Πτυχιακή Εργασία

ΕΞΑΓΩΓΗ ΓΝΩΜΗΣ ΑΠΟ ΚΕΙΜΕΝΑ ΜΕ ΑΞΙΟΠΟΙΗΣΗ ΤΗΣ ΣΥΝΤΑΚΤΙΚΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΟΚΤΩΒΡΙΟΣ 2015

Αθηνά Ιακωβίδη
ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ | ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Εξαγωγή Γνώμης από Κείμενα με Αξιοποίηση της Συντακτικής Πληροφορίας

Ιακωβίδη Αθηνά

Α.Μ.: 21112

Επιβλέπων:

Βαρλάμης Ηρακλής, Επίκουρος Καθηγητής

Μέλη τριμελούς επιτροπής:

Αναγνωστόπουλος Δημοσθένης, Καθηγητής

Τσερπές Κωνσταντίνος, Λέκτορας

Ευχαριστίες

Ἡ παροῦσα πτυχιακὴ ἐργασία δὲν εἶναι ἀποτέλεσμα ἑνὸς ἀνθρώπου ἀλλὰ ἑνὸς συνόλου ἀνθρώπων καὶ συγκυριῶν. Σὲ προσωπικὸ ἐπίπεδο, θὰ ἤθελα νὰ εὐχαριστήσω τοὺς γονεῖς μου γιὰ τὴν στήριξή τους σὲ ὅλη τὴν διάρκεια τῶν σπουδῶν μου. Ἐπιπλέον, εὐχαριστῶ τὴν πολυαγαπημένη μου ἀδερφὴ Εὔη ποὺ μὲ συμβούλεψε τόσο κατὰ τὴν διάρκεια τῶν σπουδῶν μου, ὅσο καὶ γιὰ τὴ διεξαγωγὴ τῆς παρούσας ἐργασίας. Ἐπιπροσθέτως, εὐχαριστῶ τὸν ἐπιβλέποντα καθηγητῆ κύριο Βαρλάμη γιὰ τὴν ἐξαιρετικὰ σημαντικὴ συμβολή του σὲ ὅλη τὴ διάρκεια διεξαγωγῆς αὐτῆς τῆς ἐργασίας.

Ὅσον ἀφορᾶ τὶς συγκυρίες, προφανῶς θὰ ἦταν ἀδύνατον νὰ πραγματοποιηθοῦν σπουδὲς ἀνώτατης ἐκπαίδευσης καὶ ἀντίστοιχες ἐργασίες ἂν δὲν ὑπῆρχε μιὰ κοινωνία ποὺ τό ἐπιτρέπει. Ἡ ἑλληνικὴ κοινωνία ὄχι μόνο τό ἐπιτρέπει, ἀλλὰ παρέχει δωρεὰν ποιοτικὴ ἐκπαίδευση, γι' αὐτὸ καὶ εἶμαι βαθύτατα εὐγνωμονοῦσα πρὸς τὴν κοινωνία. Εὐχαριστῶ ὅλους τοὺς ἀνθρώπους ποὺ συνέβαλαν στὴν ἀπόκτηση ὅλων τῶν θετικῶν χαρακτηριστικῶν αὐτῆς τῆς κοινωνίας, ἀλλὰ καὶ ὅλους αὐτοὺς ποὺ ἀσχολήθηκαν μὲ τὶς ἐπιστῆμες καὶ ἀνέδειξαν τὴν ἀξία τῆς ἐξέλιξής τους. Τέλος, εὐχαριστῶ ὅλους τοὺς συμπολῖτες μου ποὺ ἀναγνωρίζουν τὴν ἀξία καὶ ὑπερασπίζονται τὸ δικαίωμα τῆς δωρεὰν παιδείας ἀκόμη καὶ σὲ τόσο δύσκολες οἰκονομικὰ ἐποχές.

Περίληψη

Η παρούσα πτυχιακή εργασία εντάσσεται στην ευρύτερη ερευνητική περιοχή της ανάλυσης συναισθήματος και της εξαγωγής γνώμης από κείμενα. Πιο συγκεκριμένα εξετάζει το πρόβλημα της εξαγωγής θετικού ή αρνητικού συναισθήματος στα πλαίσια μιας πρότασης και στην απόδοσή του στις λέξεις της πρότασης. Ο αλγόριθμος που προτείνεται στην εργασία, επιτρέπει τον εντοπισμό των χαρακτηριστικών/όψεων του θέματος για το οποίο γράφτηκε μια πρόταση, καθώς και της συναισθηματικής φόρτισης αυτών των χαρακτηριστικών. Για να εξάγει αυτά τα αποτελέσματα, ο αλγόριθμος χωρίζει το κείμενο σε προτάσεις και πραγματοποιεί αποσαφήνιση της έννοιας των λέξεων, καθώς και συντακτική ανάλυση κάθε πρότασης. Έτσι, ανακτά τις συντακτικές εξαρτήσεις που υπάρχουν μεταξύ των λέξεων και τις αξιοποιεί για να αναπαραστήσει τη ροή της συναισθηματικής φόρτισης μέσα στην πρόταση χρησιμοποιώντας έναν γράφο. Τελικά, μεταδίδει τον σημασιολογικό προσανατολισμό κάθε λέξης και θεωρεί ότι όσα ουσιαστικά επηρεάστηκαν από τη συναισθηματική φόρτιση είναι χαρακτηριστικά του θέματος του κειμένου.

Ο αλγόριθμος που περιγράφεται στην παρούσα εργασία έχει υλοποιηθεί ως ένα γενικό εργαλείο που σκοπό έχει να αναλύει κείμενα ανεξαρτήτως του τομέα στον οποίο ανήκουν. Πρόκειται για έναν αλγόριθμο μη επιβλεπόμενης μάθησης, δηλαδή δε χρησιμοποιούνται δεδομένα για να εκπαιδεύσουν τον αλγόριθμο. Ωστόσο, γίνεται χρήση πρότερης γνώσης σχετικά με τη σημασία των συντακτικών εξαρτήσεων και πώς μπορούν να αξιοποιηθούν. Πιο συγκεκριμένα, για την υλοποίησή του, χρησιμοποιήθηκαν βάσεις γνώσης όπως ο λεξιλογικός θησαυρός Wordnet που παραθέτει τις διαφορετικές έννοιες μιας λέξης, ο θησαυρός SentiWordnet που αποδίδει συναίσθημα (θετικό, ουδέτερο ή αρνητικό) σε κάθε έννοια του Wordnet, λογισμικά για την συντακτική ανάλυση κειμένων, όπως ο Συντακτικός Αναλυτής του Πανεπιστημίου του Stanford καθώς και για την αποσαφήνιση της έννοιας των λέξεων μιας πρότασης, όπως το σύστημα Pythia που αναπτύχθηκε στο Χαροκόπειο Πανεπιστήμιο.

Λέξεις Κλειδιά: Εξαγωγή γνώμης, Επίπεδο χαρακτηριστικού, Συντακτική ανάλυση, Επεξεργασία φυσικής γλώσσας, Μηχανική μάθηση.

Abstract

The present Bachelor Thesis belongs in the broader research topic of Sentiment Analysis and Opinion Mining from Texts. More specifically, this Thesis looks into the problem of mining positive or negative sentiment, having a sentence as an input, and assigning this sentiment to specific words. The suggested algorithm manages to detect the aspects of the subject, about which the sentence was written, and the sentiment charge of them. To achieve this, the algorithm splits the text to sentences and disambiguates the sense of each and every word. Moreover, it performs syntactic analysis to each sentence, and thus retrieves the syntactic dependencies that exist between the words of the sentence. The algorithm utilizes this information by creating a graph that represents the sentiment flow of the sentence. Finally, it spreads each word's semantic orientation to the direction indicated by the graph, and considers every affected noun to be an aspect of the text's subject.

The algorithm described in the present Thesis was implemented as a general tool aiming to analyze texts regardless of the topic they belong to. It is an unsupervised learning algorithm, hence no data are used to train it. However, former knowledge is applied concerning the significance of the syntactic dependencies and how they can be utilized. More specifically, for the implementation of the algorithm knowledge databases were used, such as the Wordnet lexical thesaurus which quotes the various different senses of a word and the SentiWordnet thesaurus which assigns a sentiment (positive, objective or negative) to each sense of Wordnet. What's more, software was integrated so as to perform the syntactic analysis, namely Stanford's Natural Language Parser, and to disambiguate the words' senses, namely the system Pythia which was developed at Harokopio University.

Keywords: Opinion mining, Aspect Based, Syntactic Analysis, Natural Language Processing, Machine Learning.

Περιεχόμενα

Ι Εισαγωγή

Η εξαγωγή γνώμης από κείμενα έχει απασχολήσει έντονα την ερευνητική κοινότητα τα τελευταία χρόνια. Αυτό είναι απόρροια της αύξησης της πληροφορίας που μπορεί να βρεθεί σε ψηφιακή μορφή και της ανάγκης να αναπτυχθεί ένας αξιόπιστος τρόπος αυτόματης αξιοποίησης αυτής της πληροφορίας, καθώς ο όγκος της καθιστά ιδιαίτερα χρονοβόρα, αν όχι αδύνατη, την εξαγωγή γνώσης από έναν άνθρωπο.

Οι εφαρμογές της εξαγωγής γνώμης από κείμενα, ιδιαίτερα όταν η ανάλυση γίνεται σε επίπεδο χαρακτηριστικού, είναι πάρα πολλές. Μία από αυτές είναι η ανάλυση κριτικών προϊόντων και υπηρεσιών, όπου κάθε παράμετρος του προϊόντος αξιολογείται ξεχωριστά. Για παράδειγμα, ένα εστιατόριο μπορεί να αξιολογηθεί ως προς το φαγητό που προσφέρει, την εξυπηρέτηση πελατών, την προσβασιμότητά του, κ.ά.. Μερικοί πιθανοί πελάτες μπορεί να ενδιαφέρονται περισσότερο για το φαγητό και την εξυπηρέτηση πελατών, ενώ άλλοι για το φαγητό και την προσβασιμότητα. Φαίνεται, λοιπόν, ότι είναι ιδιαίτερα χρήσιμη η εξαγωγή γνώμης για κάθε ξεχωριστό χαρακτηριστικό του εστιατορίου, έτσι ώστε οι πιθανοί πελάτες να βρουν πιο εύκολα το εστιατόριο που ικανοποιεί καλύτερα τα κριτήριά τους. Άλλες εφαρμογές περιλαμβάνουν αλλά δε περιορίζονται σε εξαγωγή γνώμης από κείμενα άποψης ή σχόλια αναγνωστών για θέματα της επικαιρότητας, πολιτικά ζητήματα ή και διαχρονικά ζητήματα.

Η δυσκολία ανάπτυξης ενός εξαιρετικά ακριβούς τρόπου εξαγωγής γνώμης αυτοματοποιημένα απορρέει κυρίως από το γεγονός ότι αυτός ο ερευνητικός τομέας είναι κομμάτι της επεξεργασίας φυσικής γλώσσας. Η επεξεργασία φυσικής γλώσσας είναι ένα δυσεπίλυτο πρόβλημα, αφού η γλώσσα έχει προκύψει έπειτα από χιλιάδες χρόνια χρήσης της και δε χρησιμοποιεί μια αυστηρή δομή, ενώ πολλές φορές ποικίλλει και η ορθογραφία των λέξεων. Δεν έχει βρεθεί ακόμα ένας τρόπος έτσι ώστε οι ψηφιακοί υπολογιστές να αντιλαμβάνονται το νόημα ενός κειμένου.

Οπότε, ο υπολογιστής πρέπει μηχανικά, χωρίς να αντιλαμβάνεται το νόημα του κειμένου όπως ένας άνθρωπος, να εξάγει συμπεράσματα για τις λέξεις που φέρουν άποψη για τα χαρακτηριστικά του θέματος που αναλύει το πρόγραμμα. Πρέπει, δηλαδή, να καταλαβαίνει ποια χαρακτηριστικά αξιολογούνται και τι άποψη έχει ο συγγραφέας για αυτά. Αναπόφευκτα, λοιπόν, οι λύσεις που προτείνονται για το πρόβλημα αυτό υλοποιούν τεχνικές και μεθόδους μηχανικής μάθησης ώστε να μπορέσουν να βασιστούν σε δείγματα που είναι γνωστή η γνώμη που εκφέρουν και να εκπαιδεύσουν αλγορίθμους στη διαχείριση νέων άγνωστων δειγμάτων.

Στην παρούσα εργασία προτείνεται ένας αλγόριθμος μη επιβλεπόμενης μάθησης, δηλαδή δε χρησιμοποιούνται δεδομένα για να εκπαιδεύσουν τον αλγόριθμο. Ωστόσο, γίνεται χρήση

πρότερης γνώσης σχετικά με τη σημασία των συντακτικών εξαρτήσεων και πώς μπορούν να αξιοποιηθούν.

Ο αλγόριθμος χρησιμοποιεί έτοιμα εργαλεία για να πραγματοποιήσει αποσαφήνιση της έννοιας των λέξεων και να ανακτήσει τις συντακτικές εξαρτήσεις ανάμεσα στις λέξεις κάθε πρότασης. Η αποσαφήνιση της έννοιας των λέξεων γίνεται με χρήση πηγαίου κώδικα που αναπτύχθηκε στα πλαίσια της πτυχιακής εργασίας του Κατάκη (2014) για το σύστημα Pythia[1] (Katakis et al, 2014), ο οποίος χρησιμοποιεί το λεξικό SentiWordNet για να ανακτήσει τον σημασιολογικό προσανατολισμό των λέξεων και στη συνέχεια το ανάγει σε σημασιολογικό προσανατολισμό για ολόκληρη την πρόταση. Η συντακτική ανάλυση στην περίπτωσή μας γίνεται με έναν στατιστικό αναλυτή επιβλεπόμενης μάθησης που αναπτύχθηκε από το πανεπιστήμιο του Στάνφορντ.

Η συντακτική πληροφορία είναι θεμελιώδες συστατικό αυτής της εργασίας, αφού χρησιμοποιείται για να δημιουργηθεί ένας γράφος που αναπαριστά τη ροή της συναισθηματικής φόρτισης μέσα στην πρόταση. Τα ουσιαστικά που φορτίζονται θεωρούνται χαρακτηριστικά και μπορούν να κατηγοριοποιηθούν ως θετικά ή αρνητικά. Για τις ανάγκες της αξιολόγησης όσα χαρακτηριστικά έλαβαν πολύ μικρή φόρτιση θεωρήθηκαν ουδέτερα, ωστόσο ο αλγόριθμος δεν έχει δημιουργηθεί για να εντοπίζει ουδέτερα χαρακτηριστικά.

Το πρόβλημα της εύρεσης χαρακτηριστικών και το πρόβλημα της εύρεσης της συναισθηματικής φόρτισης κάθε χαρακτηριστικού αντιμετωπίζονται στις περισσότερες ερευνητικές μελέτες ως δύο διακριτά προβλήματα. Ωστόσο, στην πράξη (Pontiki et al. 2015) αποτελούν ένα ενιαίο πρόβλημα και στην παρούσα εργασία επιδιώχθηκε να λυθούν ταυτόχρονα. Τα αποτελέσματα δεν είναι ιδιαίτερα ενθαρρυντικά όσον αφορά την εύρεση χαρακτηριστικών, ωστόσο υπάρχει ικανοποιητική ακρίβεια (που ξεπερνά το 71%) ως προς την ορθότητα στην εύρεση συναισθήματος σε όσα χαρακτηριστικά εντοπίστηκαν επιτυχώς.

Στο επόμενο κεφάλαιο γίνεται μια συνοπτική επισκόπηση της βιβλιογραφίας που σχετίζεται με το θέμα της εργασίας. Στο τρίτο κεφάλαιο περιγράφονται τα έτοιμα εργαλεία που χρησιμοποιήθηκαν και η λογική του αλγορίθμου που αναπτύχθηκε. Στο τέταρτο κεφάλαιο αναφέρεται λεπτομερώς η δομή του προγράμματος και δίνονται εξηγήσεις για τις επιλογές που έγιναν σχετικά με την υλοποίησή του. Στο πέμπτο κεφάλαιο παρουσιάζονται και αναλύονται τα αποτελέσματα της αξιολόγησης που έγινε με βάση δύο δείγματα από τον διαγωνισμό του SemEval 2014 (Pontiki et al, 2014). Στο τελευταίο κεφάλαιο αναφέρονται επιγραμματικά

[1] http://omiotis.hua.gr/pythia/

κάποια συμπεράσματα για αυτή την εργασία και προτείνονται μερικές μελλοντικές επεκτάσεις που θα μπορούσε να επιδεχθεί ο παρών αλγόριθμος.

ΙΙ Επισκόπηση βιβλιογραφίας

2.1 Εύρεση σημασιολογικού προσανατολισμού

Καθημερινά χρησιμοποιούμε λέξεις που είναι φορτισμένες θετικά ή αρνητικά προκειμένου να εκφράσουμε τις απόψεις μας υπέρ ή κατά (αντιστοίχως) ενός ζητήματος. Παράλληλα, δεν επικοινωνούμε με μόνο στόχο τη μετάδοση των απόψεών μας, αλλά πολλές φορές γνωστοποιούμε στους συνομιλητές μας γενικώς αποδεκτές αλήθειες και τεκμηριωμένες πληροφορίες.

Η υποκειμενικότητα στη φυσική γλώσσα αναφέρεται στα στοιχεία της γλώσσας που χρησιμοποιούνται για να εκφράσουν γνώμες και αξιολογήσεις. Ετικετοποίηση της υποκειμενικότητας (subjectivity tagging) ονομάζεται η διάκριση προτάσεων που παρουσιάζουν γνώμες και άλλες μορφές υποκειμενικότητας (υποκειμενικές προτάσεις) από προτάσεις που παρουσιάζουν αντικειμενικά τεκμηριωμένες πληροφορίες (αντικειμενικές προτάσεις) (Wiebe, 2000).

Η συναισθηματική ανάλυση ορίζεται ως η εξαγωγή υποκειμενικότητας και πολικότητας από ένα κείμενο, ενώ ο σημασιολογικός προσανατολισμός ορίζεται ως τόσο η πολικότητα των λέξεων, φράσεων, ή κειμένων, όσο και η ένταση αυτής της πολικότητας (Taboada et al, , 2011).

Όπως αναφέρουν οι Taboada, et al(2011), υπάρχουν δύο μέθοδοι αυτόματης εξαγωγής συναισθήματος· η μία αφορά χρήση λεξικών (που περιέχουν πληροφορία για τον σημασιολογικό προσανατολισμό των λημμάτων), ενώ η άλλη αφορά χρήση στατιστικών μεθόδων ή μεθόδων μηχανικής μάθησης. Η διάκριση ίσως δεν είναι πάντοτε τόσο σαφής, αφού έχουν αναπτυχθεί διάφορα υβριδικά μοντέλα που κάνουν χρήση λεξικών για την εύρεση του σημασιολογικού προσανατολισμού των λέξεων, κι έπειτα χρησιμοποιούν επιβλεπόμενες μεθόδους για την ετικετοποίηση.

