ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Τμήμα Πληροφορικής και Τηλεματικής

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάπτυξη online υπηρεσίας

αποσαφήνισης λέξεων στο πλαίσιο προτάσεων

Κατάκης Ιωάννης Μανούσος

Αθήνα

Φεβρουάριος 2014

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάπτυξη online υπηρεσίας αποσαφήνισης λέξεων στο πλαίσιο προτάσεων

Κατάκης Ιωάννης Μανούσος

Α.Μ.: 20912

Επιβλέπων:

Βαρλάμης Ηρακλής, Επίκουρος Καθηγητής

Μέλη τριμελούς επιτροπής:

Μιχαήλ Δημήτριος, Λέκτορας

Τσερπές Κωνσταντίνος, Λέκτορας

Περίληψη

Τα τελευταία χρόνια, η αλματώδης εξέλιξη του παγκόσμιου ιστού άλλαξε τον τρόπο που οι άνθρωποι αλληλεπιδρούν. Η έλευση των υπηρεσιών του Web 2.0 (ιστότοποι κοινωνικής δικτύωσης, ιστολόγια ,wiki) ώθησαν την παραγωγή περιεχομένου. Το διαδίκτυο πλέον αποτελεί μια τεραστίων διαστάσεων αποθήκη πληροφοριών. Ωστόσο ο συνεχώς αυξανόμενος όγκος πληροφοριών που δημιουργείται καθημερινά από τους χρήστες (User Generated Content) και δημοσιεύεται στο διαδίκτυο καθιστά σχεδόν αδύνατη την επεξεργασία τους και την εξαγωγή γνώσης από αυτές.

Η παρούσα πτυχιακή εργασία αποσκοπεί στην δημιουργία μιας διαδικτυακής υπηρεσίας ανάλυσης συναισθήματος και αποσαφήνισης εννοιών των λέξεων ενός κειμένου. Πιο συγκεκριμένα, η υπηρεσία εφαρμόζει αλγορίθμους αποσαφήνισης εννοιών (WSD) προκειμένου να καθορίσει την έννοια των λέξεων του κειμένου εισόδου. Στη συνέχεια, αναθέτει σε κάθε λέξη του κείμενου την συναισθηματική πολικότητα που της αναλογεί σύμφωνα με ένα λεξικό συναισθημάτων. Τέλος, αναθέτει σε κάθε πρόταση του κειμένου μια καθολική συναισθηματική πολικότητα χρησιμοποιώντας ένα μοντέλο μηχανικής μάθησης.

Λέξεις Κλειδιά: Αποσαφήνιση της έννοιας των λέξεων, Κατηγοριοποίηση συναισθήματος σε επίπεδο έννοιας, Κατηγοριοποίηση συναισθήματος σε επίπεδο πρότασης, Υπηρεσία ιστού, Διεπαφή προγραμματισμού εφαρμογών.

Abstract

In recent years, the rapid development of the Web has changed the way people interact. The advent of Web 2.0 services (social networking sites, blogs, wiki) induced the generation of information. Internet is now an enormous data warehouse. However, the increasing amount of information generated every day by the users (User Generated Content) and published on the internet, makes it almost impossible to edit them and extract knowledge from them.

This thesis aims to create a word sense disambiguation (WSD) web service with sentiment analysis capabilities. More specifically, the service applies word sense disambiguation algorithms in order to determine the meaning of words in the input text. Then, it uses a sentiment lexicon in order to identify the sentiment polarity of each word. Finally, employs a machine learning model for sentence-level sentiment classification.

Keywords: word sense disambiguation, sense-level sentiment classification, sentence-level sentiment classification, web service, API.

Περιεχόμενα

Κεφάλαιο 1 Εισαγωγή

1.1 Γενικά

Τα τελευταία χρόνια, η αλματώδης εξέλιξη του παγκόσμιου ιστού άλλαξε τον τρόπο που οι άνθρωποι αλληλεπιδρούν. Η έλευση των υπηρεσιών του Web 2.0 (ιστότοποι κοινωνικής δικτύωσης, ιστολόγια ,wiki) έδωσε τη δυνατότητα στους χρήστες του διαδικτύου να δημιουργήσουν το δικό τους περιεχόμενο και να ανταλλάξουν απόψεις πάνω σε ένα ευρύ φάσμα θεμάτων.

Το διαδίκτυο πλέον αποτελεί μια τεραστίων διαστάσεων αποθήκη πληροφοριών. Ωστόσο ο συνεχώς αυξανόμενος όγκος πληροφοριών που δημιουργείται καθημερινά από τους χρήστες (User Generated Content) και δημοσιεύεται στο διαδίκτυο καθιστά σχεδόν αδύνατη την επεξεργασία τους και την εξαγωγή γνώσης από αυτές. Τη λύση στο πρόβλημα ήρθε να δώσει ένα νέο σχετικά πεδίο έρευνας που ονομάζεται Ανάλυση Συναισθήματος (Sentiment Analysis).

Με τον όρο Ανάλυση Συναισθήματος ή Εξόρυξη Γνώμης αναφερόμαστε στην αυτόματη εξαγωγή γνώσης από κείμενα που περιέχουν υποκειμενική πληροφορία. Πρόκειται για μια αυτοματοποιημένη διαδικασία που σκοπό έχει τον προσδιορισμό της στάσης, της άποψης και των συναισθημάτων του ομιλητή – συγγραφέα γύρω από ένα συγκεκριμένο αντικείμενο συζήτησης.

Η τεράστια αξία της Ανάλυσης Συναισθήματος σε πρακτικές εφαρμογές, οδήγησε σε μια εκρηκτική αύξηση τόσο της έρευνας στον ακαδημαϊκό χώρο όσο και των εφαρμογών στον τομέα της βιομηχανίας (Liu, 2012). Όλο και περισσότερες επιχειρήσεις χρησιμοποιούν την Εξόρυξη Γνώμης προκειμένου να παρακολουθήσουν την άποψη που έχει το αγοραστικό κοινό σχετικά με τα προϊόντα τους. Με αυτό τον τρόπο μπορούν να μετρήσουν την συνολική τους απόδοση, ιδίως στην διαδικτυακή αγορά, όπως και να βελτιώσουν τις στρατηγικές μάρκετινγκ που χρησιμοποιούν. Εκτός αυτού, σημαντική είναι η επίδραση του συγκεκριμένου τομέα στις πολιτικές και κοινωνικές επιστήμες καθώς είναι ένα εργαλείο που συμβάλλει στην λήψη αποφάσεων και όχι μόνο.

Η ταξινόμηση συναισθήματος μπορεί να εφαρμοστεί σε επίπεδο εγγράφου (document-level), πρότασης (sentence-level) ή χαρακτηριστικού (feature-level). Όταν η ταξινόμηση πραγματοποιείται σε επίπεδο εγγράφου γίνεται η παραδοχή ότι κάθε κείμενο εκφράζει τις απόψεις ενός μόνο ατόμου γύρω από ένα συγκεκριμένο αντικείμενο ενώ όταν πρόκειται για την ταξινόμηση μιας πρότασης θεωρείται ότι αυτή περιέχει μόνο μια άποψη. Τέλος η ταξινόμηση σε επίπεδο χαρακτηριστικού βασίζεται στην γενική

ιδέα ότι μια άποψη μπορεί να αναφέρεται στα χαρακτηριστικά μιας οντότητας είτε συνολικά είτε ξεχωριστά σε κάθε ένα από αυτά με διαφορετικό συναίσθημα (Jagtap και Pawar, 2013).

Προκειμένου να εφαρμοστούν τα παραπάνω επίπεδα ταξινόμησης συναισθήματος χρησιμοποιούνται προσεγγίσεις επιβλεπόμενης ή μη επιβλεπόμενης μάθησης που μπορούν να χωριστούν σε 2 κύριες κατηγορίες: σημασιολογικού προσανατολισμού και μηχανικής μάθησης. «Στην προσέγγιση σημασιολογικού προσανατολισμού ένα κείμενο ταξινομείται βάσει της μέσης πολικότητας των λέξεων/φράσεων που περιέχουν θετικό ή αρνητικό συναίσθημα χρησιμοποιώντας λεξικά συναισθημάτων (sentiment lexicons) και γλωσσικούς κανόνες (linguistic rules)» (Jalilvand και Salim, 2012).

Τα λεξικά συναισθημάτων αποτελούνται από έναν κατάλογο λέξεων γνώμης (opinion words) και την πολικότητάς τους (θετική, αρνητική, ουδέτερη). Λέξεις γνώμης αποκαλούνται οι λέξεις εκείνες που συχνά χρησιμοποιούνται για να εκφράσουν θετικό ή αρνητικό συναίσθημα (Liu, 2012).

Στην πραγματικότητα η παραπάνω προσέγγιση εφαρμόζεται σε επίπεδο λέξης (word-level) με σκοπό την επίτευξη ταξινόμησης επιπέδου πρότασης/εγγράφου. Ένα μειονέκτημα αυτών των μεθόδων είναι ότι δεν λαμβάνουν υπόψη τους το πλαίσιο στο οποίο εμφανίζονται οι λέξεις. Για παράδειγμα, το συναίσθημα της λέξης «break» στην πρόταση «He broke the 200m national record again.» είναι θετικό, ενώ στην πρόταση «My car broke down on the way to the interview!» είναι αρνητικό. Οι λέξεις έχουν διαφορετική έννοια ανάμεσα σε διαφορετικά συμφραζόμενα όπως και διαφορετική συναισθηματική πολικότητα (Jalilvand και Salim, 2012). Έτσι προκύπτει ότι μια καλύτερη προσέγγιση είναι η ταξινόμηση συναισθήματος σε επίπεδο έννοιας.

1.2 Στόχος

Ο στόχος της παρούσας πτυχιακής εργασίας είναι η ανάπτυξη μιας διαδικτυακής υπηρεσίας ανάλυσης συναισθήματος. Η προσέγγιση που ακολουθήθηκε αποσκοπεί στην ταξινόμηση συναισθήματος σε επίπεδο έννοιας καθώς και σε επίπεδο πρότασης. Πιο συγκεκριμένα, η υπηρεσία αναθέτει σε κάθε λέξη του εισηγμένου κείμενου την συναισθηματική πολικότητα που της αναλογεί σύμφωνα με ένα λεξικό συναισθημάτων, αφού πρώτα έχει ολοκληρώσει την αποσαφήνιση των όρων κάθε πρότασης χρησιμοποιώντας τη μέθοδο WSD που έχει επιλέξει ο χρήστης.

1.3 Συνεισφορά

Στην παρούσα πτυχιακή επιχειρείται η συναισθηματική ταξινόμηση πρώτα σε επίπεδο έννοιας και στη συνέχεια σε επίπεδο πρότασης. Κατά την εκπόνηση της, αναπτύχθηκε μια προσέγγιση η οποία στηρίζεται στη χρήση αλγορίθμων αποσαφήνισης εννοιών (Word Sense Disambiguation – WSD) και ενός λεξικού συναισθημάτων που παρέχει ξεχωριστή πολικότητα για κάθε έννοια μιας λέξης. Αρχικά, γίνεται μια προεπεξεργασία στο κείμενο εισόδου και στη συνέχεια εφαρμόζεται ένας αλγόριθμος WSD με σκοπό τον εντοπισμό της σωστής σημασίας κάθε λέξης. Η χρήση ενός αλγόριθμου WSD είναι πολύτιμη, καθώς αντιμετωπίζει το πρόβλημα της πολυσημίας των λέξεων, όπως για παράδειγμα της λέξης «bank», όπου

στην πρόταση «This is the largest bank in the world.» σημαίνει τράπεζα ενώ στην πρόταση «We walked along the bank of the river.» σημαίνει όχθη. Στη συνέχεια, ανατίθεται μια πολικότητα σε κάθε λέξη βάσει του λεξικού συναισθημάτων. Για παράδειγμα στην πρόταση «the wine is tasty but the food is bad», στην λέξη «tasty» αποδίδεται η έννοια «εύγευστο» με θετική πολικότητα ενώ στη λέξη «bad» αποδίδεται η έννοια «κακός» με έντονη αρνητική πολικότητα. Τελικά το σύστημα αποδίδει στην πρόταση μια καθολική πολικότητα βάσει των επιμέρους, χρησιμοποιώντας ένα μοντέλο μηχανικής μάθησης (Deep Learning Model) που έχει αναπτυχθεί από το Πανεπιστήμιο του Stanford.

Όλα τα παραπάνω έχουν υλοποιηθεί πίσω από ένα Web Service και μια εύχρηστη διεπαφή χρήστη που είναι διαθέσιμα στην διεύθυνση http://omiotis.hua.gr/pythia.

Κεφάλαιο 2 Θεωρητικό υπόβαθρο

2.1 Ανάλυση συναισθήματος

2.1.1 Διαφορετικά επίπεδα ανάλυσης συναισθήματος

Οι ερευνητικές προσεγγίσεις που έχουν γίνει στο πρόβλημα της ταξινόμησης συναισθήματος κατηγοριοποιούνται ανάλογα με το επίπεδο εφαρμογής τους.

Μια από αυτές είναι η ταξινόμηση σε επίπεδο εγγράφου. Αυτή η προσέγγιση θεωρεί ότι κάθε έγγραφο περιέχει τις απόψεις ενός μόνο ατόμου γύρω από ένα συγκεκριμένο θέμα και έχει ως στόχο να χαρακτηρίσει το συναίσθημα που εκφράζεται μέσα από το κείμενο ως θετικό η αρνητικό. Οι περισσότερες τεχνικές ανάλυσης συναισθήματος εγγράφων είναι επιβλεπόμενης μάθησης, ωστόσο υπάρχουν και τεχνικές μη επιβλεπόμενης μάθησης.

Η ταξινόμηση συναισθήματος είναι ουσιαστικά ένα πρόβλημα ταξινόμησης κειμένου. Έτσι, οποιαδήποτε υπάρχουσα μέθοδος επιβλεπόμενης μάθησης (π.χ. Ταξινόμηση naïve Bayes, Μηχανισμοι Διανύσματος υποστήριξης (Support Vector Machines – SVM)) μπορεί να εφαρμοστεί (Joachims, 1999; Shawe-Taylor και Cristianini, 2000) (Liu, 2012). Οι Pang κ.ά. (2002) ήταν οι πρώτοι που βασίστηκαν στην παραπάνω προσέγγιση και διαχώρισαν κριτικές ταινιών σε θετικές και αρνητικές. Έδειξαν ότι η χρήση μονογραμμάτων (unigrams) ως γνωρισμάτων (χαρακτηριστικών) για την ταξινόμηση έχει αρκετά καλά αποτελέσματα είτε με τον naïve Bayes είτε με τα SVM, ενώ παράλληλα δοκίμασαν και εναλλακτικές επιλογές χαρακτηριστικών.

Όπως και σε άλλες εφαρμογές επιβλεπόμενης μηχανικής μάθησης, το κλειδί για την ταξινόμηση συναισθήματος είναι η δημιουργία/εξαγωγή ενός συνόλου αποτελεσματικών χαρακτηριστικών (Liu, 2012). Μερικά από αυτά τα χαρακτηριστικά είναι τα ν-γράμματα μαζί με την συχνότητα εμφάνισης τους, το μέρος του λόγου κάθε λέξης, οι λέξεις και οι εκφράσεις που χρησιμοποιούνται για την έκφραση συναισθημάτων καθώς και οι εκφράσεις που χρησιμεύουν στην αλλαγή της συναισθηματικής πολικότητας.

Πέρα από τις προκαθορισμένες μεθόδους μηχανικής μάθησης, οι ερευνητές έχουν προτείνει αρκετές τεχνικές ειδικά προσαρμοσμένες στην επίλυση του προβλήματος της κατηγοριοποίησης συναισθήματος. Ο Gamon (2004) έκανε συναισθηματική ανάλυση σε δεδομένα από σχόλια πελατών εκπαιδεύοντας γραμμικά SVM με μεγάλα διανύσματα χαρακτηριστικών (large feature vectors). Στη συνέχεια οι Pang και Lee (2004) δημιούργησαν ένα γράφο μεταξύ των προτάσεων ενός κειμένου και εφάρμοσαν τον αλγόριθμο ελαχίστων κοψιμάτων (minimum cut), αφαιρώντας τις ακμές με το μικρότερο υποκειμενικό φορτίο, ώστε

να διευκολύνουν την διαδικασία της κατηγοριοποίησης, η οποία έγινε με συνήθεις τεχνικές μηχανικής μάθησης. Οι Matsumoto κ.ά. (2005) χρησιμοποίησαν τις συντακτικές σχέσεις που υπήρχαν μεταξύ των λέξεων ως χαρακτηριστικά του SVM ενώ οι Kennedy και Inkpen (2006) εξέτασαν την επίδραση των λέξεων που μπορούν να μεταλλάξουν το συναισθηματικό φορτίο (π.χ. δείκτες άρνησης) στην κατηγοριοποίηση κριτικών από ταινίες. Οι Li κ.ά. (2009) πρότειναν ένα μοντέλο που βασίζεται σε μια μη αρνητική μέθοδο παραγοντοποίησης μητρών. Οι Dasgupta και Ng (2009) εξέτασαν μια μέθοδο ημι-επιβλεπόμενης μάθησης, η οποία πρώτα εξήγαγε τις σαφείς κριτικές των πελατών με χρήση φασματικών τεχνικών και στη συνέχεια τις χρησιμοποιούσε ούτως ώστε να κατηγοριοποιήσει τις ασαφείς κριτικές συνδυάζοντας μεθόδους μάθησης που είτε αξιοποιούν την ανάδραση του χρήστη (ενεργητικές - active), είτε αποφασίζουν για τα άγνωστα δείγματα χωρίς να δημιουργήσουν ενδιάμεσο μοντέλο (μεταβιβαστικές - transductive) είτε συνδυάζουν τα αποτελέσματα επιμέρους μεθόδων (συνδυαστικές - ensemble). Το μοντέλο που προτάθηκε από τους Qiu κ.ά. (2009) περιλαμβάνει σε πρώτη φάση μια επαναληπτική διαδικασία κατηγοριοποίησης κριτικών σύμφωνα με ένα λεξικό συναισθημάτων και σε δεύτερη φάση την εκπαίδευση ενός ταξινομητή επιβλεπόμενης μάθησης με κάποια από τα δεδομένα της πρώτης φάσης. Οι Yessenalina κ.ά. (2010) χρησιμοποίησαν πολυεπίπεδα δομημένα μοντέλα για την εξαγωγή των χρήσιμων (π.χ. υποκειμενικών) προτάσεων ενός κειμένου καθώς και την πρόβλεψη του συναισθήματος αυτού, βάσει των προτάσεων που έχουν εξαχθεί. Η προσέγγιση των Wang κ.ά. (2011) προτείνει μια μέθοδο ανάλυσης συναισθήματος του περιεχομένου ενός Twitter hashtag για μια δεδομένη χρονική περίοδο, η οποία βασίζεται στους γράφους.

Αντίθετα, ο Turney (2002) πρότεινε μια μέθοδο μη επιβλεπόμενης μάθησης η οποία βασίζεται σε σταθερά συντακτικά μοντέλα, τα οποία είναι πιθανό να χρησιμοποιηθούν για την έκφραση απόψεων. Η κατηγοριοποίηση συναισθήματος γίνεται σύμφωνα με τον μέσο σημασιολογικό προσανατολισμό των φράσεων που περιέχουν επίθετα ή επιρρήματα. Η μέθοδος των Taboada κ.ά. (2011) χρησιμοποιεί λεξικά με λέξεις και φράσεις στις οποίες έχει προσημειωθεί η σημασιολογική πολικότητα (semantic polarity) και ισχύς (strength) και ενσωματώνει επίταση και άρνηση με σκοπό τον υπολογισμό μιας βαθμολογίας συναισθήματος για κάθε έγγραφο.

Μια άλλη προσέγγιση είναι η ταξινόμηση σε επίπεδο πρότασης. Σε αυτή την προσέγγιση γίνεται η παραδοχή ότι υπάρχει μόνο μια άποψη μέσα σε κάθε πρόταση. Το πρόβλημα της ταξινόμησης συναισθήματος αυτού του επιπέδου μπορεί να αντιμετωπιστεί είτε ως ένα πρόβλημα ταξινόμησης τριών κλάσεων είτε ως δύο ξεχωριστά προβλήματα ταξινόμησης.

Στην πρώτη περίπτωση, οι προτάσεις χαρακτηρίζονται απευθείας ως θετικές, αρνητικές ή ουδέτερες. Στην δεύτερη περίπτωση οι προτάσεις αρχικά διαχωρίζονται σύμφωνα με το αν εκφράζουν κάποια άποψη ή όχι (ταξινόμηση υποκειμενικότητας) και στη συνέχεια αυτές οι οποίες περιέχουν κάποια στοιχεία υποκειμενικής πληροφορίας ταξινομούνται ως θετικές ή αρνητικές.

Η ταξινόμηση υποκειμενικότητας ταξινομεί τις προτάσεις σε δύο κλάσεις, υποκειμενικές και αντικειμενικές. Αντικειμενικές ονομάζονται οι προτάσεις οι οποίες μεταφέρουν πληροφορίες με τρόπο αμερόληπτο, ενώ υποκειμενικές αυτές οι οποίες διατυπώνουν προσωπικές απόψεις (Liu, 2012). Προκειμέ-

νου να επιτευχθεί η απομάκρυνση των μη συναισθηματικά φορτισμένων προτάσεων οι Wiebe κ.ά. (1999) χρησιμοποίησαν ένα ταξινομητή naïve Bayes με ένα σύνολο δυαδικών χαρακτηριστικών όπως η παρουσία μιας αντωνυμίας ή ενός επιθέτου στην πρόταση. Στη συνέχεια, ο Wiebe (2000) πρότεινε μια μέθοδο μη επιβλεπόμενης μάθησης, η οποία βασιζόταν στην ύπαρξη ή όχι υποκειμενικών εκφράσεων με σκοπό τον καθορισμό της υποκειμενικότητας της κάθε πρότασης (Liu, 2012). Οι Yu και Hatzivassiloglou (2003) πραγματοποίησαν ταξινομήσεις υποκειμενικότητας χρησιμοποιώντας ένα ταξινομητή naïve Bayes και την ομοιότητα των προτάσεων. Η ομοιότητα των προτάσεων υπολογίζονταν από το σύστημα SIMFINDER (Hatzivassiloglou κ.ά., 2001), βάσει των κοινών τους λέξεων και φράσεων καθώς και των συνωνύμων του WordNet. Οι Raaijmakers και Kraaij (2008) έδειξαν ότι τα ν-γράμματα χαρακτήρων (μέρη λέξεων) είχαν καλύτερη απόδοση στην κατηγοριοποίηση υποκειμενικότητας σε σχέση με τα ν-γράμματα λέξεων και φωνημάτων. Τέλος, οι Benamara κ.α. (2011) αφού πραγματοποίησαν κατηγοριοποίηση υποκειμενικότητας με 4 κλάσεις, έδειξαν ότι μια υποκειμενική πρόταση μπορεί να μην είναι αξιολογήσιμη (όχι θετικό ή αρνητικό συναίσθημα) και ότι μια αντικειμενική πρόταση μπορεί να εκφράζει συναίσθημα.

