχνά μαζί. Η δεύτερη είναι η εξόρυξη διαδοχικών μοτίβων (Sequential Pattern Mining), η οποία βρίσκει και σετ δεδομένων από αντικείμενα που εμφανίζονται μαζί σε μία ακολουθία. Οι CSR εντάσσονται στην δεύτερη κατηγορία. Επειδή όμως αυτή είναι εξέλιξη της πρώτης, μελετάμε και τις δύο.
Η κλασική εφαρμογή της εξόρυξης κανόνων διασύνδεσης είναι η ανάλυση δεδομένων στις αγορές, η οποία στοχεύει να ανακαλύψει πώς αντικείμενα που αγοράζονται από τους πελάτες σε ένα σούπερ μάρκετ (ή ένα κατάστημα) συνδέονται μεταξύ τους. Για παράδειγμα, στην εικόνα 3.1 φαίνεται ένας κανόνας διασύνδεσης, ο οποίος λέει πως το 10% των πελατών αγοράζουν τυρί και μπίρα μαζί, και αυτοί που αγοράζουν τυρί, θα πάρουν και μπίρα το 80% των φορών. Οι έννοιες support και confidence είναι δύο μέτρα για την δύναμη του κανόνα, και θα αναπτυχθούν παρακάτω.

$$\text{Cheese} \rightarrow \text{Beer} \qquad [\text{support} = 10\%, \text{confidence} = 80\%]$$

Εικόνα 3.1: Παράδειγμα απλού κανόνα διασύνδεσης

Ωστόσο, η εξόρυξη κανόνων διασύνδεσης δεν θεωρεί την ακολουθία με την οποία αγοράζονται τα στοιχεία. Η εξόρυξη των διαδοχικών προτύπων φροντίζει για αυτό. Ένα παράδειγμα είναι ότι "το 5% των πελατών αγοράζουν πρώτα κρεβάτι, έπειτα στρώμα και στη συνέχεια μαξιλάρια". Τα στοιχεία δεν αγοράζονται ταυτόχρονα αλλά το ένα μετά το άλλο. Αυτά τα μοτίβα είναι χρήσιμα στην εξόρυξη χρήσης του Διαδικτύου για την ανάλυση των ροών κλικ σε αρχεία καταγραφής διακομιστών. Είναι επίσης χρήσιμα και για την εύρεση γλωσσικών προτύπων σε κείμενα φυσικής γλώσσας και για αυτό θα χρησιμοποιηθούν και σε αυτή την έρευνα.

3.5.1.2 Βασικές Έννοιες των Κανόνων Διασύνδεσης

Ο ορισμός του προβλήματος της εξόρυξης κανόνων διασύνδεσης (Association Rules Minning), είναι ο εξής.
Έχοντας ένα σετ αντικειμένων $I = \{i_1, i_2, \dots, i_m\}$ και ένα σετ από συναλλαγές $T = (t_1, t_2, \dots, t_m)$, όπου κάθε συναλλαγή t_i είναι ένα σετ αντικειμένων τέτοιο ώστε $t_i \subseteq I$. Ένας κανόνας διασύνδεσης είναι μια συνέπεια της φόρμας: $X \to Y, όπου\ X \subset I, Y \subset I\ και\ X \cap Y = \emptyset$. Το X είναι ένα σετ αντικειμένων που θα ονομάζεται Itemset. Ένα παράδειγμα για να γίνει κατανοητός ο κανόνας αυτός είναι το εξής. Θέλουμε να αναλύσουμε τον τρόπο με τον οποίο εξαρτώνται τα προϊόντα που πουλιούνται σε ένα σούπερ μάρκετ. I είναι το σετ των προϊόντων που έχουν πουληθεί. Μια συναλλαγή (transaction) είναι πολύ απλά όλα τα αντικείμενα που έχει στο καλάθι του ένας αγοραστής. Για παράδειγμα, μια τέτοια συναλλαγή μπορεί να είναι η **{Beef, Chicken, Cheese},** που σημαίνει πως ο πελάτης αγόρασε αυτά τα τρία αντικείμενα. Τώρα, ένας κανόνας διασύνδεσης μπορεί να είναι ο $Beef, Chicken \longrightarrow Cheese$. Το σύνολο {beef, chicken} είναι το itemset X και το σύνολο {Cheese} είναι το Y. Για ευκολία οι αγκύλες παραλείπονται από τους κανόνες.
Μία συναλλαγή λέμε ότι περιέχει ένα itemset X, εάν το X είναι υποσύνολο του t_i (αναφέρεται πως το X καλύπτει το t_i). Ο αριθμός υποστήριξης του κανόνα (support count) του X στο T (συμβολίζεται με **X.count**) είναι ο αριθμός των συναλλαγών μέσα στο T που περιέχουν το X. Η δύναμη του κανόνα μετράται από την υποστήριξη(support) και την αυτοπεποίθηση(confidence).
Support: Η υποστήριξη ενός κανόνα είναι είναι το ποσοστό των συναλλαγών μέσα στο T που περιέχουν $X \cup Y$, και είναι μία εκτίμηση της πιθανότητας $\mathrm{P}(X \cup Y)$. Επομένως καθορίζει πόσο συχνή είναι η εφαρμογή του κανόνα στο σύνολο των συναλλαγών T. Έστω n, ο αριθμός των συναλλαγών στο T. Η υποστήριξη του κανόνα $X \longrightarrow Y$ υπολογίζεται ως εξής:

$$\boldsymbol{support} = \frac{(\boldsymbol{X} \cup \boldsymbol{Y}).\boldsymbol{count}}{\boldsymbol{n}}$$

Η υποστήριξη είναι ένα χρήσιμο μέτρο επειδή, αν όταν είναι πολύ χαμηλό, ο κανόνας μπορεί να συμβαίνει τυχαία. Επιπλέον, σε ένα επιχειρηματικό περιβάλλον, ένας κανόνας που καλύπτει πολύ λίγες περιπτώσεις (ή συναλλαγές) μπορεί να μην είναι χρήσιμος επειδή δεν έχει επιχειρηματικό νόημα (δεν είναι κερδοφόρο).

Confidence: Η αυτοπεποίθηση ενός κανόνα, $X \longrightarrow Y$, είναι το ποσοστό των συναλλαγών μέσα στο T που περιέχουν το X, να περιέχουν και το Y. Είναι μία εκτίμηση για την υποδεικνυόμενη πιθανότητα $P(Y|X)$ και υπολογίζεται ως:

$$\boldsymbol{confidence} = \frac{(\boldsymbol{X} \cup \boldsymbol{Y}).\boldsymbol{count}}{\boldsymbol{X}.\boldsymbol{count}}$$

Επομένως, η εμπιστοσύνη καθορίζει την προβλεψιμότητα του κανόνα. Εάν η εμπιστοσύνη ενός κανόνα είναι πολύ χαμηλή, δεν μπορεί κανείς να συμπεράνει ή να προβλέψει αξιόπιστα το Y από το X. Ένας κανόνας με χαμηλή προβλεψιμότητα είναι περιορισμένης χρήσης.
Ο στόχος λοιπόν του προβλήματος της εξόρυξης κανόνων διασύνδεσης, βάση ενός καταλόγου συναλλαγών, είναι να βρεθούν όλοι οι κανόνες με υποστήριξη και αυτοπεποίθηση μεγαλύτερη ή ίση με αυτές που έχει δηλώσει ο χρήστης κατά την αρχικοποίηση του αλγορίθμου. Αυτές οι τιμές ονομάζονται minsup – minimum support και minconf – minimum confidence και είναι τα κάτω όρια των μετρικών απόδοσης των κανόνων. Ένα παράδειγμα για την καλύτερη κατανόηση του προβλήματος είναι το εξής. Στην **εικόνα 3.2** φαίνεται μια λίστα με επτά συναλλαγές. Κάθε μια από αυτές είναι ένα σετ από αντικείμενα (itemset) αγορασμένα από κάποιον πελάτη ενός καταστήματος.

t_1: Beef, Chicken, Milk
t_2: Beef, Cheese
t_3: Cheese, Boots
t_4: Beef, Chicken, Cheese
t_5: Beef, Chicken, Clothes, Cheese, Milk
t_6: Chicken, Clothes, Milk
t_7: Chicken, Milk, Clothes

Εικόνα 3.2: Λίστα με αγορές προϊόντων από πελάτες ενός καταστήματος

Δεδομένου ενός minsup = 30% και minconf = 80% καθορισμένα από τον χρήστη, μπορούν να παραχθούν οι παρακάτω κανόνες:

1) Chicken, Clothes ⟶ Milk [sup = 3/7, conf = 3/3]
2) Clothes⟶ Milk, Chicken [sup = 3/7, conf = 3/3]

Και οι δύο κανόνες είναι επιτρεπτοί. Ο πρώτος κανόνας έχει support = 3/7 = 42,86% (>30%) και confidence = 100%. Αυτό σημαίνει ότι εμφανίζεται σε τρεις από τις επτά συναλλαγές και σε αυτές είναι σίγουρο ότι θα αγοράσουν γάλα αφότου έχουν πάρει κοτόπουλο και ρούχα. Παρόμοιος είναι και ο δεύτερος κανόνας ο οποίος όμως έχει ως συνέπεια δύο αντικείμενα. Προφανώς μπορούν να υπάρξουν και άλλοι επιτρεπτοί κανόνες.
Βλέπουμε ότι αυτή η τεχνική δεν ανταποκρίνεται απόλυτα στις ανάγκες της προσέγγισης της ανάλυσης συναισθήματος, γιατί δεν λαμβάνεται υπόψη η σειρά των αγορών και δεν χρησιμοποιούνται στόχοι. Δηλαδή, οποιοδήποτε στοιχείο μπορεί να εμφανιστεί ως συνέπεια ή προϋπόθεση ενός κανόνα. Ακόμα, σε ορισμένες εφαρμογές, ο χρήστης ενδιαφέρεται μόνο για κανόνες με ορισμένα σταθερά αντικείμενα στόχου τη δεξιά πλευρά, όπως τα συναισθήματα. Τεχνικές που θα αναλυθούν παρακάτω καλύπτουν αυτά τα προβλήματα, όμως όλες τους βασίζονται πάνω στην τεχνική των κανόνων διασύδεσης και ο αλγόριθμος που χρησιμοποιείται είναι ο ίδιος - ο Apriori, με ελάχιστες παραλλαγές.