Η ετικετοποίηση μπορεί να γίνει είτε με επιβλεπόμενες, είτε με μη επιβλεπόμενες μεθόδους. Στην πρώτη περίπτωση, ένα δείγμα εκπαίδευσης πρέπει να σηματοδοτηθεί χειροκίνητα. Χρησιμοποιώντας αυτά, ο κατηγοριοποιητής ή ο συσταδοποιητής εκπαιδεύεται έτσι ώστε να ετικετοποιεί νέες, άγνωστες εισόδους. Αυτή η πρακτική εντάσσεται στο πεδίο της Μηχανικής Μάθησης, και βασίζεται σε μεγάλο βαθμό στην στατιστική επιστήμη και στο γεγονός ότι η συχνότητα εμφάνισης ενός μοτίβου στα δείγματα εκπαίδευσης θα είναι ίδια και στα πραγματικά δείγματα (δείγματα ελέγχου). Το θετικό με τις μεθόδους μηχανικής μάθησης είναι ότι τείνουν να φέρουν πολύ καλά αποτελέσματα στον τομέα για τον οποίο εκπαιδεύτηκαν (π.χ. κριτικές εστιατορίων). Ωστόσο, η επίδοσή τους πέφτει δραματικά (σχεδόν σε επίπεδο τυχαιότητας) όταν

ο κατηγοριοποιητής χρησιμοποιηθεί σε διαφορετικό τομέα (Taboada, et al, 2011). Επιπλέον, όπως σημείωσαν και οι Carenini, et al (2005), οι τεχνικές μη επιβλεπόμενης μάθησης είναι πιο οικονομικές όσον αφορά την ανάπτυξη και τη μετατροπή τους αφού δεν απαιτούν την σηματοδοσία ενός μεγάλου δείγματος εκπαίδευσης.

Όσον αφορά τα λεξικά, μπορούν να δημιουργηθούν είτε χειροκίνητα είτε αυτόματα, χρησιμοποιώντας λέξεις-σπόρους (seed words) και επεκτείνοντας τη λίστα των λέξεων. Πολλή από την έρευνα βασιζόμενη σε λεξικά επικεντρώνεται στη χρήση επιθέτων ως κατευθυντήριων του σημασιολογικού προσανατολισμού του κειμένου.

Τα αυτόματα λεξικά μπορούν να δημιουργηθούν με βάση ένα συγκεκριμένο corpus ή να είναι εξειδικευμένα σε ένα είδος κειμένου. Επιπλέον, όταν δεν υπάρχουν αρκετά στοιχεία για να υπάρξει αξιόπιστη εκτίμηση της έννοιας μιας λέξης, η πιθανότητα αυτή μπορεί να μετρηθεί ως ένα άθροισμα με βάρη των πιθανοτήτων των λέξεων που είναι όμοιες με αυτή (Lin, 1998).

Συνυφασμένοι με τη συναισθηματική ανάλυση και την εύρεση του σημασιολογικού προσανατολισμού είναι και οι αναλυτές φυσικής γλώσσας, οι οποίοι χρησιμοποιήθηκαν πολλές φορές ιστορικά ως εργαλεία για την επίτευξη των δύο πρώτων. Οι αναλυτές φυσικής γλώσσας έχουν ως στόχο τη συντακτική ανάλυση μιας πρότασης, άλλοτε χωρίζοντάς την σε υπο-φράσεις και άλλοτε βρίσκοντας ζεύγη λέξεων που συνδέονται μεταξύ τους με κάποια συγκεκριμένη σχέση συντακτικής εξάρτησης (όπως είναι, για παράδειγμα, η σχέση ενός υποκειμένου με ένα ρήμα).

Ο Lin (1994) ανέπτυξε έναν μεγάλου εύρους αναλυτή φυσικής γλώσσας που πραγματοποιούσε συντακτική ανάλυση της δομής των προτάσεων. Αργότερα, πρότεινε μια μέθοδο για αυτόματη δημιουργία λεξικού. Ο στόχος του ήταν από μεμονωμένες φράσεις οι οποίες περιείχαν μια άγνωστη λέξη, να βρει λέξεις με τις οποίες είναι όμοια αυτή. Το παράδειγμα που έδωσε περιέχει τέσσερις προτάσεις: α) ένα μπουκάλι τεσγουίνο είναι στο τραπέζι, β) το τεσγουίνο αρέσει σε όλους, γ) το τεσγουίνο σε μεθάει, δ) το τεσγουίνο φτιάχνεται από καλαμπόκι· και δήλωσε ότι στόχος της μεθόδου του ήταν να θεωρηθεί το τεσγουίνο όμοιο με τη «μπύρα», το «κρασί», τη «βότκα», κ.λπ. Η λογική της εύρεσης ομοιότητας ήταν ότι, αφού γνώριζε τον σημασιολογικό προσανατολισμό των γνωστών όμοιων λέξεων, μπορούσε να εκτιμήσει τον σημασιολογικό προσανατολισμό της άγνωστης λέξης «τεσγουίνο». Χρησιμοποιώντας τον αναλυτή που ανέπτυξε το 1994, εξήγαγε τριπλέτες συντακτικών εξαρτήσεων, όπου τα δύο μέρη ήταν δύο λέξεις και το άλλο ήταν η συντακτική εξάρτηση με την οποία ενωνόντουσαν οι δύο λέξεις. Για όλο το corpus, εξήγαγε τις τριπλέτες που περιείχαν την εκάστοτε άγνωστη λέξη και τις σύγκρινε

με τριπλέτες όπου τα υπόλοιπα δύο μέρη ήταν τα ίδια. Έτσι, μπόρεσε να εντοπίσει την ομοιότητα των άγνωστων λέξεων με τις γνωστές (Lin, 1998).

Οι Chatzivasiloglou και McKeown (1997) πρότειναν μια μέθοδο εύρεσης του σημασιολογικού προσανατολισμού των επιθέτων βασιζόμενοι στις συνδετικές λέξεις που τα ενώνουν. Έδειξαν ότι συνώνυμες λέξεις μπορεί να εκφράζουν αντίθετη συναισθηματική φόρτιση και ότι συνήθως οι επιλογές επιθέτων και συνδετικών λέξεων είναι αμοιβαίως εξαρτώμενες, φέρνοντας το εξής χαρακτηριστικό παράδειγμα:

> *Η πρόταση φόρου ήταν [απλή και αποδεκτή / απλοϊκή αλλά αποδεκτή / * απλοϊκή και αποδεκτή] από το λαό.*

Η τελευταία έκφραση δεν ηχεί οικεία και αυτό αποδεικνύει ότι οι συνδετικοί σύνδεσμοι συνδέουν συνήθως επίθετα με ίδιο σημασιολογικό προσανατολισμό, ενώ οι αντιθετικοί σύνδεσμοι συνδέουν συνήθως επίθετα με διαφορετικό σημασιολογικό προσανατολισμό. Παράλληλα, βασίστηκαν στο γεγονός ότι σχεδόν όλα τα αντώνυμα έχουν διαφορετικούς σημασιολογικούς προσανατολισμούς. Ενώ ήταν εύκολο να βρεθούν συνώνυμα και αντώνυμα, η απουσία λεξικών με συναισθηματικές πολικότητες στην εποχή τους καθιστούσε δύσκολη την κατηγοριοποίηση. Για να λύσουν αυτό το πρόβλημα, ανέπτυξαν ένα σύστημα που αρχικά εξήγαγε όλες τις συνδέσεις μεταξύ επιθέτων, δημιουργούσε ένα γράφο στον οποίο έβαζε τα επίθετα με όλες τις εκτιμηθείσες σχέσεις (συνδέσεις και αντιθέσεις), τοποθετούσε τα επίθετα σε δύο υποσύνολα με διαφορετικό προσανατολισμό και, τέλος, χαρακτήριζε το υποσύνολο με τις πιο συχνά εμφανιζόμενες λέξεις ως θετικού προσανατολισμού (επειδή είναι γνωστό ότι τα θετικά επίθετα χρησιμοποιούνται πιο συχνά από τα αρνητικά). Αυτή η προσέγγιση παρουσίαζε περισσότερη από 90% επιτυχία στην κατηγοριοποίηση της πολικότητας των επιθέτων.

Σύμφωνα, όμως, με τους Turney και Littman (2003), ο σημασιολογικός προσανατολισμός δεν έχει μόνο κατεύθυνση - ή αλλιώς πολικότητα - (θετικός, αρνητικός), αλλά και ένταση (έντονος, ήπιος). Ένα παράδειγμα που αναφέρουν και αντικατοπτρίζει αυτή την αντίθεση αφορά τα ζεύγη λέξεων καλός/καταπληκτικός (ήπιος/έντονος θετικός), καθώς και ενοχλητικός/φριχτός (ήπιος/έντονος αρνητικός).

Ο Turney (2002) ανέπτυξε μια μέθοδο κατηγοριοποίησης κριτικών ως υπέρ ή κατά του θέματος το οποίο εξετάζουν, βασιζόμενος στη μέση συναισθηματική πολικότητα των φράσεων που περιέχουν επίθετα ή επιρρήματα. Πιο αναλυτικά, εντόπιζε τις φράσεις που περιείχαν επίθετα ή επιρρήματα, έκανε μια εκτίμηση του σημασιολογικού τους προσανατολισμού, κι έπειτα υπολόγιζε το μέσο όρο όλων αυτών των φράσεων. Αφού δεν υπήρχαν λεξικά, ανέπτυξε μια

μέθοδο για να εξάγει το σημασιολογικό προσανατολισμό (πολικότητα και ένταση) των επιθέτων και επιρρημάτων. Όριζε το σημασιολογικό προσανατολισμό ως τη σημασιολογική ομοιότητα της φράσης με μια θετική λέξη-αναφορά, όπως η λέξη «άριστος» («excellent») μείον τη σημασιολογική ομοιότητά της με μια αρνητική λέξη-αναφορά, όπως η λέξη «κάκιστος» («poor»). Αφού είχε υπολογίσει τους επιμέρους σημασιολογικούς προσανατολισμούς των καίριων λέξεων υπολόγιζε τον μέσο όρο όλων όσων υπήρχαν σε κάθε κριτική. Μια κριτική με θετικό μέσο όρο χαρακτηριζόταν ως συστατική του αντικειμένου που εξέταζε, ενώ μια κριτική με αρνητικό μέσο όρο χαρακτηριζόταν ως μη συστατική. Η μέθοδος αυτή είχε μέση ακρίβεια 74% όταν εξετάστηκε σε τέσσερις διαφορετικούς τομείς (κριτικές για αυτοκίνητα, τράπεζες, ταινίες, και ταξιδιωτικούς προορισμούς).

Αργότερα, οι Turney και Littman (2003) εξέλιξαν αυτή τη μέθοδο κάνοντας διάφορες μετατροπές, μεταξύ των οποίων και η σύγκριση της προς κατηγοριοποίηση λέξης με ένα σύνολο από λέξεις αναφοράς αντί για ένα ζεύγος λέξεων (θετική και αρνητική). Έτσι, πέτυχαν 82,8% ακρίβεια, η οποία ξεπερνούσε και το 95% όταν επιτρεπόταν στον αλγόριθμο να μη κατηγοριοποιεί λέξεις με ήπιο σημασιολογικό προσανατολισμό.

Η πρακτική εξαγωγής του σημασιολογικού προσανατολισμού μιας λέξης συγκρίνοντάς την με μία ή περισσότερες λέξεις-υποδείγματα, η οποία βασίζεται στην υπόθεση ότι ο σημασιολογικός προσανατολισμός μιας λέξης τείνει να είναι ανάλογος του σημασιολογικού προσανατολισμού των γειτόνων της, ονομάζεται Σημασιολογικός Προσανατολισμός από Συσχέτιση ή SO-A (Semantic Orientation from Association) (Turney και Littman, 2003).

Η γνώση τις πολικότητας των λέξεων που χρησιμοποιούνται σε ένα κείμενο είναι χρήσιμη, όμως δεν είναι αρκετή για να εξάγουμε με ακρίβεια την πολικότητα μιας υποκειμενικής φράσης. Όπως αναφέρουν και οι Taboada, et al (2011), κατά την ανάγνωση οποιουδήποτε κειμένου, γίνεται φανερό ότι στοιχεία των συμφραζομένων πρέπει να ληφθούν υπ’ όψιν κατά την εκτίμηση του σημασιολογικού προσανατολισμού, όπως η άρνηση (π.χ. μη ευχάριστος) και η ενίσχυση (π.χ. πολύ καλός). Οι λέξεις αυτές, που μπορούν να επηρεάσουν την κατεύθυνση ή την ένταση της πολικότητας μιας φράσης, ονομάζονται μετατροπείς πολικότητας (valence shifters).

Οι Kennedy και Inkpen (2006) έλαβαν υπ’ όψιν τους τους διάφορους μετατροπείς πολικότητας κι έτσι κατάφεραν να επιφέρουν στατιστικά σημαντική αύξηση της ακρίβειας πρόβλεψης. Συγκεκριμένα, εστιάστηκαν σε τρεις κατηγορίες: τις αρνήσεις, τις ενισχύσεις και τις εξασθενήσεις. Όρισαν τους μετατροπείς πολικότητας ως όρους που αλλάζουν τον

σημασιολογικό προσανατολισμό ενός όρου· η άρνηση αντιστρέφοντάς τον, ενώ η ενίσχυση και η εξασθένηση αυξάνοντας ή μειώνοντας την έντασή του αντίστοιχα.

Οι Choi και Cardie (2008) βασίστηκαν στο ότι οι λέξεις μιας έκφρασης αλληλεπιδρούν προκειμένου να καθορίσουν την συνολική πολικότητά της. Διαμόρφωσαν τη μέθοδό τους αναγνωρίζοντας ότι ένας μετατροπέας πολικότητας θα μπορούσε να βρίσκεται σε μεγάλη απόσταση από τη λέξη που επηρεάζει. Κατάφεραν έτσι να αυξήσουν την ακρίβεια πρόβλεψης. Επιπλέον, παρατήρησαν, το εκ πρώτης όψεως παράξενο γεγονός, ότι ο συνυπολογισμός στη διαδικασία κατηγοριοποίησης συμφραζομένων των εκφράσεων που επρόκειτο να κατηγοριοποιηθούν έτεινε να μειώνει την ακρίβεια πρόβλεψης.

Οι Taboada, et al (2011) ανέπτυξαν επίσης ένα σύστημα που κατηγοριοποιούσε κείμενα ανάλογα με την άποψη του συγγραφέα για το κύριο θέμα (θετική ή αρνητική). Λάμβαναν υπ' όψιν τους μετατροπείς πολικότητας και έκαναν χρήση λεξικών για να εξάγουν το σημασιολογικό προσανατολισμό των λέξεων. Κατάφεραν έτσι να δημιουργήσουν ένα σύστημα με συνεπή επίδοση σε διάφορους τομείς με εντελώς άγνωστα δεδομένα.

Παρ' όλο που η ετικετοποίηση του όλου κείμενου ως εκφράζον θετική ή αρνητική άποψη για το κύριο θέμα είναι πολύ σημαντική, υπάρχουν περιπτώσεις όπου είναι χρήσιμο να εντοπιστούν οι απόψεις σχετικά με τα επιμέρους συστατικά του θέματος για το οποίο γράφτηκε το εξεταζόμενο κείμενο. Παραδείγματος χάριν, μια κριτική ενός εστιατορίου μπορεί να αξιολογεί διαφορετικά χαρακτηριστικά του, όπως είναι η εξυπηρέτηση πελατών, το κυρίως γεύμα, το εύρος επιλογών, κ.λπ., και να εκφράζει απόψεις διαφορετικού σημασιολογικού προσανατολισμού για κάθε ένα από αυτά τα χαρακτηριστικά. Πολλοί ερευνητές, που αναγνώρισαν την αξία αυτής της πληροφορίας, ασχολήθηκαν με τρόπους εξαγωγής των χαρακτηριστικών του κύριου θέματος από το κείμενο, καθώς και του συναισθηματικού προσανατολισμού καθ' ενός από αυτά.

Οι Hu και Liu (2004) πρότειναν μια μέθοδο εξαγωγής χαρακτηριστικών ενός προϊόντος, καθώς και της άποψης των πελατών για κάθε ένα από αυτά, αντλώντας πληροφορία από ένα μεγάλο σύνολο κριτικών για το εκάστοτε προϊόν. Το πρώτο βήμα της μεθόδου τους αφορούσε ετικετοποίηση του μέρους του λόγου κάθε λέξεως της κριτικής. Κάνοντας την παραδοχή ότι κατά κύριο λόγο τα χαρακτηριστικά πρόκεινται για ουσιαστικές λέξεις ή ονοματικές φράσεις, έφτιαχναν ένα αρχείο το οποίο περιείχε ανά γραμμή όλα τα ουσιαστικά και τις ονοματικές φράσεις της κάθε πρότασης. Από το αρχείο μπορούσαν να εξακριβώσουν τα συχνά αναφερόμενα χαρακτηριστικά του προϊόντος. Αφού έκαναν ένα ξεκαθάρισμα στα χαρακτηριστικά (για να αφαιρέσουν όσα δεν ήταν χρήσιμα ή πραγματικά), εξήγαγαν τις λέξεις

που τα χαρακτηρίζουν. Συγκεκριμένα, για κάθε πρόταση που περιείχε ένα συχνό χαρακτηριστικό, εξήγαγαν το πιο κοντινό επίθετο (το οποίο αν όντως υπήρχε θεωρούταν λέξη άποψης). Έτσι, έβρισκαν την άποψη για το κάθε χαρακτηριστικό και ταυτόχρονα συνέλεγαν τις συχνές λέξεις άποψης. Χρησιμοποιώντας τις τελευταίες, έψαχναν για προτάσεις που, ενώ δεν περιείχαν κανένα γνωστό συχνό χαρακτηριστικό, περιείχαν λέξεις άποψης. Σε αυτές τις προτάσεις εντόπιζαν ουσιαστικά και ονοματικές φράσεις, και αυτά τα θεωρούσαν μη συχνά εμφανιζόμενα χαρακτηριστικά. Με αυτή τη μέθοδο πετύχαν μέση ακρίβεια 72%.

Οι Carenini, et al (2005) παρατήρησαν ότι η μέθοδος που πρότειναν οι Hu και Liu (2004) παρήγαγε ως έξοδο πολλά χαρακτηριστικά για κάθε προϊόν, λόγω του ότι υπήρχαν περιττά χαρακτηριστικά και δεν λάμβανε υπ' όψιν τις ιεραρχικές σχέσεις των χαρακτηριστικών, οπότε κατέληγε να παρουσιάζει χαρακτηριστικά που δεν είχαν πάντα νόημα στον χρήστη. Για να λύσουν αυτό το πρόβλημα, πρότειναν μετά την εξαγωγή χαρακτηριστικών και σημασιολογικού προσανατολισμού, με κάποιον τρόπο μη επιβλεπόμενης μάθησης, να γίνεται αντιστοίχηση των χαρακτηριστικών με ταξινομήσεις του WordNet, οι οποίες παρέχονται από τους χρήστες του.

Οι ταξινομήσεις του WordNet πρόκεινται για λίστες με τα θεμελιώδη χαρακτηριστικά για κάθε τομέα. Το πρόβλημα είναι ότι κάποιος μπορεί να αξιολογήσει ένα θεμελιώδες χαρακτηριστικό σε μια κριτική εστιατορίου, όπως είναι το φαγητό, χωρίς να το αναφέρει κατ' όνομα. Παραδείγματος χάριν, μπορεί να αξιολογήσει τα επιμέρους γεύματα ενός εστιατορίου λέγοντας

«Το κοτόπουλο ήταν καλό και η σαλάτα ήταν εξαίσια.».

