ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανάλυση Συναισθήματος σε Δεδομένα Κοινωνικών Δικτύων με τον Αλγόριθμο Διάδοσης Ετικέτας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εμμανουήλ Δ. Μηλάκης

Επιβλέπουσα : Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Αθήνα, Ιούλιος 2016

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανάλυση Συναισθήματος σε Δεδομένα Κοινωνικών Δικτύων με τον Αλγόριθμο Διάδοσης Ετικέτας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εμμανουήλ Δ. Μηλάκης

Επιβλέπουσα : Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 26η Ιουλίου 2016.

.....	.....	.....
Θεοδώρα Βαρβαρίγου	Δημήτριος Ασκούνης	Βασίλειος Λούμος
Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2016

……....
Εμμανουήλ Δ. Μηλάκης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Το τεράστιο πλήθος των δεδομένων που προέρχονται από τα Κοινωνικά Δίκτυα θεωρείται μείζονος αξίας στο μάρκετινγκ και τα πλαίσια χάραξης πολιτικών και λήψης αποφάσεων. Ο κύριος στόχος τους είναι η επεξεργασία των πληροφοριών, καθώς και του κειμένου, που παρέχονται από άτομα και κοινότητες. Οι μηχανισμοί της Ανάλυσης Συναισθήματος επιτρέπουν τον εντοπισμό των θεμάτων, την ανάλυση των συναισθημάτων και την οπτικοποίηση των κοινωνικών μέσων μαζικής ενημέρωσης, έτσι ώστε να διερευνηθούν αλληλεπιδραστικά οι πτυχές των πληροφοριών.

Σε αυτά τα πλαίσια, η παρούσα εργασία μελετά την τεχνική της Διάδοσης Ετικέτας, μίας ημι-επιβλεπόμενης μεθόδου ανάλυσης, που επικεντρώνεται στην ανάδειξη της λεκτικής πολικότητας, σε πολυγνωσιακά και πολύγλωσσα κειμενικά περιβάλλοντα. Παρουσιάζει τη χρηστικότητα της αξιοποίησης των Λεξικών Συναισθήματος και τη σχετική ερευνητική βιβλιογραφία του τομέα. Εξετάζει τις εφαρμογές του αλγόριθμου Διάδοσης Ετικέτας στη δόμησή λεξικών και τις τροποποιήσεις των παραμέτρων του, προτείνοντας τις καταλληλότερες εκδοχές του, στην ανάλυση των δεδομένων από κοινωνικά δίκτυα, όπως αυτές προκύπτουν από την συγκριτική απεικόνιση των χαρακτηριστικών τους.

Λέξεις Κλειδιά

ανάλυση συναισθήματος, λεξικά συναισθήματος, Διάδοση Ετικέτας, αλγόριθμοι ημι-επιβλεπόμενης μάθησης, κατηγοριοποίηση πολικότητας, κοινωνικά δίκτυα

Abstract

The huge amount of data, which accrue from social media, are considered to be the cornerstone of marketing, strategies and decision making. Their main goal is the processing of data and texts, which are provided by individuals and communities. The Sentiment Analysis mechanisms allow the detection of issues, the analysis of sentiments and the visualization of the social media, so that the aspects of the data can be interactively examined.

In this context, this thesis deliberates the technique of the Label Propagation; a semi-supervised method of analysis, which focuses on promoting lexical polarity, to multi-topic and multi-lingual textual environments. This thesis illustrates usefulness of the utilization of Sentiment Lexicons and the according research bibliography of the field. Moreover, it examines the applications of the Label Propagation algorithm, which are used for the construction of lexicons and the modifications of parameter changes. Finally, this thesis suggests the most appropriate versions of the Label Propagation algorithm, for the analysis of data by the social media, as these come about from the comparative depiction of their characteristics.

Key words

sentiment analysis, sentiment lexicons, Label Propagation, semi-supervised learning algorithms, polarity classification, social networks

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο εργαστήριο Distributed Knowledge and Media Systems Group της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών και επισφραγίζει τις σπουδές μου στο Εθνικό Μετσόβιο Πολυτεχνείο.

Θα ήθελα να ευχαριστήσω ιδιαιτέρως την καθηγήτριά μου κα. Θεοδώρα Βαρβαρίγου για την εμπιστοσύνη που μου έδειξε και την ευκαιρία που μου παρείχε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα.

Επίσης, θα ήθελα να ευχαριστήσω τον Υποψήφιο Διδάκτορα Βρεττό Μουλό για το χρόνο που μου αφιέρωσε και την καθοδήγησή του, καθ' όλη τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Τέλος, ευχαριστώ την οικογένειά μου για τη στήριξη και βοήθεια κατά τη διάρκεια των σπουδών μου και προπάντων τον Τριαδικό Θεό, *Ὅς ἀγραμμάτους σοφίαν ἐδίδαξεν.*

Εμμανουήλ Δ. Μηλάκης

Περιεχόμενα

Περίληψη 5

Abstract 7

Ευχαριστίες 9

1. Εισαγωγή 19

1.1. Η έννοια της Ανάλυσης Συναισθήματος 19

1.2. Εφαρμογές της Ανάλυσης Συναισθήματος 20

1.3. Οργάνωση Κειμένου 21

2. Ανάλυση Συναισθήματος 23

2.1. Επίπεδα Ανάλυσης Συναισθήματος 23

2.2. Μεθοδολογικές Προσεγγίσεις 23

2.2.1. Χρήση Λεξικών 23

2.2.1.1. Ορισμός Λεξικού Συναισθήματος 23

2.2.1.2. Κατηγορίες Μεθόδων Λεξικών Συναισθήματος 24

2.2.2. Μηχανική Μάθηση 26

2.2.2.1. Ορισμός Μηχανικής Μάθησης 26

2.2.2.2. Προσεγγίσεις βασιζόμενες στη Μηχανική Μάθηση 27

2.2.2.3. Κατηγορίες Μεθόδων με βάση την είσοδο του συστήματος 27

2.2.2.4. Κατηγορίες Μεθόδων με βάση την έξοδο του συστήματος 28

2.2.3. Στατιστική Ανάλυση 30

2.2.4. Σύγκριση Προσεγγίσεων 31

2.3. Το Συναίσθημα στα Κοινωνικά Δίκτυα 33

2.3.1. Το φαινόμενο του Twitter 33

2.3.2. Ανάλυση Συναισθήματος στο Twitter 34

2.3.3. Χαρακτηριστικά Ανάλυσης των tweets 35

2.3.4. Ιδιαιτερότητες του Twitter 36

2.4. Μέθοδοι Αξιολόγησης Ανάλυσης Συναισθήματος 37
2.4.1. Διαβαθμολογική Αξιοπιστία 37
2.4.1.1. Συντελεστής Κάπα του Cohen 37
2.4.1.2. Συντελεστής Κάπα του Fleiss και Α του Robinson 39
2.4.1.3. Συντελεστής Ενδοσυσχέτισης 40
2.4.2. Στατιστικά Μέτρα Αξιολόγησης 41

3. Λεξικά Συναισθήματος 45
3.1. WordNet 45
3.2. Μέθοδοι Κατασκευής Λεξικών Συναισθήματος 47
3.3. Λεξικά Συναισθήματος με Διαβαθμίσεις 49
3.3.1. SentiWordNet 49
3.3.2. Bings Liu's Opinion Lexicon 50
3.3.3. ANEW 51
3.3.4. AFINN 51
3.4. Λεξικά Συναισθήματος χωρίς Διαβαθμίσεις 52
3.4.1. MPQA 52
3.4.2. Linguistic Inquiry and Word Count 53
3.4.3. General Inquirer 54
3.5. Λεξικά Συναισθήματος της Ελληνικής Γλώσσας 56

4. Μέθοδοι βασιζόμενες σε Λεξικά Συναισθήματος 58
4.1. Προσεγγίσεις βασιζόμενες σε Ερμηνευτικό Λεξικό 58
4.1.1. Μέθοδοι μη Επιβλεπόμενης Μάθησης 58
4.1.2. Μέθοδοι Επιβλεπόμενης Μάθησης 60
4.1.3. Bootstrapping Μέθοδοι 61
4.1.4. Στατιστικές Μέθοδοι 62
4.2. Προσεγγίσεις βασιζόμενες σε Ηλεκτρονικά Σώματα Κειμένων 63
4.2.1. Μέθοδοι Κοινού Γνωσιακού Τομέα 63
4.2.2. Μέθοδοι Διαφορετικού Γνωσιακού Τομέα 66
4.3. Μικτές Προσεγγίσεις 67

5. Διάδοση Ετικέτας 68
5.1. Προσημασμένα Δεδομένα 68
5.2. Κ-Πλησιέστεροι Γείτονες 68

5.3. Αλγόριθμος Διάδοσης Ετικέτας - LP 69
5.3.1. Διατύπωση του Προβλήματος 70
5.3.2. Δομή Αλγορίθμου 72
5.3.3. Αλγόριθμος LP/1-NN 74
5.4. Μέθοδοι Παραμετροποίησης 74
5.4.1. Ευριστική Μέθοδος 74
5.4.2. Μέθοδος Εντροπίας 75
5.4.3. Σύγκριση Μεθόδων Παραμετροποίησης 77
5.5. Εξισορρόπηση Κατανομών στις Κατηγορίες 77
5.5.1. ML-Μέθοδος 77
5.5.2. Μέθοδοι Μετεπεξεργασίας 77
5.6. Σύγκριση με Αλγορίθμους Γράφων 78
5.6.1. Αλγόριθμος Τυχαίων Περιπάτων 78
5.6.2. Παλινδρόμηση Kernel 79
5.6.3. Προσέγγιση Μέσου Πεδίου 79
5.6.4. Αλγόριθμος Ελάχιστης Τομής 80
6. Διάδοση Ετικέτας στην Ανάλυση Συναισθήματος 81
6.1. Αλγόριθμος Διάδοσης Ετικέτας Συνδυαζόμενων Σχέσεων 83
6.1.1. Δομή Αλγορίθμου CR LP 84
6.1.2. Χαρακτηριστικά Μεθόδου CR LP 85
6.2. Αλγόριθμος Διάδοσης Ετικέτας βασιζόμενος σε Χαρακτηριστικά 87
6.2.1. Δομή Αλγορίθμου AB LP 88
6.2.2. Χαρακτηριστικά Μεθόδου AB LP 88
6.3. Αλγόριθμος Διάδοσης Ετικέτας προσανατολισμένος σε Χαρακτηριστικά 91
6.3.1. Δομή Αλγορίθμου AO LP 91
6.3.2. Χαρακτηριστικά Μεθόδου AO LP 95
6.4. Αλγόριθμος Διάδοσης Γράφου 97
6.4.1. Δομή Αλγορίθμου GP 97
6.4.2. Χαρακτηριστικά Μεθόδου GP 100
6.5. Αλγόριθμος Διάδοσης Ετικέτας Τροποποιημένης Προσρόφησης 101
6.5.1. Δομή Αλγορίθμου MAD LP 102
6.5.2. Χαρακτηριστικά Μεθόδου MAD LP 106

6.6. Σύγκριση Αλγορίθμων Διάδοσης Ετικέτας 107
6.6.1. Γλωσσική Ανεξαρτησία 107
6.6.2. Ανεξαρτησία από Λεξικολογικές Πηγές 109
6.6.3. Ανεξαρτησία από το Γνωσιακό Τομέα 111
6.6.4. Ανάλυση Κοινωνικών Δικτύων 113
6.6.5. Συνολική Σύγκριση Αλγορίθμων 117
7. Σύνοψη και Συμπεράσματα 119
Βιβλιογραφία 121