3.5.1.3 Εξόρυξη με Κανόνες Διασύνδεσης

Σε αυτό το σημείο θα γίνει μια αναφορά στην λογική και στις τεχνικές των κανόνων διασύνδεσης όπως χρησιμοποιούνται και στο πρόβλημα της αναφερόμενης έρευνας. Τα μοντέλα εξόρυξης που μελετήθηκαν μέχρι στιγμής δεν χρησιμοποιούν στόχους. Δηλαδή, οποιοδήποτε στοιχείο μπορεί να εμφανιστεί ως συνέπεια ή προϋπόθεση ενός κανόνα. Ωστόσο, σε ορισμένες εφαρμογές, ο χρήστης ενδιαφέρεται μόνο για κανόνες με ορισμένα σταθερά αντικείμενα στόχου τη δεξιά πλευρά.
Ο ορισμός της τεχνικής αυτής είναι ο εξής. Ας υποθέσουμε ότι **T** είναι ένα σύνολο δεδομένων που αποτελείται από **n** συναλλαγές. Κάθε μια συναλλαγή έχει μια ετικέτα **y.** Έχοντας **I** ένα σύνολο όλων των διαφορετικών αντικειμένων μέσα στο **T**, **Y** όλων των διαφορετικών ετικετών (Class Labels) και $I \cap Y = \emptyset$. Ένας κανόνας διασύνδεσης κλάσεων (CAR) είναι μια εφαρμογή της φόρμας:

$$X \longrightarrow y, όπου\ X \subseteq I, και\ y \in Y.$$

Οι ορισμοί της υποστήριξης και της εμπιστοσύνης είναι οι ίδιοι με αυτούς που ισχύουν για τους συνήθεις κανόνες διασύνδεσης. Γενικά, ένας κανόνας διασύνδεσης κλάσης είναι διαφορετικός από τον έναν κανονικό κανόνα σύνδεσης με δύο τρόπους. Πρώτον, η συνέπεια ενός CAR έχει μόνο ένα στοιχείο, ενώ το επακόλουθο ενός κανονικού κανόνα σύνδεσης μπορεί να έχει οποιοδήποτε αριθμό στοιχείων και δεύτερον, η παραγόμενη y μπορεί να ενός CAR μπορεί να είναι μόνο από το σύνολο Y. Δεν υπάρχει στοιχείο από το I το οποίο μπορεί να εμφανιστεί ως επακόλουθο και καμία ετικέτα κλάσης να εμφανιστεί ως όρος κανόνων. Αντίθετα, μπορεί να υπάρξει ένας κανονικός κανόνας σύνδεσης ο οποίος έχει οποιοδήποτε στοιχείο ως όρο ή επακόλουθο.
Το πρόβλημα της εξόρυξης κανόνων διασύνδεσης κλάσεων είναι να δημιουργηθεί το πλήρες σύνολο από CARs που ικανοποιούν τους περιορισμούς ελάχιστης υποστήριξης (minsup) και ελάχιστης εμπιστοσύνης (minconf) που ορίζονται από τον χρήστη. Το **σχήμα 3.2** εμφανίζει ένα σύνολο δεδομένων που περιέχει επτά έγγραφα κειμένου. Κάθε έγγραφο είναι μια συναλλαγή και αποτελείται από ένα σύνολο λέξεων-κλειδιών. Κάθε συναλλαγή έχει επίσης μια ετικέτα κλάσης με θέμα την εκπαίδευση ή τον αθλητισμός.

I = {Student, Teach, School, City, Game, Baseball, Basketball, Team, Coach, Player, Spectator}
Y = {Education, Sport}.

	Transactions	Class
doc 1:	Student, Teach, School	: Education
doc 2:	Student, School	: Education
doc 3:	Teach, School, City, Game	: Education
doc 4:	Baseball, Basketball	: Sport
doc 5:	Basketball, Player, Spectator	: Sport
doc 6:	Baseball, Coach, Game, Team	: Sport
doc 7:	Basketball, Team, City, Game	: Sport

Σχήμα 3.2: Μορφή δεδομένων για την εξόρυξη κανόνων διασύνδεσης κλάσης

Υποθέτοντας πως ο χρήστης έθεσε minsup = 20% και minconf = 60%, δύο κανόνες που προκύπτουν από το παραπάνω σχήμα είναι οι εξής:

Student, School → Education [sup= 2/7, conf = 2/2]
Game → Sport [sup= 2/7, conf = 2/3]

Μια ερώτηση που μπορεί κανείς να ρωτήσει είναι: μπορούμε να εξορύξουμε τα δεδομένα απλά χρησιμοποιώντας τον Apriori αλγόριθμο και στη συνέχεια να εκτελέσει μια μετα-επεξεργασία των κανόνων που προκύπτουν και να επιλέξουμε μόνο τους κανόνες συσχέτισης τάξης που ικανοποιούν? Κατ 'αρχήν, η απάντηση είναι ναι επειδή οι CAR είναι ένας ειδικός τύπος κανόνων διασύνδεσης. Ωστόσο, στην πράξη αυτό είναι συχνά δύσκολο ή και αδύνατο λόγω του συνδυαστικού μεγέθους, επειδή ο αριθμός των παραγόμενων κανόνων από ένα σετ δεδομένων χιλιάδων γραμμών μπορεί να είναι τεράστιος. Το minsup και το minconf βοηθάνε και στην ταχύτητα εξόρυξης των κανόνων.

3.5.1.4 Βασικές Έννοιες των Διαδοχικών Μοτίβων

Η εξόρυξη κανόνων διασύνδεσης δεν λαμβάνει υπόψη την σειρά των δεδομένων. Ωστόσο, σε πολλές εφαρμογές, αυτή η τεχνική αυξάνει την απόδοση δραστικά. Για παράδειγμα, στην συγκεκριμένη μελέτη για την εύρεση συναισθήματος σε κείμενα μικρού μήκους, η πρόβλεψη της σειράς των λέξεων μας βοηθάει στην κατανόηση του συναισθηματικού νοήματος της πρότασης. Για αυτόν τον λόγο, οι CAR δεν είναι χρήσιμοι στο συγκεκριμένο πρόβλημα και χρειάζεται η χρήση των ακολουθιακών μοτίβων (Sequential Patterns).
Ορίζεται με $I = \{i_1, i_2, \dots, i_m\}$ ένα σύνολο στοιχείων. Μία ακολουθία είναι μια λίστα από itemsets σε συγκεκριμένη σειρά. Όπως έχει αναφερθεί σε προηγούμενη ενότητα, ένα itemset X είναι ένα μη-κενό υποσύνολο του I. Μία ακολουθία s συμβολίζεται με $< a_1 a_2 \dots a_r$ όπου a_i είναι ένα itemset. Καλείται και στοιχείο της s. Δηλώνουμε ένα στοιχείο (ή σύνολο αντικειμένων) μιας ακολουθίας από $\{x_1, x_2, \dots, x_k$όπου x_j □ I είναι ένα στοιχείο. Υποθέτουμε χωρίς απώλεια της γενικότητας ότι τα στοιχεία σε ένα στοιχείο μιας ακολουθίας είναι σε **λεξικογραφική σειρά**. Ενα αντικείμενο μπορεί να εμφανιστεί μόνο μία φορά σε ένα στοιχείο μιας ακολουθίας, αλλά μπορεί να εμφανιστεί πολλαπλές φορές σε διαφορετικά στοιχεία. **Το μέγεθος** μια ακολουθίας είναι ο αριθμός των στοιχείων(itemsets) μέσα στην ακολουθεία. **Το μήκος** της είναι ο αριθμός των αντικειμένων μέσα στην ακολουθία. Μια ακολουθία μήκους k ονομάζεται **k-sequence.** Εάν ένα στοιχείο εμφανίζεται πολλές φορές σε διαφορετικά στοιχεία μιας ακολουθίας, κάθε εμφάνιση συμβάλει στο μέγεθος του k. Μία ακολουθία $s_1 = < a_1 a_2 \dots a_r >$ είναι **υπο-ακολουθία (subsequence)** μιας άλλης $s_2 = < b_1 b_2 \dots b_v >$, εάν υπάρχουν ακέραιοι $1 \le j_1 < j_2 < \dots < j_{r-1} < j_r \le v$, τέτοιοι ώστε $a_1 \subseteq b_{j1}, a_2 \subseteq b_{j2}, \dots, a_r \subseteq b_{jr}$. Λέγεται ακόμα ότι το s_2 περιέχει το s_1. Ένα παράδειγμα για καλύτερη κατανόηση του ζητήματος είναι το εξής. Έχοντας, $I = \{1,2,3,4,5,6,7,8,9\}$ η ακολουθία $< \{3\}\{4,5\}\{8\} >$ είναι υποσύνολο της ακολουθίας $s = < \{6\}\{3,7\}\{9\}\{4,5,8\}\{3,8\} >$ διότι $3 \subseteq \{3,7\}$, $\{4,5\} \subseteq \{4,5,8\}, και \{8\} \subseteq \{3,8\}$. Όμως το $< \{3\}\{8\} >$ δεν περιέχεται στο $< \{3,8\} >$, ή το αντίστροφο. Το μέγεθος της ακολουθίας s είναι 5 και το μήκος 9. Επίσης σημαντικό είναι ότι η ακολουθία $< \{4,5\}\{3\}\{8\} >$ δεν είναι υποσύνολο της s, γιατί $\{4,5\} \subseteq \{4,5,8\}, 3 \subseteq \{3,8\}$ αλλά το {8} δεν αντιστοιχεί κάπου. Στο τελευταίο παράδειγμα φαίνεται η διαφορά στην σειράς της ακολουθίας.

3.5.1.5 Κανόνες διαδοχικών κλάσεων

Οι κανόνες διαδοχικών κλάσεων (CSR) είναι η τεχνική που χρησιμοποιείται από τους **Shiyang Wen και Xiaojun Wan[4]** στην συγκεκριμένη προσέγγιση του προβλήματος της ανάλυσης συναισθήματος από κείμενα μικρού μήκους. Οι CSR's βασίζονται πάνω στις τεχνικές που αναλύθηκαν παραπάνω. Συγκεκριμένα είναι ένας συνδυασμός των κανόνων διασύνδεσης (CAR) και των διαδοχικών μοτίβων. Στο συγκεκριμένο πρόβλημα υπάρχουν προτάσεις με λέξεις ως ακολουθίες και σε κάθε μια έχει ανατεθεί μια ετικέτα συναισθήματος. Αυτά τα ζεύγη D χρησιμοποιούνται ως είσοδος στον αλγόριθμο εξόρυξης.($D = \{(s_1, y_1), (s_2, y_2), \dots, (s_n, y_n)\}$, όπου s_i είναι μια ακολουθία και y_i μια ετικέτα κλάσης.) Με Y ορίζεται το σύνολο όλων των ετικετών. Στην περίπτωση αυτή, Y = {anger, disgust, fear, happiness, like, sadness, suprise, none}. Ένας κανόνας διαδοχικών κλάσεων είναι μια εφαρμογή της φόρμας:

$$X \longrightarrow y, όπου\ X\ είναι\ η\ ακολουθία\ και\ y \in Y.$$

Ο πίνακας 3.4 παρουσιάζει ένα παράδειγμα μιας ακολουθίας δεδομένων με πέντε ακολουθίες και δύο κλάσεις, c1 και c2. Έχοντας minsup 20% και minconf 40% παράγεται ο κανόνας $< \{1\}\{3\}\{7,8\} > \longrightarrow c_1$.
Από την θεωρία των διαδοχικών μοτίβων (Εν. 3.5.1.4) μπορεί να παρατηρηθεί πως ο παραγόμενος κανόνας είναι υποσύνολο των ακολουθιών 1, 2 και 5. Τότε λέμε ότι το ζεύγος καλύπτει (cover) τον κανόνα. Ακόμα λέμε πως τα ζεύγη ικανοποιούν (satisfy) τον κανόνα, εάν και μόνο εάν η ακολουθία X του κανόνα είναι υποσύνολο του ζεύγους και έχουν

τις ίδιες ετικέτες. Από την θεωρία των κανόνων διασύνδεσης (Εν. 3.5.1.2) έχουμε πως η υποστήριξη - support είναι το κλάσμα των συνολικών ζευγών που ικανοποιούν τον κανόνα και η επιστοσύνη - confidence του κανόνα είναι ο αριθμός των ζευγών που ικανοποιούν τον κανόνα, προς τον αριθμό των ζευγών που τον καλύπτουν. Έτσι ο παραγόμενος κανόνας έχει support = 2/5 και confidence = 2/3.