Όταν αναφερόμαστε στις ιεραρχικές σχέσεις των χαρακτηριστικών, για ένα εστιατόριο το φαγητό είναι ένα από τα χαρακτηριστικά που βρίσκονται στην κορυφή και εμπεριέχει άλλα χαρακτηριστικά όπως κατηγορίες φαγητών και συγκεκριμένα γεύματα.

Έρευνα έχει γίνει και για εύρεση σημασιολογικού προσανατολισμού σε επίπεδο χαρακτηριστικού κάνοντας χρήση επιβλεπόμενων μεθόδων. Οι Gamon, Aue, Corston-Oliver και Ringger (2005) παρουσίασαν ένα σύστημα για εξαγωγή γνώμης σε επίπεδο χαρακτηριστικού από κριτικές για αυτοκίνητα. Το σύστημά τους χρησιμοποιούσε συσταδοποίηση (clustering) και παρουσίαζε τα αποτελέσματα μέσω μιας εύχρηστης διεπαφής χρήστη.

Οι Zhuang, Jing και Zhu (2006) βασίστηκαν στην πρόταση των Hu και Liu (2004) και δημιούργησαν ένα σύστημα για εξαγωγή γνώμης σε επίπεδο χαρακτηριστικού από κριτικές για ταινίες στο IMDb (Internet Movie Database). Έκαναν διάφορες βελτιώσεις και προσαρμογές για να επιτύχουν αύξηση ακρίβειας κατά 8,4%. Μεταξύ αυτών, χρησιμοποίησαν ένα συντακτικό

αναλυτή για να εξακριβώσουν τις σχέσεις εξάρτησης ανάμεσα στις λέξεις κάθε πρότασης, δημιούργησαν μια λίστα κύριων και συχνών χαρακτηριστικών ταινιών και δε συμπεριλάμβαναν περιττά ή σπάνια χαρακτηριστικά στην ταξινόμηση, και εξόρυξαν πληροφορίες σχετικά με τους συντελεστές της εξεταζόμενης ταινίας, έτσι ώστε να αναγνωρίζουν τα ονόματά τους όταν αυτά αναφέρονται στις κριτικές και να διαπιστώνουν ποιο χαρακτηριστικό της ταινίας τίθεται υπό αξιολόγηση σε αυτές τις περιπτώσεις.

Οι Blair-Goldensohn, Hannan, McDonald, Neylon, Reis και Reynar (2008), στην προσπάθειά τους να αναπτύξουν μια μέθοδο για εύρεση σημασιολογικού προσανατολισμού για προϊόντα και υπηρεσίες σε επίπεδο χαρακτηριστικού, παρατήρησαν ότι σχεδόν όλες οι υπηρεσίες μοιράζονται θεμελιώδη χαρακτηριστικά και ότι ένα μεγάλο μέρος των αναζητήσεων αφορούν διαδικτυακές κριτικές ενός μικρού συνόλου ειδών υπηρεσιών. Με βάση αυτή την παρατήρηση, δημιούργησαν λίστες με τα θεμελιώδη χαρακτηριστικά κάθε τομέα υπηρεσιών και προϊόντων. Η μέθοδος που ανέπτυξαν έπαιρνε ως είσοδο ένα σύνολο κριτικών σχετικά με μια υπηρεσία. Έπειτα, πραγματοποιούσαν διαχωρισμό κάθε κριτικής σε προτάσεις ή φράσεις και υπολόγιζαν το σημασιολογικό προσανατολισμό κάθε κομματιού. Εν συνεχεία, με έναν δυναμικό εξαγωγέα έβρισκαν τα χαρακτηριστικά των κριτικών (π.χ. η σαλάτα) και με μηχανική μάθηση αντιστοιχούσαν αυτά τα χαρακτηριστικά στα θεμελιώδη χαρακτηριστικά αυτού του τύπου υπηρεσίας (π.χ. το φαγητό).

2.2 Αποσαφήνιση της έννοιας των λέξεων

Παρ’ όλο που η συναισθηματική ανάλυση ενός κειμένου ή ενός χαρακτηριστικού εξακολουθεί να παρουσιάζει ερευνητικό ενδιαφέρον, υπάρχουν πλέον πολλά λεξικά που περιέχουν τους σημασιολογικούς προσανατολισμούς κάθε λέξης. Ο πληθυντικός στην προηγούμενη πρόταση οφείλεται στο ότι μια λέξη μπορεί να έχει πολλές έννοιες και επομένως έχει διαφορετικό σημασιολογικό προσανατολισμό για κάθε έννοιά της. Για παράδειγμα, η λέξη «φυσικός» μπορεί να σημαίνει «μη τεχνητός» και επομένως να έχει θετική πολικότητα, ενώ παράλληλα μπορεί να σημαίνει «αυτός που σχετίζεται με τη φυσική» και να έχει ουδέτερη πολικότητα. Από αυτή τη πολυσημία των λέξεων έγκειται η δυσκολία εύρεσης του σωστού σημασιολογικού προσανατολισμού των μεμονωμένων λέξεων της πρότασης. Είναι φανερό ότι ένας άνθρωπος θα μπορούσε ενστικτωδώς να καταλάβει με ποια έννοια χρησιμοποιείται η κάθε λέξη διαβάζοντας μια πρόταση, ικανότητα η οποία πηγάζει από την κατανόηση των συμφραζομένων. Οι υπολογιστές δεν καταλαβαίνουν τη φυσική γλώσσα, επομένως ένας πολύ μεγάλος τομέας της επεξεργασίας φυσικής γλώσσας ασχολείται με την αποσαφήνιση της έννοιας των λέξεων. Η

χρήση των ευρημάτων του είναι απαραίτητη στη συναισθηματική ανάλυση για την ανάκτηση του σωστού σημασιολογικού προσανατολισμού των μεμονωμένων λέξεων.

Ο Lesk (1986) παρουσίασε μία μέθοδο αποσαφήνισης της έννοιας των λέξεων χρησιμοποιώντας ψηφιακά λεξικά. Τα λεξικά περιείχαν όλες τις πιθανές έννοιες κάθε λέξης και έναν ορισμό για κάθε έννοια. Ανακτούσε, λοιπόν, τους ορισμούς των γειτονικών, με την ως προς αποσαφήνιση λέξη, λέξεων της πρότασης και έβρισκε λέξεις κοινές σε αυτούς τους ορισμούς με τον ορισμό της εξεταζόμενης λέξης. Η έννοια της λέξης που περιείχε αυτές τις κοινές λέξεις επιλεγόταν ως η έννοιά της. Για παράδειγμα, η λέξη «καλλιέργεια» χρησιμοποιείται με διαφορετική έννοια στην φράση «Οι κάτοικοι ασχολήθηκαν με την καλλιέργεια της γης για να επιβιώσουν.» απ' ό,τι στην φράση «Χρειάστηκε να κάνουν καλλιέργεια του μύκητα για να βρουν τρόπους αντιμετώπισης». Αν δούμε τους ορισμούς των:

> **καλλιέργεια:** 1. το σύνολο των γεωργικών εργασιών που εκτελούνται στην *επιφάνεια* του *εδάφους*, με σκοπό την παραγωγή φυτικών προϊόντων. 2. (*βιολ.*) ανάπτυξη και πολλαπλασιασμός *μικροοργανισμών* με τεχνητά μέσα για επιστημονικούς σκοπούς. 3. ενασχόληση με κάτι, αφιέρωση του χρόνου και του ενδιαφέροντος ενός ατόμου στη μελέτη και στην ανάπτυξη μιας επιστήμης, μιας τέχνης κ.λπ..
>
> **γη**: 1. ο πλανήτης του ηλιακού μας συστήματος, τρίτος κατά σειρά σε απόσταση από τον ήλιο, που κατοικείται από τον άνθρωπο. 2. η γη ως χώρος κατοικίας και δραστηριότητας του ανθρώπου. 3α. η *επιφάνεια* πάνω στην οποία ζουν και κινούνται οι άνθρωποι και τα ζώα· *έδαφος*.
>
> **μύκητας**: (και *βιολ.*) ονομασία φυτικών οργανισμών, συνήθως *μικροοργανισμών*, που δεν έχουν χλωροφύλλη και γι' αυτό ζουν παρασιτικά, και αποτελούν μία από τις πέντε κατηγορίες στις οποίες διακρίνει η νεότερη βιολογία τα έμβια όντα.

βλέπουμε ότι η πρώτη έννοια της καλλιέργειας έχει δύο κοινές λέξεις με τη γειτονική λέξη «γη», ενώ η δεύτερη έννοια της καλλιέργειας έχει επίσης δύο κοινές λέξεις με τη γειτονική λέξη «μύκητας». Σε κάθε περίπτωση, αυτός είναι ο μέγιστος αριθμός κοινών λέξεων που βρίσκουμε μεταξύ των ορισμών και αποδεικνύει ότι έννοιες που συναντάμε μαζί συχνά μοιράζονται μερικές λέξεις στους ορισμούς τους.

Οι Banerjee και Pedersen (2002) βασίστηκαν πάνω στη μέθοδο του Lesk για να δημιουργήσουν έναν διαφορετικό τρόπο αποσαφήνισης της έννοιας των λέξεων. Αντί για το τυπικό λεξικό που

χρησιμοποίησε ο Lesk (1986) έκαναν χρήση του WordNet. Εστιάστηκαν στα synset τα οποία είναι ομάδες συνώνυμων λέξεων. Κάθε synset έχει μια περιγραφή (gloss), η οποία πρόκειται για έναν ορισμό της κοινής έννοιας που μοιράζονται αυτές οι λέξεις. Αν η ίδια λέξη ανήκει σε περισσότερα από ένα synset συνεπάγεται ότι αυτή η λέξη έχει περισσότερες από μία έννοιες. Με αυτόν τον τρόπο, οι Banerjee και Pedersen κατάφεραν να αυξήσουν την ακρίβεια αποσαφήνισης στο 32%, δηλαδή αύξηση της τάξης 9% έως 16% σε σύγκριση με την μέθοδο του Lesk.

Οι Galley και McKeown (2003) πρότειναν μια άλλη μέθοδο αποσαφήνισης της έννοιας των λέξεων (πιο συγκεκριμένα, των ουσιαστικών), η οποία βασιζόταν σε λεξικές αλυσίδες. Οι λεξικές αλυσίδες αφορούν σημασιολογικά συγγενείς λέξεις, όπως ομώνυμα, υπερώνυμα και υπώνυμα (υπερώνυμα είναι λέξεις που περιλαμβάνουν την εξεταζόμενη λέξη - για παράδειγμα, η λέξη αυτοκίνητο είναι υπερώνυμο της λέξης ταξί - ενώ υπώνυμα είναι λέξεις που ανήκουν στο υπερσύνολο της εξεταζόμενης). Οι Galley και McKeown δημιουργούσαν έναν γράφο αποσαφήνισης, ο οποίος περιείχε όλες τις συγγενικές σχέσεις όλων των εννοιών των λέξεων της πρότασης μεταξύ τους. Έπειτα πρόσθεταν τα βάρη όλων των ακμών του γράφου για κάθε πιθανή έννοια της λέξης, και η έννοια με τη μεγαλύτερη συνολική βαθμολογία επιλεγόταν. Κατάφεραν έτσι να επιτύχουν ικανοποιητική ακρίβεια, η οποία όμως μειωνόταν όσο αυξανόταν το πλήθος των πιθανών εννοιών κάθε λέξης (όμοια με τις προηγούμενες έρευνες).

2.3 Εξάπλωση της ενεργοποίησης

Κάνοντας σωστή αποσαφήνιση της έννοιας των λέξεων μιας πρότασης μπορούμε να ανακτήσουμε με μεγάλη ακρίβεια τον σημασιολογικό προσανατολισμό τους. Ωστόσο, όπως είδαμε και στο πρώτο μέρος αυτής της ενότητας, δεν αρκεί αυτό για να κάνουμε συναισθηματική ανάλυση, αλλά απαιτείται και μια μεθοδολογία έτσι ώστε να χρησιμοποιήσουμε αυτά τα δεδομένα για να εξάγουμε συμπεράσματα σε επίπεδο κειμένου ή χαρακτηριστικού. Στην παρούσα εργασία, γίνεται χρήση γράφων για αναπαράσταση της συντακτικής δομής κάθε πρότασης και η ανάλυση πραγματοποιείται μέσω της εξάπλωσης του συναισθήματος από συναισθηματικά φορτισμένες λέξεις σε όσες επηρεάζονται από τις πρώτες. Για αυτόν τον σκοπό αναπτύχθηκε μια τεχνική διάδοσης συναισθήματος εμπνευσμένη από το μοντέλο «εξάπλωση της ενεργοποίησης».

Η εξάπλωση της ενεργοποίησης (spreading activation) προτάθηκε από τον Quillian (1962) ως μια μέθοδος βελτίωσης της μηχανικής μετάφρασης σε διάφορους τομείς, συμπεριλαμβανομένου αυτού της αποσαφήνισης της έννοιας των λέξεων. Σκοπός του ήταν να αναπαραστήσει μέσω υπολογιστή τον τρόπο με τον οποίο δημιουργείται μια σημασιολογική δομή και πραγματοποιείται η ανάλυση των δεδομένων στον ανθρώπινο εγκέφαλο. Θεώρησε ότι κάθε λέξη

σχηματίζει στο νου μας μια παράσταση (concept) κι ότι κάθε παράσταση συνδέεται με άλλες παραστάσεις με μεγάλη ή μικρή ισχύ (criteriality). Επιπλέον, η ισχύς μπορεί να διαφέρει ανάλογα με την κατεύθυνση· για παράδειγμα, η λέξη «ταξί» συνδέεται με μεγάλη ισχύ με τη λέξη «αυτοκίνητο» αφού συνήθως η σκέψη της πρώτης πυροδοτεί τη σκέψη της δεύτερης, ενώ η λέξη «αυτοκίνητο» συνδέεται με τη λέξη «ταξί» με μικρότερη ισχύ. Ο Quillian (1962) πρότεινε μετά τη δημιουργία ενός γράφου, εκφράζων αυτές τις παραστάσεις και τις αλληλοσυνδέσεις τους, να επιλέγονται δύο (ή περισσότερες) λέξεις της πρότασης και να εξετάζεται αν αυτές οι λέξεις συνδέονται σημασιολογικά. Ο τρόπος που γινόταν αυτή η ανάλυση ήταν η κατανομή ενέργειας (ενεργοποίησης) στις λέξεις-κλειδιά και η εξάπλωσή της στους γειτονικούς κόμβους μέχρι να συναντηθούν οι 2 (ή περισσότερες) διαφορετικές ενέργειες, οπότε θα μπορούσε να εξαχθεί η διαδρομή και να μελετηθεί η ισχύς της.

Οι Collins και Loftus (1975) βασίστηκαν πάνω σε αυτή τη θεωρία και την εμπλούτισαν. Έκαναν διάφορες παραδοχές σχετικά με την τοπική επεξεργαστική ισχύ και, για να εξοικονομήσουν ενέργεια, προσέθεσαν τον παράγοντα φθοράς (decay). Δηλαδή, κάθε φορά που μεταφερόταν η ενέργεια, δε μεταφερόταν ολόκληρη αλλά σε ένα μεγάλο μέρος της. Όταν η ενέργεια έπεφτε κάτω από ένα κατώτατο όριο θεωρούταν ότι η απόσταση ήταν πολύ μεγάλη και δεν είχε νόημα να αναζητούνται πιο μακρινοί σημασιολογικοί συσχετισμοί.

ΙΙΙ Προτεινόμενη προσέγγιση

Στην παρούσα εργασία αναπτύχθηκε ένας αλγόριθμος για να πραγματοποιηθεί εξαγωγή γνώμης από κείμενα σε επίπεδο χαρακτηριστικού. Ο αλγόριθμος αυτός κάνει χρήση προϋπάρχοντων εργαλείων για την πραγματοποίηση της συντακτικής ανάλυσης και της αποσαφήνισης της έννοιας των λέξεων, ενώ ακολουθεί μια πρωτότυπη μέθοδο για τη μεταφορά της θετικής ή αρνητικής άποψης από μια φορτισμένη λέξη σε αυτές που χαρακτηρίζει.

3.1 Εργαλεία που χρησιμοποιήθηκαν

Ακολουθεί μια σύντομη περιγραφή των εργαλείων που προϋπήρχαν της εργασίας και που χρησιμοποιήθηκαν σε αυτή.

3.1.1 Ο Συντακτικός αναλυτής φυσικής γλώσσας του πανεπιστημίου Στάνφορντ

Ο συντακτικός αναλυτής του Στάνφορντ (Stanford NLP parser) πρόκειται για έναν στατιστικό αναλυτή που βρίσκει τη γραμματική δομή μιας πρότασης. Είναι γραμμένος σε java και παρέχεται για αγγλικά, αν και μπορεί να προσαρμοστεί και για άλλες γλώσσες (επίσημα παρέχονται ήδη αναλυτές για κινέζικα, γερμανικά και αραβικά). Λαμβάνει ως είσοδο μια πρόταση και την αναλύει (στην πραγματικότητα παρουσιάζει την πιο πιθανή ανάλυση). Κατηγοριοποιεί την κάθε λέξη της πρότασης ως ουσιαστικό, ρήμα, επίθετο ή επίρρημα, ενώ χρησιμοποιεί δύο αναλυτές για να παράξει δύο διαφορετικές εξόδους.

Η μία έξοδος του αναλυτή (Shift-reduce constituency parser) είναι το δέντρο δομής φράσεων, το οποίο πρόκειται για μια αναπαράσταση των επιμέρους φράσεων μιας πρότασης. Πιο συγκεκριμένα, διαδοχικές λέξεις που σχηματίζουν μια φράση (ονοματική, ρηματική, κ.λπ.) παρουσιάζονται ομαδοποιημένες μέσα σε ένα ζεύγος παρενθέσεων. Τελικά, στην έξοδο συμπεριλαμβάνονται όλες οι λέξεις της πρότασης με τη σειρά που εισήχθησαν, χωρισμένες σε φράσεις και υπο-φράσεις μεγαλύτερων φράσεων ή προτάσεων. Δεξιά φαίνεται ότι το σύνολο «The fried rice» είναι ονοματική φράση ενώ το «is amazing here» είναι ρηματική φράση. Επιμέρους, το the είναι άρθρο (determiner), το fried επίθετο, το rice ουσιαστικό, ενώ τα amazing και here αποτελούν υπο-φράσεις της ρηματικής φράσης στην οποία ανήκουν.

```
(ROOT
  (S
    (NP (DT The) (JJ fried)
(NN rice))
    (VP (VBZ is)
      (ADJP (JJ amazing))
      (ADVP (RB here)))
    (. .)))
```

Δέντρο δομής φράσεων για την πρόταση «The fried rice is amazing here.».