Αν μια πρόταση κατηγοριοποιηθεί ως υποκειμενική, ακολουθεί ο χαρακτηρισμός της ως θετική ή αρνητική. Για να συμβεί αυτό, εφαρμόζονται τεχνικές επιβλεπόμενης μάθησης όπως ακριβώς και στην ταξινόμηση συναισθήματος εγγράφων καθώς και τεχνικές οι οποίες βασίζονται σε λεξικά συναισθήματος.

Οι Yu και Hatzivassiloglou (2003) πρότειναν μια μέθοδο ή οποία κατηγοριοποιούσε υποκειμενικές προτάσεις και βασιζόταν σε ένα αρχικό σύνολο επιθέτων. Υπολόγισαν την συναισθηματική πολικότητα των προτάσεων χρησιμοποιώντας τη μέση βαθμολογία του λογαρίθμου πιθανοφάνειας (log-likelihood) των λέξεων (επίθετα, επιρρήματα, ρήματα, ουσιαστικά) κάθε πρότασης και στη συνέχεια επίλεξαν δυο κατώφλια βάσει των δεδομένων εκπαίδευσης προκειμένου να αποφανθούν για το αν η πρόταση είναι θετική, αρνητική ή ουδέτερη. Οι Hu και Liu (2004) πρότειναν μια μέθοδο που στηριζόταν σε ένα λεξικό συναισθήματος το οποίο δημιουργήθηκε χρησιμοποιώντας μια επαναληπτική διαδικασία με κάποιες θετικές και αρνητικές λέξεις ως σπόρους (seeds) και τις σχέσεις συνωνύμων και αντωνύμων του WordNet. Παρόλο που η μέθοδος αυτή πραγματοποιεί ταξινόμηση σε επίπεδο χαρακτηριστικού, μπορεί επίσης να λειτουργήσει και σε επίπεδο πρότασης καθώς αθροίζει τις βαθμολογίες των συναισθηματικά φορτισμένων λέξεων κάθε πρότασης. Οι McDonald κ.ά. (2007) παρουσίασαν ένα δομημένο μοντέλο για την από κοινού κατηγοριοποίηση συναισθήματος σε διαφορετικά επίπεδα λεπτομέρειας (granularity). Έτσι για παράδειγμα, σε κάθε πρόταση των δεδομένων εκπαίδευσης αποδίδεται ένα συναίσθημα και το ίδιο συμβαίνει στο κείμενο που περιέχει τις προτάσεις αυτές. Με τον τρόπο αυτό επιτυγχάνεται αύξηση της ακριβείας και στα δυο επίπεδα ταξινόμησης. Οι Davidov κ.ά. (2010) πραγματοποίησαν ταξινόμηση συναισθήματος σε δημοσιεύσεις του Twitter με χρήση δημοφιλών χαρακτηριστικών σε τεχνικές ταξινόμησης συναισθήματος γενικών κειμένων καθώς και με χρήση hashtags, smileys, σημείων στίξεως, συνδυασμών αυτών και συχνότητας εμφάνισής τους, αποφέροντας αρκετά ικανοποιητικά αποτελέσματα.

Τέλος, μια τρίτη προσέγγιση είναι η ταξινόμηση σε επίπεδο λέξης. Η προσέγγιση αυτή ουσιαστικά χρησιμοποιείται για ταξινόμηση επιπέδου πρότασης ή κειμένου και βασίζεται στην παραδοχή ότι οι πιο

σημαντικοί δείκτες συναισθημάτων είναι οι λέξεις γνώμης. Μια λίστα από τέτοιες λέξεις ονομάζεται λεξικό συναισθημάτων (Liu, 2012).

Για την δημιουργία λεξικών συναισθημάτων χρησιμοποιούνται πληροφορίες που προκύπτουν από την επεξεργασία είτε μεγάλων σωμάτων κειμένου(text corpora), είτε γλωσσολογικών πόρων όπως θησαυροί και λεξικά, με σκοπό την επέκταση μιας αρχικής λίστας με λέξεις γνώμης (seed words).

Στα λεξικά που προέρχονται από σώματα κειμένου, η επέκταση της λίστας αυτής, μπορεί να γίνει με χρήση συντακτικών μοτίβων, τα οποία ικανοποιούνται μέσα σε αυτά τα κείμενα. Ένας άλλος τρόπος είναι με τη χρήση πληροφοριών που προκύπτουν από τη συχνότητα διάφορων μοτίβων από λέξεις (Turney, 2002).

Αντίθετα, τα λεξικά που βασίζονται σε γλωσσολογικούς πόρους προσπαθούν να πραγματοποιήσουν αυτή την επέκταση χρησιμοποιώντας τα συνώνυμα, τα αντώνυμα και την ιεραρχία αυτών των λέξεων μέσα σε γλωσσολογικούς θησαυρούς όπως το WordNet. Έτσι οι Kim και Hovy (2004) χρησιμοποίησαν τις σχέσεις συνωνύμων και αντωνύμων του WordNet προκειμένου να επεκτείνουν το αρχικό σύνολο υποκειμενικών λέξεων. Αντίστοιχα, οι Hu και Liu (2004) χρησιμοποίησαν τις ίδιες σχέσεις και το θησαυρό WordNet για ένα αρχικό σύνολο 30 επιθέτων, ενώ αργότερα οι Esuli και Sebastiani (2011) βρήκαν το συναισθηματικό φορτίο για κάθε διαφορετική έννοια μιας λέξης αξιοποιώντας την ερμηνεία (gloss) της λέξης όπως αυτή δίνεται από το WordNet και ένα αρχικό σύνολο από υποκειμενικές λέξεις.

Χαρακτηριστικά παραδείγματα λεξικών συναισθημάτων είναι το Harvard General Inquirer, το Linguistic Inquiry and Word Counts (LIWC), το Bing Liu's Opinion Lexicon, το MPQA Subjectivity Lexicon και το SentiWordNet. Το Harvard General Inquirer είναι ένα λεξικό που επισυνάπτει συντακτικές, σημασιολογικές και πραγματικές πληροφορίες σε λέξεις με επισημείωση σχετικά με το μέρος του λόγου στο οποίο ανήκουν (π.χ. ουσιαστικό, ρήμα κτλ.). Το LIWC είναι μια εμπορικά διαθέσιμη βάση δεδομένων που περιέχει ένα μεγάλο αριθμό κατηγοριοποιημένων κανονικών εκφράσεων. Το Bing Liu's Opinion Lexicon αποτελείται από 6789 λέξεις οι οποίες είναι χωρισμένες σε θετικές και αρνητικές και βασίζεται στην προσέγγιση των Hu και Liu (2004) όπως αυτή περιγράφηκε στην προηγούμενη παράγραφο. Το MPQA Subjectivity Lexicon επεκτείνει μια λίστα από στοιχεία (λέξεις) υποκειμενικότητας (subjectivity clues) των Riloff και Wiebe (2003) με χρήση γλωσσολογικών πόρων, οι οποίες αφού πρώτα ομαδοποιηθούν ανάλογα με την αξιοπιστία (reliability) τους, χαρακτηρίζονται ως θετικές, αρνητικές, ουδέτερες ή ως θετικές και αρνητικές ταυτόχρονα. Τέλος, το SentiWordNet, το οποίο στηρίζεται όπως είναι προφανές στο WordNet, διαχωρίζει τη συναισθηματική πολικότητα κάθε έννοιας μιας λέξης σε 3 κατηγορίες: θετική, αρνητική και ουδέτερη, σύμφωνα με την προαναφερθείσα μέθοδο των Esuli και Sebastiani (2011).

2.1.2 Ταξινόμηση σε επίπεδο χαρακτηριστικού

Μερικές φορές η ταξινόμηση σε επίπεδο κειμένου ή πρότασης δεν είναι επαρκής για κάποιες εφαρμογές. Αυτό συμβαίνει διότι δεν είναι εφικτός ούτε ο εντοπισμός των μεταβλητών (opinion targets)

στις οποίες αναφέρεται μια γνώμη, ούτε και η ανάθεση ενός ξεχωριστού συναισθήματος σε κάθε μια από αυτές. Εκτός των παραπάνω, αν υποθέσουμε ότι ένα υποκείμενο έχει μια άποψη (θετική ή αρνητική) για μια οντότητα, δεν σημαίνει ότι θα έχει την ίδια άποψη για κάθε επιμέρους χαρακτηριστικό της (Liu, 2012). Έτσι, προέκυψε η ανάγκη για μια πιο λεπτομερή ανάλυση ούτως ώστε να είναι δυνατός ο διαχωρισμός όλων των χαρακτηριστικών στα οποία αναφέρεται μια άποψη, όπως και η εύρεση του συναισθηματικού τους φορτίου.

Η ταξινόμηση συναισθήματος σε επίπεδο χαρακτηριστικού χωρίζεται συνήθως σε δύο στάδια, στην εξαγωγή των απόψεων που περιέχονται σε ένα έγγραφο και στην ταξινόμηση συναισθήματος κάθε άποψης. Η γνώμη του συγγραφέα του εγγράφου μπορεί να διαφέρει για κάθε πλευρά (άποψη) ενός θέματος. Για παράδειγμα στην κριτική μιας φωτογραφικής μηχανής «Η ποιότητα των φωτογραφιών είναι εξαιρετική αλλά η τιμή της είναι πολύ υψηλή» ο συντάκτης εκφράζει θετική γνώμη για την ποιότητα των φωτογραφιών που τραβάει η φωτογραφική μηχανή αλλά αρνητική γνώμη για την τιμή της. Στο συγκεκριμένο παράδειγμα, ως **άποψη** ορίζεται η «ποιότητα φωτογραφιών» και η «τιμή» με αποτέλεσμα να εκφράζονται δυο γνώμες για το ίδιο αντικείμενο, μια για κάθε άποψη.

Ο στόχος του πρώτου σταδίου είναι να εξαχθούν όλες οι πτυχές του θέματος οι οποίες έχουν αξιολογηθεί από τον συντάκτη της άποψης. Για την επίτευξη αυτού του στόχου οι Hu και Liu (2004) βασίστηκαν στην διαπίστωση ότι οι άνθρωποι σχολιάζουν διαφορετικά χαρακτηριστικά μιας οντότητας με παρόμοιες λέξεις. Χρησιμοποίησαν έναν σύστημα αναγνώρισης των μερών του λόγου (Part-of-Speech Tagger) για να εξάγουν ουσιαστικά και ονοματικές φράσεις από κριτικές διαφόρων προϊόντων. Στη συνέχεια μέτρησαν την συχνότητα εμφάνισης τους και κράτησαν μόνο τα συχνά εμφανιζόμενα. Οι Blair-Goldensohn κ.ά. (2008) βελτίωσαν την παραπάνω τακτική χρησιμοποιώντας μόνο τα ουσιαστικά και τις ονοματικές φράσεις που άνηκαν σε προτάσεις με συναισθηματικό φορτίο ή σε συντακτικά μοντέλα που υποδείκνυαν την ύπαρξη συναισθήματος. Βαθμολόγησαν τις απόψεις που εξήγαγαν, υπολογίζοντας το σταθμισμένο άθροισμα της συχνότητας εμφάνισής τους σε υποκειμενικές προτάσεις, θετικές ή αρνητικές. Εκτός των παραπάνω, τη συχνότητα εμφάνισης χρησιμοποίησαν και οι Long, Zhang και Zhu (2010) προκειμένου να δημιουργήσουν ένα αρχικό σύνολο από λέξεις οι οποίες προσδιορίζουν μια άποψη (aspect words) (Liu, 2012). Έπειτα, χρησιμοποίησαν την απόσταση πληροφορίας όπως παρουσιάζεται από τους Cilibrasi και Vitanyi (2007) με σκοπό να βρουν και άλλες λέξεις οι οποίες ήταν σχετικές με μια πλευρά ενός θέματος. Το σύνολο λέξεων που προέκυψε, εφαρμόστηκε σε κριτικές ώστε να εντοπιστούν αυτές που συζητούσαν περισσότερο για ένα συγκεκριμένο χαρακτηριστικό μιας οντότητας.

Μια άλλη προσέγγιση βασίζεται στη σχέση μεταξύ μιας γνώμης και της άποψης (του χαρακτηριστικού) της οντότητας που αφορά αυτή η γνώμη. Οι Hu και Liu (2004) βρήκαν ότι οι ίδιες λέξεις χρησιμοποιούνται για να μεταβάλλουν ή να περιγράψουν διαφορετικές απόψεις ενός θέματος. Έτσι αν μια πρόταση δεν περιέχει κάποια συχνά εμφανιζόμενη άποψη αλλά έχει κάποιες λέξεις που εκφράζουν συναίσθημα, τότε αυτό αποδίδεται στο πλησιέστερο ουσιαστικό ή ονοματική φράση. Οι Zhang, Jing και Zhu

(2006) χρησιμοποίησαν ένα συντακτικό αναλυτή για σχέσεις εξάρτησης (dependency parser) προκειμένου να ανακαλύψουν σχέσεις εξάρτησης όπως οι παραπάνω.

Εκτός των προσεγγίσεων που αναφέρθηκαν, αρκετοί ήταν εκείνοι που κατέφυγαν σε λύσεις επιβλεπόμενης μάθησης για την αντιμετώπιση του προβλήματος της εξαγωγής των διαφορετικών απόψεων. Οι προσεγγίσεις αυτές βασίζονται σε μεθόδους ακολουθιακής μάθησης (sequential learning) όπως είναι τα Κρυφά Μοντέλα Markov (Hidden Markov Models - HMM) και τα Υπό Συνθήκη Τυχαία Πεδία (Conditional Random Fields - CRF). Με τον όρο Κρυφό Μοντέλο Markov αναφερόμαστε σε ένα ισχυρό στατιστικό εργαλείο που στο στοχεύει στην μοντελοποίηση συστημάτων τα οποία εικάζεται πως διέπονται από διαδικασίες Markov ενώ Υπό Συνθήκη Τυχαία Πεδία είναι μία οικογένεια διακριτών μοντέλων (discriminative model) που μπορούν να χρησιμοποιηθούν για δομημένη πρόβλεψη. Οι Jin και Ho (2009) εφάρμοσαν ένα λεξικοποιημένο μοντέλο HMM για την εκμάθηση μοτίβων που θα πραγματοποιούσε την εξαγωγή απόψεων. Οι Jakob και Gurevych (2010) εκπαίδευσαν ένα CRF με προτάσεις από κριτικές διάφορων προϊόντων με σκοπό να δημιουργήσουν ένα μοντέλο πιο γενικού σκοπού.

Τέλος, μια εναλλακτική προσέγγιση περιλαμβάνει τη χρήση μοντέλων θέματος (topic models). Πρόκειται για στατιστικά μοντέλα που στοχεύουν στην ανακάλυψη των θεμάτων που διατυπώνονται σε μια συλλογή εγγράφων. Το αποτέλεσμα της εφαρμογής των μοντέλων αυτών είναι ένα σύνολο από ομάδες λέξεων. Κάθε ομάδα αντιπροσωπεύει και ένα διαφορετικό θέμα για το οποίο γίνεται αναφορά σε ένα από τα έγγραφα της συλλογής. Τα θεματικά μοντέλα μπορούν επίσης να επεκταθούν ώστε ταυτόχρονα να παράγουν και άλλου είδους πληροφορία όπως για παράδειγμα η εξαγωγή των συναισθηματικά φορτισμένων λέξεων ενός κειμένου. Το βασικό τους μειονέκτημα είναι ότι απαιτούν τεράστια σε έκταση δεδομένα εκπαίδευσης. Δύο είναι τα βασικά μοντέλα θέματος, το pLSA (Probabilistic latent semantic analysis) και το LDA (Latent Dirichlet allocation). Οι Mei κ.ά. (2007) δημιούργησαν ένα μοντέλο το οποίο ήταν βασισμένο σε pLSA με σκοπό την εξαγωγή των απόψεων καθώς και των λέξεων συναισθήματος ενός κειμένου με χρήση πιθανοτικών μοντέλων. Αποτελούσε συνδυασμό ενός θεματικού μοντέλου, ενός μοντέλου θετικού συναισθήματος και ενός μοντέλου αρνητικού συναισθήματος. Αντίθετα, οι Brody και Elhadad (2010) χώρισαν τη διαδικασία σε δύο στάδια. Πρώτα χρησιμοποίησαν ένα θεματικό μοντέλο LDA για την εξαγωγή των απόψεων και στη συνέχεια εντόπισαν τις λέξεις συναισθήματος χρησιμοποιώντας μόνο τα επίθετα του κειμένου. Οι Zhao κ.ά. (2010) πρότειναν το υβριδικό μοντέλο MaxEnt-LDA, το οποίο είχε τη δυνατότητα να εντοπίζει ζεύγη άποψης-γνώμης με βάση συγκεκριμένες συντακτικές δομές.

Το δεύτερο στάδιο της ταξινόμησης σε επίπεδο χαρακτηριστικού περιλαμβάνει τον καθορισμό της συναισθηματικής πολικότητας κάθε άποψης που εντοπίστηκε από το πρώτο στάδιο. Αυτό μπορεί να πραγματοποιηθεί είτε με μεθόδους επιβλεπόμενης μάθησης είτε με μεθόδους που βασίζονται στη χρήση λεξικών. Κατά την πρώτη περίπτωση, όλες οι μέθοδοι επιβλεπόμενης μάθησης που χρησιμοποιήθηκαν στην ταξινόμηση επιπέδου πρότασης μπορούν να εφαρμοστούν και εδώ με ανάλογο τρόπο. Καθώς οι μέθοδοι επιβλεπόμενης μάθησης εξαρτώνται άμεσα από τα δεδομένα, ανακύπτει ένα πολύ σημαντικό πρόβλημα, που είναι γνωστό και ως αλλαγή θεματικού πεδίου (topic shift). Το πρόβλημα είναι ότι τα μο-

ντέλα και οι ταξινομητές εκπαιδεύονται ώστε να λειτουργούν με συγκεκριμένες κατηγορίες δεδομένων. Έτσι αν ένα μοντέλο έχει εκπαιδευτεί με δεδομένα από κριτικές ταινιών και εφαρμοστεί σε δεδομένα από κριτικές προϊόντων, τα αποτελέσματα είναι απογοητευτικά. Η αδυναμία γενίκευσης του χώρου δράσης των μοντέλων επιβλεπόμενης μάθησης στο πρόβλημα της εξόρυξης συναισθήματος από προτάσεις οφείλεται κυρίως στην έλλειψη αρκετών χαρακτηριστικών που μπορούν να αξιοποιηθούν στην ταξινόμηση. Αυτός είναι και ο λόγος που τα μοντέλα επιβλεπόμενης μάθησης χρησιμοποιούνται κυρίως στη ταξινόμηση σε επίπεδο εγγράφου οπού υπάρχει επαρκής αριθμός χαρακτηριστικών. Το πρόβλημα της γενίκευσης μπορεί να αντιμετωπιστεί σε ικανοποιητικό βαθμό με προσεγγίσεις που βασίζονται σε λεξικά. Οι εν λόγω μέθοδοι χρησιμοποιούν λεξικά συναισθημάτων, σύνθετες εκφράσεις, λέξεις που μπορούν να μεταβάλλουν την συναισθηματική πολικότητα μιας πρότασης (valence shifters), λέξεις που δηλώνουν αντίθεση καθώς και το συντακτικό δέντρο των προτάσεων προκειμένου να καθορίσουν το συναίσθημα κάθε άποψης. Οι Ding, Liu και Yu (2008) πρότειναν μια προσέγγιση που ανήκει στην προαναφερθείσα κατηγορία, τις οποίας τα βήματα αναλύονται στη συνέχεια. Αρχικά επισημαίνονται όλες οι λέξεων που εκφράζουν συναίσθημα και περιέχονται σε προτάσεις που διατυπώνουν μια ή παραπάνω απόψεις. Στις λέξεις με θετικό συναίσθημα αποδίδεται μια βαθμολογία «+1» ενώ στις λέξεις με αρνητικό συναίσθημα μια βαθμολογία «-1». Έπειτα, εξετάζεται η ύπαρξη λέξεων όπως το «όχι», το «ποτέ», το «κανείς» κ.α. οι οποίες προκαλούν αλλαγές της συναισθηματικής πολικότητας και γίνονται αλλαγές των βαθμολογιών όπου χρειάζεται. Επόμενη κίνηση αποτελεί η εξέταση των λέξεων που δηλώνουν αντίθεση όπως το «αλλά» και το «ωστόσο» καθώς τις περισσότερες φορές παρουσιάζουν παρόμοια συμπεριφορά με τις λέξεις του προηγούμενου βήματος. Η διαδικασία ολοκληρώνεται με την εφαρμογή μια συνάρτησης για τον τελικό υπολογισμό του συναισθήματος κάθε άποψης.