Πίνακας 3.4: Μορφή δεδομένων ακολουθιών με ετικέτες

	Data Sequence	Class
1	<{1}{3}{5}{7, 8, 9}>	c_1
2	<{1}{3}{6}{7, 8}>	c_1
3	<{1, 6}{9}>	c_2
4	<{3, 5}{6}>	c_2
5	<{1}{3}{4}{7, 8}>	c_2

3.5.2 Τα Βήματα της Προσέγγισης

Σε αυτήν την ενότητα παρουσιάζονται τα βήματα ολόκληρης της προσέγγισης. Στόχος είναι να μετατραπούν όλα τα κείμενα από τα σετ δεδομένων εκμάθησης και ελέγχου (training & test sets) σε ακολουθίες. Κάθε πρόταση σε ένα κείμενο αντιπροσωπεύεται ως ένα σύνολο στοιχείων μίας ή δύο ετικετών συναισθημάτων, καθώς και κάθε συνδετική λέξη αποτελεί και αυτή ένα itemset. Επομένως, ένα κείμενο μικρού μήκους αντιπροσωπεύεται από μια ακολουθία συνόλων στοιχείων - itemsets.
Σε προηγούμενες ενότητες παρουσιάστηκε ο τρόπος με τον οποίον χρησιμοποιούνται οι μέθοδοι με βάση το λεξικό και την μηχανική μάθηση, για την απόκτηση δύο ετικετών συναισθημάτων για κάθε πρόταση. Το πλεονέκτημα που προσφέρουν οι δύο ετικέτες συναισθημάτων είναι ότι μπορούν να χρησιμοποιηθούν και οι δύο από το διαδοχικό μοτίβο ώστε να βελτιωθεί η απόδοση. Για παράδειγμα, στον πίνακα 3.5 φαίνεται ένα παράδειγμα κειμένου μικρού μήκους από το σετ δεδομένων, με τρεις προτάσεις.

Πίνακας 3.5: Κείμενο μικρού μήκους του κινέζικου σετ δεδομένων

	Sentence	Emotion
1	今天下雨。 (Today is rainy.)	*none*
2	我有点郁闷 [流泪]! (I am a little depressed [tears]!)	*sadness*
3	但是在家里看书也不错 [嘻嘻]。 (But staying at home to read some books is also not so bad [hee hee].)	*happiness*

Αρχικά χρησιμοποιείται η μέθοδος με βάση το λεξικό για να αποκτηθούν τα συναισθήματα για κάθε πρόταση none-sadness-happiness, έπειτα με SVM αποκτούνται τα συναισθήματα sadness-sadness-happiness. Τέλος υπάρχει και μία συνδετική λέξη (but).

Με αυτόν τον τρόπο μετατρέπεται κάθε κείμενο μικρού μήκους σε ακολουθία συναισθημάτων.

$$< \{none, sadness\}\{sadness\}\{(but)\}\{happiness\} >$$

Ο λόγος για τον οποίο προσθέτουμε συνδετικές λέξεις στην ακολουθία είναι ότι αυτές αντικατοπτρίζουν το λόγο σχέσης μεταξύ προτάσεων, όπως αντιφατικές ή συντονιστικές σχέσης και συνάφειες. Οι σχέσεις λόγου μεταξύ των προτάσεων έχουν αντίκτυπο στο συναίσθημα ολόκληρου του κειμένου. Οι συζεύξεις εμφανίζονται συνήθως στην αρχή της πρότασης, και ένας μπορεί να υποδηλώνει τη σχέση μεταξύ των προτάσεων πριν και μετά από αυτήν. Ως εκ τούτου, είναι χρήσιμο αν προστεθούν και οι συζεύξεις στην ακολουθία. Συγκεκριμένα, το σετ δεδομένων ακολουθιών κατασκευάζεται με τα παρακάτω βήματα.

Πρώτον, για κάθε μια πρόταση σε ένα κείμενο χρησιμοποιούνται και οι δύο παραπάνω μέθοδοι για την απόκτηση δύο συναισθημάτων. Εάν αυτά τα συναισθήματα είναι ίδια, τότε θα ανατεθεί στην πρόταση μόνο μια ετικέτα. Σε αντίθετη περίπτωση θα ανατεθούν δύο. Για κάθε πρόταση σε ένα κείμενο στο σετ εκμάθησης (training set), χρησιμοποιούνται απευθείας οι ετικέτες που έχουν ήδη παρασχεθεί από τον άνθρωπο-σχολιαστή. Δεύτερον, συνδυάζονται οι ετικέτες των συναισθημάτων κάθε πρότασης με τις συνδετικές λέξεις και μετατρέπονται σε μία ακολουθία. Τέλος, στα κείμενα του σετ εκμάθησης προσάπτεται στο τέλος της ακολουθίας και η ετικέτα συναισθήματος ολόκληρου του κειμένου (<{none, sadness}{sadness}{但是(but)}{happiness}>, happiness).

Βάσει των δεδομένων που κατασκευάστηκαν από το το σετ εκπαίδευσης, χρησιμοποιείται ο αλγόριθμος εξόρυξης CSRs, και επιλέγονται οι κανόνες που πληρούν τις απαιτήσεις των minsup και minconf. Οι CSRs αντιπροσωπεύουν ενδεικτικά πρότυπα ειδικά για διαφορετικούς τύπους συναισθημάτων. Σε αυτό το σημείο πρέπει να αναφερθεί ότι μερικές συνδετικές λέξεις και κάποια συναισθήματα εμφανίζονται πολύ πιο σπάνια από άλλα. Για αυτό, μόνο ένα minsup για τον έλεγχο των CSRs δεν επαρκεί, γιατί για να παραχθούν κανόνες που περιλαμβάνουν σπάνιες σχέσεις και συναισθήματα απαιτεί να ορισθεί μια πολύ χαμηλή ελάχιστη τιμή υποστήριξης, η οποία θα προκαλέσει περιττούς κανόνες για συχνά συναισθήματα, κάτι που θα προκαλέσει υπερφόρτωση-overfitting. Σε αυτήν την στρατηγική, η ελάχιστη υποστήριξη ενός κανόνα προσδιορίζεται πολλαπλασιάζοντας την ελάχιστη συχνότητα εμφάνισης των συναισθημάτων του κειμένου μέσα στο σετ εκπαίδευσης, μαζί με μια παράμετρο **τ**. Έτσι, το minsup αλλάζει ανάλογα την κλάση, και για έναν κανόνα με συχνά στοιχεία το minsup θα είναι υψηλό, ενώ για έναν κανόνα με σπάνιο στοιχεία θα είναι χαμηλό.

Αφότου έχει πραγματοποιηθεί η εξόρυξη των CSRs από το σετ εκπαίδευσης, χρησιμοποιείται η ακολουθία του κάθε κανόνα σαν χαρακτηριστικό εκπαίδευσης. Δηλαδή, στην περίπτωση που υπάρχουν εκατό CSRs, ο τελικός SVM θα διδαχθεί με εκατό χαρακτηριστικά. Αυτά τα χαρακτηριστικά δημιουργούνται ως εξής. Για κάθε μια ακολουθία κειμένου του σετ δεδομένων ελέγχουμε εάν η ακολουθία του κανόνα είναι υποσύνολό της. Τότε αναθέτουμε την τιμή 1, αλλιώς την τιμή 0. Τέλος καταλήγουμε για κάθε κείμενο να έχουμε έναν πίνακα με μηδέν και ένα, ως χαρακτηριστικά. Η ίδια διαδικασία γίνεται και για τα δεδομένα ελέγχου.

3.6 Προσωπική Προσέγγιση

Σε αυτήν την ενότητα παρουσιάζεται η προσωπική μου υλοποίηση της έρευνας που παρουσιάστηκε προηγουμένως. Η βασική λογική βασίζεται πάνω στην θεωρία των **Shiyang Wen και Xiaojun Wan[4]**. Όμως υπάρχουν αρκετές διαφορές. Μια διαφορά είναι ότι η προσωπική μου υλοποίηση πραγματοποιείται σε αγγλικά δεδομένα έναντι κινέζικων. Τα βήματα της προσέγγισης αυτής είναι τα εξής. Πρώτα συλλέχθηκαν τα σετ δεδομένων, έπειτα επεξεργάστηκαν για να έρθουν σε μία όμοια και συμβατή μορφή και ακολουθεί η εξόρυξη των CSRs. Τέλος εκπαιδεύεται ο τελικός ταξινομητής και ελέγχονται τα αποτελέσματα. Αυτό το στάδιο είναι αναδρομικό, καθώς μπορούν να γίνουν αλλαγές στην εκπαίδευση του ταξινομητή για καλύτερα αποτελέσματα και γρηγορότερη εκτέλεση. Το

προγραμματιστικό μέρος της υλοποίησης μπορεί να βρεθεί μέσω του αντίστοιχου συνδέσμου στην βιβλιογραφία.