Κατάλογος Σχημάτων

2.1: Δομή XML εξόδου POS Parser 24
2.2: Διακρίσεις Ανάλυσης Συναισθήματος με Χρήση Λεξικών 25
2.3: Διακρίσεις Ανάλυσης Συναισθήματος με Μηχανική Μάθηση 30
2.4: Ετήσια Αύξηση του πλήθους των tweets 33
2.5: Παραδείγματα emoticon θετικής και αρνητικής πολικότητας 35
3.1: Παράδειγμα αποτελέσματος αναζήτησης στο WordNet 46
3.2: Διαδικασία κατασκευής Λεξικού Συναισθήματος 48
3.3: Παράδειγμα δομής SentiWordNet 50
3.4: Παράδειγμα δομής ANEW 51
3.5: Παράδειγμα διαβάθμισης πολικότητας στο AFFIN 52
3.6: Χαρακτηριστικά ουδέτερης-πολικής κατάταξης στο MPQA 53
3.7: Παράδειγμα αποτελέσματος ανάλυσης κειμένου με το LIWC 54
3.8: Παράδειγμα δομής General Inquirer 55
3.9: Παράδειγμα δομής Greek Sentiment Lexicon 57
5.1: Σύγκριση μεθόδου σήμανσης K-NN και LP 70
5.2: Δομή Πιθανολογικής Μήτρας Μετάβασης 73
5.3: Στάδια διάδοσης ετικέτας στον LP/1-NN 74
5.4: Παράδειγμα ελάχιστης ακμής υπογράφων διαφορετικής ετικέτας 75
6.1: Σύγκριση LP και SVM σε διαφορετικά μεγέθη προσημασμένων δεδομένων 82
6.2: Σύγκριση απόδοσης LP και CR LP σε ουσιαστικά 86
6.3: Παράδειγμα Ταξινομητή Πολικότητας 89
6.4: Παράδειγμα διαδικασίας ταξινόμησης σε επίπεδο πρότασης 95
6.5: Παράδειγμα διαδικασίας εξαγωγής χαρακτηριστικών 96
6.6: Σύγκριση ακρίβειας LP και MAD LP σε δεδομένα κοινωνικών δικτύων 113
6.7: Σύγκριση Ακρίβειας/Ανάκλησης AO LP και GP σε θετικές κατηγορίες 116
6.8: Σύγκριση Ακρίβειας/Ανάκλησης AO LP και GP σε αρνητικές κατηγορίες 116

Κατάλογος Πινάκων

1.1: Παραδείγματα Εφαρμογών Ανάλυσης Συναισθήματος 20

2.1: Σύγκριση Μεθόδων Ανάλυσης Συναισθήματος 32

2.2: Εξάρτηση βαθμού συμφωνίας από το συντελεστή Cohen 38

2.3: Εξάρτηση βαθμού συμφωνίας από τον ICC 40

2.4: Γενικευμένη μορφή Μήτρας Σύγχυσης 41

2.5: Παράδειγμα δομής Δυαδικής Μήτρας Σύγχυσης 42

3.1: Διαμέριση synset του WORDNET στις ομάδες του MICRO-WNOP 49

6.1: Σύγκριση βημάτων διάδοσης ετικέτας LP και GP 99

6.2: Σύγκριση βημάτων διάδοσης ετικέτας LP και MAD LP 104

6.3: Εξάρτηση Σφάλματος από τον Ταξινομητή του MAD LP 107

6.4: Συγκριτική απεικόνιση Αλγορίθμων Διάδοσης Ετικέτας 120

1. Εισαγωγή

1.1 Η έννοια της Ανάλυσης Συναισθήματος

Η Γλωσσολογία και η Επεξεργασία Φυσικής Γλώσσας (NLP) έχει μια μακρά ιστορία, ωστόσο η έρευνα για τις απόψεις και τα συναισθήματα των ανθρώπων, υπήρξε ελάχιστη ως τα τέλη του 20^{ου} αιώνα. Από τότε, το επιστημονικό αυτό πεδίο έχει γίνει μια πολύ ενεργή περιοχή έρευνας, έχοντας ένα ευρύ φάσμα εφαρμογών, σχεδόν σε κάθε τομέα και προσφέροντας πολλές προκλήσεις σε ερευνητικά προβλήματα, τα οποία ποτέ δεν είχαν μελετηθεί πριν. Η λεπτομερής ανάλυση της ανθρώπινης γνώμης και του συναισθήματος που αυτή φέρει, όπως αυτή εκφράζεται μέσα από πολυμεσικές πηγές, στις μέρες μας αποτελεί έναν ιδιαίτερο κλάδο, που ονομάζεται Ανάλυση Συναισθήματος.

Πιο συγκεκριμένα, η Ανάλυση Συναισθήματος ορίζεται ως το επιστημονικό πεδίο μελέτης, που αναλύει ανθρώπινες απόψεις, συναισθήματα, αξιολογήσεις, εκτιμήσεις, τάσεις και στάσεις ως προς κάποια οντότητα, η οποία μπορεί να εκτείνεται, από ένα προϊόν ή μία υπηρεσία, έως ένα άτομο, μία ιδεολογική θέση ή κάποια από τα χαρακτηριστικά όλων των παραπάνω. Ο όρος της Ανάλυσης Συναισθήματος σκέπει ένα ευρύτατο σύνολο επιμέρους επιστημονικών αναζητήσεων, όπως είναι η εξόρυξη συναισθήματος, η κατηγοριοποίηση γνώσης, η σύνοψη δεδομένων και η ανάλυση υποκειμενικότητας. Ως στόχο έχει να καθορίσει τη στάση του ομιλητή ή του συγγραφέα, σε σχέση με κάποιο θέμα ή τη συνολική πολικότητα των συμφραζόμενων ενός εγγράφου. Η στάση αυτή μπορεί να είναι μία κρίση ή μία αξιολόγηση, μία συναισθηματική κατάσταση ή μία προβλεπόμενη συναισθηματική επικοινωνία, δηλαδή η συναισθηματική επίδραση, που ο συγγραφέας επιθυμεί να έχει στον αναγνώστη.

Η έννοια της Εξόρυξης Γνώμης συμπλέκεται επίσης, με αυτή της Ανάλυσης Συναισθήματος, ωστόσο αυτές οι έννοιες δεν είναι πάντα ισοδύναμες. Ο όρος «γνώμη» έχει ένα ευρύτατο νοηματικό περιεχόμενο, και για αυτό το λόγο, υπάρχει μερική αντιδιαστολή από τις κεντρικές επιδιώξεις της ανάλυσης συναισθήματος, που αφορούν κυρίως εκείνες τις γνώμες οι οποίες εκφράζουν ή υπονοούν, θετικά ή αρνητικά συναισθήματα. Πιο συγκεκριμένα, η εξόρυξη γνώμης ενίοτε επικεντρώνεται στην συνολική απεικόνιση της περιεχόμενης σε ένα κείμενο άποψης, ενώ η ανάλυση συναισθήματος έγκειται στην ανάδειξη των δομικών συστατικών του, που οδηγούν σε αυτήν. Συνεπώς, η ειδοποιός διαφορά βρίσκεται στο γεγονός ότι η ανάλυση συναισθήματος εστιάζει στην διάκριση και κατηγοριοποίηση του συναισθήματος για μία οντότητα, ενώ η εξόρυξη γνώμης, στην εις βάθος περιγραφή, της ίδιας και των χαρακτηριστικών της. [77]

1.2 Εφαρμογές της Ανάλυσης Συναισθήματος

Αν και η έρευνα στο πεδίο της Ανάλυσης Συναισθήματος ξεκίνησε κυρίως από τις αρχές του 2000, υπήρχαν κάποιες παλαιότερες εργασίες σχετικά με την ερμηνεία των μεταφορών, των επίθετων συναισθήματος, της υποκειμενικότητας, και των κειμενικών σημείων που εκφράζουν άποψη, καθώς και για την επιρροή που αυτά ασκούν. Η χρήση της επεξεργασίας φυσικής γλώσσας, η ανάλυση κειμένου και η υπολογιστική γλωσσολογία συνέβαλε στον εντοπισμό και την εξαγωγή υποκειμενικής πληροφορίας, όμως η συστηματική ανάπτυξη μεθόδων εντοπισμού και ανάλυσης συναισθήματος άνοιξε ένα νέο πλαίσιο έρευνας, με μια ποικιλία εφαρμογών.

Οι παραπάνω προσεγγίσεις, μπορούν να εφαρμοστούν σε διάφορους τομείς, όπως τον επιχειρηματικό, τον πολιτικό, των δημόσιων δράσεων και των χρηματοδοτήσεων. Η ανάλυση συναισθήματος, στον τομέα των επιχειρήσεων χρησιμοποιείται κυρίως για την βολιδοσκόπηση της φωνής του καταναλωτή, τη φήμη μίας μάρκας και την διαδικτυακή διαφήμιση και τάση του εμπορίου. Σε σχέση με τον πολιτικό τομέα, η παροχή συμβουλών δημοσκοπήσεων και ψήφου, αντιπροσωπεύει μια σημαντική εφαρμογή της ανάλυσης, που αξιοποιείται επίσης για να διευκρινίσει τις θέσεις των πολιτικών και τη βελτίωση της ποιότητας των πληροφοριών στις οποίες, που οι ψηφοφόροι έχουν πρόσβαση.

Βιομηχανία	Πολιτική	Κοινωνία
Φωνή Καταναλωτή	Συμβουλευτική Δημοσκοπήσεων	Παρακολούθηση Διοργανώσεων
Φήμη Μάρκας		Δημόσια Διαβούλευση
Διαδικτυακή Διαφήμιση	Ταξινόμηση Πολιτικών Τάσεων	Συμβουλευτική Επενδύσεων
Διαδικτυακό Εμπόριο		Ευφυή Συστήματα Μεταφορών

Πίνακας 1.1: Παραδείγματα Εφαρμογών Ανάλυσης Συναισθήματος

Η ανάλυση συναισθήματος χρησιμοποιείται παράλληλα, στο πλαίσιο της διαδικασίας εφαρμογής δημόσιων κοινωνικών δράσεων. Στο πλαίσιο αυτό, συμβάλει ουσιαστικά στην παρακολούθηση του πραγματικού κόσμου και των ανθρώπινων γεγονότων. Σημαντική ακόμη εφαρμογή της ανάλυσης συναισθήματος είναι η παρακολούθηση των απόψεων που οι άνθρωποι υποβάλλουν σχετικά με τις εκκρεμείς προτάσεις πολιτικών ή κυβερνητικών ρυθμίσεων και νομοσχεδίων. Ένας νέος αναδυόμενος τομέας ανάλυσης αντιπροσωπεύεται επίσης, από τα σύγχρονα Ευφυή Συστήματα Μεταφορών (Intelligent Transportation Systems), τον τομέα Χρηματοδοτήσεων και Επένδυσης, την ανίχνευση της τάσης των τιμών, της εξέλιξης εμπορευμάτων και μετοχών και των χρηματοοικονομικών κινδύνων [2].

1.3 Οργάνωση Κειμένου

Το υπόλοιπο κείμενο διαρθρώνεται ως εξής:

Κεφάλαιο 2: Περιγράφεται διεξοδικά η Ανάλυση Συναισθήματος ως προς τις μεθόδους της, εστιάζοντας στην ταξινόμηση των τεχνικών κάθε κατηγορίας, συγκρίνοντας παράλληλα τα αντίστοιχα χαρακτηριστικά τους.

Κεφάλαιο 3: Παρουσιάζονται αναλυτικά τα Λεξικά Συναισθήματος, οι μέθοδοι κατασκευής τους και οι διακρίσεις τους σύμφωνα με τις ιδιότητες τους, τόσο στην Αγγλική όσο και στην Ελληνική γλώσσα.

Κεφάλαιο 4: Πραγματοποιείται εκτενής ανασκόπηση της βιβλιογραφίας, που σχετίζεται με προσεγγίσεις βασιζόμενες σε Λεξικά Συναισθήματος, δίνοντας έμφαση στην κατηγοριοποίηση τους και στην ανάδειξη τις αλγοριθμικής ποικιλομορφίας τους.

Κεφάλαιο 5: Παρουσιάζεται σε βάθος ο αλγόριθμος Διάδοσης Ετικέτας, οι τεχνικές παραμετροποίησης του, η δομή και τα ιδιαίτερα γνωρίσματα του, σε συνάρτηση με αντίστοιχους αλγορίθμους ανάλυσης συναισθήματος με χρήση γράφων.

Κεφάλαιο 6: Περιγράφονται οι ιδιαιτερότητες της ανάλυσης δεδομένων στα κοινωνικά δίκτυα και οι κυριότερες τροποποιήσεις του αλγόριθμου διάδοσης ετικέτας, που τον καθιστούν κατάλληλο στην ανάλυση των δεδομένων τους.

Κεφάλαιο 7: Ανακεφαλαιώνονται τα βήματα της μελέτης και συνοψίζονται τα πορίσματά της, όσον αφορά την ανάλυση συναισθήματος σε δεδομένα κοινωνικών δικτύων με τη χρήση του αλγορίθμου Διάδοσης Ετικέτας, δίνοντας έμφαση στα συμπεράσματα και τις προοπτικές της παρούσας έρευνας.