Η άλλη έξοδος του αναλυτή (Neural Network Dependency Parser) είναι μια λίστα εξαρτήσεων. Μία εξάρτηση (dependency) πρόκειται για μια γραμματική σχέση ανάμεσα σε δύο λέξεις. Για παράδειγμα, η εξάρτηση nsubj(like, I) δηλώνει ότι η λέξη «I» είναι το υποκείμενο (noun subject)

της λέξης «like». Αξίζει να σημειωθεί ότι δε συμπεριλαμβάνονται όλες οι λέξεις της πρότασης στη λίστα εξαρτήσεων, αλλά μόνο όσες αλληλεπιδρούν με μία ή περισσότερες λέξεις. Η έξοδος για το προηγούμενο παράδειγμα παρουσιάζεται δεξιά. Φαίνεται ότι το the είναι άρθρο στο rice, το fried επιθετικός προσδιορισμός στο rice, το rice υποκείμενο στο amazing κ.ο.κ.. Μια πλήρης λίστα όλων των εξαρτήσεων μπορεί να βρεθεί στο Παράρτημα.

```
det(rice-3, The-1)
amod(rice-3, fried-2)
nsubj(amazing-5, rice-3)
cop(amazing-5, is-4)
root(ROOT-0, amazing-5)
advmod(amazing-5, here-6)
```

Λίστα εξαρτήσεων για την πρόταση «The fried rice is amazing here.».

3.1.2 Pythia

Η Pythia (Katakis et al, 2014) είναι ένα διαδικτυακό εργαλείο για αποσαφήνιση της έννοιας των λέξεων και συναισθηματική ανάλυση, το οποίο αναπτύχθηκε στο πλαίσιο της πτυχιακής εργασίας του Κατάκη (2014). Δέχεται ως είσοδο μια πρόταση, επιτρέπει στον χρήστη να επιλέξει τρόπο αποσαφήνισης της έννοιας των λέξεων και μοντέλο κατηγοριοποίησης, και εμφανίζει ως έξοδο το συναίσθημα που επικρατεί σε κάθε επιμέρους πρόταση.

3.1.2.1 Αποσαφήνιση της έννοιας των λέξεων

Στο πλαίσιο της εργασίας του Κατάκη (2014) χρησιμοποιήθηκαν οι τρόποι αποσαφήνισης της έννοιας των λέξεων που παρουσιάζονται στη συνέχεια. Ο κώδικας αυτός χρησιμοποιήθηκε και για να πραγματοποιηθεί αποσαφήνιση της έννοιας των λέξεων και στην παρούσα εργασία.

3.1.2.1α Μέθοδος First Sense (FS)

Πρόκειται ουσιαστικά για την απουσία αποσαφήνισης, αφού επιλέγεται η έννοια με την οποία χρησιμοποιείται πιο συχνά η λέξη. Η μέθοδος αυτή είναι πολύ γρήγορη καθώς δεν επιλύει κάποιο πρόβλημα αποσαφήνισης με αλγοριθμικό τρόπο αλλά έχει μια μέγιστη ακρίβεια γύρω στο 60%.

3.1.2.1β Μέθοδος Weighted Degree (WDEG)

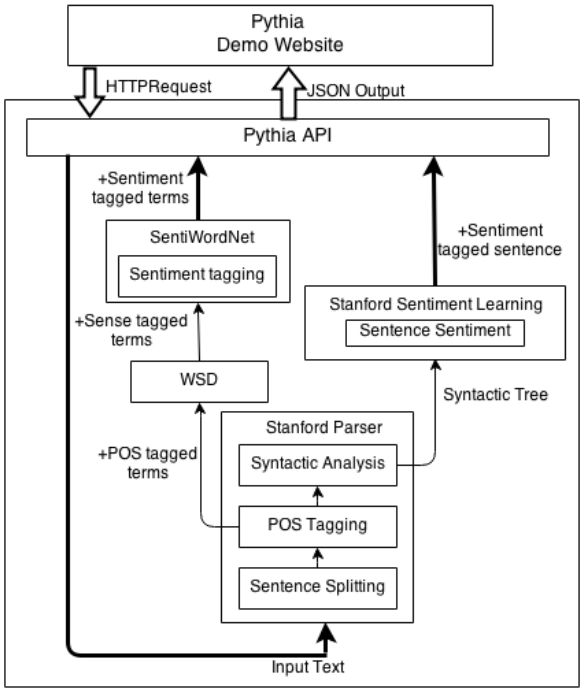
Ο αλγόριθμος WDEG δημιουργεί ένα γράφο με κόμβους τις λέξεις της πρότασης και βάρη στις ακμές (που δηλώνουν το βαθμό συσχέτισης μεταξύ των διαφορετικών εννοιών κάθε λέξης). Για κάθε έννοια υπολογίζεται το άθροισμα των βαρών και επιλέγεται η έννοια με το μεγαλύτερο συνολικό βάρος. Ωστόσο, η πληροφορία αυτή δεν αξιοποιείται στην αποσαφήνιση των εναπομεινασών λέξεων και υπάρχει πιθανότητα να συνυπολογιστούν βάρη μεταξύ μιας έννοιας της επόμενης εξεταζόμενης λέξης με μια έννοια της προηγούμενης λέξης που έχει απορριφθεί. Επομένως, αυτή η μέθοδος, αν και αρκετά ακριβής, δεν είναι ιδιαίτερα πολύπλοκη· ωστόσο είναι αρκετά γρήγορη.

3.1.2.1γ Μέθοδος Integer Linear Programming (ILP)

Ο αλγόριθμος ILP δημιουργεί έναν γράφο, όμοια με τον WDEG. Η διαφορά είναι ότι υπολογίζει τα αθροίσματα των βαρών για όλους τους πιθανούς συνδυασμούς εννοιών. Έτσι, επιλέγει το βέλτιστο σύνολο εννοιών λαμβάνοντας υπ' όψιν το λογικό συμπέρασμα ότι πρέπει να επιλέγονται έννοιες που παρουσιάζουν τη μεγαλύτερη συσχέτιση με τις επιλεγμένες (κι όχι με οποιεσδήποτε) έννοιες των υπόλοιπων λέξεων της πρότασης. Συνεπάγεται ότι αυτή η μέθοδος είναι πιο αργή από τις άλλες δύο, αλλά και πιο ακριβής.

3.1.2.2 Μέτρα συσχέτισης

Τα βάρη που αναφέρθηκαν ως μέτρα συσχέτισης των εννοιών δύο λέξεων της πρότασης εκφράζουν την πιθανότητα αυτός ο συνδυασμός εννοιών για τις δύο λέξεις να εμφανιστεί στην ίδια πρόταση. Στο Pythia προσφέρονται τρεις διαφορετικές επιλογές μέτρων συσχέτισης: Το Semantic Relatedness (SR) που βασίζεται σε πρότερη γνώση (η οποία είναι ο σημασιολογικός γράφος του WordNet)· το Pointwise Mutual Information (PMI) που βασίζεται σε σώματα κειμένου (corpora), στα οποία είναι προσημειωμένη η έννοια με την οποία χρησιμοποιείται κάθε λέξη, και αξιοποιεί τα στατιστικά συνεμφάνισης των εννοιών· και το Lesk-like (LL) που βασίζεται επίσης σε σώματα κειμένου τα οποία, όμως, δεν είναι προσημειωμένα (συγκεκριμένα, χρησιμοποιούνται οι ορισμοί κάθε έννοιας του WordNet για να βρεθούν όμοιες λέξεις ή ιδέες μεταξύ δύο εννοιών).

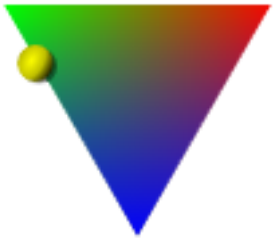


Δομή του εργαλείου Pythia. Πρώτα το κείμενο χωρίζεται σε προτάσεις, έπειτα εξακριβώνεται το μέρος του λόγου κάθε λέξης, ενώ παράλληλα γίνεται η αποσαφήνιση της έννοιάς της, στη συνέχεια πραγματοποιείται συντακτική ανάλυση και, τέλος, κατηγοριοποιείται η πρόταση ως θετική ή αρνητική με χρήση ενός κατηγοριοποιητή.

3.1.3 SentiWordNet

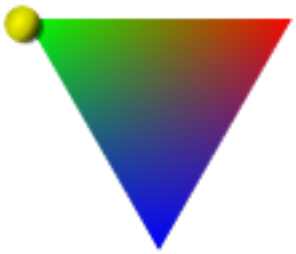
Το SentiWordNet είναι ένα λεξικό σημασιολογικών προσανατολισμών για έννοιες λέξεων. Βασίζεται στις έννοιες του λεξικού WordNet, το οποίο δημιουργήθηκε για να χρησιμοποιηθεί σε προβλήματα επεξεργασίας φυσικής γλώσσας. Το WordNet περιέχει τις πιθανές έννοιες κάθε λέξης και τοποθετεί τις έννοιες που παρουσιάζουν κάποια σχέση μεταξύ τους ή τις έννοιες που παρουσιάζουν παρόμοια σχέση με μια ορισμένη λέξη σε διάφορα υποσύνολα. Έτσι, καταφέρνει να βρει σχέσεις σε επίπεδο έννοιας κι όχι λέξης, όπως κάνουν τα παραδοσιακά λεξικά.

Το SentiWordNet περιέχει επιπλέον πληροφορία για κάθε υποσύνολο εννοιών. Συγκεκριμένα, δίνει τρεις βαθμολογίες για κάθε παράμετρο κάθε υποσυνόλου (θετικό, αρνητικό και αντικειμενικό). Η συνολική βαθμολογία είναι το 1, ενώ δε μπορεί να υπάρχει αρνητική βαθμολογία για οιαδήποτε παράμετρο. Για παράδειγμα, η λέξη «καλός» με την έννοια «επιθυμητός» θεωρείται 0,75 θετική και 0,25 αντικειμενική, με την έννοια «αναμενόμενος» θεωρείται 1 αντικειμενική, ενώ με την έννοια «ηθικός» θεωρείται 1 θετική. Αυτό αναπαρίσταται και με τη θέση μιας σφαίρας σε ένα τρίγωνο, όπου με πράσινο χρώμα απεικονίζεται η θετική παράμετρος, με μπλε η αντικειμενική και με κόκκινο η αρνητική.



P: 0.75 O: 0.25 N: 0
Καλός, επιθυμητός

P: 0 O: 1 N: 0
Καλός, αναμενόμενος



P: 1 O: 0 N: 0
Καλός, ηθικός

3.1.4 JUNG

Το JUNG (Java Universal Network/Graph Framework) είναι μια βιβλιοθήκη ανοιχτού κώδικα γραμμένη σε Java που παρέχει μεθόδους για μοντελοποίηση, ανάλυση και οπτικοποίηση δεδομένων τα οποία μπορούν να αναπαρασταθούν ως γράφοι ή δίκτυα. Χρησιμοποιήθηκε στην παρούσα πτυχιακή για την οπτικοποίηση των δεδομένων κατά την ανάπτυξη του αλγορίθμου, έτσι ώστε να εξακριβωθούν ευκολότερα κανόνες για τις συντακτικές εξαρτήσεις, αλλά και για την υλοποίηση αυτή καθ' εαυτήν του αλγορίθμου.

3.2 Ο αλγόριθμος της εργασίας

Στο πλαίσιο της παρούσας πτυχιακής εργασίας επινοήθηκε ένας αλγόριθμος προκειμένου να πραγματοποιηθεί συναισθηματική ανάλυση σε επίπεδο χαρακτηριστικού. Ο στόχος ήταν να γίνει

στο ίδιο στάδιο εύρεση των χαρακτηριστικών και της πολικότητας με την οποία είναι φορτισμένα· σε αντίθεση με τις περισσότερες παραδοσιακές μεθόδους που τείνουν να πραγματοποιούν αυτές τις δύο αναλύσεις σε δύο ξεχωριστά στάδια.

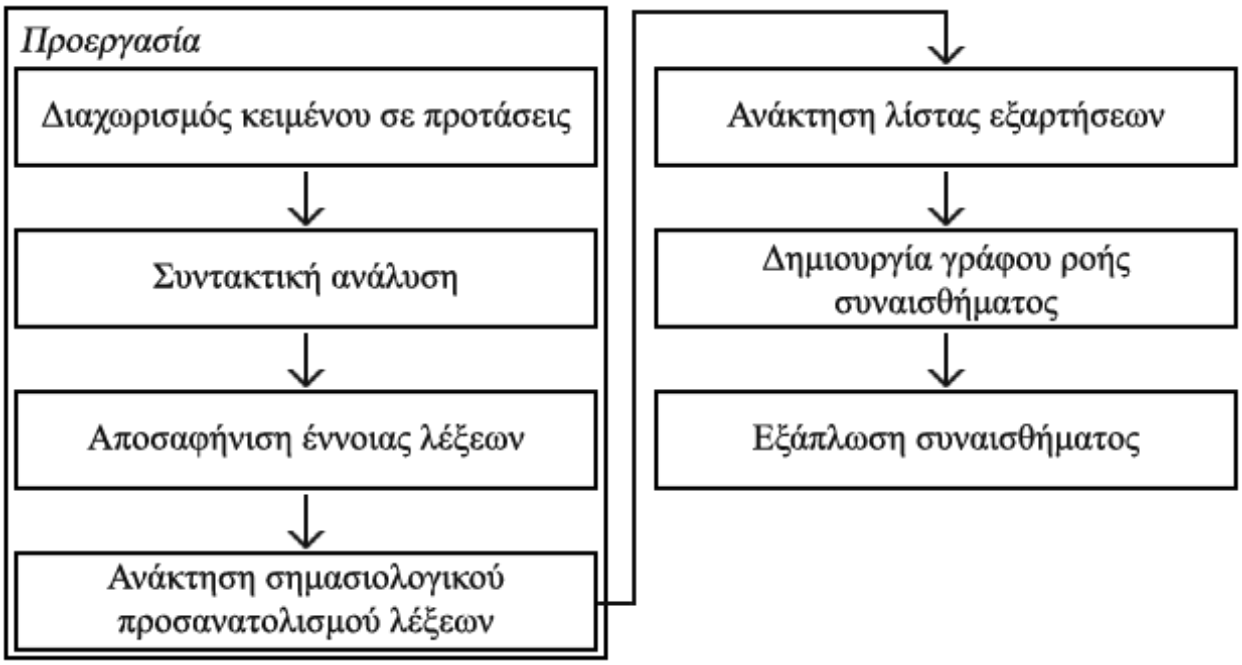
Στο πρώτο βήμα γίνεται διαχωρισμός του κειμένου σε προτάσεις, οι οποίες στη συνέχεια εξετάζονται ως οντότητες ανεξάρτητες των γειτονικών τους προτάσεων. Έπειτα πραγματοποιείται συντακτική ανάλυση της κάθε πρότασης με τον συντακτικό αναλυτή φυσικής γλώσσας του Σταντφορντ, ο οποίος δημιουργεί ένα δέντρο δομής φράσεων. Με βάση αυτήν την ανάλυση πραγματοποιείται αποσαφήνιση της έννοιας των λέξεων με τη μέθοδο FS (First Sense), αν και υποστηρίζονται και άλλες μέθοδοι. Έπειτα, ανακτώνται οι σημασιολογικοί προσανατολισμοί κάθε λέξης σύμφωνα με την έννοια με την οποία χρησιμοποιείται στην εκάστοτε πρόταση. Επομένως, σε αυτό το σημείο το κείμενο έχει διαχωριστεί σε προτάσεις, όλες οι λέξεις των οποίων έχουν συνημμένους τους σημασιολογικούς προσανατολισμούς τους.

Στο δεύτερο βήμα γίνεται ξανά συντακτική ανάλυση, αλλά αυτή τη φορά εξακριβώνεται η λίστα εξαρτήσεων που διέπουν τις λέξεις της πρότασης. Δημιουργείται ένας γράφος όπου κάθε εξάρτηση είναι μια ακμή, ενώ κόμβοι είναι οι δύο λέξεις που περιέχει αυτή η εξάρτηση. Φυσικά, αν η ίδια λέξη περιέχεται σε περισσότερες από μία εξαρτήσεις, δεν δημιουργούνται περισσότεροι του ενός κόμβοι για τη συγκεκριμένη λέξη, αλλά μόνο ένας, ο οποίος συνδέεται με περισσότερες από μία ακμές. Παράλληλα, ο γράφος περιέχει όλες τις πληροφορίες που ανακτήθηκαν στο προηγούμενο βήμα για την κάθε λέξη, όπως το μέρους του λόγου και ο σημασιολογικός προσανατολισμός της. Αξίζει να σημειωθεί ότι κάποιες εξαρτήσεις παραλείπονται έτσι ώστε να γίνει μεγαλύτερος διαχωρισμός των προτάσεων και να επιτευχθούν καλύτερα αποτελέσματα.

Επιπλέον, οι ακμές έχουν κατεύθυνση η οποία έχει αφετηρία την λέξη η οποία συνδέεται με την καθορισμένη εξάρτηση με την τερματική λέξη. Ωστόσο, αυτή η κατεύθυνση δεν δείχνει πάντα σωστά την κατεύθυνση που ακολουθεί η συναισθηματική φόρτιση. Για παράδειγμα, στην πρόταση «Το κρασί ήταν απολαυστικό» η λέξη «κρασί» είναι υποκείμενο στο ρήμα «ήταν»· εν τούτοις, η συναισθηματική φόρτιση έχει ως στόχο το κρασί (οπότε η κατεύθυνση του συναισθήματος είναι αντίθετη αυτής της εξάρτησης). Έτσι, καθορίστηκαν μερικοί κανόνες που, όταν εκπληρούνται, οδηγούν στην αντιστροφή των επικείμενων κατευθύνσεων, ούτως ώστε να εκφράζεται ορθότερα η ροή του συναισθήματος.

Στο τρίτο βήμα πραγματοποιείται εξάπλωση του συναισθήματος. Αφετηρία είναι οι εξωτερικοί κόμβοι, από τους οποίους, με βάση διάφορους κανόνες που αναλύονται στη δεύτερη

υποπαράγραφο, μεταδίδεται το συναίσθημα. Τέλος, τυπώνονται τα ουσιαστικά που επηρεάστηκαν συνοδευόμενα με την εκτιμηθείσα συναισθηματική φόρτιση, αφού γίνεται η παραδοχή ότι τα χαρακτηριστικά είναι πάντοτε ουσιαστικά. Επομένως, γίνεται και η παραδοχή ότι οι λέξεις που επηρεάστηκαν περισσότερο είναι και χαρακτηριστικά.



Τα βήματα του αλγορίθμου

3.2.1 Εκκαθάριση και αλλαγή κατεύθυνσης ακμών

Προκειμένου να επιτευχθεί καλύτερη μετάδοση συναισθήματος πραγματοποιείται διαχωρισμός των προτάσεων, όπου αυτός είναι εφικτός. Συγκεκριμένα, εξαρτήσεις τύπου conj, parataxis, root και det δε λαμβάνονται υπ’ όψιν κατά τη δημιουργία του γράφου.

Ο τύπος conj αφορά τις λέξεις που συνδέονται με συνδετικές ή αντιθετικές λέξεις (conjunctions). Συνήθως αυτή η εξάρτηση παρατηρείται μεταξύ κύριων ή δευτερευουσών προτάσεων, όπου η συναισθηματική φόρτιση της μίας δεν επηρεάζει καμμιά λέξη της άλλης. Παράλληλα, στις περιπτώσεις που η συναισθηματική φόρτιση επηρεάζει λέξη άλλης πρότασης, υπάρχουν και άλλες εξαρτήσεις που τις ενώνουν. Διαπιστώθηκε, οπότε, ότι οι εξαρτήσεις τύπου conj μπορούν να παραληφθούν κατά την ανάλυση. Για τον ίδιο λόγο παραλείπεται και η εξάρτηση parataxis, η οποία δηλώνει ότι στην πρόταση παρατάσσονται πολλές ανεξάρτητες προτάσεις, οι οποίες δεν επικοινωνούν νοηματικά μεταξύ τους.