3.6.1 Συλλογή Δεδομένων

Όπως έχει αναφερθεί σε προηγούμενα κεφάλαια, το πρώτο και σημαντικότερο στάδιο της ανάλυσης συναισθημάτων είναι η συλλογή των κατάλληλων δεδομένων. Τα σετ δεδομένων της έρευνας που μελετήθηκε είναι στα κινέζικα. Λόγο δυσκολίας στην κατανόηση της γλώσσας, επιλέχθηκε η αγγλική για την συλλογή των δεδομένων. Ακόμα είναι σχετικά δύσκολο να βρεθούν κατάλληλα open-source σετ δεδομένων, ιδιαίτερα σε κάποια άλλη γλώσσα πέρα της αγγλικής. Το πιο σημαντικό κομμάτι στα σετ εκπαίδευσης και ελέγχου (training & test set) της προσέγγισης είναι ότι κάθε κείμενο χωρίζεται σε προτάσεις, και κάθε πρόταση έχει μια ετικέτα συναισθήματος. Δυστυχώς, δεν βρέθηκε κάποιο σετ δεδομένων αυτής της μορφής, όμως αυτό το πρόβλημα αντιμετωπίζεται κατά την προεπεξεργασία των δεδομένων. Όλα τα αρχεία που θα αναφερθούν υπάρχουν στον φάκελο Data της υλοποίησης.
Αρχικά, τα σετ εκπαίδευσης και ελέγχου που χρησιμοποιήθηκαν στην υλοποίηση αυτή προέχονται από τον ιστότοπο διαγωνισμών **kaggle[6]**. Συγκεκριμένα βρέθηκαν από έναν παλιότερο διαγωνισμό με θέμα, **Sentiment Analysis: Emotion in Text**. Τα αρχεία αυτά είναι τα test_data.csv, train_data.csv και sample_submission.csv. Τα πρώτα δύο είναι τα σετ δεδομένων εκπαίδευσης και ελέγχου. Το τελευταίο περιέχει τα σωστά συναισθήματα για το test set. Με αυτό γίνεται η τελική επαλήθευση και υπολογίζεται η ακρίβεια του αλγορίθμου. Το csv είναι μια απλή μορφή αρχείου που χρησιμοποιείται για την αποθήκευση δεδομένων πίνακα, όπως ένα υπολογιστικό φύλλο ή μια βάση δεδομένων. Τα αρχεία σε μορφή csv μπορούν να εισαχθούν και να εξαχθούν από προγράμματα που αποθηκεύουν δεδομένα σε πίνακες, όπως το **Microsoft Excel** ή το **OpenOffice Calc**. Στην συνέχεια αποκτήθηκε το σετ δεδομένων για την μέθοδο με βάση το λεξικό. Το συγκεκριμένο λεξικό που βρέθηκε είναι το **NRC-Emotion-Lexicon-v0.92-In105Languages[7]**. Αυτό περιέχει λέξεις σε αλφαβητική σειρά μαζί με ετικέτες συναισθήματος. Υπάρχουν δέκα διαφορετικά συναισθήματα, Positive-Negative-Anger-Anticipation-Disgust-Fear-Joy-Sadness-Suprise-Trust. Η κάθε ετικέτα για αυτά τα συναισθήματα είναι μια τιμή 0 ή 1, με 1 εάν το συναίσθημα εκφράζει την λέξη και 0 εάν όχι. Πρέπει να σημειωθεί πως στο πρόγραμμα, το αρχικό λεξικό έχει τροποποιηθεί ώστε να περιέχει μόνο τις αγγλικές λέξεις.
Σε αυτό το σημείο υπήρξε η πρώτη δυσκολία της υλοποίησης. Οι ετικέτες του σετ εκπαίδευσης δεν είναι ίδιες με αυτές του λεξικού. Έτσι, το σετ έχει τροποποιηθεί με τον ακόλουθο τρόπο. Αρχικά το σετ περιέχει δεκατρείς διαφορετικές ετικέτες συναισθημάτων, empty-sadness-enthusiasm-neutral-worry-suprise-love-fun-hate-happiness-boredom-relief-anger. Από αυτά παραμένουν μόνο όσα είναι ίδια ή παρόμοια στο λεξικό. Δηλαδή, τα συναισθήματα sadness, anger και surprise, είναι ετικέτες και στο λεξικό και στα σετ. Το συναίσθημα joy αντιστοιχεί με το happiness. Τέλος παραμένει και η ετικέτα empty. Όταν από την λεξιλογική μέθοδο δεν θα υπάρχει κάποια συναισθηματική λέξη στην πρόταση, θα λαμβάνει την ετικέτα empty. Θα μπορούσαν να συμπεριλαμβάνονται τα δεδομένα με ετικέτα neutral, καθώς θα γίνεται καταμέτρηση των positive και negative λέξεων μέσα στο κείμενο και όταν θα είναι ίσες τότε το συναίσθημα της πρότασης θα είναι neutral. Λόγω των αποτελεσμάτων αυτό το βήμα δεν υλοποιείται στο προγραμματιστικό κομμάτι της υλοποίησης. Ακόμα πρέπει να μετατραπεί το σετ ελέγχου-test set και το σετ επιβεβαίωσης-validation set*(Υπενθύμιση: υπάρχει αντιστοιχία στις γραμμές των δύο σετ, το πρώτο κείμενο του σετ ελέγχου είναι ίδιο με το πρώτο από το σετ επιβεβαίωσης και ούτω καθεξής. Τα σετ ελέγχου δεν περιέχουν ετικέτες. Αυτές περιέχονται στο σετ επιβεβαίωσης)*. Από αυτά αφαιρούνται όλα τα κείμενα που δεν ανήκουν στις τελικές ετικέτες που αναφέρθηκαν. Το αρχείο transform.py είναι αυτό που μετατρέπει τα αρχικά σετ (train_data.csv, test_data.csv, sample_submission.csv) στα σετ που χρησιμοποιούνται από τον ταξινομητή

(train_data_red, test_data_red, val_data_red). Στον πίνακα 3.6 φαίνεται ο αριθμός των κειμένων ανά κατηγορία συναισθήματος για τα σετ δεδομένων εκπαίδευσης και ελέγχου. Παρατηρείται πως τα δεδομένα δεν είναι καλά ισομοιρασμένα.

Πίνακας 3.6: Δεδομένα κειμένου ανά συναισθηματική κατηγορία

emotion type	train dataset	test dataset
anger	*98*	*11*
happiness	*2986*	*112*
sadness	*4828*	*681*
surprise	*1613*	*166*
empty	*659*	*60*
total	*10184*	*1030*

Τελευταίο βήμα είναι η δημιουργία του λεξικού με τις συνδετικές λέξεις. Αυτό είναι το αρχείο conjunctions.txt και περιέχει τις πιο γνωστές αγγλικές συνδετικές λέξεις. Κάποιες από αυτές είναι:

For, Or, Nor, But, Yet, Since.

Συνολικά συμπεριλαμβάνονται δεκαπέντε λέξεις.

3.6.2 Προεπεξεργασία Δεδομένων

Το δεύτερο κομμάτι της υλοποίησης είναι η προ-επεξεργασία των δεδομένων για λόγους όπου αναφέρθηκαν στην ενότητα 2.1.2. Σε αυτό το σημείο, τα δεδομένα εισάγονται στον κεντρικό αλγόριθμο και τροποποιούνται για να είναι έτοιμα για το επόμενο βήμα, την εξόρυξη των χαρακτηριστικών.

Αρχικά, το πρόγραμμα διαβάζει τα δεδομένα από τα σετ εκπαίδευσης και ελέγχου και τα εκχωρεί στους πίνακες X, y. Η εισαγωγή των δεδομένων γίνεται μέσω της βιβλιοθήκης pandas, η οποία βοηθάει στην ανάγνωση των αρχείων csv. Έπειτα τα δεδομένα κειμένου αποθηκεύονται στον πίνακα χαρακτηριστικών X. Οι ετικέτες συναισθήματος θα αποθηκευτούν στον πίνακα y, αφού πρώτα μετατραπούν σε αριθμητικές τιμές. Οι αλγόριθμοι ταξινόμησης ή εκτίμησης αποτελεσμάτων δεν δέχονται γραμματοσειρές. Στην εικόνα 3.3 φαίνεται ο αλγόριθμος μετατροπής και η αντιστοίχιση των συναισθημάτων με αριθμητικές τιμές. Υπάρχει η εντολή **pd.Categoricalpd.factorize(train_da ta.sentiment)[0])**, την οποία προσφέρει η βιβλιοθήκη pandas για την αντίστοιχη λειτουργία. Όμως, η συγκεκριμένη έκδοση εμφανίζει μήνυμα λάθους, σε συνδυασμό με την **train_test_split** της βιβλιοθήκης **sklearn** που θα αναλυθεί στο τελευταίο μέρος του προγράμματος, και για αυτό δεν χρησιμοποιείται.

Όλα τα βήματα προ-επεξεργασίας των κειμένων εφαρμόζονται και για τα δύο σετ(training & test set). Τα y για το test set διαβάζονται από το validation set και χρησιμοποιούνται μόνο για την τελική εκτίμηση της απόδοσης του ταξινομητή.

```
X = train_data.content
Xtest = test_data.content
#change the value of sentiment from string to int
#y = pd.Categorical(pd.factorize(train_data.sentiment)[0])
switch = {
    'empty': 0,
    'sadness':  1,
    'neutral':  2,
    'surprise':  3,
    'happiness':  4,
    'anger': 5
}
y = []
for s in train_data.sentiment:
    y.append(switch.get(s))

ytest = []
for s in val_data.sentiment:
    ytest.append(switch.get(s))
```

Εικόνα 3.3: Τμήμα κώδικα για την αρχική απόκτηση των δεδομένων

Το δεύτερο βήμα της προ-επεξεργασίας των δεδομένων είναι ο καθαρισμός τους. Συγκεκριμένα αφαιρούνται οι άγνωστοι χαρακτήρες(non-ascii characters). Έπειτα, τα πιθανά URL's και Emoticons μέσα στα κείμενα αντικαθίστανται από αντίστοιχες ετικέτες. Η ετικέτα για τους συνδέσμους ιστοσελίδων είναι URL, για τα θετικά emoticons είναι posE και για τα αρνητικά είναι negE. Ο διαχωρισμός των emoticons σε θετικά και αρνητικά μας βοηθά στην εύρεση του συναισθήματος, ενώ τα URL's είναι πιο πολύ θόρυβος κειμένου. Ακόμα αντικαθίστανται όλα τα σημεία στίξης από μια απλή τελεία. Αυτό το βήμα είναι απαραίτητο για την διάσπαση του κειμένου σε προτάσεις και είναι σημαντικό να πραγματοποιείται μετά την αντικατάσταση των URL's και των emoticons. Σε αντίθετη περίπτωση αυτά θα αλλοιωθούν. Τέλος αντικαθίστανται hastags και @-mentions(#not, @Kostas), τα οποία είναι κάποια χαρακτηριστικά του Twitter αλλά δεν έχουν καμία επιρροή στο συναίσθημα του κειμένου.
Το τρίτο και τελευταίο βήμα πριν την εξόρυξη των συναισθημάτων είναι η μέθοδος tokenization, δηλαδή η διάσπαση του κειμένου σε ξεχωριστές λέξεις. Ταυτόχρονα, σε αυτό το βήμα βρίσκονται και αποθηκεύονται οι συνδετικές λέξεις στο κείμενο. Αρχικά χρησιμοποιείται η συνάρτηση word_tokenize της βιβλιοθήκης nltk. Αυτή θα σπάσει τα κείμενα σε ξεχωριστές λέξεις όπως φαίνεται στην εικόνα 3.4.

```
0: ['AT_MENTION', 'i', 'know', 'i', 'was', 'listenin', 'to', 'bad', 'habit', 'earlier', 'and', 'i', 'started', 'fre
akin', 'at', 'his', 'part']
1: ['Layin', 'n', 'bed', 'with', 'a', 'headache', 'ughhhh', '.', 'waitin', 'on', 'your', 'call', '.']
2: ['Funeral', 'ceremony', '.', 'gloomy', 'friday', '.']
3: ['I', 'should', 'be', 'sleep', '.', 'but', 'im', 'not', '.', 'thinking', 'about', 'an', 'old', 'friend', 'who',
'I', 'want', '.', 'but', 'he', "'s", 'married', 'now', '.', 'damn', '.', 'amp', '.', 'he', 'wants', 'me', '.', 'sca
ndalous', '.']
4: ['AT_MENTION', 'Charlene', 'my', 'love', '.', 'I', 'miss', 'you']
5: ['AT_MENTION', 'I', "'m", 'sorry', 'at', 'least', 'it', "'s", 'Friday', '.']
6: ['Ugh', '.', 'I', 'have', 'to', 'beat', 'this', 'stupid', 'song', 'to', 'get', 'to', 'the', 'next', 'rude', '.']
7: ['AT_MENTION', 'if', 'u', 'watch', 'the', 'hills', 'in', 'london', 'u', 'will', 'realise', 'what', 'tourture',
'it', 'is', 'because', 'were', 'weeks', 'and', 'weeks', 'late', 'i', 'just', 'watch', 'itonlinelol']
8: ['Got', 'the', 'news']
9: ['The', 'storm', 'is', 'here', 'and', 'the', 'electricity', 'is', 'gone']
```