2. Ανάλυση Συναισθήματος

2.1 Επίπεδα Ανάλυσης Συναισθήματος

Η Ανάλυση Συναισθήματος μπορεί να θεωρηθεί ως μια διαδικασία ταξινόμησης. Υπάρχουν τρία κύρια επίπεδα κατάταξης της. Το πρώτο από αυτά, αναφέρεται στο επίπεδο εγγράφου (document level), που στοχεύει να χαρακτηρίσει ένα έγγραφο που φέρει γνώμη, ως εκφραστή θετικού ή αρνητικού συναισθήματος, θεωρώντας ολόκληρο το έγγραφο ως μια βασική μονάδα πληροφορίας. Το δεύτερο, αναφέρεται στο επίπεδο πρότασης (sentence level), στοχεύοντας να κατατάξει το συναίσθημα που εκφράζεται σε κάθε πρόταση. Σε αυτή την προσέγγιση, το πρώτο βήμα είναι ο προσδιορισμός μίας πρότασης, ως υποκειμενική ή αντικειμενική. Αν όντως αποδειχθεί υποκειμενική, σε επίπεδο πρότασης, η περαιτέρω ανάλυση θα καθορίσει κατά πόσον η πρόταση εκφράζει θετική ή αρνητική γνώμη. Πολλές μελέτες έχουν επισημάνει βέβαια, ότι οι εκφράσεις, που είναι φορείς συναισθήματος δεν είναι κατ' ανάγκην υποκειμενικές στην φύση τους, αλλά αποκτούν κάποια χροιά στο περιβάλλον που παρατηρούνται.

Ωστόσο, δεν υπάρχει θεμελιώδης διαφορά ανάμεσα στην ταξινόμηση του επιπέδου εγγράφου και του επιπέδου πρότασης, δεδομένου ότι και οι προτάσεις μπορούν να ειδωθούν ως απλά και σύντομα έγγραφα. Οι κατηγοριοποιήσεις αυτές δεν παρέχουν την αναγκαία λεπτομερή απεικόνιση όλων των πτυχών, που απαιτούν γνωμοδοτήσεις μίας οντότητας. Σε πολλές εφαρμογές απαιτείται ένα τρίτο επίπεδο ανάλυσης, αυτό του επιπέδου χαρακτηριστικού (aspect level), που στοχεύει να κατατάξει το συναίσθημα ενός κειμένου, σε σχέση με τα συγκεκριμένα χαρακτηριστικά των οντοτήτων, που περιέχει. Αρχικά, εντοπίζονται οι οντότητες και τα χαρακτηριστικά τους, και ακολούθως οι φορείς γνώμης μπορούν να εκφράζουν διαφορετικά συναισθήματα για διαφορετικά χαρακτηριστικά, της ίδιας οντότητας.

2.2 Μεθοδολογικές Προσεγγίσεις

2.2.1 Χρήση Λεξικών

2.2.1.1 Ορισμός των Λεξικών Συναισθήματος

Η αξιοποίηση των λεξικών υπήρξε η βάση στην οποία στηρίχθηκε η ανάλυση συναισθήματος από τα πρώτα της βήματα, αφού παρείχε μία απλοϊκή αλλά αποτελεσματική πρακτική για τον διαχωρισμό υποκειμενικότητας των λέξεων (αντικειμενικές ή υποκειμενικές) και τη δυαδική ταξινόμηση τους (θετική ή αρνητική). Οι πρακτικές που βασίζονται σε τέτοια λεξικά, αντιμετωπίζουν το κείμενο το οποίο καλούνται να αναλύσουν ως μία λίστα λέξεων, που δεν υπόκεινται σε

οποιαδήποτε διάταξη, εξάρτηση ή γραμματολογική σύνδεση μεταξύ τους. Συνεπώς η κύρια στόχευσή τους έγκειται στο διαχωρισμό των λέξεων ανάμεσα σε αυτές που δεν μπορούν να αξιοποιηθούν στην ανάλυση και σε εκείνες που δηλώνουν συναίσθημα (sentiment words). Στα Λεξικά Συναισθήματος οι εισηγμένες λέξεις ταξινομούνται ανάλογα με το μέρος του λόγου στο οποίο ανήκουν και τους αποδίδονται βάρη σύμφωνα με το συναίσθημα που εκφράζουν, καθώς και την ένταση του.

Τα μέρη του λόγου στα οποία ανήκουν οι παραπάνω λέξεις παίζουν καθοριστικό ρόλο και για αυτό η κατηγοριοποίηση τους συνήθως πραγματοποιείται στο πρώτο στάδιο της επεξεργασίας του κειμένου από κατάλληλο λογισμικό (POS parsers). Τα επίθετα (adjectives) βρίσκονται στην κορυφή των POS παραγόντων για την ανάλυση συναισθήματος, αφού χαρακτηρίζουν τα ουσιαστικά και φέρουν το μεγαλύτερο συναισθηματικό φορτίο. Σε δεύτερη σημασιολογική προτεραιότητα παρουσιάζονται τα επιρρήματα (adverbs), που επηρεάζουν άμεσα τόσο τα επίθετα όσο και άλλα επιρρήματα. Περιορισμένη συναισθηματική έκταση έχουν τα ουσιαστικά (nouns) αφού κατέχουν θέση υποκειμένου ή αντικειμένου, και τις περισσότερες φορές δεν εκφράζουν από μόνα τους κάποιο συναίσθημα. Στην ίδια βάση, τα ρήματα (verbs) ως φορείς ενεργειών των υποκειμένων στα αντικείμενα, δεν μπορούν να χαρακτηρίσουν τις ίδιες τις ενέργειες και δεν λαμβάνουν σημαντικό ρόλο στην ανάλυση του κειμένου [40].

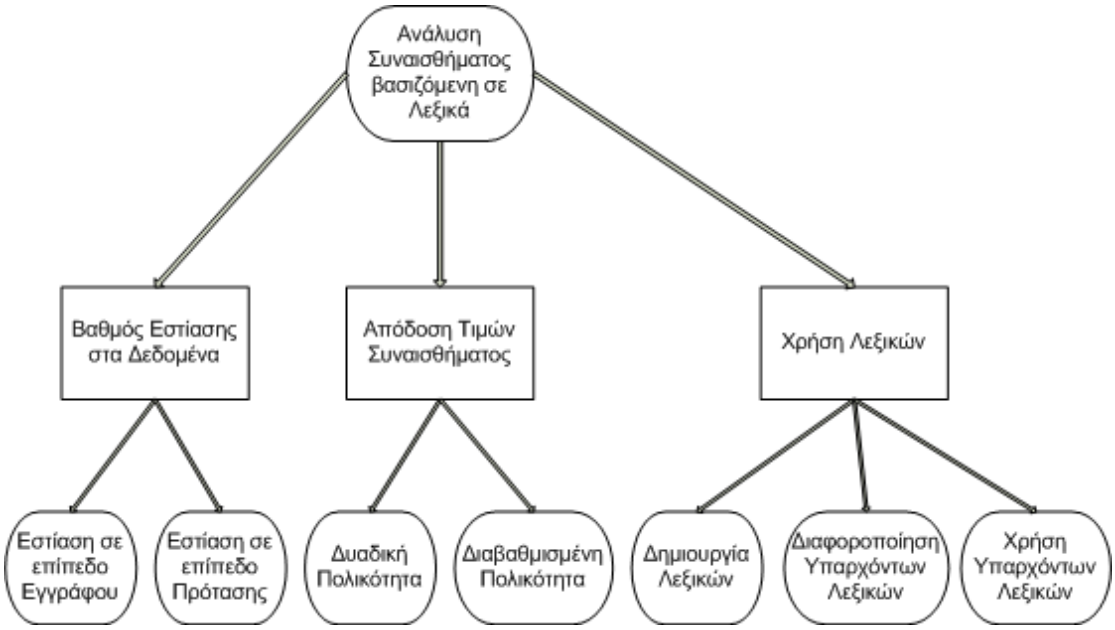
```
<S>  <NG><W C='PRP' L='SS' T='w' S='Y'> I </W></NG>

<VG> <W C='VBP'> am </W> <W C='RB'> absolutely </W></VG>

 <W C='IN'> in </W>  <NG><W C='NN'> awe </W> </NG>  <W C='IN'> of </W>

<NG> <W C='DT'> this </W>  <W C='NN'> camera</W></NG> <W C='.'> . </W> </S>
```

Σχήμα 2.1: Δομή XML εξόδου POS Parser

2.2.1.2 Κατηγορίες Μεθόδων βασιζόμενων στα Λεξικά Συναισθήματος

Καθοριστικός παράγοντας για την ανάλυση, εκτός από τις λέξεις που επιλέγονται, είναι και ο βαθμός εστίασης στο περιεχόμενο του κειμένου. Η χρήση των λεξικών μπορεί να λάβει καθολική εμβέλεια, και να χρησιμοποιηθεί για τον χαρακτηρισμό ενός εγγράφου ή ακόμα και ενός μεγαλύτερου συνόλου δεδομένων. Πρόκειται για ολιστικές προσεγγίσεις στις οποίες ο υπολογισμός της συναισθηματικής κλίσης του κειμένου, θεωρείται άμεσα εξαρτημένος από όλα τα επιμέρους στοιχεία του με αναλογικό τρόπο [34]. Από την άλλη πλευρά, η τμηματική εξαγωγή του συναισθήματος πριν τον συνολικό υπολογισμό του, αποτελεί μία εναλλακτική μέθοδο, η οποία καταρχάς υποδιαιρεί το σύνολο των δεδομένων σε υπο-μονάδες (παραγράφους, προτάσεις, φράσεις), επικεντρώνοντας έτσι την ανάλυση στα

κατάλληλα σημεία του κειμένου και αναζητώντας οποιαδήποτε σημασιολογική συσχέτιση των λέξεων που εκφράζουν συναίσθημα [24].



Σχήμα 2.2: Διακρίσεις Ανάλυσης Συναισθήματος με Χρήση Λεξικών

Στις περιπτώσεις που η κάθε λέξη αντιμετωπίζεται ως μία αυτοτελής συναισθηματική μονάδα, η συνολική εξαγωγή του συναισθήματος συνίσταται στον υπολογισμό των επιμέρους συναισθημάτων αθροιστικά, όσον αφορά τα δύο ήδη πολικότητας. Σε αυτή την τεχνική, δεν πραγματοποιείται εκπαίδευση του συστήματος αλλά χρησιμοποιούνται ήδη υπάρχοντες λεξικολογικοί πόροι (ερμηνευτικά λεξικά συνωνύμων). Εάν η διάκριση των συναισθημάτων στο λεξικό δεν είναι δυαδική (θετική ή αρνητική) αλλά μεγαλύτερης ποικιλομορφίας, ανατίθεται σε κάθε συναίσθημα μία διαβαθμισμένη τιμή της δυαδικής διάκρισης και υπολογίζεται αναλόγως το αντίστοιχο άθροισμα. Για παράδειγμα η λέξη «nice» μπορεί να θεωρηθεί 0.9 θετική και 0.1 αρνητική, ενώ η λέξη «bad» 0.9 αρνητική και 0.1 θετική. Αυτές οι διαφοροποιήσεις δεν παρεκκλίνουν από τον κεντρικό κορμό της χρήσης των λεξικών, αλλά προτείνουν έναν άλλο τρόπο προσέγγισης τους, μέσω μίας συνεχούς συναισθηματικής κλίμακας [31].

Η χρήση μεθόδων στάθμισης παρέχουν μία επιπλέον δυνατότητα προσθήκης βαρών στις λέξεις που έχουν άμεση συσχέτιση με το αντικείμενο για το οποίο καλούνται να αποφασίσουν εάν το κείμενο κείται θετικά ή αρνητικά, προσπερνώντας με αυτόν τον τρόπο τις μη σχετικές λέξεις που αναμένεται να μη είναι νοηματικά συναφείς. Η γενική μορφή του τύπου που περιέχει έναν παράγοντα διαχείρισης των αρνήσεων και της έντασης των λέξεων (modifier) και ενός παράγοντα στάθμισης τους (weight), με δεδομένο το συναίσθημα (Sent) κάθε λέξης (w), δίνεται ακολούθως:

$$TotalSent = [\sum Sent(w) \times weight(w) \times modifier(w)] \div [\sum weight(w)]$$

Η μεθοδολογία της χρήσης λεξικών στην ανάλυση συναισθήματος ακολουθεί κατά κόρον κάποια βασικά στάδια τα οποία θα μπορούσαν να διακριθούν στην δημιουργία του λεξικού, την εφαρμογή μεθόδων μείωσης, την ανάπτυξη εργαλείων για την ανάλυση συναισθήματος και την τελική εκπαίδευση του συστήματος [53].