2.2 Αποσαφήνιση της έννοιας των λέξεων (WSD)

Η αποσαφήνιση της έννοιας μιας λέξης (Word Sense Disambiguation) αναφέρεται στο πρόβλημα της εύρεσης της σωστής έννοιας μιας λέξης που βρίσκεται μέσα σε μια πρόταση ή σε ένα απόσπασμα κειμένου. Το πρόβλημα αυτό προκύπτει από την πολυσημία των λέξεων των σύγχρονων γλωσσών. Αυτό σημαίνει ότι μια λέξη μπορεί να έχει διαφορετική έννοια ανάλογα με το πλαίσιο (context) στο οποίο εμφανίζεται. Για παράδειγμα η λέξη «bank» στις προτάσεις (1) και (2) που ακολουθούν έχει διαφορετική έννοια.

«This is the largest bank in the world.» (1)

«We walked along the bank of the river.» (2)

Στην πρόταση (1) η λέξη «bank» σημαίνει τράπεζα ενώ στην πρόταση (2) σημαίνει όχθη.

Το μεγάλο εύρος εφαρμογών του συγκεκριμένου τομέα της επεξεργασίας φυσικής γλώσσας (Natural Language Processing) οδήγησε σταδιακά στην αύξηση του ενδιαφέροντος της ερευνητικής κοινότητας με αποτέλεσμα όλο και περισσότερες προσεγγίσεις να παρουσιάζονται. Η διαφοροποίηση των προ-

σεγγίσεων οφείλεται στη πηγή της πληροφορίας που χρησιμοποιούν για την επίλυση του προβλήματος της αποσαφήνισης. Οι κύριες κατηγορίες προσεγγίσεων είναι τρεις. Η περίπτωση των υβριδικών προσεγγίσεων δεν αποτελεί αυτόνομη κατηγορία καθώς ουσιαστικά πρόκειται για ένα συνδυασμό προσεγγίσεων που ανήκουν στις κύριες κατηγορίες.

Η πρώτη κατηγορία περιλαμβάνει τις προσεγγίσεις που βασίζονται σε λεξικολογικούς πόρους οι οποίοι είναι φτιαγμένοι χειροκίνητα όπως είναι τα λεξικά και οι γλωσσολογικοί θησαυροί, εξαιρώντας τη χρήση σωμάτων κειμένου (corpora). Οι προσεγγίσεις αυτές καλούνται ως προσεγγίσεις βασισμένες στη γνώση (knowledge-based approaches). Το 1986 ο Mike Lesk πρότεινε μια μέθοδο η οποία συγκρίνει τις επικαλύψεις των λέξεων που υπάρχουν στις διαφορετικές ερμηνείες της λέξης που αποσαφηνίζεται (target word) και στις ερμηνείες όλων των άλλων λέξεων που βρίσκονται στο ίδιο πλαίσιο. Η μέθοδος του Lesk αποτέλεσε βάση για αρκετές προσεγγίσεις που παρουσιάστηκαν στη συνέχεια. Οι Banerjee και Pedersen (2002) χρησιμοποίησαν τις σημασιολογικές συσχετίσεις του γλωσσολογικού θησαυρού WordNet προκειμένου να διευρύνουν το σύνολο των ερμηνειών για τις οποίες θα εξεταστεί η επικάλυψη των λέξεων που περιέχουν ενώ οι Iida κ.ά. (2008) χρησιμοποίησαν μοντέλα κοντινότερου γείτονα (nearest-neighbor models) προκειμένου να συγκρίνουν τις ερμηνείες των λέξεων. Οι σημασιολογικές συσχετίσεις του WordNet χρησιμοποιήθηκαν και από άλλες τεχνικές που βασίζονται στη γνώση. Συγκεκριμένα οι Galley και McKeown (2003) χρησιμοποίησαν τις συσχετίσεις του WordNet προκειμένου να κατασκευάσουν θεματικές αλυσίδες (lexical chains) ενώ η Mihalcea (2005) για να δημιουργήσει γραφικές δομές με στόχο την επίλυση του προβλήματος της αποσαφήνισης.

Η δεύτερη κατηγορία περιλαμβάνει τις προσεγγίσεις επιβλεπόμενης μάθησης. Οι προσεγγίσεις αυτές χρησιμοποιούν μεθόδους μηχανικής μάθησης για την εκπαίδευση ταξινομητών από μεγάλα σώματα κειμένου τα οποία έχουν επισημειωμένη την έννοια των λέξεων τους. Αφότου ολοκληρωθεί η διαδικασία της εκπαίδευσης, ο ταξινομητής εκτελεί μια εργασία ταξινόμησης προκειμένου να αναθέσει την σωστή έννοια σε κάθε λέξη. Το πλεονέκτημα αυτού του είδους των προσεγγίσεων είναι η υψηλή τους απόδοση. Αντίθετα, το μειονέκτημα τους είναι ότι προκειμένου να πετύχουν υψηλή απόδοση απαιτούν την ύπαρξη τεράστιων σωμάτων κειμένου για την εκπαίδευση των ταξινομητών. Κατά τη διάρκεια των περασμένων δεκαετιών πολλές διαφορετικές μέθοδοι επιβλεπόμενης μάθησης χρησιμοποιήθηκαν για την επίλυση του προβλήματος της αποσαφήνισης. Μια από αυτές χρησιμοποίει ένα ταξινομητή Naïve Bayes προκειμένου να υπολογίσει την υπό συνθήκη πιθανότητα κάθε έννοιας μιας λέξης ως προς τα χαρακτηριστικά του πλαισίου στο οποίο εμφανίζεται. Η έννοια που πετυχαίνει την μεγιστοποίηση της πιθανότητας επιλέγεται ως η κατάλληλη για το συγκεκριμένο πλαίσιο. Εκτός του Naïve Bayes, στις προσεγγίσεις επιβλεπόμενης μάθησης μπορούν να χρησιμοποιηθούν δέντρα απόφασης (decision trees), νευρωνικά δίκτυα (neural networks) καθώς και SVM (Navigli, 2009).

Η τρίτη κατηγορία περιλαμβάνει τις προσεγγίσεις μη επιβλεπόμενης μάθησης. Οι προσεγγίσεις αυτές δεν βασίζονται σε εξωτερικές πληροφορίες. Το πλεονέκτημα τους έναντι των παραπάνω είναι ότι μπορούν να εφαρμοστούν απευθείας σε δεδομένα τα οποία δεν εμπεριέχουν υποσημείωση για την έν-

νοια των λέξεων. Οι πιο πρόσφατες προσεγγίσεις που έχουν παρουσιαστεί βασίζονται σε γράφους, στους οποίους οι κόμβοι αναπαριστούν τις λέξεις και οι ακμές αναπαριστούν κάποιου είδους σχέση που υπάρχει ανάμεσα τους. Η Mihalcea (2005) χρησιμοποίησε τις συσχετίσεις που παρέχει το WordNet για να φτιάξει ένα γράφο με τις έννοιες των λέξεων του κειμένου. Βάσει των συσχετίσεων αυτών ανέθεσε βάρη στις ακμές που συνέδεαν τις διαφορετικές έννοιες των λέξεων του κειμένου και στη συνέχεια χρησιμοποίησε τον αλγόριθμο PageRank προκειμένου να βρει την καταλληλότερη έννοια κάθε λέξης για το συγκεκριμένο πλαίσιο. Οι Ion και Ştefănescu (2011) χρησιμοποίησαν την προσέγγιση των λεξιλογικών αλυσίδων προκειμένου να καθορίσουν το βαθμό σημασιολογικής συσχέτισης μεταξύ των διαφορετικών εννοιών των λέξεων που υπάρχουν σε ένα γράφο. Τέλος οι Panagiotopoulou κ.ά (2012) για κάθε πρόταση ενός κειμένου δημιούργησαν ένα γράφο με όλες τις έννοιες των λέξεων που υπήρχαν στο WordNet και αντιμετώπισαν το πρόβλημα της αποσαφήνισης των λέξεων ως ένα πρόβλημα γραμμικού προγραμματισμού. Η προσέγγιση αυτή χρησιμοποιήθηκε στην παρούσα πτυχιακή και αναλύεται περαιτέρω στη συνέχεια.

2.3 Υποδομές

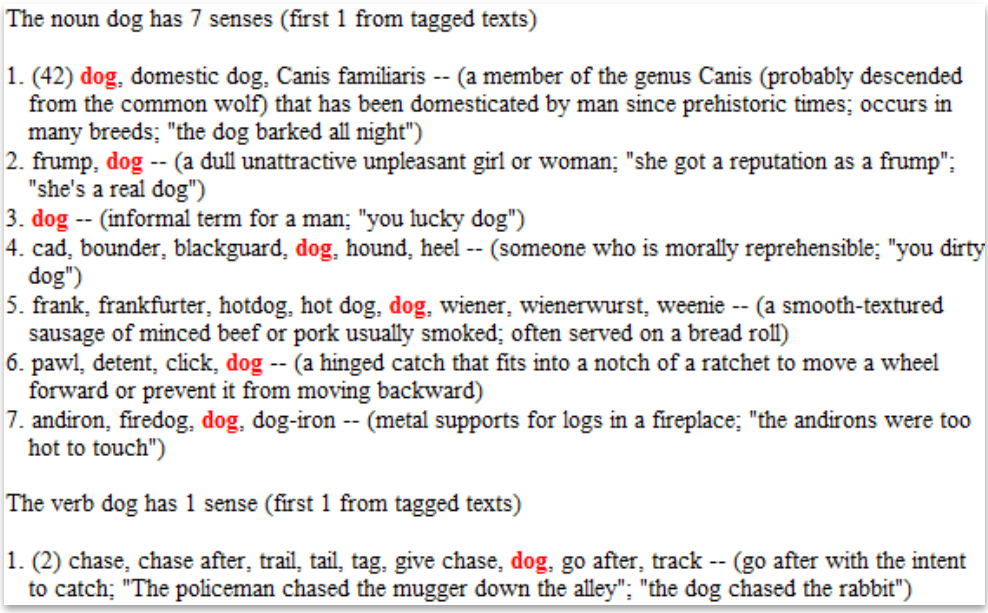
2.3.1 WordNet

Το WordNet είναι μια λεξικολογική βάση δεδομένων αγγλικών λέξεων. Δημιουργήθηκε το 1986 από το Πανεπιστήμιο του Princeton στο οποίο συνεχίζει να αναπτύσσεται1. Βασίζεται σε ψυχολογικές και γλωσσολογικές θεωρίες για τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου και στόχος της δημιουργίας του ήταν να αποτελέσει ένα συνδυασμό λεξικού με γλωσσολογικό θησαυρό ώστε να χρησιμοποιηθεί ως ένα εργαλείο αυτόματης ανάλυσης κειμένου.

Πιο συγκεκριμένα, το WordNet ομαδοποιεί τα ουσιαστικά, τα ρήματα, τα επίθετα και τα επιρρήματα σε σύνολα συνωνύμων (synsets) κάθε ένα από τα οποία αντιπροσωπεύει μια διακριτή λεξικολογική έννοια. Επιπλέον για κάθε λήμμα του, παρέχει έναν αριθμό από έννοιες. Η έννοιας μιας λέξης του WordNet αποτελείται από:

- ένα αριθμό που σηματοδοτεί τη συχνότητα εμφάνισης του όρου με τη συγκεκριμένη έννοια στα γλωσσολογικά κείμενα που έχουν χρησιμοποιηθεί από το WordNet. Βάσει αυτού μπορεί να προκύψει η πιο «δημοφιλής» έννοια για κάθε λέξη (Most Frequent Sense ή First Sense) η οποία χρησιμοποιείται συχνά ως μια γρήγορη εναλλακτική της αποσαφήνισης.
- ένα σύνολο λέξεων που σηματοδοτεί τα συνώνυμα (synsets) του συγκεκριμένης ερμηνείας της λέξης.
- ένα σύνολο από φράσεις (gloss) της καθομιλουμένης που περιέχουν τη λέξη με τη συγκεκριμένη έννοια (βλέπε Εικόνα 1).

[1] Η πιο πρόσφατη έκδοση του είναι η 3.1

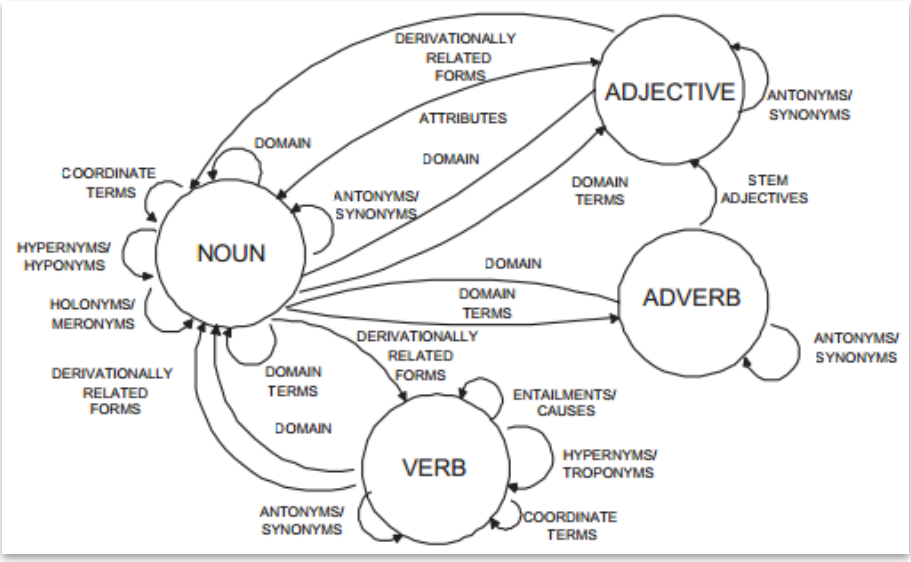


Εικόνα 1. Παράδειγμα εξόδου του WordNet για την λέξη «dog».

Πολλά από τα σύνολα συνωνύμων του WordNet συνδέονται μεταξύ τους μέσω σημασιολογικών συσχετίσεων, οι οποίες διαφοροποιούνται ανάλογα με τον τύπο των λέξεων. Οι βασικότερες συσχετίσεις που χρησιμοποιεί το WordNet, όπως φαίνονται στην Εικόνα 2 είναι οι εξής:

- **υπερώνυμο (hypernym):** ένα ουσιαστικό ή ένα ρήμα Υ είναι υπερώνυμο ενός ουσιαστικού ή ενός ρήματος Χ, όταν το Υ έχει πιο ευρεία έννοια από το Χ. Π.χ. η λέξη έντομο είναι υπερώνυμο της λέξης μέλισσα.
- **υπώνυμο (hyponym):** ένα ουσιαστικό ή ένα ρήμα Υ είναι υπώνυμο ενός ουσιαστικού ή ενός ρήματος Χ, όταν το Υ έχει πιο ειδική έννοια από το Χ. Π.χ. η λέξη μέλισσα είναι υπώνυμο της λέξης έντομο.
- **συντεταγμένοι όροι (coordinate terms):** δύο ουσιαστικά ή δυο ρήματα Χ,Υ είναι συντεταγμένοι όταν έχουν κοινό υπερώνυμο. Π.χ. η λέξη «λύκος» και η λέξη «σκύλος» είναι συντεταγμένοι όροι.
- **μερώνυμο (meronym):** ένα ουσιαστικό Υ είναι μερώνυμο του ουσιαστικού Χ, όταν το Υ αποτελεί τμήμα/μέρος του Χ. Π.χ. η λέξη παράθυρο είναι μερώνυμο της λέξης κτήριο.
- **ολώνυμο (holonym):** ένα ουσιαστικό Υ είναι ολώνυμο του ουσιαστικού Χ, όταν το Χ αποτελεί τμήμα/μέρος του Υ. Π.χ. η λέξη κτήριο είναι ολώνυμο της λέξης παράθυρο.
- **τροπώνυμο (troponym):** ένα ρήμα Υ είναι τροπώνυμο του ρήματος Χ, όταν το Υ είναι ένας τρόπος έκφρασης της δραστηριότητας του Χ. Π.χ. η λέξη «ψελλίζω» είναι τροπώνυμο της λέξης «μιλάω».

- **συνεπαγωγή (entailement):** ένα ρήμα Υ αποτελεί συνεπαγωγή του Χ, όταν η δραστηριότητα Υ εντάσσεται στη δραστηριότητα Χ. Π.χ. η λέξη «κοιμάμαι» συνεπάγεται από τη λέξη «ροχαλίζω».
- **σχετικά ουσιαστικά (related nouns):** Χρησιμοποιείται στα επίθετα. Π.χ. η λέξη «ομορφιά» σχετίζεται με την λέξη «όμορφος».
- **μετοχή ρήματος (particle of verb):** Χρησιμοποιείται στα επίθετα. Π.χ. η λέξη «τρέχοντας» είναι μετοχή του ρήματος «τρέχω».



Εικόνα 2. *Γραφική απεικόνιση όλων των συσχετίσεων του WordNet.*

2.3.2 Stanford Parser

Ο Stanford Parser είναι ένας στατιστικός συντακτικός αναλυτής φυσικής γλώσσας ο οποίος δημιουργήθηκε το 2012 από τους Dan Klein και Christopher Manning στο Stanford Natural Language Processing Group. Η αρχική του υλοποίηση είναι σε Java και παρέχει υποστήριξη για κείμενα που είναι γραμμένα στα Αγγλικά, στα Γερμανικά, στα Αραβικά και στα Κινέζικα. Τα κύρια μοντέλα που χρησιμοποιεί είναι δυο, το PCFG και το Factored. Το PCFG(Probabilistic Context-Free Grammar) είναι ένα μοντέλο το οποίο εκτελείται γρήγορα, με καλή ακρίβεια και χωρίς να απαιτεί υψηλά ποσά μνήμης. Αντίθετα το Factored είναι πιο σύνθετο και είναι πολύ απαιτητικό σε μνήμη καθώς περιέχει δύο γραμματικές και ουσιαστικά οδηγεί το σύστημα στο να τρέξει τρεις συντακτικούς αναλυτές. Παρ' όλα αυτά, δίνει η δυνατότητα στον χρήστη να χρησιμοποιήσει το δικό του μοντέλο ή ακόμα και να εκπαιδεύσει κάποιο.

2.3.3 Μέθοδοι αποσαφήνισης της έννοιας των λέξεων

2.3.3.1 Μέθοδος WDEG

Ο αλγόριθμος WDEG χρησιμοποιεί τον γράφο της που απεικονίζεται στην Εικόνα 3 και υπολογίζει για κάθε έννοια μιας λέξης (κόμβο) το άθροισμα των βαρών των ακμών που την περιέχουν (Weighted Degree) (Harris, 2008).

2.3.3.2 Μέθοδος ILP

Στόχος της μεθόδου ILP (Integre Linear Programming) είναι να αντιμετωπίσει το πρόβλημα της αποσαφήνισης στις λέξεις μιας πρότασης ως ένα πρόβλημα (μεγιστοποίησης) γραμμικού προγραμματισμού. Για το σκοπό αυτό δημιουργεί για μία πρόταση ένα πλήρη γράφο στον οποίο μετέχουν όλες οι έννοιες όλων των λέξεων της πρότασης (που υπάρχουν στο Wordnet) και οι μεταξύ τους ομοιότητες ανά δύο (Εικόνα 3).

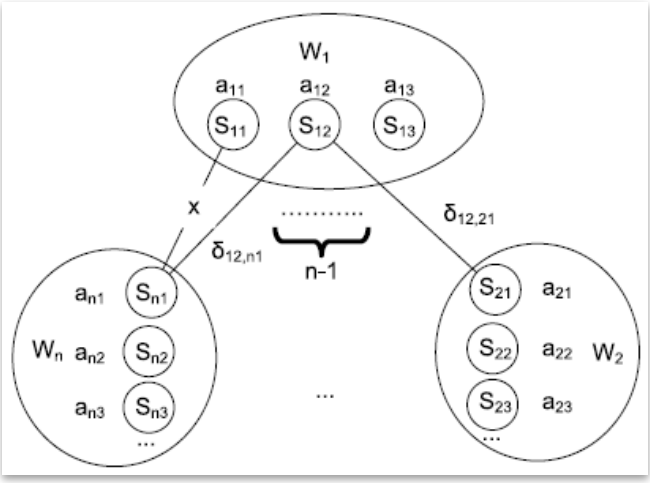
Έστω w_1, … , w_n οι εμφανίσεις μιας λέξης σε μια πρόταση, s_{ij} η j-οστή πιθανή έννοια της λέξης w_i και rel(s_{ij}, $s_{i'j'}$) η συσχέτιση μεταξύ των εννοιών s_{ij} και $s_{i'j'}$. Ο στόχος της συγκεκριμένης μεθόδου είναι να επιλέξει ακριβώς μια από τις διαθέσιμες έννοιες s_{ij} κάθε λέξης w_i, έτσι ώστε η συνολική συσχέτιση μεταξύ των ζευγών των επιλεγμένων εννοιών να είναι μέγιστη. Για κάθε έννοια s_{ij}, υπάρχει μια δυαδική μεταβλητή a_{ij} η οποία δείχνει αν είναι ενεργή (π.χ. Αν η έννοια χρησιμοποιείται τότε a_{ij} = 1 αλλιώς a_{ij} = 0). Μια πρώτη προσέγγιση θα ήταν να επιχειρηθεί να μεγιστοποιηθεί η τιμή της αντικειμενικής συνάρτησης (1), όπου σύμφωνα με τους περιορισμούς (2) απαιτεί i < i′ δεδομένου ότι το μέτρο ομοιότητας είναι συμμετρικό. Ο τελευταίος περιορισμός εξασφαλίζει ότι μόνο μια έννοια s_{ij} είναι ενεργή για κάθε $w_{ι.}$

$$\sum_{\text{i,j,i′,j′,i<}i'} rel(s_{ij}, s_{i'j'}) \cdot a_{ij} \cdot a_{i'j'} \qquad (1)$$

$$a_{ij} \in \{0,1\}, \forall i,j \;\; και \;\; \sum_{j} a_{ij} = 1 \, , \forall i. \qquad (2)$$

Η συνάρτηση (1) είναι τετραγωνική, επειδή αποτελεί το σταθμισμένο άθροισμα των γινομένων των ζευγών (aij · ai′j′). Προκειμένου το πρόβλημα να γίνει Ακέραιου Γραμμικού Προγραμματισμού (Integer Linear Programming - ILP), χρησιμοποιείται μια δυαδική μεταβλητή δij,i′j′ για κάθε ζεύγος εννοιών sij, si′j′ με i≠i′. Η εικόνα που ακολουθεί δείχνει τη νέα μορφή του προβλήματος. Κάθε ένας από τους μεγάλους κύκλους, που από εδώ και στο εξής θα καλείται «σύννεφο», περιλαμβάνει όλες τις πιθανές έννοιες (μικροί κύκλοι) μιας συγκεκριμένης λέξης wi. Θα πρέπει να υπάρχει ακριβώς μια ενεργή έννοια σε κάθε λέξη. Κάθε μεταβλητή δij,i′j′ δείχνει αν η ακμή που συνδέει δυο έννοιες sij και si′j′ από διαφορετικά σύννεφα είναι

ενεργή ($\delta_{ij,i'j'} = 1$) ή όχι ($\delta_{ij,i'j'} = 0$). Μια ακμή μπορεί να είναι ενεργή αν και μόνο αν οι έννοιες που συνδέει είναι ενεργές ($a_{ij} = a_{i'j'} = 1$).