Εικόνα 3.4: 10 παραδείγματα πριν και μετά το tokenization

Τέλος φορτώνονται οι συνδετικές λέξεις από το λεξικό σε μία λίστα. Για κάθε ένα κείμενο δημιουργούνται προτάσεις κάθε φορά που υπάρχει τελεία ή συνδετική λέξη. Ταυτόχρονα γίνεται καταμέτρηση του μήκους του κειμένου, καθώς και των θετικών και αρνητικών emoticons(αφαιρούνται μετά την καταμέτρηση από τις προτάσεις). Αυτές οι τιμές θα

χρησιμοποιηθούν ως χαρακτηριστικά εκπαίδευσης, μαζί με αυτά από τους CSR's. Στην εικόνα 3.5 φαίνεται η τελική μορφή των δεδομένων.

```
0: ['AT_MENTION i know i was listenin to bad habit earlier and i started freakin at his part']
1: ['Layin n bed with a headache ughhhh', 'waitin on your call']
2: ['Funeral ceremony', 'gloomy friday']
3: ['I should be sleep', 'but', 'but im not', 'thinking about an old friend who I want', 'but', "but he 's married
now", 'damn', 'amp', 'he wants me', 'scandalous']
4: ['AT_MENTION Charlene my love', 'I miss you']
5: ["AT_MENTION I 'm sorry at least it 's Friday"]
6: ['Ugh', 'I have to beat this stupid song to get to the next rude']
7: ['AT_MENTION', 'if', 'if u watch the hills in london u will realise what tourture it is', 'because', 'because we
re weeks and weeks late i just watch itonlinelol']
8: ['Got the news']
9: ['The storm is here and the electricity is gone']
```

Εικόνα 3.5: 10 παραδείγματα κειμένου στην τελική τους μορφή

3.6.3 Εξόρυξη Κανόνων και Δημιουργία Χαρακτηριστικών

Σε αυτό το κομμάτι θα αναλυθούν οι τρόποι με τους οποίους παράγονται τα χαρακτηριστικά εκπαίδευσης του τελικού και βασικού ταξινομητή, ο οποίος παράγει τις προβλέψεις για το συναίσθημα του κάθε κειμένου. Στην προηγούμενη ενότητα(3.5.2) αναφέρθηκε πως στην προσέγγιση των **Shiyang Wen και Xiaojun Wan[4]** χρησιμοποιούνται και η μέθοδος με βάση το λεξικό και η μέθοδος με SVM, για να παραχθούν δύο συναισθήματα για κάθε πρόταση σε κάθε κείμενο. Για την μέθοδος του SVM χρειάζεται κάθε πρόταση στο σετ δεδομένων να έχει και μία ετικέτα συναισθήματος, καθώς ο SVM ανήκει στην κατηγορία του supervised-learning. Όμως στην προσωπική μου προσέγγιση δεν υπάρχει αυτή η δυνατότητα. Όπως προαναφέρθηκε, το σετ δεδομένων περιέχει απλά κείμενα και η διάσπαση σε προτάσεις έγινε κατά την προεπεξεργασία των δεδομένων και για αυτό δεν έχει γίνει καμία ανάθεση ετικετών.
Αντί της μεθόδου με βάση τον SVM χρησιμοποιείται μόνο η μέθοδος με βάση το λεξικό και από αυτήν παράγονται δύο διαφορετικά συναισθήματα. Αυτό είναι και το πρώτο μέρος αυτού του τμήματος κώδικα. Τα δεδομένα όπως φαίνονται στην εικόνα 3.5 περνούν στην συνάρτηση lex. Αυτή είναι υπεύθυνη για την λειτουργία της μεθόδου με βάση το λεξικό. Ο λόγος για τον οποίο μπορούμε να εφαρμόσουμε αυτήν την μέθοδο είναι γιατί δεν χρειάζονται ετικέτες συναισθημάτων για τις προτάσεις ή τα κείμενα, δηλαδή ανήκει στην κατηγορία unsupervised-learning. Αυτή η διαδικασία θα πραγματοποιηθεί για τα δεδομένα εκπαίδευσης και ελέγχου. Στόχος είναι να παραχθούν αριθμητικές ακολουθίες. Κάθε αριθμός είναι ένα αντίστοιχο συναίσθημα(εικόνα 3.3 - αντιστοιχία συναισθημάτων των σετ εκπαίδευσης και ελέγχου με αριθμούς).

3.6.3.1 Χρήση της Μεθόδου με Βάση το Λεξικό

Αρχικά, με την βοήθεια της βιβλιοθήκης pandas, διαβάζεται το λεξικό. Τα μόνα συναισθήματα που χρησιμοποιούνται είναι τα 'Anger', 'Joy', 'Sadness' και 'Surprise'. Το δεύτερο βήμα είναι μια προεπεξεργασία των δεδομένων. Κάθε μια πρόταση κάθε κειμένου σπάει σε μονές λέξεις. Καταμετρούνται ώστε κάθε πρόταση να αντικατασταθεί από tuples με ξεχωριστές λέξεις και την συχνότητα εμφάνισης τους μέσα στην πρόταση. Για παράδειγμα η πρόταση "the storm is here and the electricity is gone" θα μετατραπεί σε μία λίστα από tuples [(the,2), (strom,1), (is, 2), (here,1), (and,1), (electricity,1), (gone,1)]. Το τρίτο βήμα είναι η αντικατάσταση των λέξεων με συναισθήματα. Για κάθε πρόταση βρίσκεται το άθροισμα των συναισθημάτων της κάθε λέξης. Συγκεκριμένα, ψάχνουμε κάθε λέξη μέσα στο λεξικό. Αν υπάρχει τότε επιστρέφεται μια λίστα με τις τιμές του συναισθήματος της λέξης. Τα συναισθήματα είναι τέσσερα άρα για κάθε λέξη θα έχουμε μια λίστα μήκους τέσσερα από 0 ή 1. Δηλαδή για την λέξη «the» θα έχουμε την λίστα [0,0,0,0], που σημαίνει πως κανένα συναίσθημα δεν την εκφράζει. Για κάθε μια λέξη που βρίσκεται μέσα στο λεξικό, θα

πολλαπλασιάζεται η λίστα συναισθημάτων με τον αριθμό εμφάνισής της. Για όσες λέξεις δεν υπάρχουν μέσα στο λεξικό θα τους αναθέτεται η μηδενική λίστα. Τέλος αθροίζονται όλες οι λίστες μεταξύ τους. Καταλήγουμε κάθε πρόταση να εκφράζεται από μία λίστα με τιμές εμφάνισης συναισθημάτων. Το τελευταίο βήμα είναι η εύρεση των δύο πιο συχνών συναισθημάτων σε κάθε πρόταση. Εάν είναι όλα 0 τότε καταχωρείται το κενό συναίσθημα(-

1. Χρειάζεται να γίνει μια αντιστοιχία των αριθμητικών τιμών της μεθόδου του λεξικού με αυτές των σετ δεδομένων.

Αφότου έχει πραγματοποιηθεί η εύρεση του συναισθήματος κάθε πρότασης από την μέθοδο με βάση το λεξικό ακολουθεί η μετατροπή των δεδομένων σε ακολουθίες συναισθημάτων όπως περιγράφεται στην ενότητα 3.5.2. Αυτές οι ακολουθίες συναισθημάτων και συνδετικών λέξεων ονομάζονται ruleitems για τα δεδομένα εκπαίδευσης και condition sets για τα δεδομένα ελέγχου. Αυτή είναι η κατάλληλη μορφή των δεδομένων για να χρησιμοποιηθούν από τον αλγόριθμο του CSR για την εξόρυξη των κανόνων.

3.6.3.2 CSR's mining

Ο αλγόριθμος των class sequential rules είναι παρόμοιος με αυτόν των class assosiation rules. Ονομάζεται Apriori και αποτελείται από τρία βασικά κομμάτια. Σε αυτήν την υποενότητα θα αναφερθούν όλες οι λεπτομέρειες της χρήσης του στο πρόγραμμα και οι ιδιότητες που τον κάνουν να ξεχωρίζει από τους άλλους αλγόριθμους, δηλαδή ο τρόπος με τον οποίο λαμβάνει υπόψιν την σειρά των συναισθημάτων στην ακολουθία. Τα τρία διαφορετικά κομμάτια του αλγορίθμου ονομαστικά είναι ο βασικός αλγόριθμος CSR-apriori, η εύρεση υποψήφιων κανόνων και ο έλεγχος υποσυνόλου.
Αρχικά ο αλγόριθμος CSR-apriori(CSR) δέχεται την δομή των δεδομένων ruleitems (δεδομένα εκπαίδευσης). Τα δεδομένα του σετ ελέγχου δεν περνάνε ποτέ στον αλγόριθμο. Ακόμα ορίζεται από τον χρήστη το minimum support και confidence. Στόχος του αλγορίθμου είναι να εξορύξουμε κανόνες που ικανοποιούν τα δεδομένα. Αυτοί οι κανόνες είναι υπακολουθίες των ruleitems. Στην εικόνα 3.6 φαίνεται ο ψευδοκώδικας του αλγορίθμου όπως παρουσιάζεται στο βιβλίο του Liu.
Ο ίδιος αλγόριθμος χρησιμοποιείται και στο πρόγραμμα. Η διαδικασία του CAR είναι η εξής. Πρώτα διασπώνται τα ruleitems σε condset(condition sets) και σε y(ετικέτες κάθε ακολουθίας).

Ruleitem = (condset, y).