Η αυτοματοποιημένη ή χειροκίνητη επιλογή των λέξεων, με κριτήρια που παρουσιάστηκαν παραπάνω, ακολουθείται από την τοποθέτηση παραμέτρων σήμανσης και πολικότητα σε κάθε λέξη. Η διόρθωση των ορθογραφικών λαθών και η εφαρμογή επιπλέον αφαιρετικών τεχνικών, βοηθά στην δημιουργία ενός ευκολότερα διαχειρίσιμου συνόλου λέξεων για την περαιτέρω ανάλυση. Στη συνέχεια η ανάπτυξη ενός συστήματος αξιολόγησης πραγματοποιείται σύμφωνα με κάποια υπάρχουσα πλατφόρμα στην οποία προσαρμόζονται οι κανόνες μείωσης που επιλέχθηκαν. Η εκπαίδευση του συστήματος που ακολουθεί, συνίσταται στην τροποποίηση εκείνων των παραμέτρων που θα κριθούν αναγκαίες, για την διασφάλιση της σωστής λειτουργίας του, σύμφωνα με τους αρχικούς στόχους κάθε μεθόδου.

2.2.2 Μηχανική Μάθηση

2.2.2.1 Ορισμός της Μηχανικής Μάθησης

Μηχανική Μάθηση (Machine Learning) είναι ένα υποπεδίο της Επιστήμης των Υπολογιστών, η οποία εξελίχθηκε από τη μελέτη της Αναγνώρισης Προτύπων και την υπολογιστική θεωρία μάθησης στον τομέα της Τεχνητής Νοημοσύνης. Διερευνά την μελέτη και την κατασκευή αλγορίθμων που μπορούν να μάθουν και να κάνουν προβλέψεις σχετικά με τα δεδομένα που λαμβάνουν ως είσοδο. Τέτοιοι αλγόριθμοι λειτουργούν με την οικοδόμηση ενός μοντέλου από εισόδους - παραδείγματα, προκειμένου να προβλέψουν ή να λάβουν αποφάσεις που βασίζονται σε δεδομένα, αντί να ακολουθούν αυστηρά τις στατικές οδηγίες ενός προγράμματος [9].

Η Μηχανική Μάθηση συνδέεται στενά με την υπολογιστική στατιστική, έναν πεδίο που αποσκοπεί στο σχεδιασμό αλγορίθμων για την εφαρμογή των στατιστικών μεθόδων σε υπολογιστές. Έχει ισχυρούς δεσμούς και με τη μαθηματική βελτιστοποίηση, η οποία παρέχει τις μεθόδους, τη θεωρία και τους τομείς εφαρμογής στον τομέα της. Παρόλο που εστιάζει περισσότερο στη διερευνητική ανάλυση των δεδομένων, η Μηχανική Μάθηση και η Αναγνώριση Προτύπων μπορούν να θεωρηθούν ως οι δύο όψεις του ίδιου νομίσματος. Συνεπώς η Μηχανική μάθηση χρησιμοποιείται σε μια σειρά μηχανογραφικών εργασιών, όπου το σχεδιασμός και ο προγραμματισμός ρητών αλγόριθμων είναι ανέφικτοι, όπως η Ανάλυση Συναισθήματος.

2.2.2.2 Προσεγγίσεις βασιζόμενες στη Μηχανική Μάθηση

Οι μέθοδοι της μηχανικής μάθησης στην ανάλυση συναισθήματος σε κειμενικές πηγές, αποσκοπούν στον εντοπισμό του κατάλληλου τύπου αλγορίθμου και στην εκπαίδευση του. Πιο συγκεκριμένα, αλγόριθμοι που χρησιμοποιούν κατηγοριοποιητές μηχανικής μάθησης (Naïve Bayes, Max Entropy, Support Vector Machines), εκπαιδεύονται σε σύνολα δεδομένων και εξάγουν τα χαρακτηριστικά του εκάστοτε κειμένου με τη χρήση συνδυασμών γραμμάτων, συλλαβών ή λέξεων (N-grams) και POS ετικετών. Η εκπαίδευση των αλγορίθμων κρίνεται απαραίτητη ώστε να επιτυγχάνουν υψηλότερη ακρίβεια στην κατηγοριοποίηση των άγνωστων κειμένων. Για το λόγο αυτό, καταρχάς συλλέγεται ένα σύνολο δεδομένων στο οποίο θα πραγματοποιηθεί η εκπαίδευση και το οποίο μπορεί να περιλαμβάνει σχόλια για την συναισθηματική του κατεύθυνση. Ακολούθως, κάθε επιμέρους τμήμα αυτού του συνόλου, μετουσιώνεται σε ένα διάνυσμα χαρακτηριστικών με τη χρήση κατάλληλων μεθόδων και ένας κατηγοριοποιητής εκπαιδεύεται στην αντιστοίχιση των τιμών αυτών των χαρακτηριστικών, με κάποιες συναισθηματικές καταστάσεις. Μετά το πέρας της εκπαίδευσης είναι σε θέση να πραγματοποιήσει προβλέψεις σε πηγές που δεν ανήκουν στο αρχικό σύνολο δεδομένων.

Η απόδοση αυτών των αλγορίθμων εξαρτάται άμεσα από την επιλογή των χαρακτηριστικών του κάθε διανύσματος. Μια απλή μέθοδος είναι η κωδικοποίηση όλων των όρων του συνόλου εκπαίδευσης και η παραγωγή διανυσμάτων που χαρακτηρίζονται από την ύπαρξη ή ανυπαρξία τους (δυαδικό διάνυσμα). Ένα άλλο κριτήριο επιλογής χαρακτηριστικών αποτελεί η συχνότητα με την οποία εμφανίζεται ο κάθε όρος, η οποία σηματοδοτεί και την επιρροή του στο σύνολο. Για την επιλογή των ιδανικών χαρακτηριστικών συχνά πραγματοποιούνται συγκριτικές μελέτες με πληθώρα αλγορίθμων σύμφωνα με αντικειμενικά κριτήρια που θέτουν οι ερευνητές [60].

2.2.2.3 Κατηγορίες Μεθόδων με βάση την είσοδο του συστήματος

Οι μέθοδοι μηχανικής μάθησης συνήθως κατατάσσονται σε τρεις μεγάλες κατηγορίες, ανάλογα με τη φύση του σήματος που λαμβάνουν ως είσοδο ή της ανατροφοδότησης (feedback) που παρέχουν σε ένα σύστημα μάθησης:

- Μέθοδοι Επιβλεπόμενης Μάθησης (Supervised Learning): Σύμφωνα με αυτές τις μεθόδους το σύστημα εκπαιδεύεται με εισόδους και τις αντίστοιχες επιθυμητές εξόδους που λειτουργούν ως παραδείγματα, που δίνονται από τον εκπαιδευτή, και ο στόχος τους είναι να διδάξουν στο σύστημα ένα γενικό κανόνα ο οποίος θα χαρτογραφεί τις μελλοντικές εισόδους σε επιθυμητές εξόδους.

- Μέθοδοι μη Επιβλεπόμενης Μάθησης (Unsupervised Learning): Σύμφωνα με αυτές τις μεθόδους το σύστημα δεν λαμβάνει εισόδους με ετικέτες ως είσοδο στον αλγόριθμο μάθησης, αλλά αφήνεται μόνο του να εντοπίσει τη άγνωστη δομή στα δεδομένα, στην είσοδο του. Η έξοδος του συστήματος αυτών των μεθόδων μπορεί να είναι είτε η ίδια η ευρεθείσα δομή των δεδομένων, είτε ένα στάδιο επεξεργασίας για την περαιτέρω ανάλυση τους.

- Μέθοδοι Ενισχυμένης Μάθησης (Reinforcement Learning): Σύμφωνα με αυτές τις μεθόδους το σύστημα αλληλεπιδρά με ένα δυναμικό περιβάλλον, στο οποίο πρέπει να επιτύχει ένα συγκεκριμένο στόχο, χωρίς κάποιον παράγοντα εκπαίδευσης να το ενημερώνει ρητά εάν βρίσκεται κοντά στο στόχο του. Η μάθηση σε αυτήν την περίπτωση πραγματοποιείται από την αναγνώριση των αποτελεσμάτων κάθε βήματος αλληλεπίδρασης. Αυτές οι μέθοδοι δεν έχουν ουσιώδεις εφαρμογές στην ανάλυση συναισθήματος.

Ανάμεσα στις μεθόδους Επιβλεπόμενης και μη Επιβλεπόμενης Μάθησης βρίσκονται οι μέθοδοι Ημι-Επιβλεπόμενης μάθησης (Semi-Supervised Learning), όπου το σύστημα εκπαιδεύεται με ένα μικρό ζεύγος εισόδων και τις αντίστοιχες επιθυμητές εξόδους τους, και ένα σήμα ελλιπής εκπαίδευσης δηλαδή ένα σύνολο εκπαίδευσης με κάποιες από τις επιθυμητές εξόδους να λείπουν.

2.2.2.4 Κατηγορίες Μεθόδων με βάση την έξοδο του συστήματος

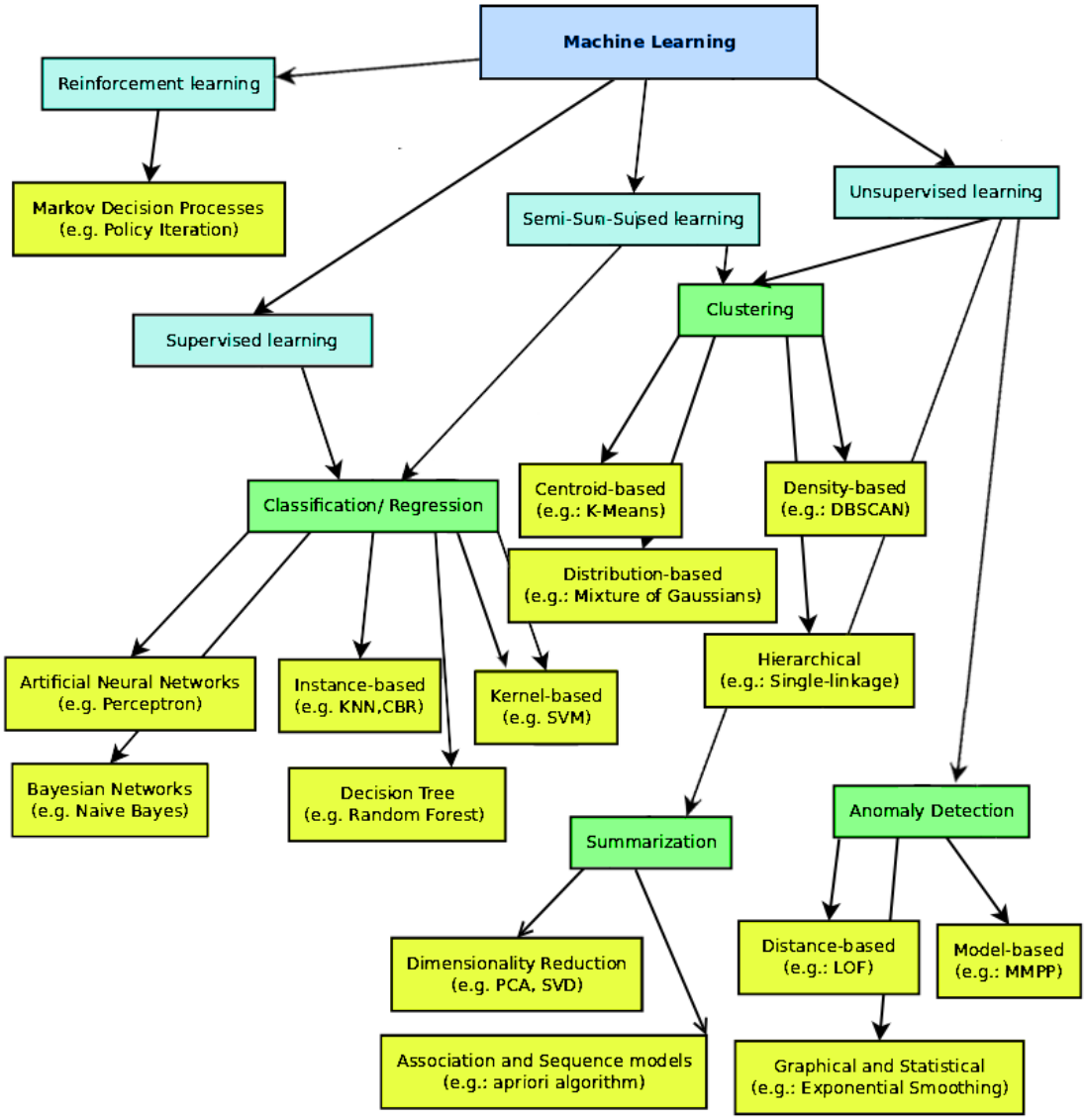
Μια άλλη κατηγοριοποίηση των μεθόδων προκύπτει όταν αναλογιστεί κανείς την επιθυμητή έξοδο ενός συστήματος μηχανικής μάθησης:

- Ανάλυσης Συστάδων (Clustering): Σύμφωνα με αυτές τις μεθόδους εκχωρείται ένα σύνολο παρατηρήσεων σε υποσύνολα (που ονομάζονται συστάδες), έτσι ώστε οι παρατηρήσεις στην ίδια ομάδα να είναι παρόμοια με κάποια έννοια, σύμφωνα με κάποια προκαθορισμένα κριτήρια, ενώ οι παρατηρήσεις που προέρχονται από διαφορετικές συστάδες είναι ανόμοιες. Διαφορετικές τεχνικές ανάλυσης συστάδων κάνουν διαφορετικές υποθέσεις σχετικά με τη δομή των δεδομένων, συχνά καθορισμένες από κάποια μετρική ομοιότητας (similarity metric), αξιολογώντας την ομοιότητα αυτή, ανάμεσα στα μέλη της ίδιας συστάδας. Άλλες μέθοδοι βασίζονται στην εκτιμώμενη πυκνότητα (estimated density) και τη συνδετικότητα γράφου (graph connectivity). Η συσταδοποίηση είναι κυρίως μια μέθοδος μάθησης χωρίς επίβλεψη, και μια συνήθης τεχνική για την ανάλυση των στατιστικών δεδομένων [28].

- Ταξινόμησης/Παλινδρόμησης (Classification/Regression): Από τη μία πλευρά, στη μέθοδο της Ταξινόμησης οι είσοδοι χωρίζονται σε δύο ή περισσότερες κατηγορίες, και το εκπαιδευόμενο σύστημα πρέπει να παράγει ως έξοδο ένα μοντέλο που αποδίδει τις άγνωστες εισόδους σε μία ή περισσότερες από αυτές τις διακριτές κατηγορίες. Από την άλλη πλευρά, στη μέθοδο της Παλινδρόμησης οι είσοδοι αντιμετωπίζονται ως ανεξάρτητες μεταβλητές και το σύστημα καλείται να τις συσχετίσει με μία εξαρτημένη μεταβλητή, και να διερευνήσει τις μορφές αυτών των σχέσεων [33]. Και οι δύο αυτές τεχνικές συνήθως εντάσσονται σε επιβλεπόμενες μεθόδους μάθησης.

- Περίληψης Κειμένου (Text Summarization): Σύμφωνα με αυτές τις μεθόδους το σύστημα παράγει ως έξοδο μια περίληψη που διατηρεί τα πιο σημαντικά σημεία του πρωτότυπου εγγράφου μέσω μίας διαδικασίας μείωσης του κειμένου, είτε εξορυκτικά (extractive summarization), δηλαδή επιλέγοντας ένα υποσύνολο των υφιστάμενων λέξεων ή φράσεων, είτε αφαιρετικά (abstractive summarization) οικοδομώντας μια εσωτερική σημασιολογική αναπαράσταση με τεχνικές δημιουργίας φυσικής γλώσσας [27].

- Ανίχνευσης Ανωμαλιών (Anomaly Detection): Σύμφωνα με αυτές τις μεθόδους το σύστημα πραγματοποιεί ταυτοποίηση των στοιχείων, των γεγονότων ή των παρατηρήσεων που λαμβάνει ως είσοδο και οι οποίες δεν ανταποκρίνονται σε έναν αναμενόμενο ρυθμό ή σε άλλα αντικείμενα μέσα στο σύνολο δεδομένων και παράγει ως έξοδο ένα σύνολο, χωρίς τα μη αποδεκτά στοιχεία. Αυτά τα μη συμμορφούμενα πρότυπα αναφέρονται συχνά σε ανωμαλίες, ακραίες τιμές, ασυμφωνίες παρατηρήσεων, εξαιρέσεις, παρεκκλίσεις, εκπλήξεις, ιδιαιτερότητες ή προσμίξεις σε διαφορετικά πεδία της εισόδου [19].



Σχήμα 2.3: Διακρίσεις Ανάλυσης Συναισθήματος με Μηχανική Μάθηση[1]

2.2.3 Στατιστική Ανάλυση

Οι στατιστικές μέθοδοι έχουν ως σκοπό την αντιμετώπιση των δυσκολιών που ανακύπτουν από την αδυναμία διαχείρισης κάποιων λέξεων ή κάποιων κειμένων από τις μεθόδους που βασίζονται σε λεξικά ή μηχανική μάθηση. Η αντιμετώπιση αυτών των προβλημάτων συνίσταται στη σύνθεση μεγάλων συνόλων εγγράφων ή αν ήταν δυνατόν ακόμα και το σύνολο των εγγράφων του διαδικτύου, και στην ένταξη τους σε ένα ευρετήριο για την δημιουργία, με μη επιβλεπόμενο τρόπο, ενός «συναισθηματικού λεξικού», που ως στοιχεία δεν έχει λέξεις, αλλά ολόκληρα έγγραφα.

[1] https://doublebyteblog.files.wordpress.com/2014/08/ml_rc-1-0.png

Στις μεθόδους Στατιστικής Ανάλυσης, κάθε λέξη αποκτά ένα σημασιολογικό προσανατολισμό σύμφωνα με τη συχνότητα με την οποία εμφανίζεται σε έγγραφα του συνόλου, που έχουν ήδη χαρακτηριστεί ως θετικού ή αρνητικού προσανατολισμού, ενώ αν οι συχνότητες είναι ταυτόσημες χαρακτηρίζεται ως ουδέτερη. Όμοια, σε άλλες τεχνικές εξετάζεται και η συχνότητα με την οποία εμφανίζεται κάθε λέξη μαζί με κάποια άλλη εγνωσμένου προσανατολισμού, υπολογίζοντας την πιθανότητα να βρεθεί σε ένα έγγραφο κοντά σε μία θετική ή αρνητική λέξη [78]. Εκτός από την κάθε λέξη ξεχωριστά, μπορεί να εξεταστούν διαδοχικές λέξεις σε σειρά. Σε αυτή την περίπτωση υπολογίζεται ο σημασιολογικός προσανατολισμός τους από την πιθανότητα να εμφανιστούν σε ένα έγγραφο θετικού ή αρνητικού προσανατολισμού, ως μία ακολουθία .

2.2.4 Σύγκριση Προσεγγίσεων

Οι μέθοδοι που βασίζονται στη χρήση ερμηνευτικών λεξικών ή λεξικών συναισθήματος χαρακτηρίζονται από έναν υψηλό βαθμό ευχρηστίας, με δυνατότητες εφαρμογής σε πολύ μεγάλα σύνολα δεδομένων και αυτόματη προσθήκη ετικετών σε κάθε λέξη τους, γεγονός που ενισχύει την ανάλυση του σημασιολογικού προσανατολισμού τους. Οι μέθοδοι που βασίζονται στη μηχανική μάθηση, αξιοποιώντας τα μικρά σύνολα δεδομένων εκπαίδευσης, επιτυγχάνουν να πραγματοποιούν προβλέψεις σε σχέση με πρωτοεμφανιζόμενα δεδομένα, μετά το πέρας της εκπαίδευσης, ακόμα και αν κάποια λέξη δεν προϋπήρχε στη βάση τους [61].

Από την άλλη πλευρά, οι παραδοσιακές στατιστικές μέθοδοι είναι συγκριτικά πιο αδύναμες σημασιολογικά, πράγμα που σημαίνει ότι με εξαίρεση την προφανή επιρροή που ασκούν λέξεις-κλειδιά γνωστού σημασιολογικού προσανατολισμού (πχ «καλός», «κακός»), οι υπόλοιπες λεξιλογικές μονάδες ή η συνύπαρξη τους σε ένα στατιστικό μοντέλο, έχει μικρή προγνωστική αξία, όταν εξετάζεται ξεχωριστά. Ως αποτέλεσμα, οι στατιστικοί ταξινομητές κείμενου λειτουργούν με αποδεκτή ακρίβεια, μόνο όταν τους δοθεί ως είσοδος μια αρκούντως μεγάλη κειμενική βάση δεδομένων. Έτσι, ενώ οι μέθοδοι αυτές μπορεί να είναι σε θέση να ταξινομήσουν συναισθηματικά το κείμενο ενός χρήστη σε επίπεδο σελίδας ή παραγράφου, δεν αποδίδουν εξίσου καλά σε μικρότερες μονάδες κειμένου, όπως το επίπεδο πρότασης [17].

Παρόλο που η κατασκευή ενός λεξικού εκ του μηδενός είναι μια δύσκολη και χρονοβόρα διαδικασία, εγγυάται την αποδοτικότητά ενός συστήματος σε διάφορα επίπεδα ανάλυσης και την δυνατότητα επέκτασης αλλά και εξειδίκευσης του, σε συγκεκριμένους γνωσιακούς τομείς, σύμφωνα με τις εκάστοτε απαιτήσεις της έρευνας στην οποία χρησιμοποιούνται. Η απουσία μεγάλων κειμενικών βάσεων δεδομένων και συνόλων εκπαίδευσης, αποτελεί το κυριότερο προσόν των μεθόδων αυτών, αφού η δημιουργία τεράστιων βάσεων είναι ασύμφορη, αλλά και τα σύνολα

εκπαίδευσης είναι συνήθως επικεντρωμένα σε συγκεκριμένα πεδία, καθιστώντας τις εφαρμογές των μεθόδων περιορισμένης εμβέλειας και εφαρμογής, με σημαντικό βαθμό ανακρίβειας στην εξαγωγή συναισθήματος, ειδικά σε σύνολα δεδομένων προς ανάλυση, μικρής συνάφειας με το αρχικό σύνολο εκπαίδευσης.

Τα λεξικά συναισθήματος έχουν την δυνατότητα να συνοδεύουν κάθε όρο τους, από ένα μεγάλο εύρος σημασιολογικών χαρακτηριστικών που σχετίζονται με ανεξάρτητα πεδία και διασφαλίζουν την ακριβέστερη αποτύπωση της έννοιας κάθε λέξης σύμφωνα με το γνωσιακό της περιβάλλον. Συν τοις άλλοις, οι μέθοδοι κατασκευής λεξικών συναισθημάτων από ήδη υπάρχουσες λεξικολογικές πηγές ή λεξικά συναισθήματος με αυτοματοποιημένο τρόπο, και η εξίσου αυτόματη απόδοση ετικετών σε κάθε λέξη, είναι σαφώς ευκολότερες τεχνικά, από την δημιουργία ενός συνόλου εκπαίδευσης και την επεξεργασία και ταξινόμηση του.