Η εξάρτηση root δεν έχει λεξικολογικό νόημα αλλά χρησιμοποιείται από τον αναλυτή για διευκόλυνσή του, γι’ αυτό και δεν υπάρχει λόγος να συμπεριληφθεί στην ανάλυση. Τέλος, η

εξάρτηση det παρατηρείται μεταξύ άρθρων και ουσιαστικών· κρίθηκε σκόπιμο να μην εισαχθούν τα άρθρα στον γράφο, αφού δεν έχουν σημασιολογικό προσανατολισμό.

Αγνοώντας αυτές τις τέσσερις εξαρτήσεις, έγινε ένας καλύτερος διαχωρισμός της κάθε πρότασης σε υπο-προτάσεις, οι οποίες αναλύονται ως ανεξάρτητες οντότητες. Επιπλέον του διαχωρισμού, καθορίστηκαν κάποιοι κανόνες για αλλαγή της κατεύθυνσης των ακμών έτσι ώστε να εκφράζουν καλύτερα την ροή της συναισθηματικής φόρτισης.

Οι ακμές των εξαρτήσεων poss και xcomp πάντοτε αντιστρέφονται. Η εξάρτηση poss δηλώνει ότι μια λέξη είναι κτητικό ενός ουσιαστικού. Για παράδειγμα, στην φράση «το ποδήλατό μου» η εξάρτηση είναι poss(μου, ποδήλατο) και κανονικά η ακμή έχει αφετηρία το ποδήλατο. Πάντοτε το κτητικό είναι τερματικός κόμβος, πράγμα που δεν έχει νόημα στην συναισθηματική ανάλυση αφού γνωρίζουμε ότι σε μια πρόταση χαρακτηρίζεται το αντικείμενο κι όχι ο ιδιοκτήτης. Όσον αφορά το xcomp, πρόκειται για συμπλήρωμα πρότασης, το οποίο επίσης δέχεται την συναισθηματική φόρτιση (αντί να τη μεταδίδει στην πρόταση την οποία συμπληρώνει). Για παράδειγμα, στην πρόταση «I am ready to leave» (μτφ.: Είμαι έτοιμος να φύγω) η εξάρτηση περιγράφεται ως xcomp(ready, leave). Οπότε, αφετηρία της εξάρτησης είναι το «φύγω», αν και σημασιολογικά αυτή η λέξη είναι ο στόχος της συναισθηματικής φόρτισης.

Άλλες εξαρτήσεις που λαμβάνονται υπ’ όψιν αλλά αντιστρέφονται ανά περίπτωση είναι οι nsubj, csubj, nsubjpass, csubjpass, dobj, iobj και mark.

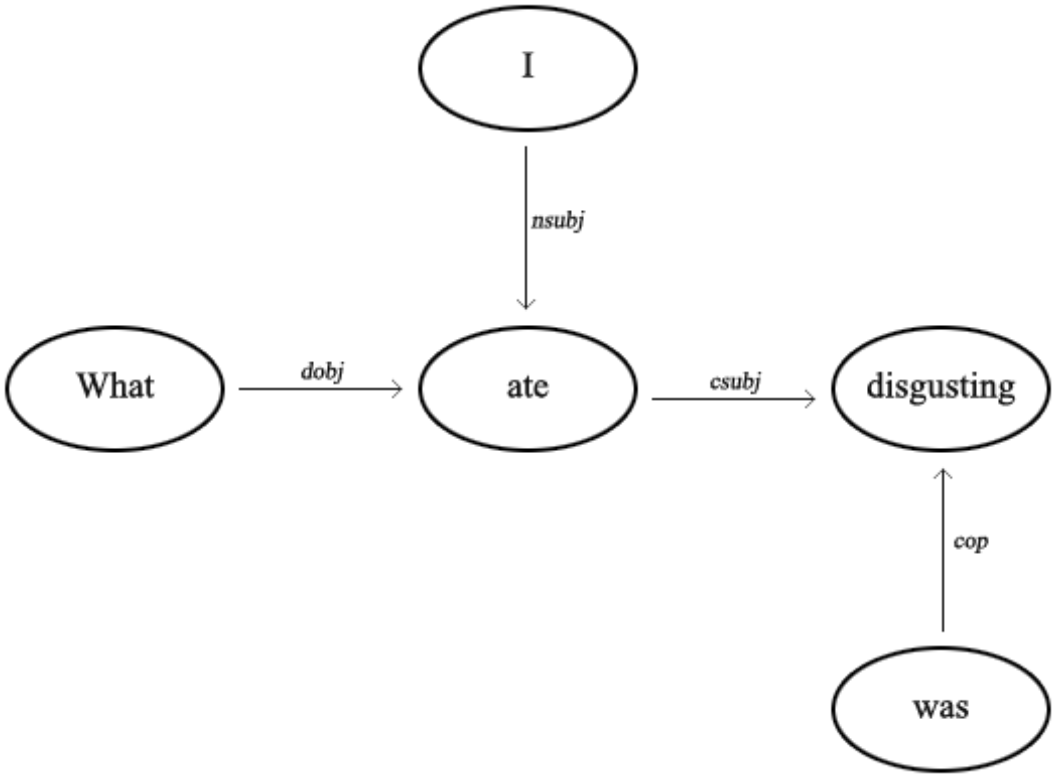
Εξάρτηση	Πλήρες όνομα	Περιγραφή
nsubj	nominal subject	υποκείμενο-ουσιαστικό
csubj	clausal subject	υποκείμενο-πρόταση
nsubjpass	passive nominal subject	υποκείμενο-ουσιαστικό σε παθητική φωνή
csubjpass	passive clausal subject	υποκείμενο-πρόταση σε παθητική φωνή
dobj	direct object	άμεσο αντικείμενο
iobj	indirect object	έμμεσο αντικείμενο
mark	marker	συνδετική λέξη που συνδέει μια δευτερεύουσα με μια άλλη πρόταση

Ο αλγόριθμος εξετάζει μία-μία όλες τις λέξεις του γράφου, αρχικά διαπιστώνοντας ποιες από τις παραπάνω εξαρτήσεις δέχεται ως εισερχόμενες ακμές ο κάθε κόμβος-λέξη. Έπειτα, ακολουθεί τους κανόνες που περιγράφονται παρακάτω για να αντιστρέψει την κατεύθυνση των ακμών όπου χρειάζεται.

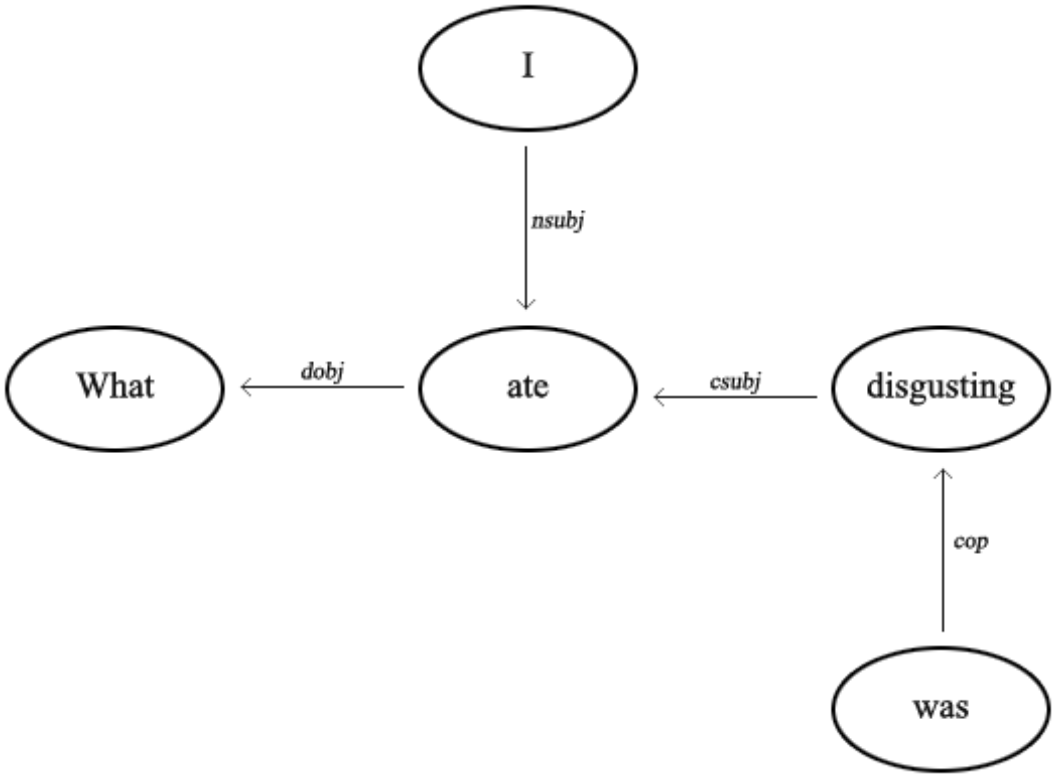
1. Αν βρέθηκε άμεσο αντικείμενο, υποκείμενο και έμμεσο αντικείμενο, τότε αντιστρέφεται η ακμή του υποκείμενου.
2. Αν βρέθηκε άμεσο αντικείμενο και marker, τότε δεν αντιστρέφεται καμμία ακμή.
3. Αν βρέθηκε άμεσο αντικείμενο, αλλά όχι marker, τότε αντιστρέφεται η ακμή του αντικειμένου.
4. Αν δε βρέθηκε άμεσο αντικείμενο αλλά βρέθηκε υποκείμενο ή υποκείμενο σε παθητική φωνή, τότε αντιστρέφεται η ακμή του υποκειμένου.

Παρακάτω περιγράφεται η εκτέλεση ενός ενδεικτικού παραδείγματος και αναφέρονται επιγραμματικά μερικά ακόμη, έτσι ώστε να φανεί σχηματικά η λογική με την οποία καθορίστηκαν αυτοί οι κανόνες.

Στο παράδειγμα «What I ate was disgusting» (δηλαδή «Αυτό που έφαγα ήταν αηδιαστικό»), υπάρχουν οι εξής συντακτικές εξαρτήσεις: το What είναι άμεσο αντικείμενο στο ate, το I υποκείμενο-ουσιαστικό στο ate, το ate υποκείμενο-πρόταση στο disgusting και το disgusting κατηγορούμενο στο was. Παρακάτω φαίνονται αυτές οι εξαρτήσεις σχηματικά.

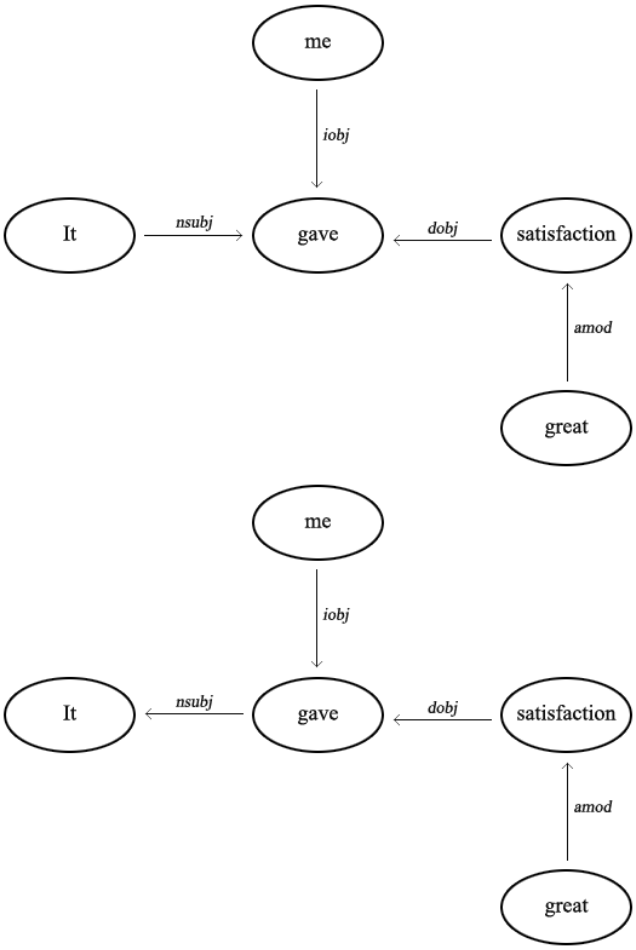


Με αυτό το παράδειγμα γίνεται ξεκάθαρο ότι οι συντακτικές εξαρτήσεις δεν περιγράφουν σωστά τη ροή της συναισθηματικής φόρτισης, αφού η λογική επιτάσσει ότι η λέξη που χαρακτηρίζεται στο προαναφερθέν παράδειγμα είναι η «What». Ακολουθώντας τους κανόνες που περιγράφηκαν προηγουμένως, θα έπρεπε να εξετάσουμε όλους τους κόμβους που έχουν εισερχόμενες ακμές για την πιθανότητα να πρέπει να αντιστραφούν κάποιες από αυτές. Εξετάζοντας τον κόμβο «ate» βλέπουμε ότι δέχεται ακμές dobj και nsubj· πρόκειται για την περίπτωση 3, οπότε η ακμή dobj πρέπει να αλλάξει κατεύθυνση. Εξετάζοντας τον κόμβο «disgusting» βλέπουμε ότι δέχεται ακμές csubj και cop· επομένως, πρέπει να αλλάξει η κατεύθυνση της csubj σύμφωνα με τον τέταρτο κανόνα. Το τελικό σχήμα που προκύπτει από τον αλγόριθμο, που έχει ως στόχο να εκφράσει τη ροή του συναισθήματος μέσα στην πρόταση είναι το εξής:

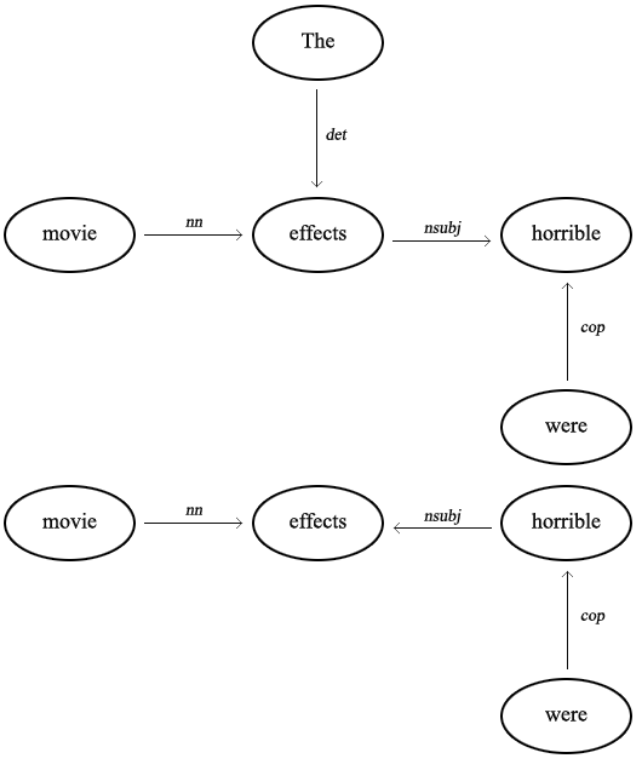


Παρακάτω αναφέρονται σχηματικά δύο ακόμη παραδείγματα που επιδεικνύουν τους λόγους που επιλέχθηκαν οι τέσσερις κανόνες. Στην πρώτη εικόνα κάθε παραδείγματος φαίνονται οι συντακτικές σχέσεις και στη δεύτερη η ροή του συναισθήματος, όπως αποτυπώνεται μετά την εφαρμογή των επιλεγμένων κανόνων. Το παράδειγμα 2 αφορά τον πρώτο κανόνα σχετικά με την ύπαρξη άμεσου αντικειμένου, ενώ το παράδειγμα 3 σχετίζεται με τον τέταρτο κανόνα.

Παράδειγμα 2: It gave me great satisfaction.



Παράδειγμα 3: The movie effects were horrible.



3.2.2 Μετάδοση συναισθήματος

Αφού δημιουργηθεί ο γράφος που περιέχει όλες τις χρήσιμες ακμές με τις κατευθύνσεις που περιγράφουν ορθά τη ροή του συναισθήματος, πραγματοποιείται η μετάδοσή του. Η συναισθηματική φόρτιση μεταφέρεται από τους κόμβους που έχουν μία ή περισσότερες εξερχόμενες ακμές, αλλά καμμία εισερχόμενη, προς όλους τους κόμβους με τους οποίους συνδέονται. Αφού μεταφερθεί το συναίσθημα, η ακμή διαγράφεται (οπότε οι εξωτερικοί κόμβοι παύουν να είναι μέρος του γράφου). Αφού διαγράφεται η ακμή, όσοι κόμβοι τη δεχόντουσαν ως τη μοναδική εισερχόμενη δεν έχουν πια εισερχόμενες ακμές, επομένως μπορεί να συνεχιστεί η μετάδοση του συναισθήματος από αυτούς.

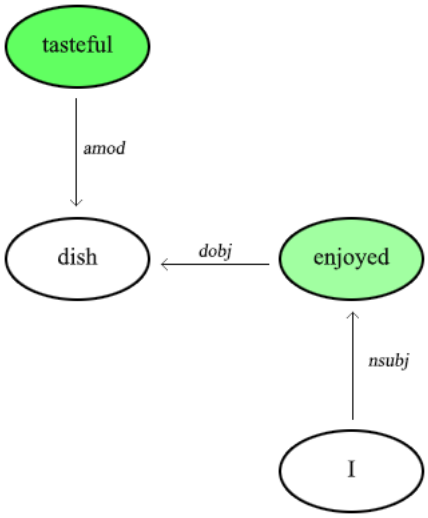
Κάθε λέξη έχει μία τιμή για το θετικό και μια για το αρνητικό συναίσθημα. Όταν πραγματοποιείται μετάδοση του συναισθήματος μεταφέρεται μόνο το 85% του συναισθήματος της λέξης-αφετηρίας στην λέξη-στόχο. Αυτό συμβαίνει για να μη μεταδίδεται το συναίσθημα μιας λέξης σε όλη την πρόταση, αλλά να περιορίζεται στις λέξεις που επηρεάζει πιο άμεσα. Υπάρχει ένα όριο ελάχιστης μεταβολής που πρέπει να έχει υποστεί στη συναισθηματική φόρτιση μιας λέξης έτσι ώστε να θεωρηθεί θετικά ή αρνητικά φορτισμένη. Οι πολύ μακρινές λέξεις επηρεάζονται σε πολύ μικρό βαθμό οπότε δεν πληρούν αυτό το κριτήριο.

Μια ακόμη ιδιοτροπία που αξίζει να σημειωθεί είναι ότι σε περίπτωση εξάρτησης neg (negation, δηλαδή άρνηση) δε μεταδίδεται το συναίσθημα αλλά πραγματοποιείται ανταλλαγή των δύο συναισθημάτων (θετικού και αρνητικού) της λέξης που δέχεται αυτή την ακμή ως εισερχόμενη. Αυτό συμβαίνει διότι η άρνηση δηλώνει ότι η λέξη αποκτά τη διαμετρικά αντίθετη σημασία· δηλαδή, μια θετική λέξη αποκτά αρνητική σημασία και το αντίστροφο.