Όπως και ο Aproiri, έτσι και ο CAR παράγει όλα τα συχνά(frequent) ruleitems, κάνοντας πολλαπλά περάσματα πάνω από τα δεδομένα. Στο πρώτο πέρασμα υπολογίζεται ο αριθμός υποστήριξης του κάθε κανόνα πρώτου επιπέδου(1-ruleitem). Αυτοί περιέχουν μόνο ένα στοιχείο στο condset. Ταυτόχρονα δημιουργούνται και οι πρώτοι υποψήφιοι(1-candidates). Αυτοί είναι της μορφής:

$$C_1 = \{(\{i\}, y) | \, i \in I, και \, y \, \in Y\}$$

Έπειτα βρίσκονται εάν οι 1-candidates είναι frequent, δηλαδή όσους candidates έχουνε support πάνω από το minimum. Από τους πιο συχνούς υποψήφιους κανόνες παράγονται οι πρώτοι 1-CSR κανόνες, δηλαδή όσοι frequent candidates έχουν confidence πάνω από το ελάχιστο. Το επόμενο βήμα είναι το επαναληπτικό περάσματα από τα δεδομένα(level-wise search). Για το k πέρασμα χρησιμοποιούμε τους k-1 frequent rules για να παραχθούν οι k-candidates. Αυτό γίνεται μέσο της συνάρτησης CSRcandidate_gen (Ο τρόπος του συνδυασμού των υποψήφιων είναι αυτός που διαφοροποιεί τον CAR από τον CSR και θα αναλυθεί παρακάτω). Μετά την εύρεση των k-candidates υπολογίζεται το support και το confidence για τους νέους κανόνες και ξαναβρίσκονται οι πιθανοί k-frequent κανόνες και οι παραγόμενοι k-CSR's. Εάν δεν υπάρχουν k-1frequent κανόνες στην αρχή της επανάληψης ή εάν δεν παραχθεί κανένας k-candidate από τους k-1frequent, τότε η διαδικασία σταματάει. Σύμφωνα με το λήμμα, κάθε υποψήφιος κ-1 όπου δεν είναι frequent

δεν θα έχει και k-candidate που να είναι frequent. Με αυτή τη συνθήκη, οι υποψήφιοι κλαδεύονται σε κάθε επίπεδο και έτσι εγγυάται ο τερματισμός του αλγορίθμου.

```
Algorithm CAR-Apriori(T)
1   C_1 ← init-pass(T);                                  // the first pass over T
2   F_1 ← {f | f ∈ C_1, f. rulesupCount / n ≥ minsup};
3   CAR_1 ← {f | f ∈ F_1, f.rulesupCount / f.condsupCount ≥ minconf};
4   for (k = 2; F_{k-1} ≠ ∅; k++) do
5       C_k ← CARcandidate-gen(F_{k-1});
6       for each transaction t ∈ T do
7           for each candidate c ∈ C_k do
8               if c.condset is contained in t then  // c is a subset of t
9                   c.condsupCount++;
10                  if t.class = c.class then
11                      c.rulesupCount++
12          endfor
13      end-for
14      F_k ← {c ∈ C_k | c.rulesupCount / n ≥ minsup};
15      CAR_k ← {f | f ∈ F_k, f.rulesupCount / f.condsupCount ≥ minconf};
16  endfor
17  return CAR ← ⋃_k CAR_k;
```

Εικόνα 3.6: Ο αλγόριθμος CAR

Η συνάρτηση CSRcandidate_gen χρησιμοποιείται για την εύρεση των k-candidates κανόνων. Αρχικά ο αλγόριθμος δέχεται του k-1 συχνούς κανόνες μαζί με το επίπεδο της επανάληψης k. Αρχικά, όλοι οι κανόνες πρέπει να είναι σε λεξικογραφική σειρά, δηλαδή ταξινομούνται με βάση το μήκος τους και τα περιεχόμενα τους. Έπειτα ακολουθούν δύο επαναληπτικά βήματα. Πρώτο βήμα είναι αυτό του συνδυασμού. Οι υποψήφιες ακολουθίες δημιουργούνται συνδυάζοντας δύο k-1frequent κανόνες. Μια ακολουθία s1 συνδέεται με το s2 εάν η υπακολουθία που λαμβάνεται με την απομάκρυνση του πρώτου στοιχείου του s1 είναι η ίδια με την υπακολουθία που λαμβάνεται με την απόσπαση του τελευταίου στοιχείου του s2. Η υποψήφια ακολουθία που δημιουργείται με την ένωση s1 με s2 είναι η ακολουθία s1 που επεκτείνεται με το τελευταίο στοιχείο στο s2. Υπάρχουν δύο περιπτώσεις:

1. το πρόσθετο στοιχείο αποτελεί ξεχωριστό στοιχείο αν ήταν ξεχωριστό στοιχείο στο s2, και προσαρτάται στο τέλος του s1 στην συγχωνευμένη ακολουθία,
2. αλλιώς το στοιχείο που προστέθηκε είναι μέρος του τελευταίου στοιχείου του s1 στην συγχωνευμένη ακολουθία.

Όταν το k = 1 και συνδέουμε το F1 με το F1, πρέπει να προσθέσουμε το στοιχείο στο s2 και ως μέρος ενός itemset και ως ξεχωριστό στοιχείο. Δηλαδή η ένωση του <{x}> με το <{y}> μας δίνει και <{x,y}> και <{x}{y}>. Σημαντικό είναι ότι τα x και y πρέπει να τηρούν την ακολουθία. Στον πίνακα 3.7 βλέπουμε ένα παράδειγμα συνένωσης 3-candidate σε 4-candidate. Κάθε αριθμός αναπαριστά ένα συναίσθημα. Ταυτόχρονα φαίνεται και η ταξινόμηση των συχνών κανόνων. Αυτές οι ακολουθίες είναι μόνο τα condset χωρίς τα labels.

Πίνακας 3.7: Παράδειγμα candidate generation

Frequent 3-sequences	Candidate 4-sequences	
	after joining	after pruning
⟨{1, 2} {4}⟩	⟨{1, 2} {4, 5}⟩	⟨{1, 2} {4, 5}⟩
⟨{1, 2} {5}⟩	⟨{1, 2} {4} {6}⟩	
⟨{1} {4, 5}⟩		
⟨{1, 4} {6}⟩		
⟨{2} {4, 5}⟩		
⟨{2} {4} {6}⟩		

Πέρα από τον τρόπο παραγωγής των υποψήφιων κανόνων, άλλος ένας λόγος για τον οποίο τηρείται η σειρά της ακολουθίας στους κανόνες είναι η συνάρτηση subset. Η συνάρτηση αυτή χρησιμοποιείται στο στάδιο του κλαδέματος και πιο συγκεκριμένα κατά τον υπολογισμό του support ενός υποψηφίου. Ουσιαστικά ελέγχουμε για κάθε μια ακολουθία εάν οι υποψήφιοι είναι υπακολουθίες της και όχι απλά υποσύνολα. Δηλαδή μας ενδιαφέρει και η σειρά των εμφανίσεων (εξηγείται στην ενότητα 3.5.1.4).
Ανάλογα με τις ελάχιστες τιμές υποστήριξης και σιγουριάς, παράγεται διαφορετικό πλήθος κανόνων. Γενικά, όσο χαμηλότερες είναι αυτές οι τιμές τόσο περισσότεροι κανόνες παράγονται. Όμως αυτό δεν είναι απαραίτητα καλό. Στόχος είναι να βρεθεί το πλήθος κανόνων όπου θα μας προσφέρουν σίγουρες απαντήσεις, χωρίς να συμβαίνει overfitting ή underfitting στον ταξινομητή. Περισσότερες λεπτομέρειες σχετικά με τις αλλαγές των τιμών και πως αυτές επηρεάζουν τον αλγόριθμο θα αναφερθούν στο κεφάλαιο των αποτελεσμάτων. Οι optimal τιμές που χρησιμοποιούνται από το πρόγραμμα είναι minsupp = 1% και minconf = 1%(λόγο του πλήθους των δεδομένων), και παράχθηκαν 53 διαφορετικοί κανόνες. Στην εικόνα 3.7 παρουσιάζονται κάποια παραδείγματα των ruleitems(δεδομένα εισαγωγής στον CSR) και στην εικόνα 3.8 οι κανόνες του αλγορίθμου. Πρέπει να σημειωθεί πως ο αλγόριθμος του CSR επιστρέφει μόνο το condset κάθε παραγόμενου κανόνα, χωρίς την ετικέτα y, καθώς αυτή δεν χρειάζεται για την δημιουργία των χαρακτηριστικών.

```
ruleitems:
[['0'], 1]
[['0', '0'], 4]
[[('4', '3'), ('1', '0')], 1]
[[('4', '0')], 1]
[[('4', '0'), '0', ('5', '0'), '0'], 1]
[['0', 'but', ('1', '0')], 1]
[[('5', '0'), 'so', '0', 'but', '0', 'so', '0', '0', '0', '0'], 1]
[['0', 'for', ('1', '0'), '0', '0', 'either', 'either', '0', 'so', '0'], 1]
[[('1', '0')], 1]
```

Εικόνα 3.7: Παραδείγματα δεδομένων εισαγωγής του αλγορίθμου CSR

Τέλος απομένει μόνο να δημιουργηθούν τα τελικά χαρακτηριστικά που θα εκπαιδεύσουν τον αλγόριθμο του SVM. Αυτή η διαδικασία είναι η ίδια με αυτή που περιγράφεται στην τελευταία παράγραφο της ενότητας 3.5.2. Τα τελικά χαρακτηριστικά αποτελούνται από αυτά των κανόνων, από τα θετικά και αρνητικά emoticons και από το μήκος του κειμένου. Τα δεδομένα εκπαίδευσης είναι 10184 και του ελέγχου 1030. Στον CSR εισάγονται όλα τα δεδομένα εκπαίδευσης και παράγονται 24 κανόνες. Μαζί με τα 3 επιπλέον χαρακτηριστικά καταλήγουμε να έχουμε δύο πίνακες μεγέθους 10184x27 και 1030x27 χαρακτηριστικών εκπαίδευσης και ελέγχου αντίστοιχα. Στην εικόνα 3.9 παρουσιάζεται η τελική μορφή των χαρακτηριστικών X. Τα πρώτα 24 στοιχεία είναι οι κανόνες και τα τελευταία τρία είναι το μήκος του κειμένου, τα θετικά και τα αρνητικά emoticons.

```
but
('3', '0')
so
('4', '3')
0
when
('5', '1')
for
for
('5', '0')
as
('5', '4')
so
('1', '0')
('4', '1')
('4', '3')
0
('4', '0')
('4', '0')
but
(('1', '0'), '0')
(('5', '0'), '0')
(('5', '1'), '0')
```

Εικόνα 3.8: Οι ακολουθίες των παραγόμενων κανόνων από τον αλγόριθμο CSR

```
[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 8, 0, 0]
[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 7, 0, 0]
[0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 18, 0, 0]
[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 7, 0, 0]
[0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 27, 0, 0]
[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 7, 0, 0]
[0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 37, 0, 0]
[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 28, 0, 0]
[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 8, 0, 0]
```