Εικόνα 3. *Αναπαράσταση του ILP μοντέλου για την αποσαφήνιση της έννοιας των λέξεων.*

Το πρόβλημα μπορεί τώρα να διατυπωθεί από τη μεγιστοποίηση της σχέσης (3), όπου i, i′ ∈ {1,…, n}, i ≠ i′ και (i, j), (i′, j′) κυμαίνεται πάνω από τους δείκτες των πιθανών εννοιών s_{ij} και $s_{i'j'}$ αντίστοιχα.

$$\sum_{\mathrm{i,j,i',j',i<}i'} rel(s_{ij}, s_{i'j'}) \cdot \delta_{ij,i'j'} \qquad (3)$$

$$a_{ij} \in \{0,1\}, \forall i,j \ \ και \ \sum_{j} a_{ij} = 1\ , \forall i. \qquad (4)$$

$$\delta_{ij,i'j'} \in \{0,1\}\ και\ \delta_{ij,i'j'} = \ \delta_{i'j',ij}\ \ , \forall i,j,i',j' \quad (5)$$

$$και \sum_{j'} \delta_{ij,i'j'} = a_{ij}\ , \forall i,j,i'. \qquad (6)$$

Το δεύτερο μέρος του περιορισμού (5) αντικατοπτρίζει το γεγονός ότι οι ακμές (και οι ενεργοποιήσεις του) δεν είναι κατευθυνόμενες. Ο περιορισμός (6) μπορεί να γίνει αντιληπτός λαμβάνοντας υπόψη χωριστά τις πιθανές τιμές του a_{ij}.

- Αν $a_{ij} = 0$ (s_{ij} είναι ανενεργό), $\sum_{j'} \delta_{ij,i'j'} = 0, \forall i'$
- Αν $a_{ij} = 1$ (s_{ij} είναι ενεργό), $\sum_{j'} \delta_{ij,i'j'} = 1, \forall i'$

Ένα πλεονέκτημα των ILP solver είναι ότι εγγυώνται την εύρεση της καλύτερης λύσης, αυτή υπάρχει. Το πρόβλημα που πρέπει να λυθεί είναι NP-δύσκολο αλλά με την χρήση αποδοτικών solver και με ένα μικρό αριθμό μεταβλητών και περιορισμών η λύση του μπορεί να γίνει αρκετά γρήγορα. Στα πλαίσια της υλοποίησης της μεθόδου ILP, αναπτύχθηκε και μια επιλογή κλαδέματος εννοιών (sense pruning) η οποία αφαιρεί από το γράφο κάθε έννοια s_{ij} που η WordNet ερμηνεία της δεν περιέχει καμία από τις λέ-

ξεις που αποσαφηνίζονται εκείνη τη στιγμή. Με την επιλογή αυτή μειώνει σημαντικά τον αριθμό των υποψήφιων εννοιών καθώς και τον χρόνο επίλυσης του προβλήματος.

2.3.4 Μέτρα συσχέτισης των λέξεων

Τα μέτρα συσχέτισης των λέξεων καθορίζουν τον βαθμό που μια λέξη μοιάζει με μια άλλη. Ο όρος «μέτρο συσχέτισης» αποτελεί γενίκευση του όρου «μέτρο ομοιότητας» ο οποίος αναφέρεται στη σύγκριση δύο λέξεων στα πλαίσια της συνωνυμίας. Στη συνέχεια της παρούσας πτυχιακής θα χρησιμοποιείται ο όρος «μέτρο ομοιότητας» προς αποφυγή συγχύσεων.

Τα μέτρα ομοιότητας μπορούν να διαχωριστούν σε τρεις κύριες κατηγορίες. Η πρώτη κατηγορία αποτελείται από τα μέτρα που βασίζονται σε πρότερη γνώση όπως είναι αυτά που βασίζονται σε λεξικά, γλωσσολογικούς θησαυρούς και υπερσύνδεσμους της Wikipedia. Η δεύτερη κατηγορία αποτελείται από τα μέτρα που βασίζονται σε σώματα κειμένου (corpora). Αυτά προκειμένου να καθορίσουν την ομοιότητα των λέξεων χρησιμοποιούν τα στατιστικά συνεμφάνισης των λέξεων ή των εννοιών τους σε ένα κείμενο (π.χ. το μέτρο PMI). Η τελευταία κατηγορία αποτελείται από τα υβριδικά μέτρα, τα οποία αποτελούν συνδυασμό των δύο προηγούμενων. Τέλος, αξίζει να σημειωθεί ότι παρά του ότι μερικά μέτρα ομοιότητας έχουν υλοποιηθεί για να υπολογίζουν ομοιότητα μεταξύ λέξεων ή φράσεων, μπορούν να τροποποιηθούν ούτως ώστε να λειτουργούν σε επίπεδο έννοιας.

2.3.4.1 Μέτρο ομοιότητας SR

Το μέτρο ομοιότητας SR ανήκει στα μέτρα που βασίζονται σε πρότερη γνώση. Απαιτεί την ύπαρξη ενός γλωσσολογικού θησαυρού $\boldsymbol{\Theta}$ με λεξικολογικές συσχετίσεις και ενός σταθμισμένου συστήματος για αυτές. Στα πλαίσια της παρούσας πτυχιακής χρησιμοποιήθηκε ο γλωσσολογικός θησαυρός WordNet. Δεδομένου ενός ζεύγους εννοιών s_1, s_2 και ενός μονοπατιού (ακολουθίας) $P = <p_1,\ldots, p_k>$ εννοιών οι οποίες συνδέουν το $s_{1=}\ p_1$ στο $s_{2=}\ p_k$ μέσω λεξικολογικών συσχετίσεων, η «σημασιολογική πυκνότητα» του μονοπατιού P (Semantic Compactness – SCM), η οποία υποδηλώνει το πόσο σχετικές μεταξύ τους είναι οι έννοιες που περιέχονται στο μονοπάτι, και υπολογίζεται από τον παρακάτω τύπο

$$SCM(P) = \prod_{i=1}^{k-1} w_{i \to i+1}$$

Όπου $w_{i \to i+1}$ είναι τα βάρη των λεξικολογικών συσχετίσεων από έννοια σε έννοια του P. Εκτός από τον υπολογισμό του SCM το SR απαιτεί τον υπολογισμού της «σημασιολογικής σαφήνειας μονοπατιού» του P (Semantic Path Elaboration - SPE) που στην ουσία αντιπροσωπεύει πόσο ειδικές ή γενικές είναι οι έννοιες που περιέχονται στο μονοπάτι P.

$$SPE(P) = \prod_{i=1}^{k} \frac{2d_i \cdot d_{i+1}}{d_i + d_{i+1}} \cdot \frac{1}{d_{max}}$$

Όπου di το βάθος του pi στην ιεραρχία του ***Θ*** και το dmax το μέγιστο βάθος του Ο. Η ομοιότητα μεταξύ του εννοιών s1 και s2 βάσει του SR ορίζεται από την παρακάτω σχέση, οπου το P κυμαίνεται πάνω από όλα τα μονοπάτια που συνδέουν τα s1 και s2. Αν δεν υπάρχει κανένα μονοπάτι, τότε το SR(s1, s2) = 0.

$$\mathrm{SR}(s_1, s_2) = \max_{\mathrm{P} = <s_1,\ldots,s_k>}\{SCM(P) \cdot SPE(P)\}$$

Αντί για την χρήση του SR(s_1, s_2), στην παρούσα υλοποίηση χρησιμοποιείται το $e^{\mathrm{SR}(s_1, s_2)}$ καθώς οδηγεί σε λίγο καλύτερα αποτελέσματα.

2.3.4.2 Μέτρο ομοιότητας PMI

Το μέτρο PMI ανήκει στην κατηγορία των μέτρων ομοιότητας που υπολογίζουν την ομοιότητα δύο λέξεων βάσει των στατιστικών συνεμφάνισης του σε ένα κείμενο. Για δύο λέξεις w_1, w_2 το PMI σκορ τους είναι $\mathrm{PMI}(w_1, w_2) = \log\frac{P(w_1,\ w_2)}{P(w_1)\cdot P(w_2)}$, όπου το P(w_1, w_2) εκφράζει την πιθανότητα να συνεμφανίζονται (π.χ. στην ίδια πρόταση). Αν τα w_1, w_2 είναι ανεξάρτητα τότε το PMI σκορ τους είναι ίσο με το 0 ενώ αν τα w_1, w_2 συνεμφανίζονται πάντα το PMI σκορ τους μεγιστοποιείται και είναι ίσο με $-\log P(w_1) = -\log P(w_2)$. Σε περίπτωση που το κείμενο περιέχει επισημάνσεις σχετικά με την έννοια των λέξεων του, το PMI σκορ δυο εννοιών s_1, s_2 υπολογίζεται παρομοίως. Στην παρούσα υλοποίηση αντί για λέξεις ή έννοιες χρησιμοποιούνται οι ερμηνείες g(s_1) και g(s_2) που παρέχει το WordNet για τις έννοιες s_1, s_2 και τα PMI σκορ όλων των ζευγών λέξεων w1, w2 από τα g(s_1) και g(s_2) αντίστοιχα, εξαιρώντας τα σημεία στίξης.

$$PMI(s_1, s_2) = \frac{\sum_{w_1 \in g(s_1),\, w_2 \in g(s_2)} PMI(w_1 \cdot w_2)}{|g(s_1)| \cdot |g(s_2)|}$$

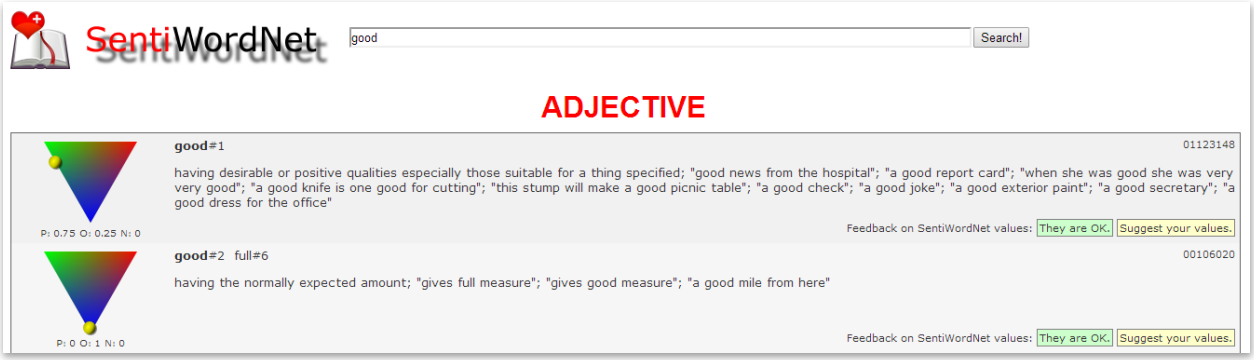
Όπου $|g(s)|$ είναι το μήκος του g(s) σε λέξεις. Με βάση τα παραπάνω γίνει η υπόθεση ότι αν τα s_1 και s_2 σχετίζονται τότε οι λέξεις που χρησιμοποιούνται στις ερμηνείες τους θα συνεμφανίζονται συχνά.

2.3.4.3 Μέτρο ομοιότητας LL

Το μέτρο Lesk-like (LL) ανήκει στην κατηγορία των μέτρων ομοιότητας που υπολογίζουν την ομοιότητα δύο λέξεων βάσει των στατιστικών συνεμφάνισης του σε ένα κείμενο. Υπολογίζει την ομοιότητα μεταξύ των ερμηνειών g(s_1) και g(s_2) δυο εννοιών s_1 και s_2 του WordNet χρησιμοποιώντας τη Σημειακή Αμοιβαία Πληροφορία (Pointwise Mutual Information - PMI) και τα στατιστικά συνεμφάνισης των λέξεων σε ένα μεγάλο γλωσσολογικό κείμενο (corpus) το οποίο δεν περιέχει επισημάνσεις για την έννοια των λέξεων του.

2.3.5 SentiWordNet

Το SentiWordNet είναι ένας λεξιλογικός πόρος ο οποίος προορίζεται για την υποστήριξη συστημάτων ταξινόμησης συναισθήματος και εξόρυξης δεδομένων. Δημιουργήθηκε το 2006 από τους αρχείου κειμένου αλλά παράλληλα μπορεί να χρησιμοποιηθεί μέσα από την ιστοσελίδα της εφαρμογής[2] (βλέπε Εικόνα 4).



Εικόνα 4. *Στιγμιότυπο από την ιστότοπο του SentiWordNet.*

Το SentiWordNet είναι το αποτέλεσμα της αυτόματης επισήμανσης των συνόλων συνωνύμων του WordNet σύμφωνα με τις έννοιες της «θετικότητας», της «αρνητικότητας» και της «ουδετερότητας» (Baccianella, 2010). Κάθε σύνολο συνωνύμων συσχετίζεται με τρείς τιμές οι οποίες δείχνουν πόσο θετικοί, αρνητικοί και ουδέτεροι είναι οι όροι που ανήκουν στο σύνολο. Παρ' όλα αυτά, οι διαφορετικές έννοιες του ίδιου όρου μπορεί να έχουν διαφορετικές συναισθήματος. Οι τιμές που υποδεικνύουν το συναίσθημα κάθε συνόλου συνωνύμων (Εικόνα 5) κυμαίνονται στο διάστημα [0.0, 1.0] και το άθροισμα τους θα πρέπει να είναι πάντα ίσο με 1.

```
a   01123148    0.75    0   good#1  having desirable or positive qualities especially those suitable for a thing specified; "good news from the
hospital"; "a good report card"; "when she was good she was very very good"; "a good knife is one good for cutting"; "this stump will make a good
picnic table"; "a good check"; "a good joke"; "a good exterior paint"; "a good secretary"; "a good dress for the office"
a   00106020    0   0   good#2 full#6   having the normally expected amount; "gives full measure"; "gives good measure"; "a good mile from here"
```

Εικόνα 5. *Στιγμιότυπο από το αρχείο του SentiWordNet.*

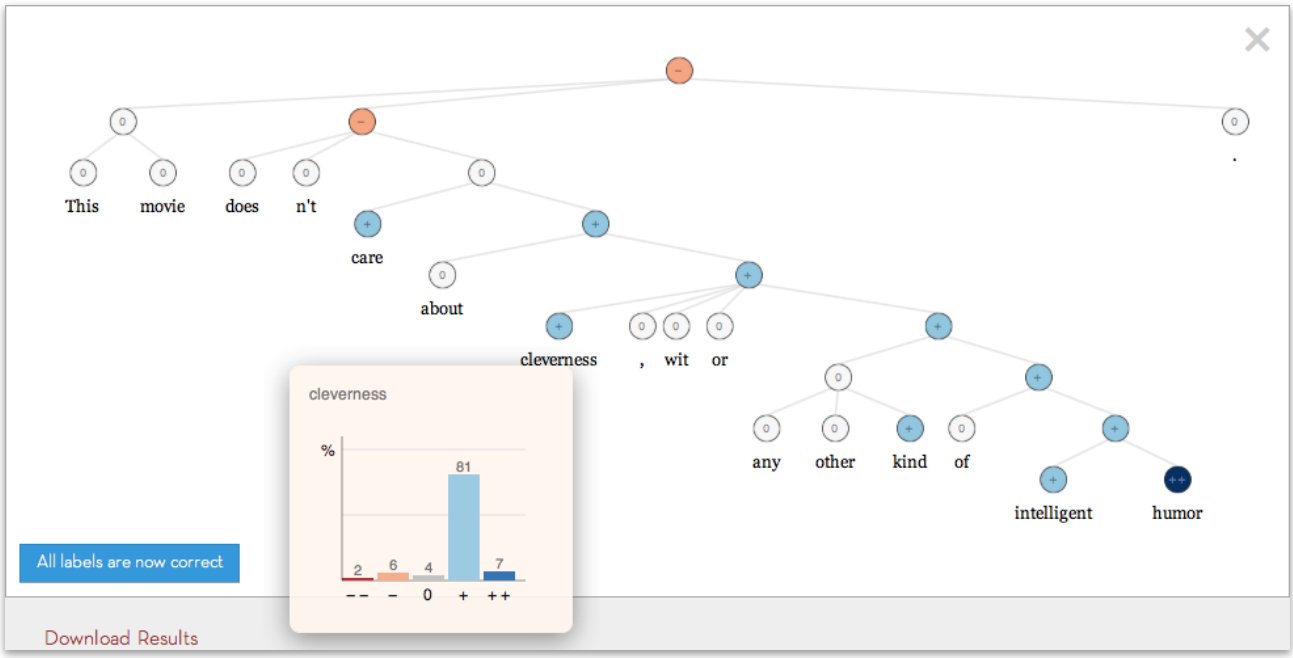
Η διαδικασία δημιουργίας του ξεκινάει με την δημιουργία δυο ομάδων με όλα τα σύνολα συνωνύμων του WordNet που περιέχουν εφτά ξεκάθαρα θετικούς και αρνητικούς όρους αντίστοιχα. Στη συνέχεια τα σετ αυτά επεκτείνονται βάσει των δυαδικών συσχετίσεων του WordNet. Επόμενο βήμα αποτελεί η εκπαίδευση ενός ταξινομητή τριών κλάσεων. Ως δεδομένα εκπαίδευσης του ταξινομητή χρησιμοποιούνται

[2] http://sentiwordnet.isti.cnr.it/

οι ερμηνείες των όρων που περιέχονται στα σετ του πρώτου βήματος μαζί με τις ερμηνείες ενός σετ με όρους που θεωρούνται ουδέτεροι. Έπειτα ακολουθεί η ταξινόμηση όλων των συνόλων συνωνύμων σε τρεις κατηγορίες (θετικά, αρνητικά, ουδέτερα) χρησιμοποιώντας τον ταξινομητή του προηγούμενου βήματος. Αφού ολοκληρωθεί η ταξινόμηση τα βήματα 2 και 3 επαναλαμβάνονται άλλες εφτά φορές με τη διαφορά ότι ο ταξινομητής υλοποιείται με διαφορετικές παραμέτρους. Οι τιμές που προκύπτουν από τις 8 ταξινομήσεις για κάθε σύνολο συνωνύμων αθροίζονται και η μέση τιμή τους δίνει το τελικό σκορ για κάθε σύνολο. Για την ολοκλήρωση της διαδικασίας απαιτείται η εκτέλεση του αλγορίθμου Τυχαίου Περιπάτου πάνω στο γράφο του WordNet χρησιμοποιώντας ως αρχικές τιμές τις τιμές που προέκυψαν στο προηγούμενο βήμα.

2.3.6 Stanford Deep Learning Model for Sentiment Analysis

Το Stanford Deep Learning Model for Sentiment Analysis είναι ένα μοντέλο το οποίο αναπτύχθηκε από τους Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng and Christopher Potts στο Πανεπιστήμιο του Stanford προκειμένου να παρέχει μια πιο λεπτομερή προσέγγιση στο θέμα της απόδοσης συναισθήματος σε επίπεδο πρότασης. Η διαφοροποίηση της συγκεκριμένης προσέγγισης έγκειται στο γεγονός ότι αξιοποιεί μεγαλύτερου εύρους συντακτική πληροφορία από τις προτάσεις όπως για παράδειγμα η σειρά των λέξεων. Γι' αυτό το λόγο οι Socher κ.ά. (2013) πραγματοποίησαν συντακτική ανάλυση στο σώμα κειμένου που παρουσίασαν οι Pang και Lee το 2005. Με αυτό τον τρόπο εξήγαγαν 215.154 μοναδικές φράσεις στις οποίες επισήμαναν χειροκίνητα το συναίσθημα που μετέφεραν. Το νέο σύνολο δεδομένων τους έδωσε τη δυνατότητα να συλλάβουν και να περιγράψουν πιο σύνθετα γλωσσολογικά φαινόμενα καθώς και περιπτώσεις που ξεγελούσαν τα παραδοσιακά συστήματα εξαγωγής συναισθήματος. Προκειμένου να αποθηκεύσουν όλη αυτή την πληροφορία, δημιούργησαν ένα είδος Παλινδρομικού Νευρωνικού Δικτύου (Recursive Neural Network – RNN) το οποίο ονόμασαν Recursive Neural Tensor Network (RNTN). Τα RNTN έχουν ως είσοδο φράσεις οποιουδήποτε μήκους τις οποίες αναπαριστούν χρησιμοποιώντας τα διανύσματα των λέξεων τους και ένα συντακτικό δέντρο. Με αυτό τον τρόπο είναι σε θέση να υπολογίσουν το συντακτικό δέντρο ολόκληρης της πρότασης και να καθορίσουν το τελικό της συναίσθημα. Η Εικόνα 1 *που ακολουθεί απεικονίζει το συντακτικό δέντρο που δημιούργησε το Stanford Deep Learning Model για την πρόταση «This movie doesn't care about cleverness, wit or any other kind of intelligent humor.». Τα φύλλα του δέντρου συμβολίζουν τις λέξεις τις πρότασης ενώ οι ενδιάμεσοι συμβολίζουν τις φράσεις που δημιουργούνται. Αξίζει να σημειωθεί ότι παρά του ότι η πρόταση περιέχει αρκετές λέξεις με θετικό συναίσθημα, το τελικό συναίσθημα της πρότασης σωστά ορίζεται ως αρνητικό.