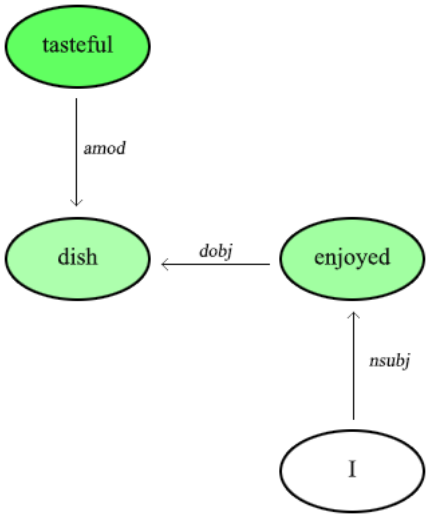
Συνεπώς, το συναίσθημα μεταφέρεται από τους εξωτερικούς στους εσωτερικούς κόμβους, αφού υποστεί μια μικρή φθορά της τάξης του 15% και οι αρνήσεις υπολογίζονται πάντα τελευταίες για κάθε κόμβο που έχει περισσότερες της μίας εισερχόμενες ακμές. Σε περίπτωση που δεν υπάρχουν εξωτερικοί κόμβοι, διότι ο γράφος περιέχει ένα κυκλικό σχήμα, εντοπίζεται ο κύκλος και αποθηκεύονται προσωρινά τα συναισθήματα όλων των λέξεων. Έπειτα, μεταφέρονται όλα τα συναισθήματα μεταξύ όλων των λέξεων και διαγράφονται όλες οι ακμές. Με αυτή τη μέθοδο αποφεύγεται η παγίδευση σε έναν ατέρμονα βρόγχο (με αέναο πολλαπλασιασμό του συναισθήματος).

Ένα παράδειγμα εξάπλωσης συναισθήματος με φθορά περιγράφεται παρακάτω. Στην πρόταση «I enjoyed the tasteful dish» (δηλαδή «Μου άρεσε το εύγευστο πιάτο) η λέξη enjoyed έχει σημασιολογικό προσανατολισμό 0,375 θετικό και καθόλου αρνητικό, η λέξη tasteful 0,75 θετικό

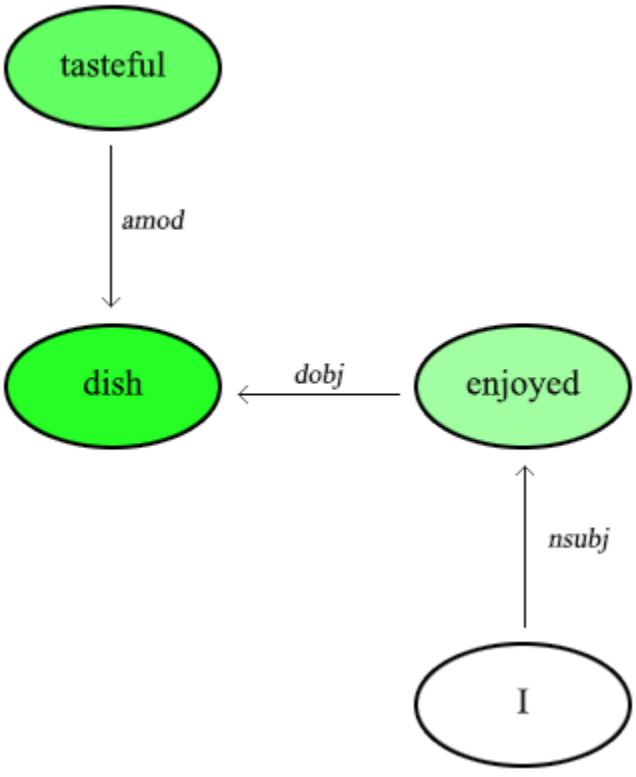
και καθόλου αρνητικό, και οι dish και I δεν έχουν καθόλου θετικό, ούτε αρνητικό σημασιολογικό προσανατολισμό. Στο πρώτο στάδιο μεταδίδεται το συναίσθημα από το I στο enjoyed, οπότε δεν αλλάζει καμμία συναισθηματική φόρτιση. Έπειτα μεταδίδεται από το enjoyed στο dish, επομένως η θετική συναισθηματική φόρτιση του dish αποκτάει την τιμή 0,319 (από 0 που ήταν). Στη συνέχεια, μεταδίδεται το συναίσθημα από το tasteful στο dish, οπότε η συναισθηματική φόρτιση του dish αυξάνεται κατά 0,531 και φτάνει το 0,85.



Ο σημασιολογικός προσανατολισμός των λέξεων.



Η συναισθηματική φόρτιση μετά τη μετάδοση του συναισθήματος της λέξης enjoyed.



Η συναισθηματική φόρτιση μετά τη μετάδοση του συναισθήματος της λέξης tasteful.

IV Υλοποίηση

Ο αλγόριθμος που περιγράφηκε στην προηγούμενη ενότητα υλοποιήθηκε σε java. Το πρόγραμμα δομήθηκε σε 2 πακέτα: το πακέτο που περιέχει τις κλάσεις με τη λογική του προγράμματος, και το πακέτο που περιέχει τις κλάσεις-αντικείμενα του προγράμματος. Επιπλέον, συμπεριλήφθηκε έτοιμος κώδικας που παρασχέθηκε από το πανεπιστήμιο και αφορά την αποσαφήνιση της έννοιας των λέξεων (ο οποίος προέρχεται από το εργαλείο Pythia).

4.1 Δομές δεδομένων

Τα κλάσεις που δημιουργήθηκαν για τις ανάγκες της εργασίας και εκφράζουν δομές δεδομένων είναι τα Word, Dependency, SentiGraph, και WordWithDependecy. Η κλάση Word (δλδ. «λέξη») χρησιμοποιείται για να περιγράψει μια λέξη της πρότασης και περιέχει τα εξής γνωρίσματα:

- αυθεντική λέξη: η λέξη με την ορθογραφία και κεφαλαιοποίηση όπως βρέθηκε στην πρόταση·
- θέση: η θέση της λέξης στην πρόταση (π.χ. 2 αν είναι $2^{η}$ λέξη της πρότασης)·
- μέρος του λόγου: το μέρος του λόγου της λέξης, που μπορεί να είναι ουσιαστικό, ρήμα, επίθετο ή επίρρημα (αποκτάται μετά τη συντακτική ανάλυση)·
- ορισμός: ο ορισμός της λέξης (αποκτάται μετά την αποσαφήνισή της μαζί με τον σημασιολογικό προσανατολισμό της)·
- θετικός προσανατολισμός: έχει εύρος 0-1 ανάλογα με την ένταση της θετικής σημασίας αυτής της έννοιας της λέξης·
- αρνητικός προσανατολισμός: αντίστοιχα με τον θετικό προσανατολισμό για την αρνητική σημασία·
- διαφορά θετικού προσανατολισμού: περιέχει την τιμή της επιπλέουσας θετικής συναισθηματικής φόρτισης που δέχθηκε η λέξη από άλλες λέξεις της πρότασης· και
- διαφορά αρνητικού προσανατολισμού: όμοια με τον θετικό για την αρνητική συναισθηματική φόρτιση.

Η κλάση Dependency (δλδ. «εξάρτηση») χρησιμοποιείται για να εκφράσει την εξάρτηση μεταξύ δύο λέξεων και περιέχει σαν πληροφορία μόνο το γνώρισμα του τύπου της εξάρτησης.

Η κλάση SentiGraph (δλδ. «γράφος συναισθήματος») περιέχει μόνο έναν γράφο, με κόμβους τύπου Word και ακμές τύπου Dependency, ως γνώρισμα.

Οι τρεις αυτές κλάσεις χρησιμοποιούνται από όλο το πρόγραμμα για διαχείριση του γράφου, μετάδοση της συναισθηματικής φόρτισης και εξαγωγή συμπερασμάτων.

Η κλάση DependencyWithSource (δλδ. εξάρτηση μαζί με τον αφετηριακό κόμβο) δημιουργήθηκε αποκλειστικά για την επίλυση ενός τμήματος του προβλήματος, το οποίο περιγράφεται στη συνέχεια. Το πρόβλημα σχετίζεται με την αμφίδρομη μετάδοση συναισθήματος μεταξύ δύο ή περισσότερων κόμβων και σκοπός της κλάσης είναι να αποθηκευτεί πληροφορία σχετικά με τον σημασιολογικό προσανατολισμό και την τρέχουσα συναισθηματική φόρτιση κάθε κόμβου, έτσι ώστε να μην επιστρέψει ένας σημασιολογικός προσανατολισμός στον αφετηριακό κόμβο ως συναισθηματική φόρτιση. Η κλάση αυτή περιέχει τα εξής γνωρίσματα: εξάρτηση (τύπου Dependency), αφετηριακός κόμβος (τύπου Word), θετικό συνολικό συναίσθημα (άθροισμα σημασιολογικού προσανατολισμού με συναισθηματική φόρτιση) και αρνητικό συνολικό συναίσθημα.

4.2 Λογική προγράμματος

Τρεις είναι οι κύριες κλάσεις που παρέχουν όλα τα εργαλεία για τη συναισθηματική ανάλυση: η WSD, η SyntaxAnalyser και η SpeadingActivation.

4.2.1 Η κύρια μέθοδος

Αυτές τις κλάσεις τις ενώνει η κλάση SyntaxSentimentAnalysis (δλδ. συναισθηματική ανάλυση με αξιοποίηση του συντακτικού). Περιέχει μόνο την κύρια μέθοδο του προγράμματος (main). Η κύρια μέθοδος αρχικά δημιουργεί από ένα αντικείμενο WSD και SyntaxAnalyser. Έπειτα, ανοίγει το αρχείο (που είναι σε μορφή .txt) στο οποίο περιέχονται οι προτάσεις που πρόκειται να αναλυθούν, και δημιουργεί το αρχείο όπου θα γράψει τα αποτελέσματα. Στη συνέχεια, καλεί τη μέθοδο wsd για κάθε γραμμή του αρχείου ανάγνωσης, η οποία χωρίζει τη γραμμή σε προτάσεις, αποσαφηνίζει την έννοια κάθε λέξης και ανακτάει το συναίσθημα κάθε έννοιας από το SentiWordNet. Έπειτα, η κύρια μέθοδος μετατρέπει την έξοδο της wsd σε μορφή που να μπορεί να διαβαστεί από τη μέθοδο createGraph, δηλαδή μετατρέπει τη λίστα από SentiSentence σε λίστα από String. Για κάθε πρόταση της λίστας καλεί την createGraph του αντικειμένου τύπου SyntaxAnalyser, η οποία δημιουργεί το SentiGraph κάθε πρότασης. Τέλος, για κάθε γράφο πραγματοποιεί μετάδοση του συναισθήματος δημιουργώντας ένα αντικείμενο SpreadingActivation και καλώντας τη μέθοδο του ονόματι evaluate.

Όταν ολοκληρωθεί αυτή η διαδικασία, οι λέξεις του γράφου περιέχουν όλες τις χρήσιμες πληροφορίες· συγκεκριμένα, την αυθεντική λέξη, το μέρος του λόγου και τη διαφορά θετικού και αρνητικού προσανατολισμού. Αυτή τη πληροφορία χρησιμοποιεί μέθοδος getAffectedNouns

του SentiGraph, την οποία καλεί η κύρια μέθοδος προκειμένου να τυπώσει τα αποτελέσματα κάθε γραμμής.

Η κύρια συνάρτηση γράφει σε κάθε γραμμή του αρχείου το κείμενο που εξετάστηκε μέσα σε διπλά αγγλικά εισαγωγικά και έπειτα όλα τα συναισθηματικά φορτισμένα ουσιαστικά που βρήκε, καλώντας την getAffectedNouns, η οποία τοποθετεί πρώτα ένα διαχωριστικό (ελληνικό ερωτηματικό), έπειτα το ουσιαστικό, ένα ακόμη διαχωριστικό και, τέλος, τη διαφορά του θετικού με το αρνητικό συναίσθημα. Προϋπόθεση για να θεωρηθεί ένα ουσιαστικό φορτισμένο είναι να έχει είτε θετική φόρτιση, είτε αρνητική, είτε και τις δύο.

```
output = "\"" + input + "\"";
for (SentiGraph sentence : graphSentences) {
        output += sentence.getAffectedNouns();
}
bufferedWriter.append(output);
```

Κώδικας από την κύρια μέθοδο

```
for (Word w : g.getVertices()) {
        if ((w.getPositDif() != 0 || w.getNegatDif() != 0) &&
w.getPosTag().equals("NOUN")) {
                sentiment = w.getPositDif() - w.getNegatDif();
                words += ";" + w.getOriginal() + ";" +
df.format(sentiment);
        }
}
```

Κώδικας από τη συνάρτηση getAffectedNouns

4.2.2 Η κλάση WSD

Η κλάση WSD αναλαμβάνει την αποσαφήνιση της έννοιας των λέξεων. Προέρχεται από τον πηγαίο κώδικα της κλάσης WSDWS του εργαλείου Pythia και αφέθηκε σχεδόν αυτούσια. Έγιναν μερικές τροποποιήσεις έτσι ώστε να εξυπηρετήσει καλύτερα τις ανάγκες του προγράμματος. Αφαιρέθηκαν πολλές μέθοδοι που δε χρειαζόντουσαν για την υλοποίηση του προγράμματος και άλλαξε ο τύπος εξόδου της καλούμενης συνάρτησης (wsd) έτσι ώστε να επιστρέφει αντικείμενα τύπου SentiSentence αντί για τύπου Output.

4.2.3 Η κλάση SyntaxAnalyser

Ο συντακτικός αναλυτής αναλαμβάνει όλες τις εργασίες που αφορούν το στήσιμο του γράφου· από τη δημιουργία του μέχρι τη βελτιστοποίησή του έτσι ώστε να αντιπροσωπεύει τη ροή της συναισθηματικής φόρτισης στην πρόταση.

Περιέχει μόνο μία μέθοδο, την createGraph, που δέχεται ως είσοδο ένα String (που είναι η πρόταση προς ανάλυση) και ένα SentiSentence (που είναι η πρόταση χωρισμένη σε λέξεις, η κάθε μία από τις οποίες συνοδεύεται με πληροφορίες που ανακτήθηκαν από την wsd). Η createGraph ακολουθεί τα εξής βήματα:

1. πραγματοποιεί συντακτική ανάλυση του ακατέργαστου String και ανακτά μία λίστα με τις εξαρτήσεις (dependencies) μεταξύ των λέξεων·
2. δημιουργεί έναν κενό γράφο·
3. για κάθε λέξη της πρότασης δημιουργεί ένα αντικείμενο τύπου Word, συμπεριλαμβάνοντας σε αυτό τις πληροφορίες από το στιγμιότυπο SentiSentence, και προσθέτει το Word ως κόμβο στον γράφο·
4. πραγματοποιεί μια επανάληψη για όλες τις εξαρτήσεις της πρότασης, τις οποίες προσθέτει στον γράφο ως ακμές με κατεύθυνση από την πηγαία λέξη στην τερματική, εκτός από τις περιπτώσεις εξαρτήσεων poss και xcomp όπου δημιουργεί τις ακμές με την αντίστροφη κατεύθυνση, και τις περιπτώσεις conj, parataxis, root και det, τις οποίες αγνοεί·
5. δημιουργεί ένα αντικείμενο SentiGraph με τον γράφο που δημιούργησε ως γνώρισμα έτσι ώστε να μπορεί να αξιοποιήσει τις σχετικές μεθόδους της κλάσης·
6. αντιστρέφει τη κατεύθυνση των ακμών σύμφωνα με τις συνθήκες που αναλύθηκαν στο κεφάλαιο 3,
7. επιστρέφει το στιγμιότυπο SentiGraph.

Από τα παραπάνω αξίζει να αναλυθεί η υλοποίηση του έκτου βήματος. Αρχικά, δεσμεύεται χώρος για μια λίστα από εξαρτήσεις. Έπειτα, γίνεται επανάληψη ανάμεσα σε όλες τις λέξεις της πρότασης έτσι και εξετάζονται όλες οι εισερχόμενες σε αυτή ακμές (δηλαδή οι εξαρτήσεις). Αφού εξεταστούν όλες οι εισερχόμενες εξαρτήσεις της λέξης, εξετάζεται αν πρέπει να αντιστραφούν μία ή περισσότερες ακμές. Αν ναι, προστίθενται στη λίστα με τις εξαρτήσεις (οι εξαρτήσεις ανακτώνται με τη μέθοδο lookupIncoming του Sentigraph, η οποία επιστρέφει όλες τις εξαρτήσεις που έχουν έναν από τους τύπους που της δίνονται ως γνωρίσματα). Αφού εξεταστούν όλες οι λέξεις του γράφου, πραγματοποιούνται οι αντίστοιχες αντιστροφές. Αυτός ο χρονισμός είναι πολύ σημαντικός καθώς, αν γίνει αντιστροφή στο τέλος της ανάλυσης κάθε κόμβου, τα αποτελέσματα θα είναι απρόβλεπτα. Αυτό συμβαίνει διότι ο αλγόριθμος βασίζεται στη λογική ότι μια ακμή έχει μόνο έναν τερματικό κόμβο (στον οποίο είναι εισερχόμενη). Αν αντιστραφεί, θα μπορούσε να επηρεάσει την ανάλυση για τον αφετηριακό κόμβο, αν αυτή πραγματοποιηθεί πριν την ανάλυση του τερματικού κόμβου.

```
ArrayList<Dependency> DepsToBeReversed =
new ArrayList<>();
for (Word w : sg.getG().getVertices()) {
        [...]
        // if should be reversed
        sg.lookupIncoming(DepsToBeReversed
, w, depType1, depType2);
}
if (DepsToBeReversed.size() > 0) {
        sg.reverseEdges(DepsToBeReversed);
}
```

Το αν πρέπει να αντιστραφούν ακμές εξαρτάται από το σύνολο των τύπων ακμών που «δείχνουν» στον εκάστοτε κόμβο-λέξη. Όταν βρίσκεται nsubj ή csub, σημειώνεται ότι έχει βρεθεί ένα υποκείμενο, όταν βρίσκεται nsubjpass ή csubjpass σημειώνεται ότι έχει βρεθεί ένα παθητικό υποκείμενο, όταν βρίσκεται dobj σημειώνεται ότι έχει βρεθεί ένα άμεσο αντικείμενο, iob έμμεσο αντικείμενο και mark ένα marker.

```
switch (inDep.getType()) {
        case "nsubj":
        case "csubj":
                foundSubj = true;
                break;
        case "nsubjpass":
        case "csubjpass":
                foundSubjpass = true;
                break;
        case "dobj":
                foundDobj = true;
               break;
        case "iobj":
                foundIobj = true;
               break;
        case "mark":
                foundMark = true;
               break;
}
```

Αφού βρεθούν όλοι οι τύποι των εισερχόμενων ακμών, πραγματοποιείται έλεγχο για το αν πρέπει να αντιστραφούν κάποιες από αυτές (και, αν ναι, ποιες). Αν έχει βρεθεί άμεσο αντικείμενο, εξετάζεται αν έχει βρεθεί υποκείμενο σε ενεργητική φωνή και έμμεσο αντικείμενο. Αν ναι, αντιστρέφεται η κατεύθυνση των υποκειμένων. Αν όχι, εξετάζεται αν έχει βρεθεί marker. Αν έχει βρεθεί, δεν αντιστρέφεται καμμία ακμή, διαφορετικά αντιστρέφονται όσες ακμές είναι τύπου άμεσου αντικειμένου. Τέλος, αν δε βρέθηκε άμεσο αντικείμενο, αλλά βρέθηκε υποκείμενο (είτε σε ενεργητική, είτε σε παθητική φωνή), τότε αντιστρέφεται κάθε ακμή τύπου υποκειμένου.

```
if (foundDobj) {
        if (foundSubj && foundIobj) {
                sg.lookupIncoming(DepsToBeReversed, w, "nsubj", "csubj");
        } else if (!foundMark) {
                sg.lookupIncoming(DepsToBeReversed, w, "dobj");
        }
} else if (foundSubj || foundSubjpass) {
        sg.lookupIncoming(DepsToBeReversed, w, "nsubj", "csubj",
"nsubjpass", "csubjpass");
}
```

Σχετικά με την lookupIncoming· πρόκειται για μια μέθοδο της κλάσης SentiGraph, η οποία δέχεται ως παραμέτρους μια λίστα εξαρτήσεων, μια λέξη, και άπειρο αριθμό String, τα οποία πρόκεινται για ονόματα εξαρτήσεων. Αναλαμβάνει να βρίσκει όλες τις εξαρτήσεις που στον γράφο έχουν ως τερματικό τους κόμβο την ορισμένη λέξη, οι οποίες έχουν έναν από τους ορισμένους τύπους, και να τις προσθέτει στη λίστα με τις εξαρτήσεις.