Σύγκριση Μεθόδων	Χρήση Λεξικών	Μηχανική Μάθηση	Στατιστική Ανάλυση
Πολυεπίπεδη Ανάλυση	✓	✓	✗
Πολυπεδιακή Ανάλυση	✓	✓	✓
Ανεξαρτησία από Σύνολα Δεδομένων	✓	✗	✗
Διαχείριση Άγνωστων Όρων	✗	✓	✓
Αυτόματη Απόδοση Ετικετών	✓	✗	✗

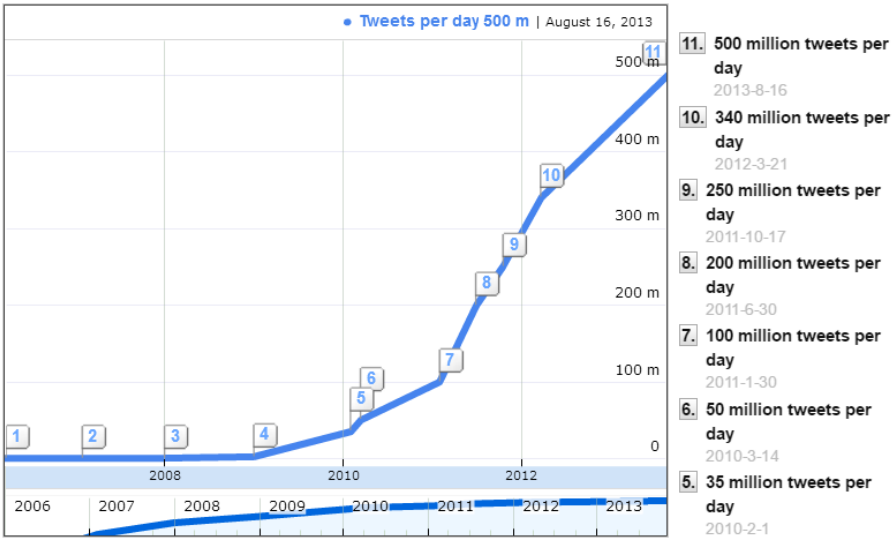
Πίνακας 2.1: Σύγκριση Μεθόδων Ανάλυσης Συναισθήματος

2.3 Το Συναίσθημα στα Κοινωνικά Δίκτυα

Για πρώτη φορά στην ανθρώπινη ιστορία, έχουμε ένα τεράστιο όγκο δεδομένων, στα μέσα κοινωνικής δικτύωσης. Χωρίς αυτά τα δεδομένα, ένα μεγάλο μέρος της έρευνας στην Ανάλυση Συναισθήματος δεν θα ήταν δυνατό. Δεν αποτελεί έκπληξη, ότι η έναρξη και η ταχεία ανάπτυξη της ανάλυσης συναισθήματος συμπίπτουν με αυτή του διαδικτύου και ειδικότερα της κοινωνικής δικτύωσης. Στην πραγματικότητα, η ανάλυση συναισθήματος βρίσκεται ακριβώς στο κέντρο της κοινωνικής έρευνας, που εξετάζει τα μέσα ενημέρωσης και το δημόσιο διαδικτυακό διάλογο. Ως εκ τούτου, η έρευνα στην ανάλυση συναισθήματος δεν έχει μόνο σημαντικό αντίκτυπο στη NLP, αλλά μπορεί επίσης να έχει μια βαθιά επίδραση στις επιστήμες της διαχείρισης, της πολιτικής επιστήμης, της οικονομίας και των κοινωνικών επιστημών, όπως είναι όλα επηρεάζονται από τις απόψεις των ανθρώπων.

2.3.1 Το φαινόμενο του Twitter

Το Twitter είναι μια υπηρεσία μικρο-blogging , όπου οι χρήστες δημοσιεύουν tweet, που δεν υπερβαίνουν τους 140 χαρακτήρες. Με περίπου 500 εκατομμύρια χρήστες, που παράγουν 350 χιλιάδες tweet ανά λεπτό, το Twitter αποτελεί ένα από τα μεγαλύτερα και πιο δυναμικά σύνολα δεδομένων, με περιεχόμενο παραγόμενο από χρήστες. Μαζί με άλλες ιστοσελίδες κοινωνικής δικτύωσης, όπως το Facebook, το περιεχόμενο στο Twitter δημιουργείται σε πραγματικό χρόνο. Συνεπώς, παρατηρούνται tweet για κάθε φύσεως γεγονός, από τα γενέθλια ενός φίλου, έως έναν καταστροφικό σεισμό, που μπορεί να αναρτηθούν κατά τη διάρκεια και αμέσως μετά το εν λόγω γεγονός[2].



Σχήμα 2.4: Ετήσια Αύξηση του πλήθους των tweets[3]

[2] https://en.wikipedia.org/wiki/Twitter
[3] http://www.internetlivestats.com/twitter-statistics

Αυτό το τεράστιο ρεύμα δεδομένων, πραγματικού χρόνου, έχει μείζονες επιπτώσεις για όσους ενδιαφέρονται για την κοινή γνώμη ή ακόμη και ενεργούν με βάση τί μαθαίνει και αλληλεπιδρά με το κοινό άμεσα. Οι εταιρείες έχουν την ευκαιρία να εξετάσουν τι λένε οι πελάτες και οι δυνητικοί πελάτες τους, για τα προϊόντα και τις υπηρεσίες τους, χωρίς δαπανηρές και χρονοβόρες έρευνες ή συγκεκριμένα αιτήματα για σχόλια. Πολιτικές οργανώσεις και υποψήφιοι μπορούν να προσδιορίζουν ποια θέματα ενδιαφέρουν περισσότερο το κοινό, καθώς και ποιά θα είναι η θέση τους σ' αυτά τα θέματα. Η χειροκίνητη διερεύνηση των tweet μπορεί να είναι χρήσιμη για πολλές από αυτές τις αναλύσεις, αλλά πολλές εφαρμογές και ερωτήματα απαιτούν ανάλυση πραγματικού χρόνου σε τεράστιο όγκο περιεχομένου των κοινωνικών μέσων. Έτσι, είναι σε υψηλή ζήτηση τα υπολογιστικά εργαλεία που είναι σε θέση αυτομάτως να εξάγουν και να αναλύουν σχετικές πληροφορίες, για μία γνώμη που εκφράζεται στο Twitter και άλλες πηγές κοινωνικών δικτύων.

2.3.2 Ανάλυση Συναισθήματος στο Twitter

Οι κοινωνικές συνδέσεις, των χρηστών, μπορούν να χρησιμοποιηθούν στην ταξινόμηση της πολικότητας, κάθε μεμονωμένου tweet. Για το λόγο αυτό, μπορεί να κατασκευαστεί για κάθε χρήστη, ο γράφος ακολούθων του, στο Twitter (Twitter follower graph), χρησιμοποιώντας δημοσίως διαθέσιμες δεδομένα από το Twitter API. Από την πλήρη λίστα των ακολούθων του καθενός χρήστη, διατηρούνται μόνο οι ακόλουθοι, που βρίσκονται στα εκάστοτε σύνολα δεδομένων κάθε μεθόδου. Αυτό περιορίζει άγνωστους χρήστες που δεν έκαναν tweet για το εξεταζόμενο θέμα και έτσι είναι απίθανο να παρέχουν χρήσιμες πληροφορίες. Τέτοιες μέθοδοι δίνουν σχεδόν πλήρεις γράφους, αλλά έχουν δύο βασικά μειονεκτήματα. Πρώτον, πολλοί χρήστες με την πάροδο του χρόνου έχουν αυξήσει το επίπεδο προστασίας προσωπικών δεδομένων τους, πράγμα που δυσχεραίνει τη δυνατότητα πρόσβασης στο γράφο ακολούθων τους και είναι γνωστές μόνο οι πληροφορίες των tweet τους. Δεύτερον, λόγω του γρήγορου ρυθμού ανάπτυξης του Twitter, οι γράφοι χρήστη τείνουν να αναπτύσσονται εξίσου γρήγορα. Έτσι ο κατασκευασμένος γράφος τείνει να αναπαριστά τον τρέχοντα κοινωνικό γράφο του χρήστη, και όχι τον ακριβή γράφο, που υπήρχε κατά τη στιγμή του tweet.

Στην επιβλεπόμενη ή την ήμι-επιβλεπόμενη μηχανική μάθηση, σε συστήματα που ειδικεύονται στο Twitter, καθοριστικό ρόλο παίζουν τα σύνολα εκπαίδευσης. Βαρύνουσας σημασίας είναι τα σύνολα δεδομένων με σχόλια πολικότητας, προγενέστερων εργασιών. Μία τέτοια συλλογή αποτελεί και το Σώμα Κειμένου του Στάντφορντ (Stanford Twitter Sentiment – STS Corpus)[4], με 216 σχολιασμένα tweets διαφόρων θεμάτων. Οι Shamma et al. (2009) χρησιμοποίησαν το Amazon Mechanical Turk[5] για να σχολιάσουν 3.269 tweet, που αναρτήθηκαν στη διάρκεια της προεκλογικής αντιπαράθεσης της 26 Σεπτεμβρίου 2008, μεταξύ Μπαράκ Ομπάμα και John McCain (Obama-McCain Debate - OMD). Κάθε tweet σχολιάστηκε από έναν ή περισσότερους, σχολιαστές για τις κατηγορίες θετικό, αρνητικό, μικτό, ή άλλο. Οι Speriusu et al. (2011) [71] κατάρτισαν μία βάση tweets που αναφέρονται στην υγειονομική μεταρρύθμιση της Αμερικής (Health Care Reform – HCR) επιλέγοντας εκείνα, με hashtag «#hcr», δημιουργώντας ένα σύνολο πέντε υποκειμενικών

[4] http://twittersentiment.appspot.com
[5] https://www.mturk.com/mturk

διαβαθμίσεων (θετικά, αρνητικά, ουδέτερα, άσχετα, αβέβαια). Στο ίδιο μήκος κύματος κινούνται και τα σύνολα δεδομένων, από χιλιάδες tweets που συντάσσονται από την ετήσια διοργάνωση του Semantic Evaluation of System challenge (SemEval Datasets)[6].

Παράλληλα, αξιοποιούνται και τα emoticon, συνήθως ως θορυβώδεις δείκτες πολικότητας, συμπεριλαμβανομένης της προχωρημένης αναζήτησης του Twitter, με θετική / αρνητική διάθεση. Αν και ατελής, υπάρχει δυνατότητα, για εκατομμύρια tweet που περιέχουν emoticon, να χρησιμεύσουν ως πηγή θορυβώδους εκπαιδευτικού υλικού για έναν επιβλεπόμενο ταξινομητή.

+	:) :D =D =) :] =] :-) :-D :-] ;) ;D ;] ;-) ;-D ;-]
−	:(=(:[=[:-(:-[:'(:'[D:

Σχήμα 2.5: Παραδείγματα emoticon θετικής και αρνητικής πολικότητας

2.3.3 Χαρακτηριστικά Ανάλυσης των tweets

Ένα από τα βασικά σημεία της ανάλυσης των μηνυμάτων του Twitter, αποτελεί η αποσαφήνιση της συμφραστικής πολικότητας τους, όπου δεδομένου ενός μηνύματος, που περιέχει ένα σημαντικό λεκτικό ή φραστικό στιγμιότυπο, απαιτείται η κατηγοριοποίηση του, ως θετικό, αρνητικό ή ουδέτερο, μέσα στο πλαίσιο στο οποίο παρουσιάζεται. Έπειτα, ακολουθεί η συνήθης ταξινόμηση πολικότητας, σύμφωνα με την οποία, ένα μήνυμα, κατατάσσεται αναλόγως, εάν είναι θετικού, αρνητικού ή ουδέτερου συναισθήματος, στην αρμόζουσα κατηγορία, για περεταίρω επεξεργασία. Για τα μηνύματα που μεταφέρουν τόσο θετικό και αρνητικό συναίσθημα, πρέπει να επιλεγεί το ισχυρότερο συναίσθημα.

Η ταξινόμηση πολικότητας κάθε μηνύματος μπορεί να βασίζεται επίσης, σε κάθε θέμα συζήτησης στο Twitter (hashtag chat), στην οποία εντοπίζεται, κατατάσσοντάς το, ως θετικό, αρνητικό ή ουδέτερο συναισθήματος, ως προς το συγκεκριμένο θέμα. Για τα μηνύματα που μεταφέρουν τόσο θετικό όσο και αρνητικό κλίμα, προς το συγκεκριμένο θέμα, πρέπει να επιλεγεί το πιο έντονο από τα δύο. Επίσης, συχνό αντικείμενο μελέτης αποτελεί η ανίχνευση τάσεων (trends) προς μια θεματική ενότητα (topic) στο Twitter. Λαμβάνοντας υπόψη μια σειρά από μηνύματα για ένα συγκεκριμένο θέμα, της ίδιας χρονικής περιόδου, απαιτείται ο καθορισμός του κυρίαρχου συναισθήματος ως προς το στοχευόμενο θέμα σε αυτά τα μηνύματα, (έντονα θετικό, ασθενώς θετικό, ουδέτερο, ασθενώς αρνητικό, ή έντονα αρνητικό).