Εικόνα 3.9: Δείγμα μορφής των τελικών χαρακτηριστικών εκπαίδευσης και ελέγχου

3.6.4 Εκπαίδευση του Τελικού Ταξινομητή

Το τελευταίο μέρος του προγράμματος είναι η εκπαίδευση του αλγορίθμου SVM και η εξόρυξη συναισθημάτων των κειμένων μικρού μήκους.
Είναι αρκετά σημαντικό να χωρίσουμε τα χαρακτηριστικά εκπαίδευσης ώστε να δημιουργηθεί ένα σετ ενδιάμεσου ελέγχου(cross-validation), το οποίο βοηθάει στους ελέγχους των αποτελεσμάτων. Με την συνάρτηση train_test_split της βιβλιοθήκης sklearn, χωρίζουμε τα χαρακτηριστικά εκπαίδευσης(τα X και τα Y). Στο πρόγραμμα, το 80% των αρχικών δεδομένων παραμένει ως training features και χρησιμοποιείται για την εκπαίδευση του ταξινομητή και το υπόλοιπο 20% για τον ενδιάμεσο έλεγχο. Τα δεδομένα χωρίζονται με τυχαίο τρόπο ώστε να διατηρηθεί η ποικιλομορφία των χαρακτηριστικών. Στην συνέχεια χρησιμοποιούμε το μοντέλο SVC από την βιβλιοθήκη sklearn.svm για την εξόρυξη των συναισθημάτων. Στον SVM μπορούν να χρησιμοποιηθούν διάφοροι πυρήνες για την καλύτερη λειτουργία του. Μια γενική λογική επιλογής πυρήνα είναι η εξής:
Ανάλογα με τον αριθμό των χαρακτηριστικών n και τον αριθμό των δεδομένων εκπαίδευσης m,

1) Εάν το n είναι μεγάλο σε σχέση με το m(π.χ. n=10.000, m=10-1000, χρησιμοποιείται svm χωρίς πυρήνα(linear-kernel),
2) εάν το n είναι μικρό και το m μεσαίο(π.χ. n=1-1000, m=10-10.000), χρησιμοποιείται svm με Gaussian kernel και
3) εάν το n είναι μικρό και το m μεγάλο(π.χ. n=1-1000, m=50.000+), χρησιμοποιείται πάλι svm χωρίς πυρήνα.

Η προσέγγιση αυτή ανήκει στην δεύτερη περίπτωση καθώς τα δεδομένα εκπαίδευσης είναι 10.184 και τα χαρακτηριστικά 27. Έπειτα από την εκπαίδευση του SVM με την εντολή fit, ακολουθεί η εξόρυξη των συναισθημάτων για το cross-validation set και έπειτα για το test set. Η συνάρτηση predict δέχεται ως όρισμα νέα χαρακτηριστικά δεδομένων και παράγει μια πρόβλεψη για το συναίσθημα των δεδομένων με βάση το μοντέλο εκπαίδευσης. Αυτή η πρόβλεψη είναι η αντίστοιχη αριθμητική τιμή του συναισθήματος. Τέλος, με το μοντέλο f1_score συγκρίνονται τα αποτελέσματα με τις ετικέτες Y και παράγεται το ποσοστό επιτυχίας του αλγορίθμου(F-score).
Η παραπάνω διαδικασία είναι επαναληπτική καθώς παρακολουθώντας τα αποτελέσματα, βγάζουμε συμπεράσματα πάνω σε αυτά και με κατάλληλες αλλαγές στα χαρακτηριστικά, στους κανόνες ή και στον ταξινομητή, επαναλαμβάνουμε την διαδικασία για πιθανώς καλύτερα αποτελέσματα.

4. ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Σε αυτό το κεφάλαιο παρουσιάζονται τα αποτελέσματα των διαφορετικών εκτελέσεων της προσέγγισης που μελετήθηκε. Αρχικά θα παρουσιαστούν στατιστικές από την μέθοδο των **Shiyang Wen και Xiaojun Wa[4]** και στη συνέχεια θα συγκριθούν με τα αποτελέσματα της προσωπικής μου υλοποίησης.
Αρχικά η προσέγγιση αυτή συγκρίνεται με διάφορες άλλες προσεγγίσεις ώστε να αποδειχθεί η ανωτερότητα της. Πρέπει να σημειωθεί πως υπάρχουν δύο διαφορετικές μετρήσεις για την προσέγγιση. Η πρώτη είναι μόνο με την χρήση των CSR's ως χαρακτηριστικά και η δεύτερη με την χρήση όλων των χαρακτηριστικών. Κάποιες άλλες μέθοδοι με τις οποίες συγκρίνεται η προσέγγιση είναι:

1) Best CMSAE Team: Η καλύτερη μέθοδος που προτάθηκε στην οργάνωση CMSAE.
2) Lexicon: Μόνο η μέθοδος με βάση το λεξικό.
3) SVM: Μόνο η μέθοδος του SVM.
4) Lexicon-vote και SVM-vote: Η μέθοδος με βάση το λεξικό ή με βάση τον SVM, μαζί με την μέθοδο simple majority vote.
5) FirstSen(Lexicon) και FirstSen(SVM): Χρησιμοποιείται η μέθοδος του λεξικού ή του SVM μόνο για την πρώτη πρόταση του κάθε κειμένου
7) LasrSen(Lexicon) και LastSen(SVM): Το αντίθετο από την FirstSen, δηλαδή μόνο για την τελευταία πρόταση.

Οι παράμετροι του minconf=0.01, δηλαδή 1% και του minsup=0.05, δηλαδή 5%. Στον πίνακα 4.1 παρουσιάζονται τα αποτελέσματα όλων των μεθόδων.

Πίνακας 4.1: Ποσοστά F-Score όλων των μεθόδων

Method	macro average			micro average		
	precision	recall	F-measure	precision	recall	F-measure
Best CMSAE Team	0.2842	0.348	0.3129	0.3232	0.3868	0.3521
Lexicon*#	0.4150	0.3648	0.3883	0.4131	0.4142	0.4137
SVM*#	0.3671	0.3289	0.3469	0.4252	0.4085	0.4167
Lexicon vote*#	0.4202	0.349	0.3813	0.414	0.3919	0.4026
SVM vote*#	0.3505	0.2482	0.2906	0.4225	0.3415	0.3777
FirstSen(Lexicon) *#	0.4235	0.2104	0.2811	0.3962	0.2302	0.2912
FirstSen(SVM) *#	0.3745	0.1515	0.2157	0.4326	0.2009	0.2744
LastSen(Lexicon) *#	0.4264	0.2407	0.3077	0.4164	0.277	0.3327
LastSen(SVM) *#	0.369	0.1737	0.2362	0.4479	0.2383	0.3111
Our Method (CSRs only)	0.4064	0.4267	0.4163	0.3924	0.4882	0.4351
Our Method (All features)	0.4133	0.4283	**0.4207**	0.3991	0.495	**0.4419**

Είναι φανερό πως η συγκεκριμένη προσέγγιση είναι καλύτερη από τις υπόλοιπες. Όμως πρέπει να αναφερθεί πως η έρευνα αυτή πραγματοποιήθηκε την χρονιά του 2014. Ακόμα παρατηρείται μία διαφορά του 2% ανάμεσα στην μεθόδο με όλα τα χαρακτηριστικά και αυτής μόνο με τα CSR's. Τέλος, οι τιμές των minsup και minconf μεταβάλλονται από 0.005 έως 0.05(0.5-5%). Οι βαθμολογίες του F-score είναι σχεδόν σταθερές. Συνολικά, οι επιδόσεις της μεθόδου δεν επηρεάζονται σημαντικά από την αλλαγή αυτόν τον τιμών, τουλάχιστον σε τόσο μικρά ποσοστά.
Στην προσωπική μου προσέγγιση, κατά την εκπαίδευση του αλγορίθμου, το training set έχει 8.147 δεδομένα, το cross-validation έχει 2.037(80-20% διάσπαση του αρχικού training dataset) και το test set έχει 1030. Ακόμα υπάρχουν πέντε διαφορετικές τάξεις συναισθημάτων. Πραγματοποιήθηκαν εκτελέσεις με διαφορετικά minsup και minconf. Οι τιμές αυτές κυμαίνονται από 1%-5% για το minsup και 1%-20% για το minconf. Πάνω από αυτές τις τιμές είτε δεν παράγονται κανόνες για τα συγκεκριμένα δεδομένα, είτε τα χαρακτηριστικά είναι πολύ λίγα και υπάρχει underfitting. Τα ποσοστά κάθε εκτέλεσης είναι τα εξής:

Minsup/minconf = 1/1%
53 rules generated
F-score Wrong Examples
Train Set
0.5027 4051
Cross-Validation Set
0.5135 991
Test set F-score: 0.555

Minsup/minconf = 1/5%
48 rules generated
F-score Wrong Examples
Train Set
0.502 4055
Cross-Validation Set
0.5144 989
Test set F-score: 0.5582

Minsup/minconf = 1/20%
27 rules generated
F-score Wrong Examples
Train Set
0.4974 4094
Cross-Validation Set
0.5076 1003
Test set F-score: 0.5631

Minsup/minconf = 5/1%
15 rules generated
F-score Wrong Examples
Train Set
0.4929 4131
Cross-Validation Set
0.5081 1002
Test set F-score: 0.564

Minsup/minconf = 5/20%
13 rules generated
F-score Wrong Examples
Train Set
0.4940 4122
Cross-Validation Set
0.5081 1002
Test set F-score: 0.5621

Με βάση αυτά τα παραδείγματα παρατηρούνται κάποιες ομοιότητες και διαφορές με την θεωρητική προσέγγιση. Μία πρώτη παρατήρηση είναι η επιρροή των τιμών minsup και minconf σε σχέση με το πλήθος των παραγόμενων κανόνων. Όπως είναι λογικό, το support επηρεάζει το πλήθος των κανόνων περισσότερο από το confidence, καθώς η υποστήριξη ενός κανόνα είναι γενικά ένα μικρότερο νούμερο από την εμπιστοσύνη. Παρατηρούμε πως με μια αύξηση του minsup από 1% σε 5%, οι κανόνες μειώνονται κατά