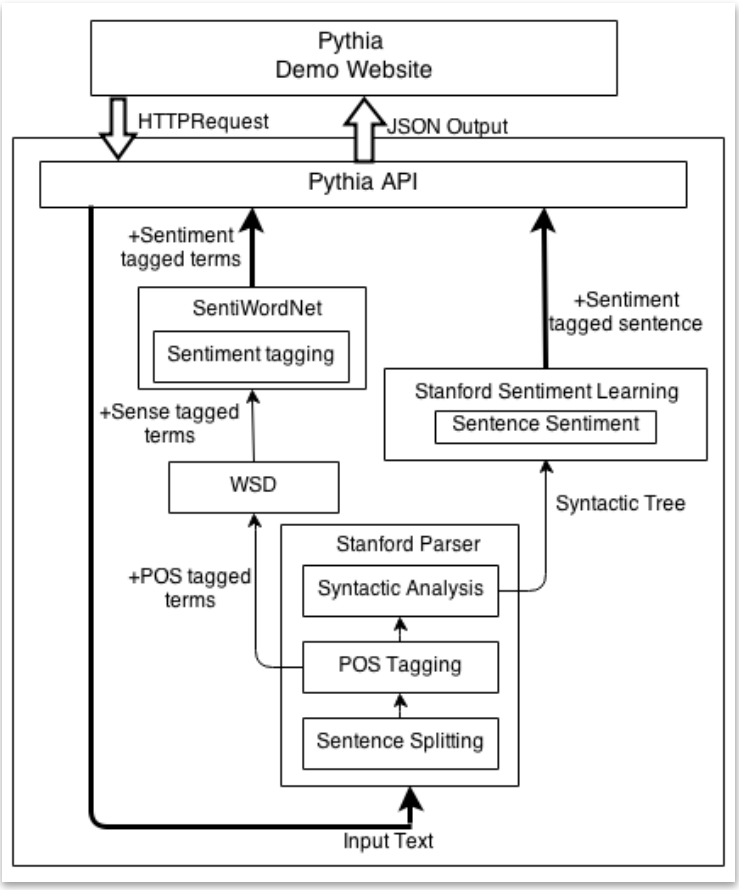
Εικόνα 6. *Δενδρική δομή της πρότασης «This movie doesn't care about cleverness, wit or any other kind of intelligent humor.» που δημιουργείται από το Stanford Deep Learning Model*

Κεφάλαιο 3 Δομή του συστήματος Pythia

Στο παρόν κεφάλαιο περιγράφεται η δομή του συστήματος Pythia που υλοποιήθηκε κατά την εκπόνηση της παρούσας πτυχιακής εργασίας. Παρουσιάζονται κατά σειρά όλα τα δομικά του στοιχεία καθώς και ο τρόπος που διασυνδέονται μεταξύ τους.

3.1 Αρχιτεκτονική

Το σύστημα που αναπτύχθηκε αποτελείται από τρία βασικά συστατικά στοιχεία (components), την υπηρεσία ιστού (Web Service) του συστήματος, τη Διεπαφή Προγραμματισμού Εφαρμογών (Application Programming Interface - API) της υπηρεσίας (Pythia API) και το γραφικό περιβάλλον / διεπιφάνεια χρήστη (Pythia Demo Website). Η Εικόνα 7 παρουσιάζει τα συστατικά στοιχεία που συνθέτουν το σύστημα καθώς και τις μεταξύ τους διασυνδέσεις.



Εικόνα 7. *Η αρχιτεκτονική του συστήματος Pythia.*

3.1.1 Web Service

Η υπηρεσία ιστού (Web Service) είναι το στοιχείο το οποίο πραγματοποιεί όλες τις απαραίτητες διεργασίες στα δεδομένα εισόδου με σκοπό την παραγωγή της πληροφορίας (συντακτική ανάλυση, αποσαφήνιση έννοιας λέξεων, συναισθηματική ανάλυση σε επίπεδο έννοιας λέξης, συναισθηματική ανάλυση σε επίπεδο πρότασης) που υπόσχεται το σύστημα. Αποτελείται από πέντε επιμέρους στοιχεία:

- το στοιχείο που πραγματοποιεί συντακτική ανάλυση στο κείμενο εισόδου
- το στοιχείο που πραγματοποιεί αποσαφήνιση της έννοιας των λέξεων του κειμένου
- το στοιχείο που αποδίδει σε κάθε λέξη τη συναισθηματική της πολικότητα
- το στοιχείο που αποδίδει σε κάθε πρόταση τη συναισθηματική της πολικότητα

Η πρώτη διεργασία που πραγματοποιεί το Web Service είναι η συντακτική ανάλυση του κειμένου εισόδου με χρήση του συντακτικού αναλυτή που αναπτύχθηκε από το Πανεπιστήμιο του Stanford. Έχουν ενσωματωθεί δύο ειδών συντακτικοί αναλυτές, ένας για μικρά και ανεπίσημου χαρακτήρα κείμενα (informal texts π.χ. tweets) και ένας για επίσημα έγγραφα μεγάλης έκτασης. Η διαφορά τους έγγυται στο τρόπο που εκμεταλλεύονται τα μορφολογικά χαρακτηριστικά του κειμένου και όχι στην διαδικασία ανάλυσης που ακολουθούν. Αρχικά ο συντακτικός αναλυτής εντοπίζει τις προτάσεις του κειμένου και τις διαχωρίζει. Στη συνέχεια για κάθε λέξη του κειμένου σημειώνει το μέρος του λόγου στο οποίο ανήκει. Το σύνολο ετικετών για την επισήμανση του μέρους του λόγου των λέξεων βασίζεται στο Penn Treebank και παρουσιάζεται στον παρακάτω πίνακα.

Tag	Περιγραφή
CC	conjunction, coordinating
CD	cardinal number
DT	determiner
EX	existential there
FW	foreign word
IN	conjunction, subordinating or preposition
JJ	adjective
JJR	adjective, comparative
JJS	adjective, superlative
LS	list item marker
MD	verb, modal auxillary
NN	noun, singular or mass
NNS	noun, plural
NNP	noun, proper singular
NNPS	noun, proper plural
PDT	predeterminer
POS	possessive ending

PRP	pronoun, personal
PRP$	pronoun, possessive
RB	adverb
RBR	adverb, comparative
RBS	adverb, superlative
RP	adverb, particle
SYM	symbol
TO	infinitival to
UH	interjection
VB	verb, base form
VBZ	verb, 3rd person singular present
VBP	verb, non-3rd person singular present
VBD	verb, past tense
VBN	verb, past participle
VBG	verb, gerund or present participle
WDT	wh-determiner
WP	wh-pronoun, personal
WP$	wh-pronoun, possessive
WRB	wh-adverb
.	punctuation mark, sentence closer
,	punctuation mark, comma
:	punctuation mark, colon
(	contextual separator, left paren
)	contextual separator, right paren

Πίνακας 1. Πίνακας μερών λόγου σύμφωνα με το *Penn Treebank*

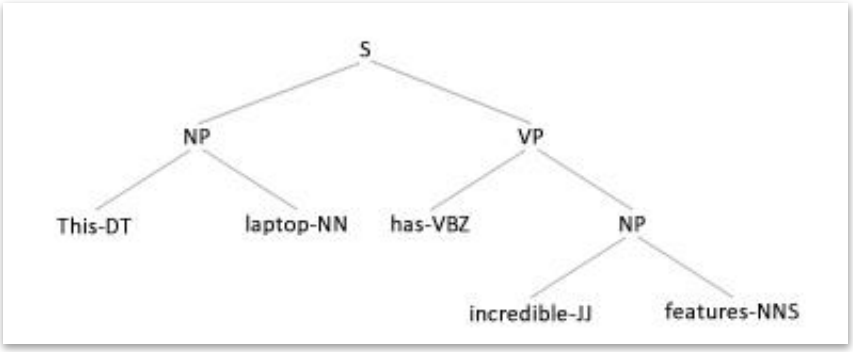
Αυτή η λειτουργία είναι πολύ σημαντική καθώς υποδεικνύει ποιες λέξεις θα χρησιμοποιηθούν στο στάδιο της αποσαφήνισης της έννοιας κάθε λέξης του κειμένου. Παράλληλα με την επισήμανση του μέρους του λόγου των λέξεων, ο συντακτικός αναλυτής ομαδοποιεί τις λέξεις σε μικρές, αυτόνομες φράσεις. Στις φράσεις που δημιουργούνται αποδίδονται ετικέτες όπως και στην περίπτωση των λέξεων. Παρακάτω παρουσιάζεται ο πίνακας με τις ετικέτες που χρησιμοποιούνται για την επισήμανση των φράσεων.

Tag	**Περιγραφή**
ADJP	Adjective phrase
ADVP	Adverb phrase
NP	Noun phrase
PP	Prepositional phrase

S	Simple declarative clause
SBARQ	Direct question introduced by wh-element
SINV	Declarative sentence with subject-aux inversion
SQ	Yes/no questions and subconstituent of SBARQ excluding wh-element
VP	Verb phrase
WHADVP	Wh-adverb phrase
WHNP	Wh-noun phrase
WHPP	Wh-prepositional phrase
X	Constituent of unknown or uncertain category
*****	"Understood" subject of infinitive or imperative
0	Zero variant of that in subordinate clauses
T	Trace of wh-Constituent

Πίνακας 2. Πίνακας συντακτικών ετικετών σύμφωνα με το *Penn Treebank*

Τέλος, δημιουργεί το συντακτικό δέντρο του κειμένου, το οποίο είναι χρήσιμο για την εξαγωγή του συναισθήματος κάθε πρότασης. Το συντακτικό δέντρο απεικονίζει τη συντακτική δομή της πρότασης. Συγκεκριμένα τα φύλλα του δέντρου συμβολίζουν τις λέξεις της πρότασης και οι υπόλοιποι κόμβοι συμβολίζουν τον συντακτικό κανόνα που ακολουθούν οι απόγονοι του. Στην Εικόνα 8 απεικονίζεται το συντακτικό δέντρο της πρότασης «This laptop has incredible features».



Εικόνα 8. Παράδειγμα: *Συντακτικό δέντρο της πρότασης «This laptop has incredible features»*

Μόλις ολοκληρωθεί η συντακτική ανάλυση του κειμένου, ακολουθεί το στάδιο της αποσαφήνισης της έννοιας των λέξεων που υπάρχουν στο κείμενο. Στο στάδιο αυτό χρησιμοποιούνται μόνο τα ουσιαστικά, τα ρήματα, τα επίθετα και τα επιρρήματα. Αφού γίνει ο παραπάνω διαχωρισμός, εφαρμόζεται στις λέξεις κάθε πρότασης μια από τις τρεις προσφερόμενες μεθόδους επίλυσης του προβλήματος της αποσαφήνισης, First Sense (FS) ή Weighted Degree (WDEG) ή ILP. Πιο συγκεκριμένα, η μέθοδος FS δεν λύνει κάποιο πρόβλημα αποσαφήνισης αλλά επιστρέφει την πιο συχνά χρησιμοποιούμενη έννοια κάθε λέξης όπως αυτή παρέχεται από το WordNet. Αντίθετα οι μέθοδοι WDEG και ILP βασίζονται σε γράφους όπου με την βοήθεια ενός μέτρου ομοιότητας (SR, PMI, LL) επιλέγουν τη σωστή έννοια για κάθε λέξη.

Επόμενο βήμα αποτελεί η εύρεση του συναισθήματος που μεταφέρει κάθε λέξη. Προκειμένου να γίνει αυτό εφικτό, χρησιμοποιείται το λεξικό συναισθήματος SentiWordNet. Σε κάθε λέξη της οποίας η έννοια έχει βρεθεί, αντιστοιχίζονται τρείς τιμές βάσει του λεξικού συναισθήματος. Η πρώτη αντιστοιχεί στο ποσό θετικού συναισθήματος που εκτιμάται ότι μεταφέρει, η δεύτερη στο ποσό αρνητικού συναισθήματος και η τρίτη στο ποσό ουδέτερου συναισθήματος. Το άθροισμα των τριών αυτόν τιμών ισούται με 1.

Τελευταίο βήμα αποτελεί η εξαγωγή του συναισθήματος κάθε πρότασης του κειμένου. Για το σκοπό αυτό ενσωματώθηκε η Deep Learning προσέγγιση του Πανεπιστημίου του Stanford, η οποία βασίζεται στο συντακτικό δέντρο που δημιούργησε ο συντακτικός αναλυτής. Ως αποτέλεσμα λαμβάνεται μια τιμή που εκφράζει σε μια κλίμακα από το 0 έως το 5 πόσο αρνητική ή θετική είναι η κάθε πρόταση.

3.1.2 API

Το στοιχείο της Διεπαφής Προγραμματισμού Εφαρμογών (API) είναι ένα κρίσιμο συστατικό κάθε υπηρεσίας ιστού. Αυτό συμβαίνει διότι πρώτον περιγράφει το σύνολο των λειτουργιών που παρέχει η υπηρεσία και δεύτερον καθορίζει τον τρόπο με τον οποίο οι λειτουργίες της υπηρεσίες θα χρησιμοποιηθούν. Έτσι δίνει τη δυνατότητα στους ενδιαφερόμενους να κάνουν χρήση της υπηρεσίας χωρίς να έχουν πρόσβαση στον πηγαίο κώδικα της.

Το Pythia API βασίζεται στην αρχιτεκτονική Representational State Transfer (REST) και μπορεί να χρησιμοποιηθεί αποστέλλοντας αιτήσεις HTTP. Παρέχει ξεχωριστές κλήσεις για κάθε μια από τις τρεις υποστηριζόμενες μεθόδους αποσαφήνισης της έννοιας των λέξεων και τα δεδομένα που επιστρέφονται είναι σε μορφή JavaScript Object Notation (JSON) ούτως ώστε η επεξεργασία τους να γίνεται με απλό και αποτελεσματικό τρόπο.

3.1.3 Web Interface

Το Web Interface είναι το μέσο με το οποίο ο χρήστης αλληλεπιδρά με την υπηρεσία. Αποτελεί τον τρόπο με τον οποίο ο χρήστης θα εξερευνήσει τις δυνατότητες της υπηρεσίας και θα χρησιμοποιήσει την ίδια την υπηρεσία. Έτσι καθίσταται αναγκαία η κατασκευή ενός ιστότοπου που θα είναι προσβάσιμος και εξίσου λειτουργικός από όλους τους περιηγητές διαδικτύου (web browsers).

Καθώς ο όγκος της πληροφορίας που επιστρέφεται στον χρήστη είναι μεγάλος απαιτείται η σωστή δόμηση του περιεχομένου. Ο σχεδιασμός του είναι λιτός και ξεκάθαρος και δίνει στο στον χρήστη την πληροφορία με εύληπτο τρόπο. Επίσης παρουσιάζει αναλυτικά όλες τις δυνατότητες του ιστότοπου και συνάμα παρέχει την απαραίτητη καθοδήγηση προκειμένου η χρήση της υπηρεσίας να γίνει ακόμα ευκολότερη.

Κεφάλαιο 4 Υλοποίηση

Στο παρόν κεφάλαιο περιγράφεται η υλοποίηση του συστήματος Pythia, το οποίο αναπτύχθηκε στα πλαίσια της παρούσας πτυχιακής. Συγκεκριμένα, παρατίθενται αναλυτικές πληροφορίες σχετικά με την υλοποίηση του web service του συστήματος, του Pythia API καθώς και του web interface που δημιουργήθηκε.

4.1 Web Service

Το Web Service του συστήματος είναι το στοιχείο το οποίο πραγματοποιεί όλες τις απαραίτητες διεργασίες με σκοπό να προσφέρει πληροφορίες σχετικές με την έννοια των λέξεων και το συναισθηματικό φορτίο αυτών και των προτάσεων του κειμένου εισόδου. Έχει υλοποιηθεί σε Java και χρησιμοποιεί όλες τις βασικές δομές της για να επιτύχει την απαιτούμενη λειτουργικότητα.

4.1.1 Parser

Η πρώτη διεργασία του WS είναι η εφαρμογή του συντακτικού αναλυτή στο κείμενο εισόδου. Ο συντακτικός αναλυτής που χρησιμοποιήθηκε έχει δημιουργηθεί από το πανεπιστήμιο του Stanford και περιέχεται σε ένα αρχείο μορφής jar. Εξαιτίας του μεγάλου μεγέθους του και προκειμένου να αποφευχθεί η μεγάλη κατανάλωση μνήμης, ο συντακτικός αναλυτής φορτώνεται μια φορά κατά την εκκίνηση του WS και από εκεί στο εξής όταν χρειάζεται να κληθεί, επαναχρησιμοποιείται το αντικείμενο στο οποίο έχει ανατεθεί. Αφού ολοκληρωθεί ο διαχωρισμός των προτάσεων και η επισήμανση του μέρους του λόγου των λέξεων τα δεδομένα αποθηκεύονται σε ArrayLists. Ένα αντικείμενο ArrayList της Java είναι ουσιαστικά μια συλλογή αντικειμένων. Ο λόγος που προτιμήθηκε η συγκεκριμένη δομή δεδομένων είναι η ικανότητα που έχει να αυξομειώνει το μέγεθος της κατά την διάρκεια εκτέλεσης. Έτσι κάθε πρόταση είναι ένα ArrayList από αντικείμενα μιας κλάσης Term. Κάθε ένα αντικείμενο Term αντιστοιχεί σε ένα όρο της πρότασης και διατηρεί πληροφορίες όπως το μέρος του λόγου στο οποίο ανήκει και η ρίζα του. Εκτός των παραπάνω, αποθηκεύει πληροφορίες σχετικά με την επεξήγηση του και το συναίσθημα που μεταφέρει, οι οποίες όμως συμπληρώνονται κατά την ολοκλήρωση των διεργασιών που ακολουθούν.

Όπως αναφέρθηκε στο Κεφάλαιο 3 η επισήμανση του μέρους του λόγου στον οποίο ανήκει κάθε όρος της πρότασης γίνεται χρησιμοποιώντας τις ετικέτες Penn Treebank. Οι ετικέτες αυτές διαχωρίζουν λεπτομερώς τον χώρο που ανήκει κάθε λέξη. Για παράδειγμα αν ένα ουσιαστικό είναι στον ενικό (π.χ. dog)

λαμβάνει την ετικέτα NN ενώ αν είναι στον πληθυντικό (π.χ. dogs) λαμβάνει την ετικέτα NNS. Στο σύστημα Pythia αυτές οι περιπτώσεις ομαδοποιούνται στην αντίστοιχη γενική κατηγορία. Συγκεκριμένα, οι όροι με ετικέτες NN, NNP, NNPS, NNS εντάσσονται στην κατηγορία των ουσιαστικών, οι όροι με ετικέτες VB, VBD, VBG, VBN, VBP, VBZ εντάσσονται στην κατηγορία των ρημάτων, οι όροι με ετικέτες JJ, JJR, JJS εντάσσονται στην κατηγορία των επιθέτων και οι όροι με ετικέτες RB, RBR, RBS, WRB εντάσσονται στην κατηγορία των επιρρημάτων. Αυτό συμβαίνει διότι οι αλγόριθμοι αποσαφήνισης της έννοιας των λέξεων καθώς και ο μηχανισμός συναισθηματικής ανάλυσής τους, που ενσωματώθηκαν στο WS, έχουν βασιστεί στο WordNet το οποίο ορίζει συσχετίσεις μόνο μεταξύ αυτών των τεσσάρων κατηγοριών λέξεων. Επίσης η ανάκτηση της ρίζας των λέξεων γίνεται μέσω του WordNet.

Η ανάκτηση των λεξιλογικών πληροφοριών του WordNet γίνεται μέσω της εφαρμογής που αναπτύχθηκε από τους δημιουργούς του. Προκειμένου το WS να αποκτήσει πρόσβαση στα δεδομένα της εφαρμογής χρησιμοποιήθηκε η βιβλιοθήκη JWNL (Java WordNet Library).

4.1.2 WSD

Στη συνέχεια ακολουθεί το στάδιο της αποσαφήνισης της έννοιας των λέξεων. Για κάθε ένα διαφορετικό τρόπο αποσαφήνισης υπάρχει και η αντίστοιχη μέθοδος που πρέπει να κληθεί. Όλες οι διαθέσιμες μέθοδοι (wsdFSonlySentence, wsdWDEGSentence, wsdSentence) επεξεργάζονται το κείμενο εισόδου σε επίπεδο πρότασης. Κατά την εκκίνηση της διαδικασίας, διαχωρίζονται οι λέξεις οι οποίες είναι εφικτό να συμμετέχουν στην διαδικασία της αποσαφήνισης (ουσιαστικά, ρήματα, επίθετα και επιρρήματα) και αποθηκεύονται σε ένα ArrayList. Έπειτα εφαρμόζεται σε αυτές ο επιλεγμένος τρόπος αποσαφήνισης. Στην περίπτωση που η πρόταση έχει μόνο έναν όρο, τότε ανεξαρτήτως της μεθόδου αποσαφήνισης που έχει επιλεχθεί, η έννοια του καθορίζεται βάσει της πιο χρησιμοποιούμενης έννοιας όπως αυτή παρέχεται από το WordNet.