Ένα πρόσφατο πεδίο έρευνας συνίσταται στον προσδιορισμός της δύναμης, της σύνδεσης των όρων του Twitter με θετικό συναίσθημα ή του βαθμού προϋπάρχουσας πολικότητας (prior polarity degree). Δεδομένης μιας λέξης ή μιας φράσης,

[6] http://alt.qcri.org/semeval2016

αναζητείται μια βαθμολογία, μεταξύ του μηδενός και της μονάδος, που λειτουργεί ως ένδειξη της ισχύος της σύνδεσης με το θετικό κλίμα. Ο βαθμός 1 υποδηλώνει μέγιστη συσχέτιση με θετικό συναίσθημα (ή τουλάχιστον ελάχιστη ταύτιση με κάποιο αρνητικό συναίσθημα), και ένας βαθμός κοντά στο 0 δείχνει την μέγιστη συσχέτιση με κάποιο αρνητικό συναίσθημα, αντίστοιχα. Εάν μια λέξη είναι πιο θετική από μία άλλη, τότε απαραίτητα θα πρέπει να έχει και μια υψηλότερη βαθμολογία.

2.3.4 Ιδιαιτερότητες του Twitter

Όπως στα περισσότερα κοινωνικά δίκτυα, έτσι και στο twitter, η συνεχής μεταβολή των γλωσσικών ιδιωμάτων και διαλεκτικών χαρακτηριστικών του προφορικού λόγου, όπως αυτός αποτυπώνεται στα βραχύσωμα μηνύματα του διαδικτύου, αποτελεί το μείζον πρόβλημα στην ανάλυση συναισθήματος. Χαρακτηριστικά όπως ο σαρκασμός και η ειρωνεία, οι συντομογραφίες λέξεων, η ρητορική χρήση ερωτηματικού, αλλά και η επανάληψη συλλαβών, είναι μόνιμα φαινόμενα που δυσχεραίνουν την ανάδειξη της κατηγοριοποίησης μίας λέξης ή μίας φράσης. Η ίδια η δομή των μηνυμάτων με το περιορισμένο μέγεθος των 140 χαρακτήρων προδιαγράφει την όλο και αυξανόμενη ανάγκη λεκτικής συμπύκνωσης του νοηματικού περιεχομένου, γεγονός που αναπόδραστα λειτουργεί ως τροχοπέδη στους επίδοξους αναλυτές αυτών των σύντομών κειμένων.

Οι παραπάνω εκφάνσεις, έχουν αντίκτυπο τόσο στις μεθόδους που βασίζονται σε λεξικά συναισθήματος, όσο και εκείνες που χρησιμοποιούν μηχανική μάθηση. Στην πρώτη περίπτωση, τα λεξικά είναι αναγκαίο να ανανεώνουν συνεχώς το περιεχόμενό τους είτε αυτά είναι δομημένα χειρωνακτικώς, μέσα από την επέκτασή του με καινοφανείς νεολογισμούς, είτε κατασκευάζονται αυτοματοποιημένα, αξιοποιώντας μεγάλης κλίμακας σύνολα δεδομένων tweet. Στην περίπτωση της μηχανικής μάθησης, η προεπεξεργασία των μηνυμάτων, με την απομάκρυνση των σημείων στίξης, των αριθμών, των υπερσυνδέσμων και την επισήμανση των emoticon με δείκτες πολικότητας κρίνεται ως ένα καθοριστικής σημασίας στάδιο για την περεταίρω εκπαίδευση κάθε συστήματος. Συνεπώς, το συνεχώς μεταβαλλόμενο γλωσσικό περιβάλλον του διαδικτύου έχει ένα ευρύ φάσμα επιρροής σε κάθε μέθοδο ανεξάρτητα προσέγγισης.

2.4 Μέθοδοι Αξιολόγησης Ανάλυσης Συναισθήματος

2.4.1 Διαβαθμολογική Αξιοπιστία

Η αξιολόγηση ενός συστήματος Εξόρυξης Γνώμης και Ανάλυσης Συναισθήματος επιτυγχάνεται μέσα από τη σύγκριση των αποτελεσμάτων κατηγοριοποίησης των ταξινομητών του, έχοντας ως σημείο αναφοράς τις αντίστοιχες ταξινομήσεις που προέρχονται από την χειρονακτική ταξινόμηση, ενός ή περισσοτέρων βαθμολογητών. Θα περίμενε κανείς, οι παραπάνω βαθμολογήσεις να συγκλίνουν, αρκετές φορές όμως παρατηρούνται αποκλίσεις στην ανθρώπινη κατηγοριοποίηση, γεγονός που δυσχεραίνει την τελική αξιολόγηση της αυτόματης, επιβλεπόμενης ή μη, κατηγοριοποίησης των ταξινομητών Μηχανικής Μάθησης, ή των συστημάτων που βασίζονται σε Λεξικά Συναισθήματος. Συνεπώς, κρίνεται ως αναγκαία η διασφάλιση της διαβαθμολογικής αξιοπιστίας (inter-rater reliability), δηλαδή ομοιογένειας των αξιολογήσεων των βαθμολογητών, για την υπερκέραση της ανομοιογένειας, της ανθρώπινης κρίσης, η οποία αποτελεί και το ακρότατο όριο, απόδοσης και επιτυχίας, της ανάλυσης κάθε συστήματος. Σε κάθε περίπτωση, η εφαρμογή μίας ή και παραπάνω από αυτές τις μετρικές μεθόδους, δίνει τη δυνατότητα αξιολόγησης της λειτουργίας αλλά και σύγκρισης μεταξύ των συστημάτων.

2.4.1.1 Συντελεστής Κάπα του Cohen

Ο Συντελεστής Κάπα του Cohen (Cohen's Kappa Coefficient) είναι ένα στατιστικό μέτρο συμφωνίας μεταξύ αξιολογητών για ποιοτικά στοιχεία ενταγμένα σε κλάσεις, που μετρά την ταύτιση δύο βαθμολογητών, οι οποίοι ταξινομούν n στοιχεία σε C αμοιβαία αποκλειόμενες κατηγορίες. Γενικά πιστεύεται ότι είναι πιο ισχυρός από τον απλό υπολογισμό της εκατοστιαίας συμφωνίας, δεδομένου ότι λαμβάνει υπόψη και τη συμφωνία που βασίζεται στην τυχαιότητα [66].

Ο Δείκτης Κάπα υπολογίζεται σύμφωνα με την ακόλουθη σχέση:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

όπου p_o είναι η σχετική παρατηρούμενη συμφωνία μεταξύ των εκτιμητών, και p_e είναι η υποθετική πιθανότητα τυχαίας συμφωνίας, χρησιμοποιώντας τα παρατηρούμενα δεδομένα για τον υπολογισμό των πιθανοτήτων κάθε παρατηρητή, τυχαία για κάθε κατηγορία. Εάν οι βαθμολογητές είναι σε πλήρη συμφωνία, τότε $\kappa = 1$. Αν δεν υπάρχει συμφωνία μεταξύ των βαθμολογητών, εκτός από αυτή που θα αναμενόταν από την τύχη, όπως δίνεται από το p_e, τότε ο συντελεστής παίρνει τιμές στο διάστημα $(-\infty, 0]$.

Ο Σταθμισμένος Κάπα επιτρέπει τον υπολογισμό των διαφωνιών με διαφορετικό τρόπο και είναι ιδιαίτερα χρήσιμος όταν οι υπάρχει διάταξη στα δυαδικά δεδομένα. Οι τρεις πίνακες που εμπλέκονται στον καθορισμό του, είναι η μήτρα των παρατηρούμενων βαθμολογιών, η μήτρα των αναμενόμενων αποτελεσμάτων βάσει τυχαίας συμφωνίας, και η μήτρα βάρους. Τα κελιά της μήτρας βάρους που βρίσκονται στη διαγώνιο της, αντιπροσωπεύουν τη συμφωνία και ως εκ τούτου περιέχουν μηδενικά. Τα μη διαγώνια κελιά περιέχουν τα βάρη που δείχνουν τη σοβαρότητα της εν λόγω διαφωνίας [22].

Ο Σταθμισμένος Δείκτης Κάπα προκύπτει ως εξής:

$$\kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\, x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\, m_{ij}}$$

όπου κ αριθμός των δυαδικών δεδομένων και w_{ij} , x_{ij} , m_{ij} είναι τα στοιχεία της μήτρας βάρους, της παρατηρούμενης, και της αναμενόμενης μήτρας, αντίστοιχα.

κ	Ερμηνεία
<0	Συμφωνία Μικρότερη της Τυχαίας
0.01-0.20	Μικρή Συμφωνία
0.21-0.40	Ισότιμη Συμφωνία
0.41-0.60	Μέτρια Συμφωνία
0.61-0.80	Ουσιώδης Συμφωνία
0.81-0.99	Σχεδόν Πλήρης Συμφωνία
1	Τέλεια Συμφωνία

Πίνακας 2.2: Εξάρτηση βαθμού συμφωνίας από το συντελεστή Cohen

Όταν τα στοιχεία της διαγωνίου περιέχουν μηδενικά βάρη και όλα τα μη διαγώνια στοιχεία, μοναδιαία βάρη, ο σταθμισμένος συντελεστής Κάπα ταυτίζεται με τον αντίστοιχο μη σταθμισμένο. Η συμφωνία των μέτρων Κάπα του Cohen, χρησιμοποιείται μόνο μεταξύ δύο βαθμολογητών. Ο συντελεστής συνήθως χρησιμοποιείται για να συγκριθούν επιδόσεις των μεθόδων Μηχανικής Μάθησης, αλλά η κατευθυντική έκδοση του (Informedness) εκτιμάται ως καταλληλότερη, κυρίως για τις επιβλεπόμενες μεθόδους Μηχανικής Μάθησης.

2.4.1.2 Συντελεστής Κάπα του Fleiss και A του Robinson

Ένα παρόμοιο μέτρο συμφωνίας, το οποίο χρησιμοποιείται όταν υπάρχουν περισσότεροι των δύο βαθμολογητών, είναι ο συντελεστής Κάπα του Fleiss (Fleiss' Kappa Coefficient), ο οποίος σε αντίθεση με άλλους συντελεστές Κάπα, δεν εφαρμόζεται μόνο κατά την εκτίμηση της συμφωνίας μεταξύ δύο βαθμολογητών. Αν και λειτουργεί παρεμφερώς με το συντελεστή Cohen, δεν αποτελεί παράγωγό του, αλλά είναι ένα πολυ-εκτιμητής, γενίκευση της π στατιστικής του Scott.

Ο συντελεστής Κάπα του Fleiss είναι ένα στατιστικό μέτρο για την αξιολόγηση της αξιοπιστίας της συμφωνίας μεταξύ ενός σταθερού αριθμού βαθμολογητών, κατά την ανάθεση βαθμολογιών σε κατηγορίες, σε μια σειρά από στοιχεία ή την ταξινόμηση αντικειμένων. Ο συντελεστής υπολογίζει το βαθμό συμφωνίας στην κατάταξη, του υποθετικά αναμενόμενου τυχαίου αποτελέσματος, και μπορεί να χρησιμοποιηθεί μόνο για αξιολογήσεις δυαδικής κλίμακας. Μπορεί να ερμηνευθεί, επίσης, ως έκφραση του βαθμού στον οποίο η παρατηρούμενη ποσότητα της συμφωνίας μεταξύ βαθμολογητών υπερβαίνει την αναμενόμενη, εάν όλοι οι βαθμολογητές έκαναν τις αξιολογήσεις τους, εντελώς τυχαία.

Ως συμφωνία νοείται το μέτρο της συνέπειας, που εκφράζει ο συντελεστής, υποθέτοντας ότι ένας σταθερός αριθμός αξιολογητών εκχωρήσει αριθμητικές βαθμολογίες σε μια σειρά από στοιχεία. Ο Δείκτης Κάπα του Fleiss υπολογίζεται σύμφωνα με την ακόλουθη σχέση:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Ο συντελεστής $1 - \bar{P}_e$ δίνει το βαθμό συμφωνίας που είναι εφικτή με τη συνδρομή της τύχης, και ο $\bar{P} - \bar{P}_e$ βαθμό συμφωνίας που πράγματι επιτεύχθηκε στην αντίστοιχη περίπτωση. Αν οι βαθμολογητές είναι σε πλήρη συμφωνία, τότε $\kappa = 1$. Αν δεν υπάρχει συμφωνία μεταξύ των βαθμολογητών, τότε ο συντελεστής παίρνει τιμές στο διάστημα $(-\infty, 0]$.