Η reverseEdges είναι επίσης μέθοδος της κλάσης SentiGraph, που δέχεται μια λίστα από εξαρτήσεις ως παράμετρο και καλεί για κάθε μία από αυτές τη μέθοδο reverseEdge. Η reverseEdge διαγράφει την ακμή και δημιουργεί μια καινούργια με αφετηρία τον τερματικό κόμβο και στόχο την αφετηρία της διαγεγραμμένης.

4.3.4 Η κλάση SpreadingActivation

Η κλάση SpreadingActivation αναλαμβάνει τη διάδοση της συναισθηματικής φόρτισης στον γράφο. Έχει ως γνωρίσματα έναν γράφο και μια σταθερά, η οποία ορίζει την εξασθένηση που υφίσταται η συναισθηματική φόρτιση πριν μεταφερθεί. Αυτή έχει οριστεί ως 0,85· δηλαδή, το 85% του αθροίσματος σημασιολογικού προσανατολισμού με τη συναισθηματική φόρτιση μεταφέρεται από τον αφετηριακό κόμβο στον τερματικό.

Η διάδοση της συναισθηματικής φόρτισης διαχειρίζεται από τη μέθοδο evaluate. Αυτή, όσο υπάρχουν ακμές στον γράφο, μεταδίδει το συναίσθημα, καλώντας την ιδιωτική μέθοδο propagateSentiment, από τους εξωτερικούς κόμβους (που δεν έχουν εισερχόμενους κόμβους) στους εσωτερικούς. Κάθε φορά που πραγματοποιείται μια μετάδοση φόρτισης διαγράφεται η αντίστοιχη ακμή, οπότε δημιουργούνται νέοι εξωτερικοί κόμβοι. Ωστόσο, ενδέχεται να υπάρχει κύκλος μέσα στον γράφο κι έτσι κάποια στιγμή ο γράφος δε θα έχει κανέναν εξωτερικό κόμβο. Σε αυτή τη περίπτωση, δε θα πραγματοποιηθεί καμμία αλλαγή από την propagateSentiment, και ως συνέπεια θα κληθεί η resolveLoops. Η ιδιωτική μέθοδος resolveLoops εντοπίζει κύκλους στον γράφο και μεταδίδει ταυτόχρονα το συναίσθημα όλων των κόμβων, διαγράφοντας τις αντίστοιχες ακμές. Οπότε, πραγματοποιείται μια αλλαγή στον γράφο και, εφόσον υπάρχουν και άλλες ακμές (δηλαδή αν οι ακμές του κύκλου δεν ήταν οι τελευταίες) συνεχίζεται η διάδοση συναισθήματος από την evaluate. Ωστόσο, το πρόβλημα εύρεσης κύκλων έχει υψηλή πολυπλοκότητα κι έτσι προτιμήθηκε μια πιο αφελής προσέγγιση η οποία, αν και καταφέρνει να εντοπίσει τους κύκλους στις περισσότερες περιπτώσεις, μερικές φορές αποτυγχάνει, με αποτέλεσμα να μη πραγματοποιείται καμμία αλλαγή. Τότε, σταματάει η διάδοση συναισθηματικής φόρτισης και επιστρέφει ο έλεγχος στη μέθοδο που κάλεσε την evaluate (δηλαδή στην κύρια μέθοδο).

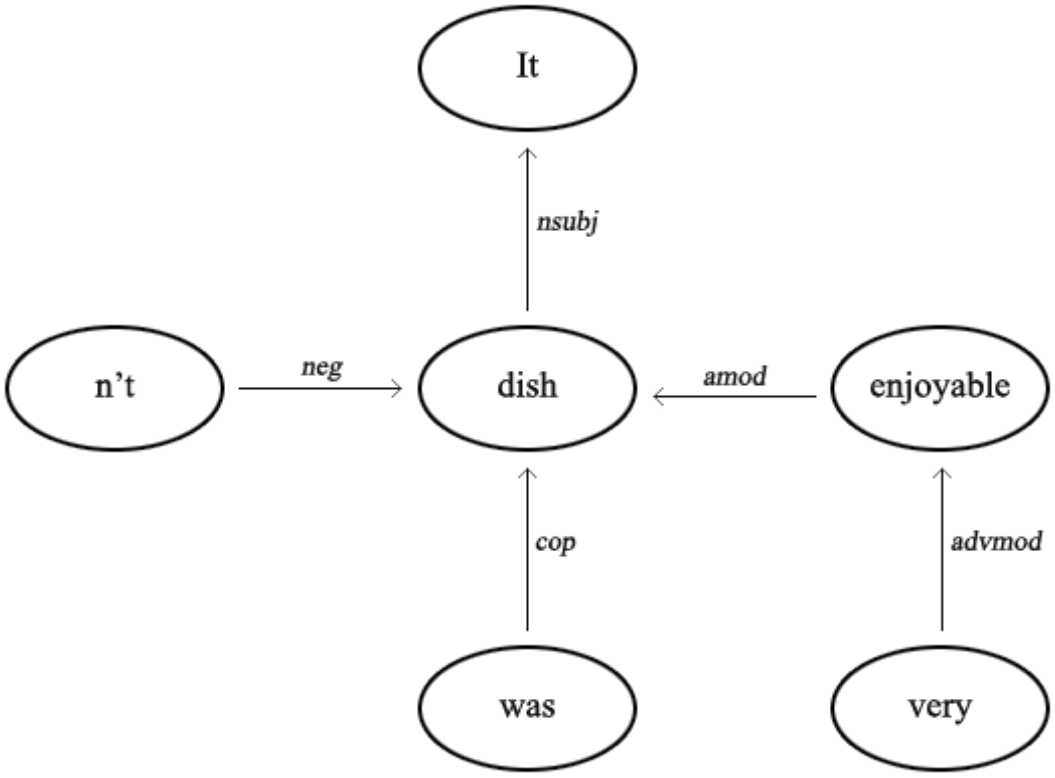
```
public void evaluate() {
    int edgeCount = g.getEdgeCount();
    boolean changeOccurred;

    while (edgeCount > 0) {
        changeOccurred = propagateSentiment();
        if (!changeOccurred) {
            changeOccurred = resolveLoops();
            if (!changeOccurred) {
                // Could not resolve loop
                break;
            }
        }
        edgeCount = g.getEdgeCount();
    }
}
```

4.3.4.1 Η μέθοδος propagateSentiment

Η μέθοδος propagateSentiment πραγματοποιεί μια επανάληψη για κάθε ακμή του γράφου και εξετάζει αν ο αφετηριακός κόμβος (από τον οποίο εξέρχεται) έχει εισερχόμενες ακμές. Αν όχι,

τότε αυτός ο κόμβος είναι εξωτερικός, οπότε πραγματοποιείται η διάδοση του συναισθήματος. Ωστόσο, αν η εξάρτηση είναι τύπου neg (negation, άρνηση), τότε αναβάλλεται η διαχείριση αυτής της ακμής μέχρι να μην υπάρχει καμμία άλλη ακμή εισερχόμενη στον τερματικό κόμβο. Αυτό συμβαίνει διότι οι αρνήσεις αντιστρέφουν το συναίσθημα του τερματικού κόμβου, άρα πρέπει πρώτα να αποκτήσει τη συνολική φόρτιση από τις άλλες εξαρτήσεις και έπειτα να υπολογιστεί η άρνηση. Κάθε φορά που εντοπίζει έναν εξωτερικό κόμβο πραγματοποιεί τη διάδοση του συναισθήματος (με την καθορισμένη εξασθένιση) και επιστρέφει αλήθεια (true). Αν τελειώσει η επανάληψη για κάθε εξάρτηση και δεν έχει βρει εξωτερικό κόμβο επιστρέφει ψέμα (false).



Παράδειγμα «It wasn't a very enjoyable dish», το οποίο δείχνει σχηματικά γιατί η άρνηση εξετάζεται τελευταία.

Αξίζει να σημειωθεί ότι για να μεταδώσει το συναίσθημα καλεί την getPositiveSentiment (και έπειτα τη getNegativeSentiment) της πηγαίας λέξης, η οποία επιστρέφει το άθροισμα του σημασιολογικού προσανατολισμού με τη συναισθηματική φόρτιση (που έχει επιβληθεί από άλλες λέξεις). Αντίστοιχα, η ανάθεση της φόρτισης πραγματοποιείται με τη μέθοδο setPositDif (και έπειτα τη setNegatDif), η οποία προσθέτει τη νέα συναισθηματική φόρτιση (λαμβάνοντας υπ' όψιν την εξασθένιση) στην προηγούμενη. Επομένως, ο σημασιολογικός προσανατολισμός της λέξης, ο οποίος ανακτήθηκε από το λεξικό, δεν αλλάξει ποτέ.

Όταν αντιμετωπίζεται εξάρτηση τύπου άρνησης, ο σημασιολογικός προσανατολισμός της πηγαίας λέξης δεν επηρεάζει καθόλου την τερματική λέξη. Συγκεκριμένα, καλείται η μέθοδος swapSentiments η οποία αναλαμβάνει να αντιστρέψει το συνολικό θετικό συναίσθημα της λέξης με το αρνητικό, πάντοτε επηρεάζοντας μόνο τη θετική και αρνητική διαφορά. Σημειώνεται ότι αυτή είναι η μόνη περίπτωση που ενδέχεται η συναισθηματική φόρτιση να παρουσιάσει αρνητική τιμή.

```
public void swapSentiments() {
        Double oldps = getPositiveSentiment();
        psDifference = getNegativeSentiment() - positiveSentiment;
        nsDifference = oldps - negativeSentiment;
    }
```

4.3.4.2 Η μέθοδος resolveLoops

Η μέθοδος resolveLoops αναλαμβάνει την εύρεση και διευθέτηση των κύκλων μέσα στον γράφο. Πραγματοποιεί μια επανάληψη για κάθε εξάρτηση και κρατάει μια λίστα με τις ακμές και τους κόμβους που οδηγούν σε αδιέξοδο (δηλαδή έχουν μόνο εισερχόμενους κι όχι εξερχόμενες ακμές, επομένως αποκλείεται να είναι μέρη του κύκλου). Επιπλέον, κρατάει μια λίστα από εξαρτήσεις, συνοδευόμενες από τους αφετηριακούς κόμβους (στιγμιότυπα DependencyWithSource). Αυτό το κάνει έτσι ώστε να έχει αποθηκευμένο το μονοπάτι που ακολούθησε όταν φτάσει να ανιχνεύσει έναν κύκλο. Ανιχνεύει το μονοπάτι ψάχνοντας τον πηγαίο κόμβο-λέξη της τρέχουσας εξάρτησης στη λίστα από εξαρτήσεις που αποθήκευσε, καλώντας τη συνάρτηση findPos. Αν δε βρει τη λέξη στη λίστα, την τοποθετεί και εξετάζει την επόμενη. Αν τη βρει, αφαιρεί όσες εξαρτήσεις δεν είναι μέρος του κύκλου (δηλαδή τις εξαρτήσεις που βρίσκονται πριν την εξάρτηση από την οποία αρχίζει ο κύκλος) και έπειτα καλεί την propagateLoopSentiment για να μεταδώσει τη συναισθηματική φόρτιση μεταξύ όλων των κόμβων.

Η propagateLoopSentiment είναι μια ιδιωτική μέθοδος της κλάσης SpreadingActivation. Δέχεται ως είσοδο τη λίστα από τις συναρτήσεις που σχηματίζουν κύκλο. Αποθηκεύει το συνολικό συναίσθημα της πρώτης λέξης και έπειτα πραγματοποιεί τη μετάδοση συναισθήματος από την τελευταία λέξη της λίστας. Στη συνέχεια, πραγματοποιεί μια επανάληψη για κάθε λέξη, αρχίζοντας από τη δεύτερη. Αποθηκεύει το συνολικό συναίσθημα της τρέχουσας λέξης και πραγματοποιεί τη μετάδοση από την προηγούμενη. Κάθε φορά που ολοκληρώνει μια μετάδοση, διαγράφει και την αντίστοιχη ακμή από τον γράφο. Έτσι, μεταδίδει το συναίσθημα όλων των λέξεων του κύκλου σε έναν κόμβο μόνο, χωρίς να επηρεάζονται οι επόμενοι.

V Αξιολόγηση

Η μέθοδος αξιολογήθηκε στα δύο δείγματα που παρασχέθηκαν από το SemEval του 2014 για το Task 4: Aspect Based Sentiment Analysis[2] (δλδ. Συναισθηματική ανάλυση βάσει χαρακτηριστικού) (Pontiki et al, 2014). Τα ένα δείγμα αφορά προτάσεις που προέρχονται από κριτικές εστιατορίων και το άλλο αφορά προτάσεις που προέρχονται από κριτικές φορητών υπολογιστών. Ωστόσο, αυτά τα δείγματα περιέχουν μερικές ιδιαιτερότητες που δε λήφθηκαν υπ' όψιν κατά τη δημιουργία του προγράμματος.

5.1 Ιδιαιτερότητες των δειγμάτων αξιολόγησης

Πρώτον, εκτός από τις λέξεις της πρότασης που χαρακτηρίζονται, τις οποίες ονομάζουν χαρακτηριστικούς όρους (aspect terms), υπάρχει πληροφορία και για την κατηγορία του αντικειμένου που χαρακτηρίζεται. Για παράδειγμα, σε μια πρόταση που η λέξη «σαλάτα» φορτίζεται θετικά, αναφέρεται στο δείγμα η λέξη «σαλάτα» ως όρος της πρότασης με θετικό συναίσθημα και η κατηγορία «φαγητό» η οποία επίσης έχει θετικό συναίσθημα. Στη μέθοδο της παρούσας εργασίας, δε γίνεται περαιτέρω ομαδοποίηση των όρων σε κατηγορίες. Επομένως δε θα μπορούσε να γίνει αξιολόγηση βάσει κατηγορίας.

Δεύτερον, σε κάποιες προτάσεις δεν έχουν κανένα χαρακτηριστικό όρο, αλλά έχουν κατηγορία που χαρακτηρίζεται. Δηλαδή, ενώ δεν υπάρχει κανένας όρος της πρότασης που χαρακτηρίζεται, υπάρχει φόρτιση και η κατηγορία στην οποία κατευθύνεται εξάγεται από τα συμφραζόμενα. Αφού η παρούσα μέθοδος δε βρίσκει κατηγορίες, παραλήφθηκαν από την αξιολόγηση όλες οι προτάσεις που δεν περιείχαν κανέναν χαρακτηριστικό όρο.

Τρίτον, στα δείγματα ποτέ δε θεωρούνται ως χαρακτηριστικοί όροι οι γενικοί όροι ή οι αντωνυμίες. Για παράδειγμα, στην κριτική «Το εστιατόριο ήταν εξαιρετικό.» η λέξη «εστιατόριο» δε θεωρείται χαρακτηριστικό. Παράλληλα, μέρη του λόγου εκτός των ουσιαστικών, όπως ρήματα, σε κάποιες περιπτώσεις θεωρούνται χαρακτηριστικοί όροι. Για παράδειγμα, στην πρόταση «Μάς σέρβιραν με επιδεξιότητα.» η λέξη «σέρβιραν» θεωρείται χαρακτηριστικός όρος. Η παρούσα προσέγγιση προσπαθεί να εντοπίσει μόνο ουσιαστικά τα οποία δέχονται συναισθηματική φόρτιση και συμπεριλαμβάνονται σε αυτή την ανάλυση και οι αντωνυμίες και οι γενικοί όροι. Πρόκειται, δηλαδή, για διαφορετική προσέγγιση του τι θα πρέπει να θεωρηθεί ότι είναι χαρακτηριστικό. Πάρα ταύτα, όλες οι προτάσεις των δειγμάτων που έχουν χαρακτηριστικούς όρους, ακόμα κι αυτές που χαρακτηρίζουν διαφορετικές λέξεις ως

[2] http://alt.qcri.org/semeval2014/task4/

χαρακτηριστικά από αυτές που λήφθηκαν υπ' όψιν κατά τη δημιουργία του προγράμματος, συμπεριλήφθηκαν στην αξιολόγηση.

Τέλος, τα δείγματα περιείχαν τέσσερεις διαφορετικές κατηγοριοποιήσεις για τους όρους: θετικός, αρνητικός, ουδέτερος και σύγκρουση (όταν ο όρος λαμβάνει και θετική και αρνητική φόρτιση). Ωστόσο, ο αλγόριθμος φτιάχτηκε για να εντοπίζει μόνο θετικά ή αρνητικά φορτισμένα χαρακτηριστικά. Κάποια χαρακτηριστικά που δέχθηκαν σχεδόν ίση θετική με αρνητική φόρτιση (|θετική-αρνητική| < 0.06) θεωρήθηκαν ουδέτερα. Ωστόσο, το πιο πιθανό είναι ότι τα περισσότερα ουδέτερα ουσιαστικά δεν εντοπίστηκαν καθόλου, αφού δεν έλαβαν ούτε θετική ούτε αρνητική φόρτιση. Όσον αφορά την περίπτωση της σύγκρουσης, αυτή δε λήφθηκε καθόλου υπ' όψιν στην παρούσα προσέγγιση.

5.2 Τα αποτελέσματα

Υπάρχουν δύο ξεχωριστά στοιχεία που αξιολογούνται: η σωστή εύρεση χαρακτηριστικών και η ορθή πρόβλεψη του συναισθήματος, όταν βρίσκονται τα σωστά χαρακτηριστικά.