Ιδιαίτερη προσοχή απαιτείται κατά την ολοκλήρωση της διαδικασίας, όπου το ArrayList που περιέχει όλους τους όρους της πρότασης πρέπει να εμπλουτιστεί με την πληροφορία των αποσαφηνισμένων όρων. Η έξοδος συγκεκριμένων μεθόδων αποσαφήνισης (wsdWDEGSentence, wsdSentence) είναι ένα ArrayList από Terms που όμως δεν ακολουθούν την ίδια σειρά με τα Terms του ArrayList που δέχτηκαν οι μέθοδοι ως είσοδο. Για παράδειγμα, όταν δοθεί η πρόταση «This laptop has incredible features» σε μια από τις παραπάνω μεθόδους θα δημιουργηθούν τρία ArrayLists. Το πρώτο θα περιέχει ένα Term για κάθε όρο της πρότασης, το δεύτερο θα περιέχει μόνο τους όρους που θα συμμετάσχουν στη διαδικασία αποσαφήνισης («laptop», «has», «incredible», «features») και το τρίτο θα περιέχει τους όρους του δεύτερου, αποσαφηνισμένους, αλλά σε διαφορετική σειρά. Υπό αυτές τις συνθήκες η αντιστοιχία των όρων του πρώτου ArrayList με αυτούς του τρίτου γίνεται ακόμα δυσκολότερη. Προκειμένου να αντιμετωπιστεί αυτό το πρόβλημα, χρησιμοποιήθηκε η δομή HashMap της Java.

Το HashMap είναι στην ουσία ένας πίνακας που αντιστοιχεί κλειδιά σε τιμές. Κάθε κλειδί είναι μοναδικό και αντιστοιχίζεται σε μια μόνο τιμή. Επίσης επιτρέπει την αναζήτηση μιας τιμής βάσει του κλειδιού της.

Αφού δημιουργηθεί το ArrayList με τις αποσαφηνισμένες λέξεις, δημιουργείται ένα HashMap με κλειδί τη ρίζα των λέξεων που υπάρχουν στο ArrayList και τιμή τα αντικείμενα Term που αντιστοιχούν σε κάθε ρίζα. Έπειτα, γίνεται μια διάσχιση του αρχικού ArrayList και αν η ρίζα κάποιου όρου της πρότασης βρίσκεται μέσα στο HashMap συμπληρώνονται οι πληροφορίες που σχετίζονται με την ετυμολογία του.

4.1.3 SentiWordNet

Ο προσδιορισμός του συναισθήματος των λέξεων του κειμένου εισόδου γίνεται με τη χρήση του λεξικού συναισθήματος SentiWordNet, το οποίο όπως μαρτυράει και το όνομα του βασίζεται στο WordNet. Αυτό σημαίνει ότι για να προσδιοριστεί το συναίσθημα ενός λήμματος πρέπει πρώτα να έχει καθοριστεί η έννοια του βάσει του WordNet. Το SentiWordNet ουσιαστικά αποτελείται από ένα έγγραφο κειμένου που κάθε εγγραφή του αντιστοιχίζει το μοναδικό αναγνωριστικό κάθε έννοιας του WordNet με τρεις τιμές. Η πρώτη τιμή αναφέρεται στο θετικό συναίσθημα που μεταφέρει το λήμμα, η δεύτερη στο ουδέτερο και η τρίτη στο αρνητικό.

Υπάρχουν δύο εκδόσεις του SentiWordNet, η 1.0 η οποία βασίζεται στο WordNet 2.0 και η 3.0 η οποία βασίζεται στο WordNet 3.0 και συνάμα παρέχει αρκετές βελτιώσεις έναντι της πρώτης. Ως εκ τούτου, στο σύστημα που υλοποιήθηκε επιλέχθηκε να ενσωματωθεί η τελευταία έκδοση του SentiWordNet. Αυτή η επιλογή είχε ως αποτέλεσμα το αρχείο του SentiWordNet να μην μπορεί να χρησιμοποιηθεί απευθείας διότι οι μέθοδοι αποσαφήνισης βασίζονταν στο WordNet 2.0.

Η ασυμβατότητα αυτή αντιμετωπίστηκε με την εφαρμογή WordNet SQL Builder. Πρόκειται για μια Java εφαρμογή, η οποία δημιουργεί μια SQL βάση δεδομένων με δύο πίνακες. Ο ένας πίνακας αντιστοιχεί τα μοναδικά αναγνωριστικά των εννοιών του WordNet 2.0 σε αυτά του WordNet 2.1, ενώ ο άλλος αντιστοιχεί τα μοναδικά αναγνωριστικά των εννοιών του WordNet 2.1 σε αυτά του WordNet 3.0. Όπως είναι προφανές, η λύση του προβλήματος της ασυμβατότητας προήλθε από την συνένωση των δύο πινάκων και έτσι η έκδοση 3.0 του SentiWordNet έγινε συμβατή με την έξοδο των μεθόδων αποσαφήνισης.

Το τελευταίο πρόβλημα που έπρεπε να αντιμετωπιστεί ήταν ο τρόπος προσπέλασης του αρχείου του SentiWordNet. Το μεγάλο του μέγεθος καθιστούσε απαγορευτική από άποψη απόδοσης τη σειριακή αναζήτηση. Γι’ αυτό το λόγο προτιμήθηκε η χρήση τεσσάρων HashMap, ενός για κάθε ξεχωριστή κατηγορία λημμάτων (ουσιαστικά, ρήματα, επίθετα, επιρρήματα). Κατά την εκκίνηση του συστήματος γίνεται μια διάσχιση του αρχείου του SentiWordNet και οι τιμές που υποδεικνύουν τη συναισθηματική πολικότητα κάθε έννοιας αποθηκεύονται στο κατάλληλο HashMap με κλειδί το μοναδικό αναγνωριστικό της έννοιας στο WordNet. Έτσι όταν χρειαστεί να ενημερωθούν οι πληροφορίες που σχετίζονται με το συναίσθημα

ενός όρου, το WS ανατρέχει στο κατάλληλο HashMap και χρησιμοποιώντας το αναγνωριστικό της έννοιας του ανακτά την ζητούμενη πληροφορία.

4.1.4 Stanford Sentiment

Η τελευταία διεργασία που επιτελεί το WS είναι ο καθορισμός του συναισθήματος που μεταφέρει κάθε πρόταση του κειμένου. Η προσέγγιση που χρησιμοποιήθηκε έχει υλοποιηθεί από το Πανεπιστήμιο του Stanford και προκειμένου να λειτουργήσει απαιτούνται μια σειρά από αρχεία jar.

4.2 API

Ο τρόπος με τον οποίο έχει υλοποιηθεί το WS του συστήματος Pythia δίνει τη δυνατότητα σε όποιον ενδιαφέρεται να χρησιμοποιήσει απομακρυσμένα την υπηρεσία καθώς και να ενσωματώσει την λειτουργικότητα της σε άλλες εφαρμογές. Η αρχιτεκτονική REST στην οποία βασίζεται, επιτρέπει την μεταβολή της κατάστασης των πόρων του συστήματος μέσω του πρωτοκόλλου HTTP από διάφορους πελάτες (clients) ανεξαρτήτως της γλώσσας προγραμματισμού που χρησιμοποιούν. Για να γίνει αυτό εφικτό χρησιμοποιήθηκε το Jersey Framework 1.13. Το Jersey Framework είναι μια βιβλιοθήκη ανοικτού κώδικα η οποία διευκολύνει την ανάπτυξη RESTful Web Services και βασίζεται στο JAX-RS (Java API for RESTful Web Services) της Java.

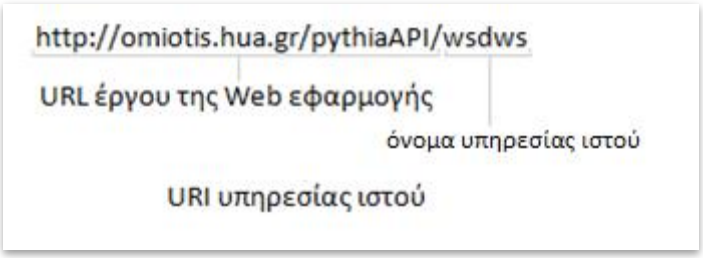
Το γεγονός ότι η βάση όλων των λειτουργιών που προσφέρει το σύστημα Pythia είναι η αποσαφήνιση της έννοιας των λέξεων οδήγησε στην επιλογή να διαχωριστούν οι πόροι του συστήματος βάσει των μεθόδων αποσαφήνισης που ενσωματώθηκαν σε αυτό. Στα πλαίσια της αρχιτεκτονικής REST όλοι οι πόροι της υπηρεσίας θα πρέπει να χαρακτηρίζεται από ένα URI μέσω του οποίου θα μπορούν να είναι προσβάσιμοι. Έτσι κάθε μέθοδος αποσαφήνισης του συστήματος έχει και ένα ξεχωριστό URI, στο οποίο μπορούν να αποστέλλονται αιτήματα HTTP GET ή HTTP POST.

Ο καθορισμός του URI για πρόσβαση στην υπηρεσία ιστού γίνεται καταρχήν με τον καθορισμό του ονόματος της υπηρεσίας. Το όνομα ορίζεται με την επισήμανση (Annotation) **@Path** του JAX-RS η οποία τοποθετείται πριν από την κλάση που υλοποιεί το WS. Στην περίπτωση του συστήματος Pythia το όνομα της υπηρεσίας είναι το «wsdws» όπως φαίνεται παρακάτω.

```
@Path("wsdws")
public class WSDWS {
    .
    ..
}
```

Εικόνα 9. Δήλωση ονόματος υπηρεσίας ιστού

Αυτό το όνομα στη συνέχεια προσαρτάται στο URL του έργου της Web εφαρμογής και έτσι πλέον είναι δυνατή η αναφορά στην υπηρεσία ιστού με το URI που προκύπτει. Παρακάτω παρουσιάζεται ο τρόπος σύνθεσης του URI του Pythia WS.



Εικόνα 10. Ορισμός του URI της υπηρεσίας ιστού

Εκτός από τον ορισμό της διαδρομής του WS, ο σχολιασμός **@Path** μπορεί να χρησιμοποιηθεί και για τον ορισμό της διαδρομής προς μια συγκεκριμένη μέθοδο της κλάσης. Μέρη της διαδρομής που περικλείονται σε άγκιστρα υποδεικνύουν τις παραμέτρους οι οποίες περνούν στην υπηρεσία ιστού. Στην περίπτωση που η μέθοδος προσπελάζεται μέσω ενός HTTP αιτήματος GET και χρησιμοποιεί ως ορίσματα, τα ορίσματα του URL της, τότε αυτά θα πρέπει να αντιστοιχηθούν. Για να συμβεί αυτό απαιτείται η προσθήκη του σχολιασμού **@PathParam** στα ορίσματα της μεθόδου. Αντίθετα όταν μια μέθοδος προσπελάζεται μέσω ενός HTTP αιτήματος POST στα ορίσματα της προστίθεται ο σχολιασμός **@FormParam**. Αξίζει να σημειωθεί ότι ο διαχωρισμός μεταξύ μιας μεθόδου που λαμβάνει αιτήματα HTTP GET και μιας μεθόδου που λαμβάνει αιτήματα HTTP POST πραγματοποιείται με το σχολιασμό **@GET** και **@POST** αντίστοιχα ενώ ο προσδιορισμός του τύπου των δεδομένων που επιστρέφει μια μέθοδος στον πελάτη γίνεται με το σχολιασμό **@Produces**. Οι πιο συνηθισμένοι τύποι μορφοποίησης δεδομένων στις REST υπηρεσίες είναι JSON (application/json), XML (application/xml) και XHTML (application/xhtml+xml). Στις εικόνες που ακολουθούν παρουσιάζεται η ίδια μέθοδος υλοποιημένη την πρώτη φορά να δέχεται αιτήματα HTTP GET και την δεύτερη να δέχεται αιτήματα HTTP POST.

```
@GET //ορισμός του τύπου των αιτήσεων που λαμβάνει η μέθοδος
@Path("/welcome/{name}") //πέρασμα του name ως παράμετρο της διαδρομής
@Produces("application/json")   //καθορισμός τύπου δεδομένων επιστροφής
public String welcome (@PathParam("name") String name) {
    ..
    .
}
```

Εικόνα 11. *Μέθοδος η οποία λαμβάνει αιτήματα HTTP GET.*

```
@POST //ορισμός του τύπου των αιτήσεων που λαμβάνει η μέθοδος
@Path("/welcome") //ορισμός της διαδρομής χωρίς ορίσματα
@Produces("application/json")   //καθορισμός τύπου δεδομένων επιστροφής
public String welcome (@FormParam("name") String name) {
    ..
    .
}
```

Εικόνα 12. *Μέθοδος η οποία λαμβάνει αιτήματα HTTP POST.*

Προκειμένου να είναι η διαδικασία κλήσης κάθε μεθόδου του Pythia WS πιο ξεκάθαρη, αποφασίστηκε κάθε μέθοδος να δέχεται αιτήματα HTTP σε ξεχωριστό URL. Οι τύποι των αιτημάτων που μπορεί να λάβει μια μέθοδος είναι δύο, HTTP GET και HTTP POST. Στον πίνακα που ακολουθεί φαίνονται τα URL των μεθόδων του Pythia WS.

Μέθοδος	**URL**	**HTTP GET**	**HTTP POST**
FS	http://omiotis.hua.gr/pythiaAPI/wsdws/wsdFS	✓	✓
WDEG	http://omiotis.hua.gr/pythiaAPI/wsdws/wsdWDEG	✓	✓
ILP	http://omiotis.hua.gr/pythiaAPI/wsdws/wsdILP	✓	✓

Πίνακας 3. Τ*α URL των μεθόδων του Pythia WS.*

Ο λόγος ενσωμάτωσης της δυνατότητα κλήσης μιας μεθόδου μέσω HTTP POST, οφείλεται στους περιορισμούς που θέτει ο Apache Tomcat, στον οποίο φιλοξενείται το WS, σχετικά με τους ειδικούς χαρακτήρες που μπορούν να περιέχονται σε ένα URL. Συγκεκριμένα, μια από τις παραμέτρους όλων των μεθόδων είναι το κείμενο εισόδου. Όταν αυτό αποστέλλεται μέσω του URL σε ένα αίτημα HTTP GET και περιέχει κάποιο ειδικό χαρακτήρα που έχει αποκλείσει ο Apache τότε ο εξυπηρετητής επιστρέφει μήνυμα λάθους.

Στη συνέχεια παρουσιάζονται οι παράμετροι που απαιτούνται για την κλήση των μεθόδων του Pythia WS.

	Παράμετροι			
Μέθοδος	**text**	**parserType**	**metric**	**prn**
FS	✓	✓		
WDEG	✓	✓	✓	
ILP	✓	✓	✓	✓

Πίνακας 4. Οι παράμετροι *των μεθόδων του Pythia WS.*

Συνοπτικά:

- Η παράμετρος text αναφέρεται στο κείμενο εισόδου.
- Η παράμετρος parserType αναφέρεται στο είδος του συντακτικού αναλυτή που θα χρησιμοποιηθεί. Οι επιτρεπτές τιμές είναι «short» και «long».

- Η παράμετρος metric αναφέρεται στην μετρική που θα χρησιμοποιηθεί από τη μέθοδο αποσαφήνισης. Οι επιτρεπτές τιμές είναι «sr», «pmi» και «ll».
- Η παράμετρος prn αναφέρεται στην περικοπή των εννοιών που δεν περιέχουν κάποιο από τους όρους της πρότασης μέσα στον ορισμό τους. Οι επιτρεπτές τιμές είναι «prune» και «full».

Τέλος, παρουσιάζεται ένα δείγμα των δεδομένων που επιστρέφονται από το Pythia WS. Τα δεδομένα επιστροφής είναι μορφοποιημένα σε JSON.


```
{
  "outputSentences": [
    {
      "sentence": [
        {
          "positiveSentiment": 0.25,
          "negativeSentiment": 0.125,
          "original": "is",
          "posTag": "VBZ",
          "definition": "have the quality of being; (copula, used with an adjective or a predicate noun);",
          "lemma": "be",
          "offsets": [
            2526983,
            2538502,
            2576058,
            2526084
          ],
          "numSenses": 13,
          "wsdoffset": 2526983
        }
      ],
      "sentenceSentiment": 2
    }
  ],
  "message": "OK"
}
```


Εικόνα 13. *Δείγμα των δεδομένων που επιστρέφονται από το Pythia WS.*

Συνοπτικά:

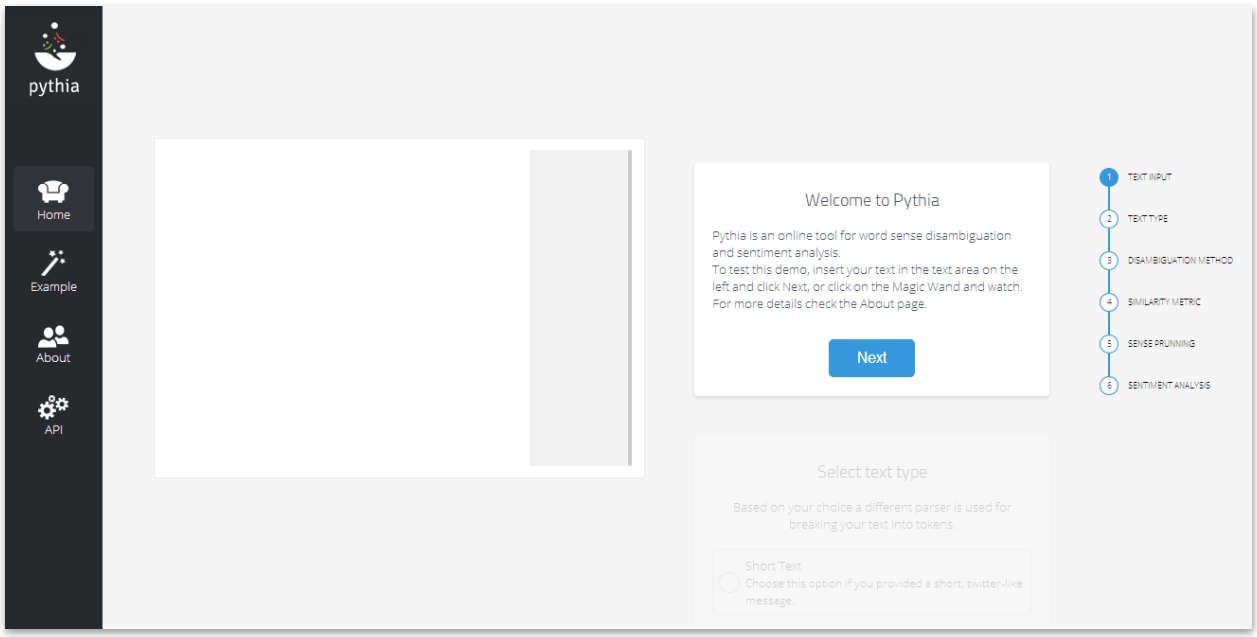
- Το γνώρισμα outputSentences περικλείει όλα τα αντικείμενα που αναπαριστούν τις προτάσεις του κειμένου εισόδου.
- Το γνώρισμα message αναφέρεται στην κατάσταση του HTTP αιτήματος. Αν αυτό ολοκληρώθηκε με επιτυχία έχει την τιμή «OK». Σε αντίθετη περίπτωση η τιμή του αναγράφει το πρόβλημα που παρουσιάστηκε.
- Το γνώρισμα sentence περικλείει όλα τα αντικείμενα που αναπαριστούν τις λέξεις της πρότασης.
- Το γνώρισμα positiveSentiment υποδηλώνει το ποσό του θετικού συναισθήματος που μεταφέρει η λέξη.
- Το γνώρισμα negativeSentiment υποδηλώνει το ποσό του αρνητικού συναισθήματος που μεταφέρει η λέξη
- Το γνώρισμα original υποδηλώνει την αρχική μορφή της λέξης.
- Το γνώρισμα posTag υποδηλώνει το μέρος του λόγου στο οποίο ανήκει η λέξη.

- Το γνώρισμα definition υποδηλώνει τον ορισμό της λέξης.
- Το γνώρισμα lemma υποδηλώνει τη ρίζα της λέξης.
- Το γνώρισμα offsets περικλείει τα αναγνωριστικά των εννοιών που έχει η λέξη στο WordNet.
- Το γνώρισμα numSenses υποδηλώνει τον αριθμό των διαφορετικών εννοιών που έχει η λέξη στο WordNet.
- Το γνώρισμα wsdoffset υποδηλώνει το αναγνωριστικό της έννοιας του WordNet που επιλέχτηκε για τη συγκεκριμένη λέξη.

Σημείωση: Το ποσό του ουδέτερου συναισθήματος που μεταφέρει η λέξη προκύπτει από την πράξη objectiveSentiment = 1 - positiveSentiment – negativeSentiment.

4.3 Web Interface

Αφού ολοκληρώθηκε η δημιουργία του Pythia WS, ακολούθησε η ανάπτυξη του γραφικού περιβάλλοντος της υπηρεσίας μέσω του οποίου δίνεται η ευκαιρία στον χρήστη να εξερευνήσει και να δοκιμάσει τις λειτουργίες τις υπηρεσίας. Η κατασκευή σελίδων του ιστότοπου βασίζεται στις τεχνολογίες HTML, CSS και JavaScript. Επίσης προκειμένου να επιτευχθεί το απαιτούμενο επίπεδο λειτουργικότητας του ιστότοπου χρησιμοποιήθηκαν οι παρακάτω βιβλιοθήκες JavaScript: jQuery, Chart, jquery-contenteditable, rangy, highlight, iCheck και slimScroll. Στη συνέχεια απεικονίζεται η αρχική σελίδα του ιστότοπου Pythia.



Εικόνα 14. *Αρχική σελίδα του ιστότοπου Pythia.*

Βασικός στόχος του Web Interface είναι η επίδειξη των δυνατοτήτων του συστήματος Pythia. Για την επίτευξη του παραπάνω στόχου απαιτείται η επικοινωνία του Web Interface και του Pythia

WS. Η επικοινωνία μεταξύ αυτών των δύο συστατικών στοιχείων πραγματοποιείται με την αποστολή HTTP αιτημάτων POST.