30+. Αντίθετα με μία αύξηση του minconf από 1% σε 5% υπάρχει μία μείωση των 5 κανόνων, και από 1% σε 20% παράγονται μόνο 20 λιγότεροι. Ακόμα παρατηρούμε πως αυτές οι δύο μονάδες είναι αλληλένδετες. Με minsup 5% και minconf 1%, παράγονται 15 κανόνες. Όταν αυξήσουμε το minconf σε 20% θα παραχθούν 13 κανόνες. Εδώ βλέπουμε πως η μείωση είναι πολύ μικρή, γιατί το minsup έχει κλαδέψει ήδη πολλούς μη-συχνούς κανόνες. Ακόμα, σε αυτήν την υλοποίηση η διαφορά των αποτελεσμάτων, σε όλες τις εκτελέσεις, μεταξύ της μεθόδου μόνο με τα χαρακτηριστικά CSR's και της μεθόδου με όλα, είναι 2% υπέρ της μεθόδου με όλα τα χαρακτηριστικά. Άρα είναι ίδια με αυτή της προτεινόμενης προσέγγισης.
Αν και αυτές οι αλλαγές στις τιμές επηρεάζουν το πλήθος των κανόνων, δεν επηρεάζουν τα ποσοστά επιτυχίας. Άλλη μια ομοιότητα των προσεγγίσεων είναι η διαφορά των F-Score μεταξύ των εκτελέσεων. Παρατηρείται λοιπόν, πως το εύρος των σκορ των διαφορετικών εκτελέσεων είναι το πολύ 1%(0.01). Σε αυτό το σημείο μπορεί να παρατηρηθεί και μια διαφορά. Το F-Score του Training-Set είναι μικρότερο από το σκορ των CV και Test sets. Όμως στην συγκεκριμένη περίπτωση δεν είναι κάτι ανησυχητικό. Γενικά γνωρίζουμε ότι όταν το ποσοστό επιτυχίας εκπαίδευσης είναι πολύ μεγαλύτερο από το ποσοστό επιτυχίας του CV, τότε εμφανίζεται overfitting, και όταν και τα δύο ποσοστά είναι χαμηλά τότε undefitting. Στην περίπτωση αυτή, η διαφορά των ποσοστών είναι αμελητέα(<1%). Ακόμα παρατηρούμε πως όσο περισσότερα τα χαρακτηριστικά τόσο πιο πολύ συγκλίνουν τα ποσοστά. Αυτό είναι λογικό καθώς με περισσότερα χαρακτηριστικά, δηλαδή περισσότερη εκπαίδευση του ταξινομητή, το ποσοστό επιτυχίας εκπαίδευσης θα μεγαλώνει και θα ξεπεράσει το ποσοστό CV, κάτι που οδηγεί σε ovefitting. Ο κύριος λόγος που το σκορ εκπαίδευσης είναι ελάχιστα μικρότερο από τα άλλα, είναι το μέγεθος των dataset. Τα σετ ελέγχου και CV είναι πολύ μικρά. Όμως, αν παρατηρήσουμε το πλήθος των λαθών σε κάθε εκτέλεση, βλέπουμε πως τα λάθη στο training set είναι περίπου 4 φορές περισσότερα από αυτά του CV, κάτι που είναι αναμενόμενο καθώς το training set είναι 4 φορές μεγαλύτερο του CV(χωρίστηκε το αρχικό training set κατά 80-20). Τέλος θέλουμε να επιλέξουμε την καλύτερη εκτέλεση(optimal). Αυτή όπως προαναφέρθηκε είναι η εκτέλεση με minsup και minconf 1%. Ο τρόπος επιλογής εξαρτάται κυρίως από τα F-score, την περίπτωση λάθους και μερικές φορές και από την ταχύτητα εκπαίδευσης. Στο πρόγραμμα επιλέχθηκε η παραπάνω εκτέλεση ως η ιδανική, καθώς εμφανίζει τις λιγότερες λανθασμένες προβλέψεις, και τα F-Score του training και CV έχουν την μικρότερη διαφορά.
Αφότου έχουμε ελέγξει την εκπαίδευση του ταξινομητή και έχουμε επιλέξει ένα ικανοποιητικό μοντέλο, παρατηρούμε το F-Score πάνω στο σετ ελέγχου. Αυτό της optimal εκτέλεσης είναι 55,5%. Μπορεί να μην είναι το υψηλότερο(σε άλλες εκτελέσεις έφτανε το 56%), όμως είναι το "πιο σωστό". Η πιο σημαντική παρατήρηση είναι πως αυτό το αποτέλεσμα ξεπερνά το σκορ της προσέγγισης των **Wen και Xiaojun Wa[4]**(44,2%). Το μειονέκτημα της προσωπικής μου υλοποίησης είναι ότι το αρχικό σετ εκπαίδευσης ήταν διαφορετικό, χωρίς ετικέτες για κάθε πρόταση. Οπότε δεν έχει χρησιμοποιηθεί η μέθοδος με βάση τον SVM για την εύρεση δεύτερου συναισθήματος για τις ακολουθιακές προτάσεις. Ακόμα ο αλγόριθμος εξόρυξης κανόνων δεν χρησιμοποιεί την τεχνική των πολλαπλών minimum support. Όμως, η διάσπαση του σετ δεδομένων σε προτάσεις και η χρήση του lexicon based method για την εύρεση δύο συναισθημάτων έναντι ενός, μαζί με την σωστή αντιστοίχιση των συναισθημάτων των σετ δεδομένων με αυτών του λεξικού, καλύπτει αυτά τα προβλήματα. Ακόμα, στην υλοποίηση μου, υπάρχουν πέντε διαφορετικές τάξεις συναισθημάτων έναντι των επτά της αρχικής προσέγγισης. Τέλος, ο τελικός ταξινομητής SVM που χρησιμοποιείται ως μοντέλο εκπαίδευσης και εξόρυξης των συναισθημάτων, καθώς και οι τεχνικές προετοιμασίας των δεδομένων, έχουν βελτιωθεί πολύ σε σχέση με αυτά του 2014(η χρονολογία που υλοποιήθηκε η προσέγγιση των Wen και Xiaojun Wa). Για αυτούς τους λόγους είναι καλύτερη η απόδοση αυτής της υλοποίησης.
Η υλοποίηση για την ανάλυση συναισθήματος πάνω σε κείμενα μικρού μήκους μπορεί να εξελιχθεί με πολλούς τρόπους. Οι πρώτοι και πιο εμφανείς είναι η χρήση της τεχνικής των

πολλαπλών minimum support στον αλγόριθμο του CSR καθώς και η συλλογή μεγαλύτερων σετ δεδομένων, με το σετ εκπαίδευσης να περιέχει ετικέτες συναισθήματος για κάθε πρόταση.

Άλλος ένας τρόπος βελτίωσης είναι η μελέτη του προβλήματος της ανάλυσης συναισθημάτων πάνω στο επίπεδο οντότητας και ανάλυσης. Μια τέτοια τεχνική είναι η ανάλυση του λόγου ή μελέτη του λόγου (Discourse analysis). Τα αντικείμενα της ανάλυσης του λόγου ορίζονται με διαφορετικό τρόπο από την άποψη συνεκτικών ακολουθιών προτάσεων ή ομιλίας . Σε αντίθεση με την πλειονότητα της παραδοσιακής γλωσσολογίας, οι αναλυτές λόγου όχι μόνο μελετούν τη χρήση της γλώσσας «πέρα από το όριο της φράσης» αλλά προτιμούν να αναλύουν τη «γλωσσική χρήση» που δεν έχει επινοηθεί. Η ουσιαστική διαφορά μεταξύ της ανάλυσης του λόγου και της γλωσσολογίας του κειμένου είναι ότι η ανάλυση του λόγου αποσκοπεί στην αποκάλυψη κοινωνικο-ψυχολογικών χαρακτηριστικών ενός ατόμου/ατόμων και όχι δομή κειμένου.

Τέλος μπορούν να προστεθούν και άλλα διαφορετικά συναισθήματα στα λεξικά και στα σετ δεδομένων, όπως αυτό του ουδέτερου και έτσι η υλοποίηση θα προβλέπει ακόμα περισσότερα συναισθήματα. Αυτός ο τρόπος δεν θα βελτιώσει απαραίτητα το ποσοστό επιτυχίας του αλγορίθμου, όμως θα τον κάνει πολύ πιο χρήσιμο.

Ο σχεδιασμός ενός μοντέλου με όλες τις παραπάνω βελτιώσεις θα ήταν μια ιδιαίτερη πρόκληση.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
Class Sequential Rules	Ακολουθιακοί κανόνες κλάσεων
Natural Language Processing	Επεξεργασία φυσικής γλώσσας
Blogs	Ιστολόγια
Dataset	Σετ δεδομένων
Label	Ετικέτα
Feature Extraction	Εξαγωγή χαρακτηριστικών
Overfitting	Υπερπροσαρμογή
Underfitting	Υποπροσαρμογή
Low/High Bias	Χαμηλή/Υψηλή μεροριψία
Low/Hight Variance	Χαμηλή/Υψηλή διακύμανση
Neural Networks	Νευρονικά δίκτυα
Maximum Entropy	Μέγιστη εντροπία
Hyperplane	Υπερπλαίσιο
Support Vector Machine	Μηχανή υποστήριξης διανυσμάτων
Kernel	Πυρήνας
Semantic Orientation	Σημασιολογικός Προσανατολισμός
Blog Post	Ανάρτηση Ιστολογίου
Clustering	Συσταδοποίηση/Συσσωμάτωση
Microblog text	Κείμενο μικρού μήκους/μικρο-ιστολογίου
Emoticons	Προσωπάκια/Εικονίδια
Association Rules Mining	Εξόρυξη κανόνων διαδύνδεσης
Support	Υποστήριξη
Confidence	Αυτοπεποίθηση
Itemset	Σετ αντικειμένων
Class label	Ετικέτα κλάσεων
Sequence	Ακολουθία
Subsequence	Υπο-ακολουθία
Sequential patterns	Ακολουθιακά μοτίβα
Covers	Καλύπτει
Satisfy	Ικανοποιεί
Training set	Σετ εκπαίδευσης
Test set	Σετ ελέγχου
Cross-Validation set	Σετ ενδιάμεσου ελέγχου
Ruleitem	Αντικείμενο κανόνα
Frequent	Συχνό
Candidates	Υποψήφιο
Lexicon	λεξικό
Discourse Analysis	Ανάλυση του λόγου
Condition	Συνθήκη

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

CSR	Class Sequential Rules
NLP	Natural Language Processing
SVM	Support Vector Machine
CAR	Class Association Rules
CV	Cross-Validation set
SO	Semantic Orientation
Minsup	Minimum support
Minconf	Minimum confidence
Condset	Condition set

ΑΝΑΦΟΡΕΣ

[1] Pang, B., Lee, L., & Vaithyanathan, S. 2002, July. Thumbs up?: *sentiment classification using machine learning techniques*. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume
10 (pp. 79-86). Association for Computational Linguistics.
[2] Mishne, G. 2005, August. *Experiments with mood classification in blog posts*. In Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access (Vol. 19).
[3] Go, A., Bhayani, R., & Huang, L. 2009. *Twitter sentiment classification using distant supervision*. CS224N Project. Report, Stanford, 1-12.
[4] Xiaojun Wan, Shiyang Wen, 2014, Emotion Classification in Microblog Texts Using Class Sequential Rules, Proc. of 28th AAAI Conference on Artificial Intelligence(AAAI'14),(p. 187-193). Institute of Computer Science and Technology, The MOE Key Laboratory of Computational Linguistics.
[5] benchmark dataset from the 2013 Chinese Microblog Sentiment Analysis Evaluation (CMSAE); http://tcci.ccf.org.cn/conference/2013/pages/page04_eva.html [Προσπελάστηκε 15/10/2019]
[6] competition datasets from the competition Sentiment Analysis: Emotion in Text; Identify emotion in text using sentiment analysis(Kaggle);
https://www.kaggle.com/c/sa-emotions/data[Προσπελάστηκε 15/10/2019]
[7] Saif Mohammad, NRC Word-Emotion Association Lexicon(aka EmoLex), Version 0.92(10 July 2011), 2011 National Research Council Canada;
https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm [Προσπελάστηκε 15/10/2019]
[8] Το πρόγραμμα της υλοποίησης; https://github.com/KonstantinosMaragkos/Emotion-Classification-in-Microblog-Texts-Using-Class-Sequential-Rules