5.2.1 Εύρεση χαρακτηριστικού

Όπως φαίνεται και στον διπλανό πίνακα, από τα 3657 χαρακτηριστικά του SemEval βρέθηκαν τα 2424, ενώ βρέθηκαν και 2168 επιπλέουσες λέξεις που δεν ήταν χαρακτηριστικά του SemEval. Το ποσό των αληθώς ψευδών (true negative) λέξεων υπολογίστηκε από το δεδομένο ότι το σύνολο των λέξεων του δείγματος είναι 31002 και αναφέρεται ενδεικτικά.

Εργασία \ Δείγμα	Χαρακτηριστικό	Μη χαρακτηριστικό
Χαρακτηριστικό	2424	2169
Μη χαρακτηριστικό	1233	(25176)

Πίνακας σύγχυσης για εστιατόρια.

Για την αξιολόγηση χρησιμοποιούνται οι τύποι:

1. $precision = \frac{tp}{tp+fp}$, δηλαδή $ακρίβεια = \frac{αληθώς\ θετικά}{αληθώς\ θετικά + ψευδώς\ θετικά}$· και
2. $recall = \frac{tp}{tp+fn}$, δηλαδή $ανάκληση = \frac{αληθώς\ θετικά}{αληθώς\ θετικά + ψευδώς\ αρνητικά}$

Επομένως, η ακρίβεια πρόβλεψης χαρακτηριστικού για το δείγμα των εστιατορίων είναι 52,8%, ενώ η ανάκληση είναι 66,3%. Αντίστοιχα, για τις κριτικές φορητών υπολογιστών η ακρίβεια πρόβλεψης είναι 35,6% και η ανάκληση 53,9%.

Γενικά, εξάγεται το συμπέρασμα ότι η ακρίβεια πρόβλεψης και ανάκλησης χαρακτηριστικών δεν είναι αρκετά καλή. Η μεγάλη μείωση και των δύο παραμέτρων αξιολόγησης στο δείγμα με τις κριτικές φορητών υπολογιστών υποδεικνύει είτε ότι η μέθοδος δε λειτουργεί τόσο καλά σε όλους τους τομείς κριτικών (επειδή, για παράδειγμα, σε κάποιους χρησιμοποιούνται πιο συχνά ρήματα ως χαρακτηριστικά), είτε ότι οι γραφείς χρησιμοποιούν διαφορετικό είδος γλώσσας και δε μπορεί να γίνει τόσο καλή συντακτική ανάλυση.

Εργασία \ Δείγμα	Χαρακτηριστικό	Μη χαρακτηριστικό
Χαρακτηριστικό	1242	2248
Μη χαρακτηριστικό	1064	(21714)

Πίνακας σύγχυσης για φορητούς υπολογιστές.

5.2.2 Εύρεση συναισθήματος

Η εύρεση συναισθήματος αξιολογείται ως προς την επιτυχία εύρεσης ότι τα χαρακτηριστικά, που βρέθηκαν από τον αλγόριθμο, είναι θετικά, αρνητικά ή ουδέτερά. Όσον αφορά τη σύγχυση, απλά αναγράφεται ως αποτυχία εύρεσης του σωστού τύπου χωρίς να εξετάζονται στατιστικά στοιχεία για το αν προβλέφθηκε ως θετικό, αρνητικό ή ουδέτερο. Εξ άλλου, υπάρχουν μόνο 57 χαρακτηριστικά που βρέθηκαν και χαρακτηρίζονται ως σύγχυση στο δείγμα των εστιατορίων και 15 στο δείγμα των φορητών υπολογιστών.

Εργασία \ Δείγμα	Θετικό	Αρνητικό	Ουδέτερο
Θετικό	1256	216	140
Αρνητικό	128	205	45
Ουδέτερο	38	25	21

Εργασία \ Δείγμα	Θετικό	Αρνητικό	Ουδέτερο
Θετικό	438	112	56
Αρνητικό	36	202	23
Ουδέτερο	19	18	13

Από τον πίνακα σύγχυσης των εστιατορίων μπορεί να υπολογιστεί, με χρήση των προαναφερθέντων τύπων, η ακρίβεια και η ανάκληση για κάθε χαρακτηριστικό. Συγκεκριμένα, η ακρίβεια πρόβλεψης ενός θετικού χαρακτηριστικού είναι 77,9%, ενός αρνητικού 54% και ενός ουδέτερου 25%, ενώ η ακρίβεια ανάκλησης είναι 88,3%, 46% και 10,2% αντίστοιχα. Φαίνεται εύκολα η δυσκολία του αλγορίθμου να εντοπίσει ουδέτερα χαρακτηριστικά, η οποία είναι αναμενόμενη, δεδομένης της απώλειας μέριμνας για τα ουδέτερα χαρακτηριστικά κατά τη δημιουργία του αλγορίθμου. Φαίνεται επίσης ότι λίγα από όσα κατηγοριοποιεί ως ουδέτερα είναι πράγματι ουδέτερα, το οποίο οφείλεται στον ίδιο λόγο· δηλαδή στο γεγονός ότι για να βρεθεί

ένα χαρακτηριστικό θα πρέπει να έχει μια, έστω και μικρή, συναισθηματική φόρτιση (γεγονός που δε συμβαδίζει με τη λογική των ουδέτερων χαρακτηριστικών, τα οποία θα πρέπει να μην είναι καθόλου φορτισμένα). Ωστόσο, η ολική ακρίβεια του αλγορίθμου είναι ικανοποιητική αφού ανέρχεται στο 71,46%. Επιπλέον, η μετρική Κάππα είναι 0,327.

Σημειώνεται ότι η μετρική Κάππα υπολογίζεται ως:

$kappa = \frac{total\ accuracy - random\ accuracy}{1 - random\ accuracy}$, δηλαδή $καππα = \frac{ολική\ ακρίβεια - τυχαία\ ακρίβεια}{1 - τυχαία\ ακρίβεια}$, όπου:

$$random\ accuracy = \frac{(TP+FNegP+FneuP)*(TP+FP)+(TNeg+FPNEG+FNeuNEG)*(TNeg+FNeg)+(TNeu+FPNEU+FNegNEU)*(TNeu+FNeu)}{total*total}$$

, όπου:

- TP: αληθώς θετικά, TNeg: αληθώς αρνητικά, TNeu: αληθώς ουδέτερά, δηλαδή χαρακτηριστικά που προβλέφθηκαν σωστά·
- FNegP: ψευδώς αρνητικό στη θέση θετικού, δηλαδή χαρακτηριστικά που προβλέφθηκαν ως αρνητικά ενώ ήταν θετικά – αντίστοιχα για FneuP, FPNeg, FNeuNeg, FPNeu, FNegNeu· και
- FP: ψευδώς θετικά, δηλαδή χαρακτηριστικά που προβλέφθηκαν ως θετικά ενώ ήταν αρνητικά ή ουδέτερα – αντίστοιχα FNeg και Fneu.

Στη πραγματικότητα, η μετρική Κάππα συγκρίνει την ακρίβεια, ως προς όλες τις κλάσεις, του συστήματος με αυτή ενός τυχαίου συστήματος. Όσο πιο μεγάλο είναι, τόσο πιο καλά θεωρείται ότι αποδίδει το σύστημα και η μέγιστη τιμή που μπορεί να πάρει είναι το 1.

Όσον αφορά το δείγμα των φορητών υπολογιστών, αυτό σημείωσε ακρίβεια 72,3% για τα θετικά χαρακτηριστικά, 77,4% για τα αρνητικά και 26% για τα ουδέτερα. Αντίστοιχα, η ανάκληση είναι 88,8%, 60,8% και 14,1%, ενώ η συνολική ακρίβεια 71,21%. Παρατηρείται ότι ενώ η ακρίβεια για τα θετικά είναι λίγο χαμηλότερη, η ακρίβεια για τα αρνητικά είναι πολύ υψηλότερη. Επιπλέον, η ανάκληση είναι εμφανώς αυξημένη για τα αρνητικά χαρακτηριστικά. Έτσι, ενώ η συνολική ακρίβεια είναι σχεδόν ίση με αυτή των κριτικών των εστιατορίων, το Κάππα παρουσιάζει αύξηση και φτάνει το 0,463. Αυτό οφείλεται στο γεγονός ότι υπάρχει μεγαλύτερη αρμονία στην επιτυχία πρόβλεψης και ανάκλησης μεταξύ των κατηγοριών θετικό και αρνητικό.

Από αυτήν την ανάλυση εξάγεται το συμπέρασμα ότι ο παρών αλγόριθμος μπορεί να εντοπίσει θετικά φορτισμένα χαρακτηριστικά με καλή ακρίβεια και ανάκληση, αρνητικά φορτισμένα χαρακτηριστικά με μέτρια ακρίβεια, ενώ δεν είναι επιτυχής στην εύρεση μη συναισθηματικά

φορτισμένων χαρακτηριστικών. Συνολικά, η ακρίβειά του είναι ικανοποιητική αφού ξεπερνά το 71%.

Δεν μπορούν να εξαχθούν συμπεράσματα σε σχέση με το κατά πόσο το είδος κειμένου και ο τομέας στον οποίο ανήκει επηρεάζουν την ακρίβεια του αλγορίθμου, καθώς εξετάστηκε μόνο για κριτικές και μόνο σε δύο τομείς. Ωστόσο, υπάρχουν ενδείξεις ότι η εύρεση των αρνητικών χαρακτηριστικών εξαρτάται σε κάποιον βαθμό από τον τομέα του κειμένου.

VI Επίλογος

Συμπεράσματα

Για την παρούσα πτυχιακή εργασία αναπτύχθηκε ένας αλγόριθμος που σκοπό έχει να πραγματοποιήσει στο ίδιο στάδιο εύρεση χαρακτηριστικών και συναισθηματική ανάλυση. Ο πυρήνας του αλγορίθμου ήταν η αξιοποίηση της συντακτικής πληροφορίας, και συγκεκριμένα των γραμματικών εξαρτήσεων ανάμεσα στις λέξεις μιας πρότασης.

Χρησιμοποιήθηκαν έτοιμα εργαλεία για την πραγματοποίηση της συντακτικής ανάλυσης, της αποσαφήνισης της έννοιας των λέξεων και την ανάκτηση του σημασιολογικού προσανατολισμού των εννοιών. Ωστόσο, αναπτύχθηκε πρωτότυπος κώδικας για τη μετάδοση του συναισθήματος, η οποία πραγματοποιήθηκε με αναπαράσταση κάθε πρότασης ως γράφου.

Πραγματοποιήθηκε αξιολόγηση του αλγόριθμου και για τις δύο εργασίες που σκόπευε να πραγματοποιήσει. Ο αλγόριθμος παρουσιάζει αδυναμία στην εύρεση των χαρακτηριστικών, αλλά είναι αρκετά επιτυχής στην ταξινόμησή τους (ειδικά όσων είναι θετικά ή αρνητικά).

Μελλοντικές Επεκτάσεις

Ο αλγόριθμος αυτός θα μπορούσε να επωφεληθεί από έναν αυτόνομο τρόπο εύρεσης χαρακτηριστικών. Θα μπορούσαν πρώτα να εντοπίζονται τα χαρακτηριστικά κι έπειτα να ανακτάται η συναισθηματική τους φόρτιση από τον παρών αλγόριθμο. Επιπλέον, θα μπορούσε να αξιοποιείται η σχετική πληροφορία για να πραγματοποιηθεί αλλαγή της κατεύθυνσης κάποιων ακμών του γράφου.

Τέλος, στην παρούσα εργασία αξιοποιούνται μόνο οι γραμματικές αρνήσεις. Μία ακόμη επέκταση που θα μπορούσε να γίνει είναι να χρησιμοποιηθούν λεξικά μετατροπέων πολικότητας έτσι ώστε να εντοπίζονται περισσότερες αρνήσεις, αλλά και ενισχύσεις και εξασθενήσεις, και να πραγματοποιείται ειδική μεταχείριση για αυτές τις λέξεις.

Βιβλιογραφία

Banerjee, S., και Pedersen, T. 2003, Αύγουστος. Extended gloss overlaps as a measure of semantic relatedness. *IJCAI*, Τόμος 3, 805-810.

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., & Reynar, J. 2008, Απρίλιος. Building a sentiment summarizer for local service reviews. *WWW Workshop on NLP in the Information Explosion Era*, 14.

Carenini, G. Ng, R.T και Zwart, E. 2005. Extracting knowledge from evaluative text. *Proc. Third International Conference on Knowledge Capture*.

Choi, Y. και Claire, C. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 793–801, Honolulu, HI.

Collins, A., και Loftus, E. 1975. A spreading activation theory of semantic processing. *Psychological Review*, 82: 407-428.

Galley, M. και McKeown, K. 2003, Αύγουστος. Improving word sense disambiguation in lexical chaining. IJCAI, Τόμος 3, 1486-1488.

Gamon, M. Aue, A. Corston-Oliver, S. και Ringger, E. 2005. Pulse: Mining customer opinions from free text. *Proceedings of the 6th International Symposium on Intelligent Data Analysis (IDA)*.

Κατάκης, I. M. 2014. Ανάπτυξη online υπηρεσίας αποσαφήνισης λέξεων στο πλαίσιο προτάσεων. *ΕΣΤΙΑ – Ψηφιακό Αποθετήριο Βιβλιοθήκης και Κέντρου Πληροφόρησης Χαροκοπείου Πανεπιστημίου*.

Katakis, I. Varlamis, I., Tsatsaronis, G., "PYTHIA: Employing Lexical and Semantic Features for Sentiment Analysis", Demo paper presented in the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2014), Nancy, France, September 15-19, 2014.

Kennedy, A. και Inkpen, D. 2006. Sentiment classification of movie and product reviews using contextual valence shifters. *Computational Intelligence*, 22(2): 110–125.

Lesk, M. 1986, Aytomatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of SIGDOG Conference 1986*.

Lin, D. 1994. Principar–an efficient, broad-coverage, principle-based parser. *Proc. COLING '94*, 482–488.

Lin, D. 1998. Automatic retrieval and clustering of similar words. *Proc. COLING-ACL '98*, 768–773.

Pontiki, M., Papageorgiou, H., Galanis, D., Androutsopoulos, I., Pavlopoulos, J., & Manandhar, S. (2014, August). Semeval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014) (pp. 27-35).

Pontiki, M., Galanis, D., Papageogiou, H., Manandhar, S., & Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. In*Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado*.

Quillian, M. R. 1962. A revised design for an understanding machine. *Mechanical Translation*, 7: 17-29.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., και Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2): 267-307.

Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th annual meeting on association for computational linguistics*, 417-424, Association for Computational Linguistics.

Turney, P. D. και Michael L. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4): 315–346.

Wiebe, J. 2000. Learning subjective adjectives from corpora. *Proceedings of 17th National Conference on Artificial Intelligence (AAAI)*, 735–740, Austin, TX.

Χατζηβασιλόγλου, Β., και McKeown, K.R. 1997. Predicting the semantic orientation of adjectives. *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL*, 174-181, New Brunswick, NJ: ACL.

Zhuang, L. Jing, F. και Zhu, X. 2006. Movie review mining and summarization. *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*

Παράρτημα

Οι εξαρτήσεις του συντακτικού αναλυτή του πανεπιστημίου Στάνφορντ

- acomp: adjectival complement (επιθετικό συμπλήρωμα)
- advcl: adverbial clause modifier (επιρρηματικός προσδιορισμός πρότασης)
- advmod: adverb modifier (προσδιορισμός επιρρήματος) π.χ.: advmod(often,less)
- agent: agent (ποιητικό αίτιο)
- amod: adjectival modifier (επιθετικός προσδιορισμός)
- appos: appositional modifier (παρενθετικός προσδιορισμός)
- aux: auxiliary (βοηθητικό ρήμα)
- auxpass: passive auxiliary (βοηθητικό ρήμα σε παθητική φωνή)
- cc: coordination (σύζευξη)
- ccomp: clausal complement (προτασιακό συμπλήρωμα)
- conj: conjunct (ένωση/συσχέτιση)
- cop: copula (συνδετικό ρήμα)
- csubj: clausal subject (πρόταση-υποκείμενο)
- csubjpass: clausal passive subject (πρόταση-υποκείμενο σε παθητική φωνή)
- dep: dependent (εξάρτηση) – όταν δε μπορεί να βρεθεί πιο συγκεκριμένη συσχέτιση από τον συντακτικό αναλυτή
- det: determiner (άρθρο συνήθως)
- discourse: discourse element (επιφώνημα κ.λπ.)
- dobj: direct object (άμεσο αντικείμενο)
- expl: expletive (το υπαρξιακό «there») – π.χ.: There is a ghost in the room. expl(is,There)
- goeswith: goes with (ενώνει λέξεις που γράφονται ως δύο ενώ κανονικά είναι μία)
- iobj: indirect object (έμμεσο αντικείμενο)
- mark: marker (λέξη που εισάγει μια πρόταση μέσα σε μια άλλη πρόταση)
- mwe: multi-word expression (έκφραση με πολλές λέξεις)
- neg: negation modifier (αρνητικός προσδιορισμός)
- nn: noun compound modifier (ουσιαστικό ως προσδιορισμός)
- npadvmod: noun phrase as adverbial modifier (ονοματική πρόταση ως επιρρηματικός προσδιορισμός)
- nsubj: nominal subject (ονοματική πρόταση ή ουσιαστικό ως υποκείμενο)
- nsubjpass: passive nominal subject (ονοματική πρόταση ή ουσιαστικό ως υποκείμενο σε πρόταση παθητικής φωνής)

- num: numeric modifier (αριθμητικός προσδιορισμός)
- number: element of compound number (μέρος σύνθετου αριθμού) – π.χ.: number(thousand,four)
- parataxis: parataxis (παράταξη) – Ένωση μεταξύ δύο (κύριων) προτάσεων
- pcomp: prepositional complement (προθεσιακό συμπλήρωμα)
- pobj: object of a preposition (αντικείμενο πρόθεσης)
- poss: possession modifier (κτητική αντωνυμία)
- possessive: possessive modifier (κτητικός προσδιορισμός)
- preconj: preconjunct – Σχέση μεταξύ της ονοματικής φράσης και της εμφατικής λέξης π.χ.: Both the boys and the girls are here. preconj(boys,both)
- predet: predeterminer – π.χ.: All the boys are here. predet(boys,all)
- prep: prepositional modifier (προσδιορισμός που εισάγεται με πρόθεση)
- prepc: prepositional clausal modifier (πρόταση-προσδιορισμός που εισάγεται με πρόθεση)
- prt: phrasal verb particle (συνήθως για να ενώσει phrasal verb)
- punct: punctuation (θαυμαστικά κ.λπ.)
- quantmod: quantifier phrase modifier (αριθμός ως προσδιορισμός)
- rcmod: relative clause modifier – π.χ.: I saw the man you love. rcmod(man,love)
- ref: referent (αναφορά)
- root: root (μια εξάρτηση που βάζει ο αναλυτής για να δείξει τη ρίζα της πρότασης)
- tmod: temporal modifier (χρονικός προσδιορισμός)
- vmod: reduced non-finite verbal modifier
- xcomp: open clausal complement (συνήθως δείχνει ένα ρήμα που δεν έχει υποκείμενο) – π.χ.: He says that you like to swim. xcomp(like,swim)
- xsubj: controlling subject (συνδέεται με τη σχέση xcomp, αφού αφορά του υποκείμενο που εννοείται για αυτό το ρήμα) – π.χ. στο προηγούμενο παράδειγμα θα είχαμε xsubj(swim,you).