Όταν ο χρήστης καθορίσει τις παραμέτρους σύμφωνα με τις οποίες θα πραγματοποιηθεί η αποσαφήνιση των εννοιών και η συναισθηματική ανάλυση του κειμένου που έχει εισάγει, καλείται η συνάρτηση CallService που έχει υλοποιηθεί με την συνδρομή της JavaScript βιβλιοθήκης jQuery. Κατά την εκτέλεση της συνάρτησης CallService αποστέλλει στον εξυπηρετητή του Pythia WS ένα αίτημα HTTP POST χρησιμοποιώντας την τεχνολογία AJAX. Η εν λόγω τεχνολογία επιτρέπει την αποστολή και λήψη δεδομένων στο παρασκήνιο χωρίς να χρειάζεται η επανάληψη της φόρτωσης της σελίδας. Ακολουθεί το κομμάτι του κώδικα που πραγματοποιεί την αποστολή του αιτήματος προς τον εξυπηρετητή.

```
$.ajax({
    type: 'POST',
    url: serviceUrl,
    data: inputData,
    dataType: "json",
    success: function(data) {
        ..
        .
    }, error: function(jqXHR, textStatus, errorThrown)
    {
        ..
        .
    }
});
```

Εικόνα 15. Παράδειγμα *αιτήματος HTTP POST με χρήση της τεχνολογίας AJAX.*

Συνοπτικά:

- η παράμετρος **type** υποδηλώνει τον τύπο του αιτήματος που θα σταλεί
- η παράμετρος **url** υποδηλώνει το URL του εξυπηρετητή στον οποίο θα αποσταλεί το αίτημα
- η παράμετρος **data** αναφέρεται στα δεδομένα που θα πρέπει να σταλούν στον εξυπηρετητή
- η παράμετρος **dataType** υποδηλώνει τον τύπο των δεδομένων που θα επιστραφούν από τον εξυπηρετητή
- η συνάρτηση **success** περιλαμβάνει τον κώδικα που θα εκτελεστεί αν η κλήση ολοκληρωθεί με επιτυχία
- η συνάρτηση **error** περιλαμβάνει τον κώδικα που θα εκτελεστεί σε περίπτωση λάθους.

Σε περίπτωση που το αίτημα εξυπηρετήθηκε επιτυχώς, γίνεται η απεικόνιση των δεδομένων που επιστράφηκαν από το Pythia WS. Τα δεδομένα επιστροφής είναι αποθηκευμένα σε ένα δισδιάστατο πίνακα, όπου οι γραμμές του συμβολίζουν τις προτάσεις και οι στήλες τις λέξεις του κειμένου.

Η διαδικασία της απεικόνισης ξεκινάει με τον καθαρισμό του χώρου που εισάγεται το κείμενο του χρήστη. Αυτό γίνεται διότι το κείμενο πρέπει να επανεγγραφεί με τρόπο που θα εξυπηρετεί την ενσωμάτωση της πληροφορίας που επιστράφηκε. Στη συνέχεια, πραγματοποιείται μια διάσχιση του πίνακα που επιστράφηκε από το WS και κάθε λέξη αφού περικλειστεί σε ένα στοιχείο span της HTML με id της μορφής wordXY, επανεγγράφεται στο χώρο του κειμένου. Η ιδιότητα id ενός HTML στοιχείου είναι μοναδική σε κάθε ιστοσελίδα. Ο χαρακτήρας X του id είναι ένας ακέραιος που υποδηλώνει τον αριθμό της πρότασης στην οποία ανήκει η λέξη ενώ ο χαρακτήρας Y είναι ένας ακέραιος που υποδηλώνει τη σειρά που έχει η λέξη μέσα στην πρόταση. Η ίδια τακτική ακολουθείται και για τον διαχωρισμό των προτάσεων του κειμένου. Όλα τα στοιχεία span με τις λέξεις κάθε πρότασης περικλείονται σε ένα άλλο span με id της μορφής sentenceX, όπου X ένας ακέραιος που υποδηλώνει τον αριθμό της πρότασης. Επιπρόσθετα στα span των προτάσεων καθορίζεται η ιδιότητα class, η οποία παίρνει μια από τις τιμές «positive-sentence», «negative-sentence», «neutral-sentence» ανάλογα με το συναίσθημα που μεταφέρουν.

Εφόσον ολοκληρωθεί η επανεγγραφή του κειμένου, επόμενο στάδιο αποτελεί η παρουσίαση των πληροφοριών που αφορούν τις λέξεις του κειμένου. Οι πληροφορίες απεικονίζονται σε ένα πλαίσιο που εμφανίζεται όταν ο χρήστης μετακινήσει το δείκτη του ποντικιού πάνω από μια λέξη. Για να συμβεί αυτό ανατίθεται σε κάθε span λέξης ένα jQuery Event Handler. Μέσω αυτού είναι δυνατή η ανίχνευση της κίνησης του δείκτη του ποντικιού πάνω από το εκάστοτε στοιχείο. Έτσι όταν ο χρήστης τοποθετήσει τον δείκτη πάνω από μια λέξη, δημιουργείται ένα event που πυροδοτεί την εκτέλεση μιας ακολουθίας εντολών JavaScript. Μια από αυτές πραγματοποιεί την ανάκτηση του id του span και κατά συνέπεια των συντεταγμένων που έχει η λέξη μέσα στον πίνακα που επιστράφηκε από το WS. Αφού συμπληρωθούν στο πλαίσιο οι πληροφορίες της λέξης καλείται η μέθοδος slideDown() προκειμένου να εμφανιστεί το πλαίσιο με ένα ωραίο εφέ κύλισης. Σε περίπτωση που ο χρήστης απομακρύνει το δείκτη από τη λέξη δημιουργείται ένα event το οποίο αποκρύπτει το πλαίσιο. Η παρακάτω εικόνα απεικονίζει το πλαίσιο που εμφανίζεται.

Εικόνα 16. *Το πλαίσιο με τις πληροφορίες μιας λέξης.*

Το πλαίσιο που εμφανίζεται αποτελείται από δύο περιοχές. Η αριστερή περιοχή παρέχει πληροφορίες σχετικά με τη ρίζα, το μέρος του λόγου και τον ορισμό της λέξης ενώ η δεξιά παρέχει πληροφορίες για το συναίσθημα που μεταφέρει. Η πληροφόρηση του χρήστη σχετικά με τη συναισθηματική πολικότητα της λέξης γίνεται με ένα γράφημα δακτυλίου, το οποίο δημιουργήθηκε με την βιβλιοθήκη Chart.js.

Το τελικό στάδιο της απεικόνισης των δεδομένων που επιστράφηκαν περιλαμβάνει την παρουσίαση των πληροφοριών που σχετίζονται με το συναίσθημα κάθε πρότασης. Οι πληροφορίες αυτές απει-

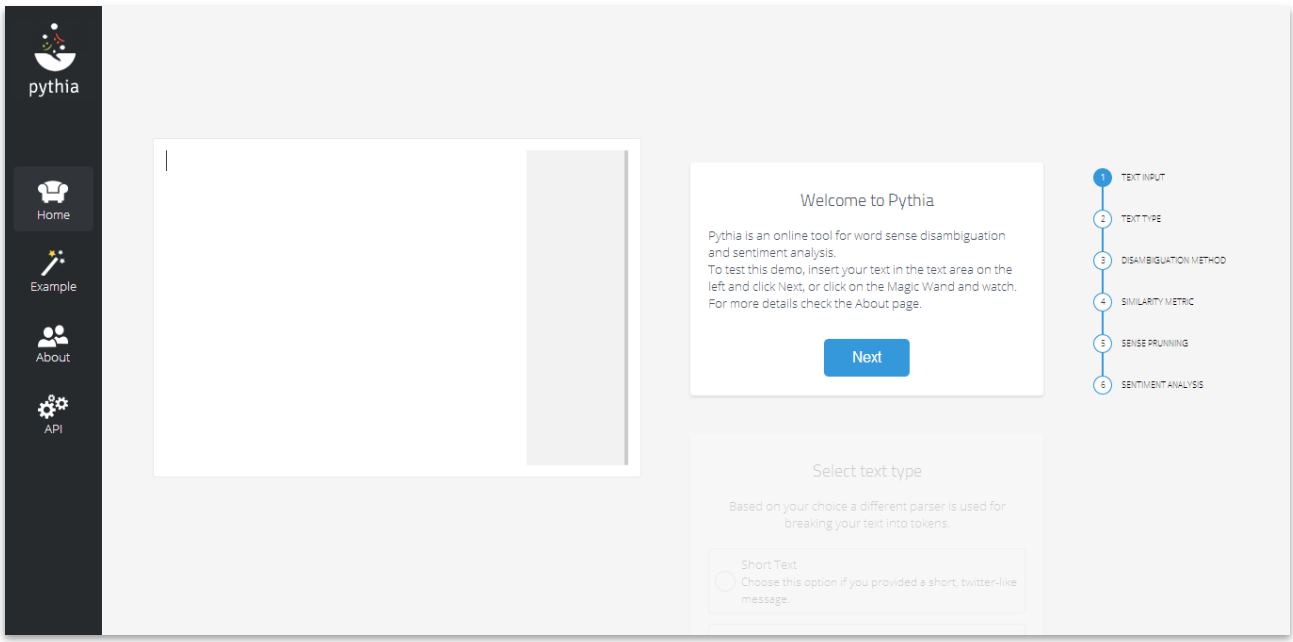
κονίζονται σε ένα πλαίσιο που εφάπτεται του χώρου στον οποίο ο χρήστης εισάγει το κείμενο. Το συγκεκριμένο πλαίσιο αποτελεί μια μικρογραφία του μορφοποιημένου κειμένου με τη διαφορά ότι η κάθε πρόταση απεικονίζεται ως μια χρωματιστή μπάρα. Ο καθορισμός του χρώματος των μπαρών γίνεται με βάση την τιμή της ιδιότητας class που περιέχεται στο span των προτάσεων. Η Εικόνα 17 παρουσιάζει το χώρο στον οποίο ο χρήστης εισάγει το κείμενο και το πλαίσιο στο οποίο απεικονίζεται το συναίσθημα κάθε πρότασης.

Εικόνα 17. *Ο χώρος εισαγωγής κειμένου και το πλαίσιο στο οποίο απεικονίζεται το συναίσθημα κάθε πρότασης.*

Εκτός των παραπάνω, στα πλαίσια κατασκευής του Web Interface υλοποιήθηκε μια μικρής διάρκειας ξενάγηση στην αρχική σελίδα του ιστότοπου η οποία βασίζεται στην JavaScript βιβλιοθήκη Trip. Η ξενάγηση ξεκινάει με το πάτημα του κουμπιού «Example» που υπάρχει στην μπάρα πλοήγησης και παρέχει τις κατάλληλες διευκρινήσεις προκειμένου να κάνει την χρήση της υπηρεσίας ακόμα πιο εύκολη.

Κεφάλαιο 5 Σενάριο χρήσης υπηρεσίας

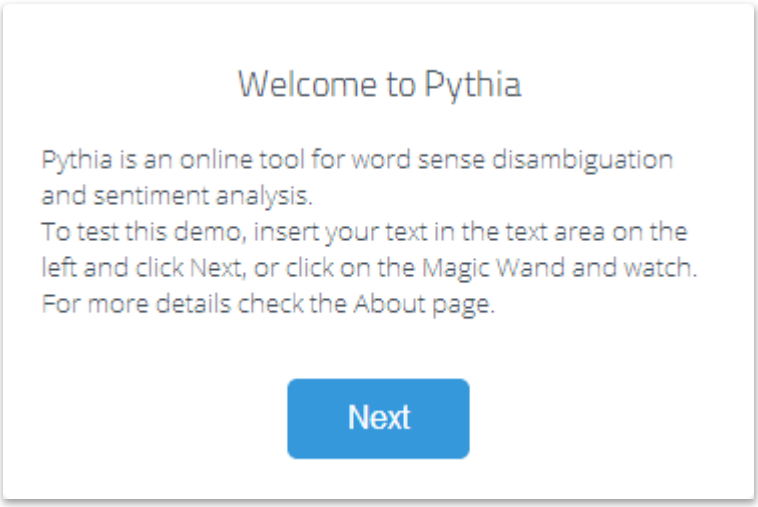
Στο παρόν κεφάλαιο παρουσιάζεται ο τρόπος χρήσης του Web Interface της υπηρεσίας Pythia. Σκοπός του κεφαλαίου είναι να παρουσιάσει την βασική ακολουθία ενεργειών που απαιτούνται από τον χρήστη προκειμένου να χρησιμοποιήσει την υπηρεσία. Στο συγκεκριμένο σενάριο, ως μέθοδος αποσαφήνισης επιλέγεται η ILP - PMI - PRUNE. Η Εικόνα 18 απεικονίζει την αρχική σελίδα της υπηρεσίας.



Εικόνα 18. *Αρχική σελίδα του ιστότοπου της υπηρεσίας*

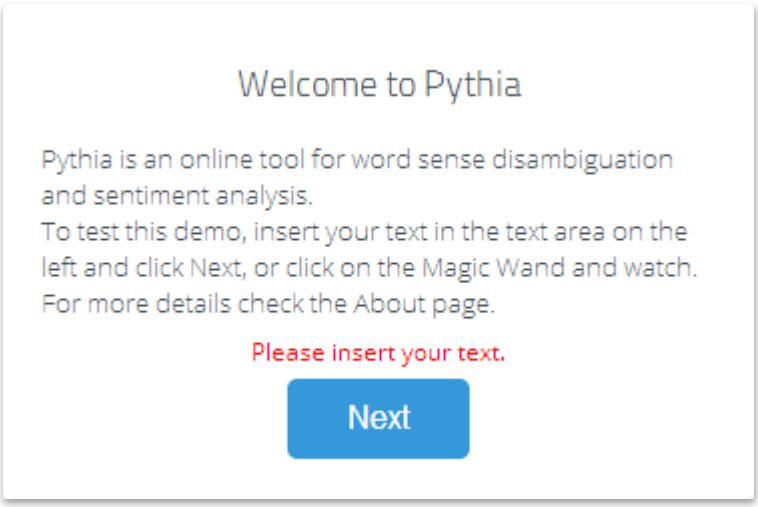
Στο αριστερό μέρος υπάρχει η μπάρα πλοήγησης, η οποία περιέχει το λογότυπο της υπηρεσίας και τέσσερα κουμπιά. Το κουμπί «Home» αναφέρεται στην αρχική σελίδα. Το κουμπί «Example» αναφέρεται στην εκτέλεση ενός σεναρίου χρήσης της υπηρεσίας που στοχεύει στο να δείξει στον χρήστη τον τρόπο λειτουργίας του ιστοτόπου. Το κουμπί «About» αναφέρεται στην σελίδα που περιέχει πληροφορίες σχετικά με την υπηρεσία Pythia ενώ το κουμπί «API» αναφέρεται στην σελίδα που περιέχει πληροφορίες σχετικά με το API της υπηρεσίας και τον τρόπο με τον οποίο μπορεί να το χρησιμοποιήσει. Μετά την μπάρα πλοήγησης ακολουθεί ο χώρος που ο χρήστης εισάγει το κείμενο του. Τέλος το δεξί μέρος της σελίδας αποτελείται από τον χώρο παραμετροποίησης της υπηρεσία καθώς και τον χώρο που απεικονίζει το στάδιο στο οποίο βρίσκεται η διαδικασία. Ο χώρος παραμετροποίησης της υπηρεσίας αποτελείται από κάποιες «κάρτες» που κατευθύνουν τον χρήστη έτσι ώστε να διευκολύνουν την χρήση της υπηρεσίας. Η πρώτη κάρτα

(βλέπε Εικόνα 19) εξηγεί στον χρήστη ότι για να ξεκινήσει την διαδικασία θα πρέπει να εισάγει το κείμενο του.



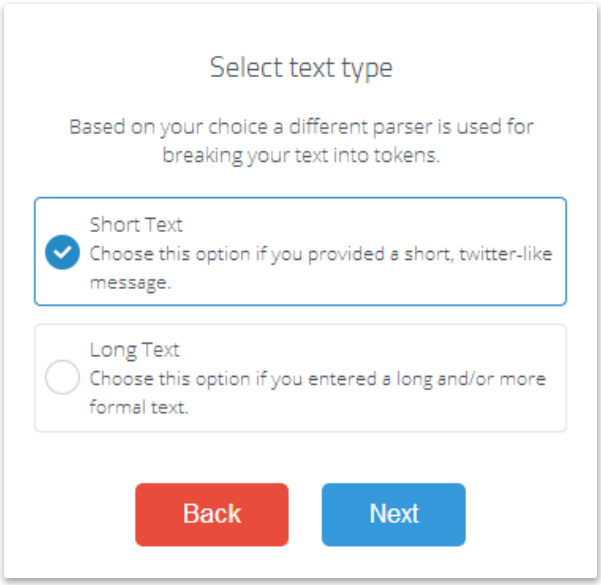
Εικόνα 19. *Κάρτα του χώρου παραμετροποίησης της υπηρεσίας που υποδεικνύει στο χρήστη ότι το πρώτο βήμα της διαδικασίας απαιτεί την εισαγωγή του κειμένου του*

Σε περίπτωση που ο χρήστης προσπαθήσει να προχωρήσει στο δεύτερο βήμα χωρίς να έχει εισάγει το κείμενο του, η διαδικασία σταματάει και εμφανίζεται ένα μήνυμα λάθους με κόκκινα γράμματα πάνω από το κουμπί «Next» (βλέπε Εικόνα 20). Αντίστοιχα μηνύματα εμφανίζονται και στα επόμενα βήματα της διαδικασίας.



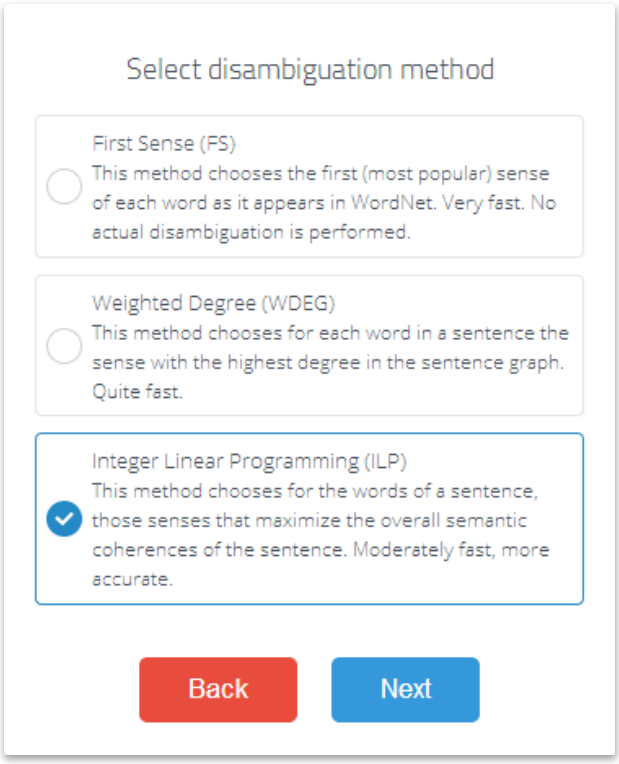
Εικόνα 20. Εμφάνιση μηνύματος λάθους πάνω από το κουμπί «Next»

Αφού ο χρήστης εισάγει επιτυχώς το κείμενο του, καλείται να επιλέξει τον συντακτικό αναλυτή που επιθυμεί να χρησιμοποιήσει (βλέπε Εικόνα 21). Στη συγκεκριμένη περίπτωση επιλέγεται ο συντακτικός αναλυτής που προορίζεται για ανεπίσημα κείμενα μικρού μήκους.



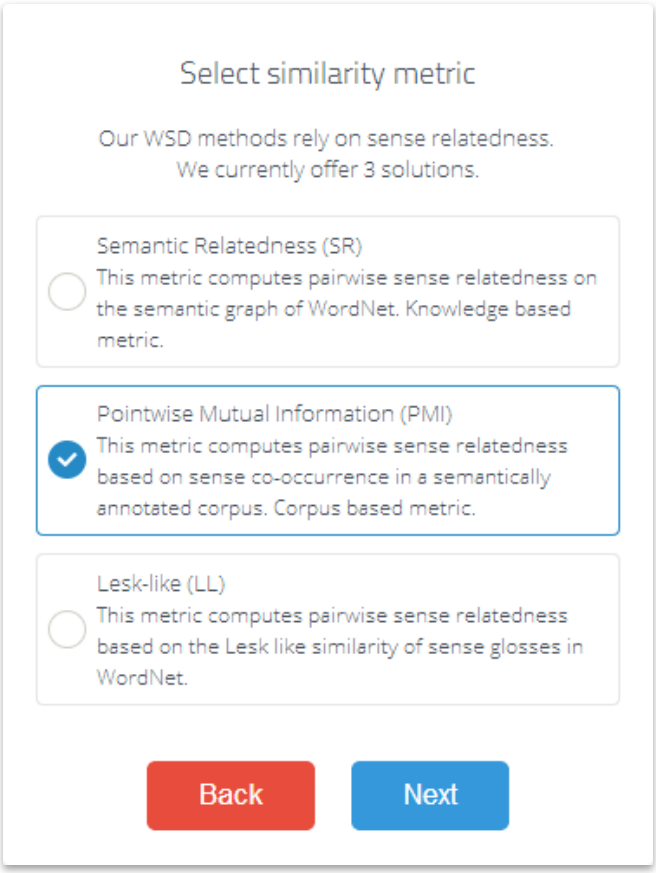
Εικόνα 21. *Κάρτα επιλογής συντακτικού αναλυτή*

Στη συνέχεια ο χρήστης καλείται να επιλέξει τη μέθοδο αποσαφήνισης που επιθυμεί να χρησιμοποιήσει (βλέπε Εικόνα 22). Στη συγκεκριμένη περίπτωση επιλέγεται η μέθοδος ILP.



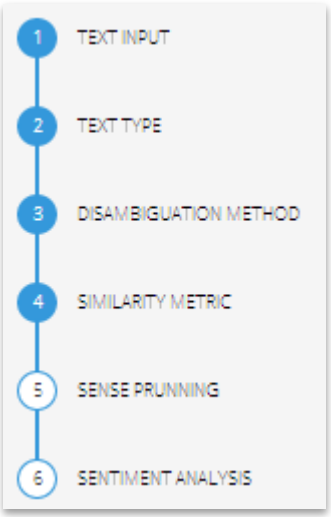
Εικόνα 22. *Κάρτα επιλογής μεθόδου αποσαφήνισης των λέξεων του κειμένου*

Έπειτα ο χρήστης καλείται να επιλέξει τη μετρική ομοιότητας που επιθυμεί να χρησιμοποιήσει (βλέπε Εικόνα 23). Στη συγκεκριμένη περίπτωση επιλέγεται η μετρική PMI.



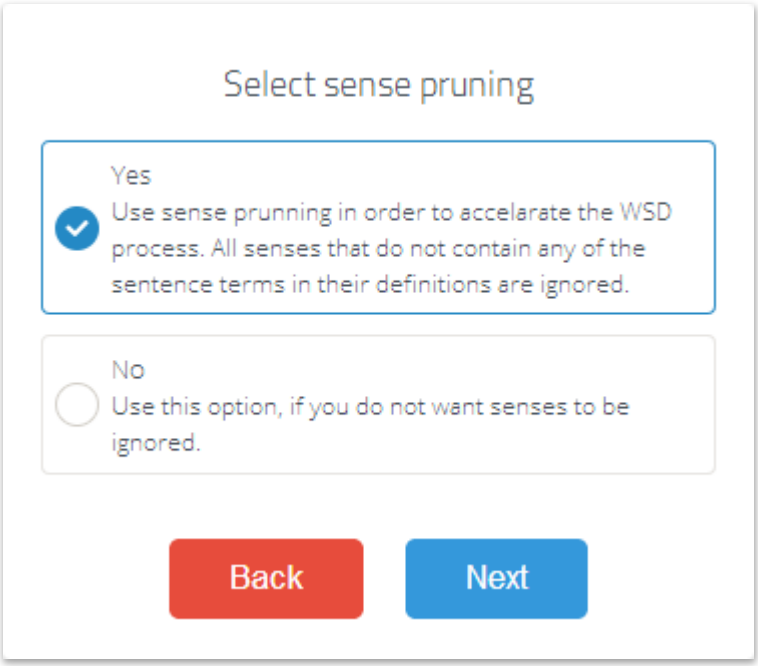
Εικόνα 23. *Κάρτα επιλογής μετρικής ομοιότητας*

Κατά τη διάρκεια παραμετροποίησης της υπηρεσίας ο χρήστης μπορεί να ενημερώνεται για την εξέλιξη της διαδικασίας (βλέπε Εικόνα 24).



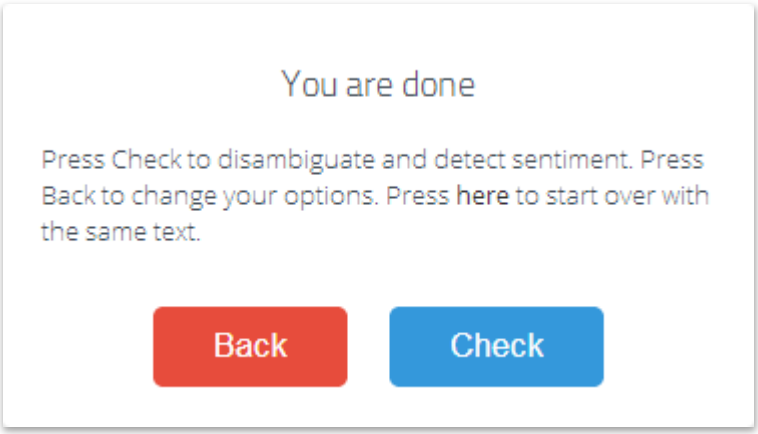
Εικόνα 24. *Ενημέρωση του χρήστη για την εξέλιξη της διαδικασίας*

Μετά την επιλογή της μετρικής ομοιότητας, ο χρήστης καλείται να επιλέξει αν θέλει να χρησιμοποιήσει την επιλογή κλαδέματος που υλοποιείται από την μέθοδο αποσαφήνισης ILP (βλέπε Εικόνα 25). Στην συγκεκριμένη περίπτωση γίνεται χρήση της επιλογής κλαδέματος.



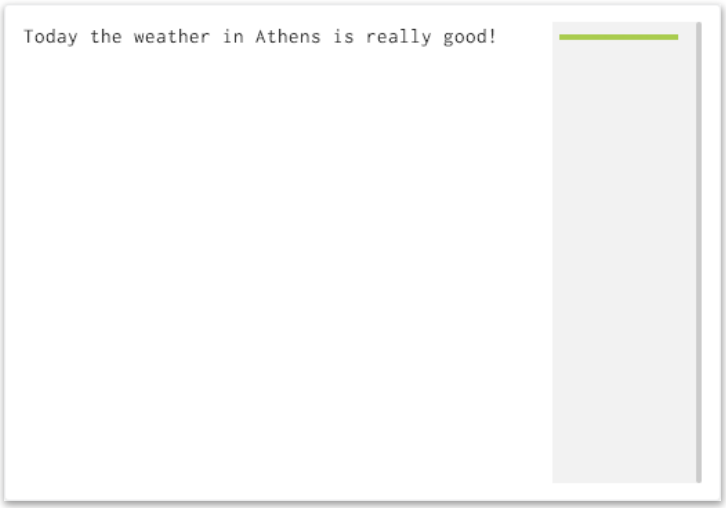
Εικόνα 25. *Κάρτα επιλογής για το αν θα εφαρμοστεί κλάδεμα των εννοιών ή όχι από τη μέθοδο ILP*

Αφού ο χρήστης επιλέξει αν θα εφαρμοστεί κλάδεμα των εννοιών ή όχι από τη μέθοδο ILP ολοκληρώνεται η διαδικασία παραμετροποίησης και ο χρήσης καλείται να πατήσει το κουμπί «Check» προκειμένου να ξεκινήσει η συντακτική ανάλυση και η αποσαφήνισης των εννοιών του κειμένου που έχει εισάγει (βλέπε Εικόνα 26).



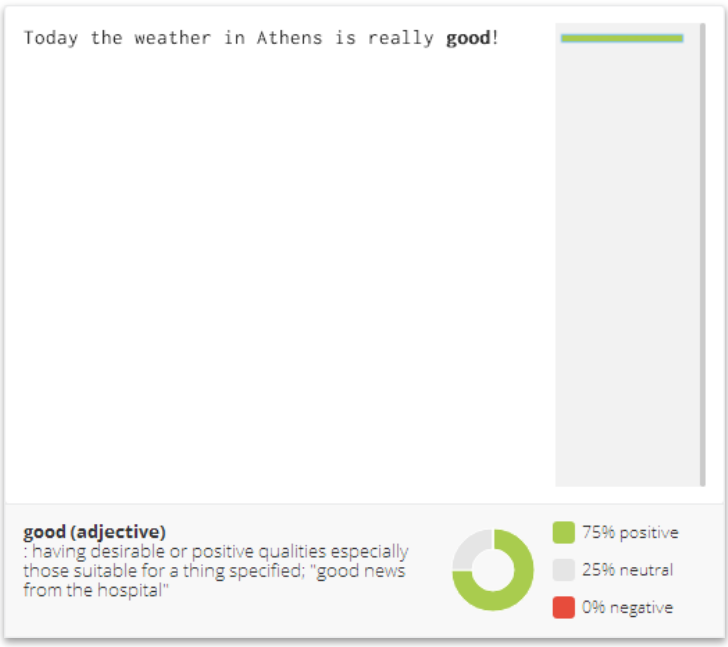
Εικόνα 26. *Κάρτα που υποδεικνύει την ολοκλήρωση της διαδικασίας παραμετροποίησης της υπηρεσίας*

Αφού ολοκληρωθεί η επεξεργασία των δεδομένων που εισήγαγε ο χρήστης, η περιοχή εισαγωγής κειμένου διαμορφώνεται όπως φαίνεται στην Εικόνα 27. Μέσα στο γκρι πλαίσιο απεικονίζεται το συναίσθημα που μεταφέρουν οι προτάσεις του κειμένου.



Εικόνα 27. *Η περιοχή εισαγωγής κειμένου μετά την ολοκλήρωση της συναισθηματικής ανάλυσης και της αποσαφήνισης των εννοιών του κειμένου*

Όταν ο χρήστης τοποθετήσει τον δείκτη του ποντικιού του πάνω από μια λέξη, εμφανίζεται ένα πλαίσιο κάτω από την περιοχή εισαγωγής κειμένου το οποίο περιέχει πληροφορίες σχετικά με την ερμηνεία και το συναισθηματικό της φορτίο. Παράλληλα στο γκρι πλαίσιο φωτίζεται η μπάρα που αντιστοιχεί στην πρόταση που περιέχει την συγκεκριμένη λέξη (βλέπε Εικόνα 28).



Εικόνα 28. *Εμφάνιση της ερμηνείας και του συναισθηματικού φορτίου της επιλεγμένης λέξης*

Κεφάλαιο 6 Αξιολόγηση

Μετά την ολοκλήρωση της υλοποίησης του συστήματος Pythia, ακολούθησε η αξιολόγηση της απόδοσής του. Η αξιολόγηση του συστήματος έγινε με βάση το Task 2 του SemEval 2013[3], ενός διεθνούς διαγωνισμού για συστήματα σημασιολογικής ανάλυσης. Το συγκεκριμένο task επικεντρώνεται στην συναισθηματική ανάλυση δημοσιεύσεων του κοινωνικού δικτύου Twitter και περιέχει δύο σκέλη. Το πρώτο, στο οποίο έγινε και η αξιολόγηση, αναφέρεται στην εύρεση του συναισθήματος που μεταφέρει μια λέξη ή μια φράση μιας δημοσίευσης, ενώ το δεύτερο στην εύρεση του συναισθήματος που εκφράζεται από ολόκληρη τη δημοσίευση. Ο λόγος για τον οποίο αποφασίστηκε να μην γίνει αξιολόγηση στο δεύτερο σκέλος είναι ότι η εύρεση του συναισθήματος σε επίπεδο πρότασης βασίζεται στο Stanford's Recursive Deep Model.

Η εξέταση των επιδόσεων του συστήματος Pythia έγινε πάνω στα τέσσερα σύνολα δεδομένων που διατέθηκαν από τους διοργανωτές. Τα τρία πρώτα σύνολα (δεδομένα εκπαίδευσης, ανάπτυξης και επαλήθευσης) αποτελούνταν από μηνύματα που προέρχονταν από το Twitter ενώ το τέταρτο (δεδομένα επαλήθευσης) αποτελούνταν από SMS. Αυτό συνέβη προκειμένου να διαπιστωθεί σε τι βαθμό ένα σύστημα συναισθηματικής ανάλυσης το οποίο προορίζεται για χρήση σε μηνύματα από το Twitter μπορεί να χρησιμοποιηθεί και σε άλλου είδους μηνύματα. Ο πίνακας που ακολουθεί παρουσιάζει την κατανομή των μηνυμάτων κάθε συνόλου ως προς το συναίσθημα που μετέφεραν.

DataSet	Positive	Negative	Neutral
Twitter - Dev	404	247	27
Twitter - Training	3714	1952	301
Twitter - Test	2734	1541	160
SMS - Test	1071	1104	159

Πίνακας 5. *Κατανομή των μηνυμάτων κάθε συνόλου ως προς το συναίσθημα που μετέφεραν.*

Η αξιολόγηση του συστήματος έγινε χρησιμοποιώντας την μέθοδο ILP. Ως μέτρο ομοιότητας επιλέχθηκε το PMI ενώ δεν έγινε χρήση κλαδέματος εννοιών. Η επιλογή της συγκεκριμένης μεθόδου έγινε βάσει των υψηλών επιδόσεων που είχε σημειώσει σε προηγούμενα πειράματα. Εκτός της επιλεγμένης μεθόδου, στα μηνύματα εφαρμόστηκαν και κάποιοι ευριστικοί μέθοδοι προκειμένου να γίνει καλύτερη διαχείριση των ιδιομορφιών (ιδιωματικές λέξεις, σύνδεσμοι, emoticons) τους. Για παράδειγμα αν εντοπι-

[3] http://www.cs.york.ac.uk/semeval-2013/task2/

ζόταν μια ιδιωματική λέξη μέσα σε ένα μήνυμα, αυτή αντικαθιστούνταν με τη ρίζα της. Επιπλέον αν η λέξη της οποίας έπρεπε να βρεθεί το συναίσθημα, ξεκινούσε με «http://», θεωρούταν σύνδεσμος και της αναθέτονταν απευθείας ουδέτερο συναίσθημα. Τέλος, προκειμένου να αντιμετωπιστεί η αδυναμία της μεθόδου να καθορίσει το συναίσθημα μερικών φράσεων, δημιουργήθηκε μια λίστα με θετικές λέξεις και emoticons και μια με αρνητικές. Αν ένας όρος της φράσης άνηκε σε μια από τις παραπάνω λίστες τότε στη φράση αναθέτονταν το αντίστοιχο το αντίστοιχο συναίσθημα.

Η αξιολόγηση του συστήματος έγινε βάσει της μετρικής F-Measure. Η μετρική F-Measure δίνει μια εκτίμηση της επίδοσης μιας εργασίας, συνδυάζοντας την ακρίβεια (precision) και την ανάκληση (recall). Πιο συγκεκριμένα, αποτελεί τον αρμονικό μέσο όρο (harmonic mean) της ακρίβειας και της ανάκλησης. Η μετρική αυτή είναι γνωστή και σαν F_1, επειδή προσδίδει την ίδια βαρύτητα στην ακρίβεια και την ανάκληση ενώ ορίζεται από την ακόλουθη εξίσωση:

$$F_1 = 2 \cdot \frac{presicion \cdot recall}{precision + recall}$$

Η μετρική F_1 υπολογίστηκε ξεχωριστά για τις θετικές και τις αρνητικές κλάσεις ενώ η συνολική απόδοση καθορίστηκε από τον μέσο αυτών των δύο. Στη συνέχεια παρουσιάζεται η απόδοση του συστήματος Pythia μαζί με την καλύτερη, χειρότερη και μέση απόδοση που σημειώθηκε στα πλαίσια του διαγωνισμού.

DataSet	Worst	Avg	Best	Pythia
Twitter - Dev	51.2	74.7	87.8	69.8
Twitter - Training	64.7	82.4	90.8	68.9
Twitter - Test	68.8	83.6	90.9	71.1
SMS - Test	66.5	78.5	81.2	69.7

Πίνακας 6. Συγκριτικά αποτελέσματα της απόδοσης (μετρική F_1) του συστήματος Pythia, της καλύτερης, χειρότερης και μέσης απόδοσης που σημειώθηκε στα πλαίσια του διαγωνισμού.

Κεφάλαιο 7 Επίλογος

7.1 Συμπεράσματα

Κατά τη διάρκεια εκπόνησης της πτυχιακής αναπτύχθηκε μια online υπηρεσία συναισθηματικής ανάλυσης και αποσαφήνισης της έννοιας των λέξεων. Επιπρόσθετα αναπτύχθηκε μια γραφική διεπαφή προκειμένου να κάνει ευκολότερη τη χρήση της.

Η υπηρεσία που αναπτύχθηκε επιχειρεί συναισθηματική ταξινόμηση σε επίπεδο έννοιας, αφού πρώτα έχει πραγματοποιήσει αποσαφήνιση των εννοιών της κάθε πρότασης του εισηγμένου κειμένου. Το γεγονός αυτό προσφέρει ευελιξία μεταξύ των επιμέρους συστατικών και δίνει τη δυνατότητα για περαιτέρω πειραματισμό και εξέλιξη αυτών. Παρ' όλα αυτά προσθέτει σαφείς περιορισμούς στην τελική απόδοση του συστήματος οι οποίοι προέρχονται από την αλληλεξάρτηση των συστατικών στοιχείων της υπηρεσίας. Έτσι, αν ένα συστατικό του συστήματος δεν έχει την αναμενόμενη απόδοση, σίγουρα θα επισκιάσει και τα υπόλοιπα.

7.2 Μελλοντικές επεκτάσεις

Στην παρούσα υλοποίηση η ανάθεση συναισθήματος σε επίπεδο πρότασης γίνεται χρησιμοποιώντας το Deep Learning Model που αναπτύχθηκε το από Πανεπιστήμιο του Stanford. Η μελέτη του τρόπου με τον καθορίζεται το συναισθηματικό φορτίο της πρότασης θα μπορούσε να αποτελέσει το έναυσμα για την υλοποίηση μια προσέγγισης που θα βασίζεται στην πληροφορία που προκύπτει από την αποσαφήνιση της έννοιας των λέξεων της πρότασης, σε συνδυασμό με το συντακτικό της δέντρο.

Βιβλιογραφία

1. Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In*LREC* (Vol. 10, pp. 2200-2204).
2. Banerjee, S., & Pedersen, T. (2003, August). Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI* (Vol. 3, pp. 805-810).
3. Benamara, F., Chardon, B., Mathieu, Y. Y., & Popescu, V. (2011). Towards Context-Based Subjectivity Analysis. In *IJCNLP* (pp. 1180-1188).
4. Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., & Reynar, J. (2008, April). Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era* (p. 14).
5. Brody, S., & Elhadad, N. (2010, June). An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 804-812). Association for Computational Linguistics.
6. Cilibrasi, R. L., & Vitanyi, P. M. (2007). The google similarity distance.*Knowledge and Data Engineering, IEEE Transactions on*, *19*(3), 370-383.
7. Dasgupta, S., & Ng, V. (2009, August). Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2* (pp. 701-709). Association for Computational Linguistics.
8. Davidov, D., Tsur, O., & Rappoport, A. (2010, August). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 241-249). Association for Computational Linguistics.
9. Esuli, A., & Sebastiani, F. (2011). Enhancing opinion extraction by automatically annotated lexical resources. In *Human Language Technology. Challenges for Computer Science and Linguistics* (pp. 500-511). Springer Berlin Heidelberg.
10. Galley, M., & McKeown, K. (2003, August). Improving word sense disambiguation in lexical chaining. In *IJCAI* (Vol. 3, pp. 1486-1488).
11. Gamon, M. (2004, August). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In*Proceedings of the 20th international conference on Computational Linguistics*(p. 841). Association for Computational Linguistics.
12. Harris, J. M., Hirst, J. L., & Mossinghoff, M. J. (2008). *Combinatorics and graph theory* (Vol. 2). Heidelberg: Springer.
13. Hatzivassiloglou, V., Klavans, J. L., Holcombe, M. L., Barzilay, R., Kan, M. Y., & McKeown, K. (2001). Simfinder: A flexible clustering tool for summarization. Proceedings of the NAACL Workshop on Automatic Summarizatio.

14. Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In*Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.
15. Iida, R., McCarthy, D., & Koeling, R. (2008, January). Gloss-Based Semantic Similarity Metrics for Predominant Sense Acquisition. In *IJCNLP* (pp. 561-568).
16. Ion, R., & Ştefănescu, D. (2011). Unsupervised word sense disambiguation with lexical chains and graph-based context formalization. In *Human Language Technology. Challenges for Computer Science and Linguistics* (pp. 435-443). Springer Berlin Heidelberg.
17. Jagtap, V. S., & Pawar, K. (2013). Analysis of different approaches to Sentence-Level Sentiment Classification. International Journal of Scientific Engineering and Technology (ISSN: 2277-1581) Volume, 2, 164-170.
18. Jakob, N., & Gurevych, I. (2010, October). Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*(pp. 1035-1045). Association for Computational Linguistics.
19. Jalilvand, A., & Salim, N. (2012). Sentiment Classification Using Graph Based Word Sense Disambigution. In *Advanced Machine Learning Technologies and Applications* (pp. 351-358). Springer Berlin Heidelberg.
20. Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, *22*(2), 110-125.
21. Kim, S. M., & Hovy, E. (2004, August). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics* (p. 1367). Association for Computational Linguistics.
22. Li, T., Zhang, Y., & Sindhwani, V. (2009, August). A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1* (pp. 244-252). Association for Computational Linguistics.
23. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1-167.
24. Matsumoto, S., Takamura, H., & Okumura, M. (2005). Sentiment classification using word subsequences and dependency sub-trees. In *Advances in Knowledge Discovery and Data Mining* (pp. 301-311). Springer Berlin Heidelberg.
25. McDonald, R., Hannan, K., Neylon, T., Wells, M., & Reynar, J. (2007, June). Structured models for fine-to-coarse sentiment analysis. In *Annual Meeting-Association For Computational Linguistics* (Vol. 45, No. 1, p. 432).
26. Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007, May). Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web* (pp. 171-180). ACM.
27. Mihalcea, R. (2005, October). Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In*Proceedings of the conference on Human*

Language Technology and Empirical Methods in Natural Language Processing (pp. 411-418). Association for Computational Linguistics.

28. Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, *41*(2), 10.
29. Panagiotopoulou, V., Varlamis, I., Androutsopoulos, I., & Tsatsaronis, G. (2012). Word sense disambiguation as an integer linear programming problem. In *Artificial Intelligence: Theories and Applications* (pp. 33-40). Springer Berlin Heidelberg.
30. Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics* (p. 271). Association for Computational Linguistics.
31. 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.
32. Qiu, L., Zhang, W., Hu, C., & Zhao, K. (2009, November). Selc: a self-supervised model for sentiment classification. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 929-936). ACM.
33. Raaijmakers, S., & Kraaij, W. (2008). A Shallow Approach to Subjectivity Classification. In *ICWSM*.
34. Riloff, E., & Wiebe, J. (2003, July). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 105-112). Association for Computational Linguistics.
35. Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1631-1642).
36. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.
37. Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417-424). Association for Computational Linguistics.
38. Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011, October). Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1031-1040). ACM.
39. Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999, June). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 246-253). Association for Computational Linguistics.
40. Yessenalina, A., Yue, Y., & Cardie, C. (2010, October). Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 1046-1056). Association for Computational Linguistics.
41. Yu, H., & Hatzivassiloglou, V. (2003, July). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 confe-*

rence on Empirical methods in natural language processing (pp. 129-136). Association for Computational Linguistics.

42. Zhao, W. X., Jiang, J., Yan, H., & Li, X. (2010, October). Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 56-65). Association for Computational Linguistics.