Αντίστοιχα με τον Κάπα του Fleiss, λειτουργεί και ο συντελεστής A του Robinson (Robinson's A Coefficient). Ο Δείκτης A είναι δημοφιλής για την μέτρηση της αξιοπιστίας βαθμολογητών ποσοτικών μεταβλητών. Για τον Robinson, ως D δηλώνεται η διαφωνία μεταξύ των βαθμολογητών και ομοίως ως D_{max} η μεγαλύτερη δυνατή τιμή της. Ο Δείκτης A προκύπτει ως εξής:

$$A = \frac{D_{max} - D}{D_{max}} = 1 - \frac{D}{D_{max}}$$

Ο συντελεστής A του Robinson είναι καταλληλότερος στις περιπτώσεις, όπου η διαφορά των βαθμολογητών αναμένεται να είναι μεγάλη, και αντίστοιχα η συμφωνία των αποτελεσμάτων τους μικρή.

2.4.1.3 Συντελεστής Ενδοσυσχέτισης

Ο Συντελεστής Ενδοσυσχέτισης (Intraclass Correlation Coefficient - ICC), μετρά την αξιοπιστία των αξιολογήσεων συγκρίνοντας τις διακυμάνσεις των διαφόρων βαθμολογητών για το ίδιο θέμα με τη συνολική διακύμανση, μεταξύ όλων των βαθμολογητών και όλων των θεμάτων. Είναι ένα περιγραφικό στατιστικό μέτρο που χρησιμοποιείται όταν οι μετρήσεις σε ποσοτικές μεταβλητές έχουν γίνει σχετικά με οντότητες που είναι οργανωμένες σε σύνολα. Περιγράφει πόσο έντονα μονάδες στο ίδιο σύνολο μοιάζουν μεταξύ τους. Ενώ θεωρείται ως ένα είδος συσχέτισης, σε αντίθεση με τα περισσότερα άλλα μέτρα συσχέτισης, λειτουργεί σε στοιχεία που διαρθρώνονται σε σύνολα, αντί για ζεύγη παρατηρήσεων. Ο Δείκτης Ενδοσυσχέτισης υπολογίζεται σύμφωνα με την ακόλουθη σχέση:

$$ICC = \frac{\sigma_\alpha}{\sigma_\alpha + \sigma_\varepsilon}$$

όπου ως σ_α συμβολίζεται η διαφορά μεταξύ των βαθμολογητών και ως σ_ε η διακύμανση μεταξύ των βαθμολογητών λόγω του θορύβου. Ο ICC μπορεί να υπολογιστεί για τη συνοχή, δηλαδή την ανεξαρτησία των συστηματικών διαφορών μεταξύ των βαθμολογητών, και την απόλυτη συμφωνία, δηλαδή την σημαντικότητα των συστηματικών διαφορών. Και οι δύο αναδεικνύουν τυχόν τάσεις προς την αρνητικότητα ή θετικότητα, μεμονωμένων βαθμολογητών [74].

ICC	Ερμηνεία
<0.40	Μικρή Συμφωνία
0.40-0.59	Ισότιμη Συμφωνία
0.60-0.74	Ουσιώδης Συμφωνία
0.75-1.00	Πλήρης Συμφωνία

Πίνακας 2.3: Εξάρτηση βαθμού συμφωνίας από τον ICC

2.4.2 Στατιστικά Μέτρα Αξιολόγησης

- Μήτρα Σύγχυσης

Στον τομέα της Μηχανικής Μάθησης και συγκεκριμένα σε προβλήματα στατιστικής ταξινόμησης, ως Μήτρα Σύγχυσης, (Confusion Matrix), ορίζεται ένας ειδικός πίνακας που επιτρέπει την απεικόνιση της απόδοσης ενός αλγορίθμου επιβλεπόμενης μάθησης. Κάθε στήλη του πίνακα αντιπροσωπεύει τις περιπτώσεις μιας προβλεπόμενης κατηγορίας, ενώ κάθε σειρά αντιπροσωπεύει τις περιπτώσεις μίας πραγματικής κατηγορίας, ή το αντίστροφο. Η δομή αυτή καθιστά εύκολο να δούμε αν το σύστημα συγχέει δύο κατηγορίες, δηλαδή συνήθως εσφαλμένη επισήμανση το ένα ως το άλλο). Κατά αναλογία, στις μεθόδους μη επιβλεπόμενης μάθησης χρησιμοποιείται η Μήτρα Ταιριάσματος (Matching Matrix).

Θεωρώντας ότι κάθε κελί της μήτρας C περιγράφει το πλήθος των στοιχείων, που ανήκουν στην πραγματικότητα σε μία κλάση i, και ο ταξινομητής το απέδωσε στην κατηγορία j, οι σωστές κατηγοριοποιήσεις βρίσκονται στη διαγώνιο της μήτρας, ενώ οι εσφαλμένες στα υπόλοιπα κελιά της. Ως εκ τούτου η βέλτιστη απόδοση ενός ταξινομητή, έγκειται στον μηδενισμό των μη διαγώνιων στοιχείων της μήτρας του.

	C_j	C_{j+1}	….	C_m
C_i	$C_{i,j}$	$C_{i,j}$	….	$C_{i,m}$
C_{i+1}	$C_{i+1,j}$	$C_{i+1,j+1}$	….	$C_{i+1,m}$
….	….	….	….	
C_n	$C_{n,j}$	$C_{n,j+1}$	….	$C_{n,m}$

Πίνακας 2.4: Γενικευμένη μορφή Μήτρας Σύγχυσης

Στην δυαδική ταξινόμηση, οι περιπτώσεις προβλέψεων που προκύπτουν, είναι οι ακόλουθες:

1. ***Αληθώς Θετικά (True Positives - TP)***: Πρόκειται για περιπτώσεις στις οποίες έχουμε θετική πρόβλεψη, και η πρόβλεψη επιβεβαιώνεται.
2. ***Αληθώς Αρνητικά (True Negatives - TN)***: Πρόκειται για περιπτώσεις στις οποίες έχουμε αρνητική πρόβλεψη, και η πρόβλεψη επιβεβαιώνεται.
3. ***Ψευδώς Θετικά (False Positives - FP)***: Πρόκειται για περιπτώσεις στις οποίες έχουμε θετική πρόβλεψη, και η πρόβλεψη δεν επιβεβαιώνεται.
4. ***Ψευδώς Αρνητικά (False Negatives - FN)***: Πρόκειται για περιπτώσεις στις οποίες έχουμε αρνητική πρόβλεψη, και η πρόβλεψη δεν επιβεβαιώνεται.

		ΚΑΤΗΓΟΡΙΑ ΤΑΞΙΝΟΜΗΤΗ	
		ΘΕΤΙΚΗ	ΑΡΝΗΤΙΚΗ
ΠΡΑΓΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΑ	ΘΕΤΙΚΗ	TP	FN
	ΑΡΝΗΤΙΚΗ	FP	TN

Πίνακας 2.5: Παράδειγμα δομής Δυαδικής Μήτρας Σύγχυσης

- Ορθότητα και Λόγος Σφάλματος

Ως Ορθότητα (Accuracy) ορίζεται το ποσοστό των επιτυχημένων προβλέψεων ενός ταξινομητή, σε σχέση με το συνολικό χώρο δειγμάτων του:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Στην περίπτωση που περιγράφεται το σφάλμα στην κατηγοριοποίηση ενός ταξινομητή, προτιμότερο μέτρο αποτελεί ο Λόγος Σφάλματος (Error Rate). Εάν η κατηγοριοποίηση πραγματοποιείται πάνω στα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του ταξινομητή ο Λόγος Σφάλματος ταυτίζεται με το Σφάλμα Επαναληπτικής Αντικατάστασης (Resubsitution Error).

$$error\ rate = 1 - \ accuracy = \frac{FP + FN}{TP + TN + FP + FN}$$

- Ακρίβεια και Ανάκληση

Το μέτρο της Ακρίβειας (Precision) εστιάζει σε μία συγκεκριμένη κατηγορία, εκφράζοντας το ποσοστό επιτυχίας, όλων των ταξινομήσεων, σε αυτή. Είναι ουσιαστικά ένας δείκτης πιστότητας της κατηγοριοποίησης κάθε κλάσης. Στην

Ανάλυση Συναισθήματος ως ακρίβεια θεωρείται ο λόγος των σωστά επισημασμένων, ως αληθή, στιγμιότυπων προς το σύνολό τους.

$$precision = \frac{TP}{TP + FP}$$

Αντίστοιχα, το μέτρο της Ανάκλησης (Recall) είναι ένας δείκτης πληρότητας της κατηγοριοποίησης. Όπως και η Ακρίβεια, αναφέρεται σε μία συγκεκριμένη κατηγορία, αλλά στη συγκεκριμένη περίπτωση περιγράφεται το ποσοστό των εύστοχα ταξινομημένων στοιχείων εν συγκρίσει με το συνολικό τους πλήθος. Η απόδοσή της, είναι αντιστρόφως ανάλογη της Ακρίβειας και η αύξησή της, συνεπάγεται μείωση της Ανάκλησης.

$$recall = \frac{TP}{TP + FN}$$

- Ευαισθησία και Εξειδίκευση

Η Ευαισθησία (Sensitivity) ως μέτρο εκφράζει το ποσοστό των θετικών προβλέψεων που έχουν αναγνωριστεί σωστά, και ο ταξινομητής τις έχει αποδώσει στην κατάλληλη κατηγορία. Στη στατιστική ταυτίζεται με τον Λόγο Αναγνώρισης Αληθώς Θετικών Στιγμιότυπων (True Positive Recognition Rate).

$$sensitivity = \frac{TP}{TP + FN}$$

Η Εξειδίκευση (Specificity), σε πλήρη αντιστοιχία με την Ευαισθησία, εκφράζει το ποσοστό των αρνητικών προβλέψεων που έχουν αναγνωριστεί σωστά, και ο ταξινομητής τις έχει αποδώσει στην κατάλληλη κατηγορία. Στη στατιστική ταυτίζεται με τον Λόγο Αναγνώρισης Αληθώς Αρνητικών Στιγμιότυπων (True Negative Recognition Rate).

$$specificity = \frac{TN}{TN + FP}$$

Μετά τον παραπάνω ορισμό, η Ορθότητα μπορεί να εκφραστεί επίσης, ως γραμμικός συνδυασμός της Ευαισθησίας και της Εξειδίκευσης, ως ακολούθως:

$$accuracy = \frac{TP + FN}{TP + TN + FP + FN} \times sensitivity + \frac{TN + FP}{TP + TN + FP + FN} \times specificity$$

- Σταθμισμένος Αρμονικός Μέσος

Η εξάρτηση των μεγεθών της Ευαισθησίας και της Εξειδίκευσης συχνά οδηγεί στην χρήση ενός ακόμη στατιστικού μέτρου, του Σταθμισμένου Αρμονικού Μέσου (F-Measure). Παρόλα αυτά, η μη αξιοποίηση του πλήθους των αληθώς αρνητικών προβλέψεων, στον υπολογισμό του, μειώνει την απόδοση του στην δυαδική ταξινόμηση, συγκριτικά με τους συντελεστές Κάπα.

$$Fmeasure = \frac{2 \times precision \times recall}{precision + recall} = \frac{2 \times TP}{(2 \times TP + FP + FN)}$$

- Συντελεστής Συνάφειας του Matthews

Ο συντελεστής Συνάφειας του Matthews (Matthews Correlation Coefficient) χρησιμοποιείται στη Μηχανική Μάθηση, ως μέτρο της ποιότητας δυαδικών ταξινομήσεων. Λαμβάνει υπόψη τόσο τις αληθείς, όσο και τις ψευδείς, θετικές και αρνητικές προβλέψεις και θεωρείται ως ένα ισορροπημένο μέτρο, το οποίο μπορεί να χρησιμοποιηθεί ακόμα και αν οι κατηγορίες είναι διαφορετικού μεγέθους. Στην ουσία, είναι ένας συντελεστής συσχέτισης μεταξύ των παρατηρούμενων και των προβλεπόμενων ταξινομήσεων, επιστρέφοντας μια τιμή μεταξύ -1 και $+1$, ανάμεσα στην πλήρη διαφωνία με την πραγματικότητα και την τέλεια πρόβλεψη, αντίστοιχα.

$$MCC = \frac{TP \times TP - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

3. Λεξικά Συναισθήματος

3.1 WordNet

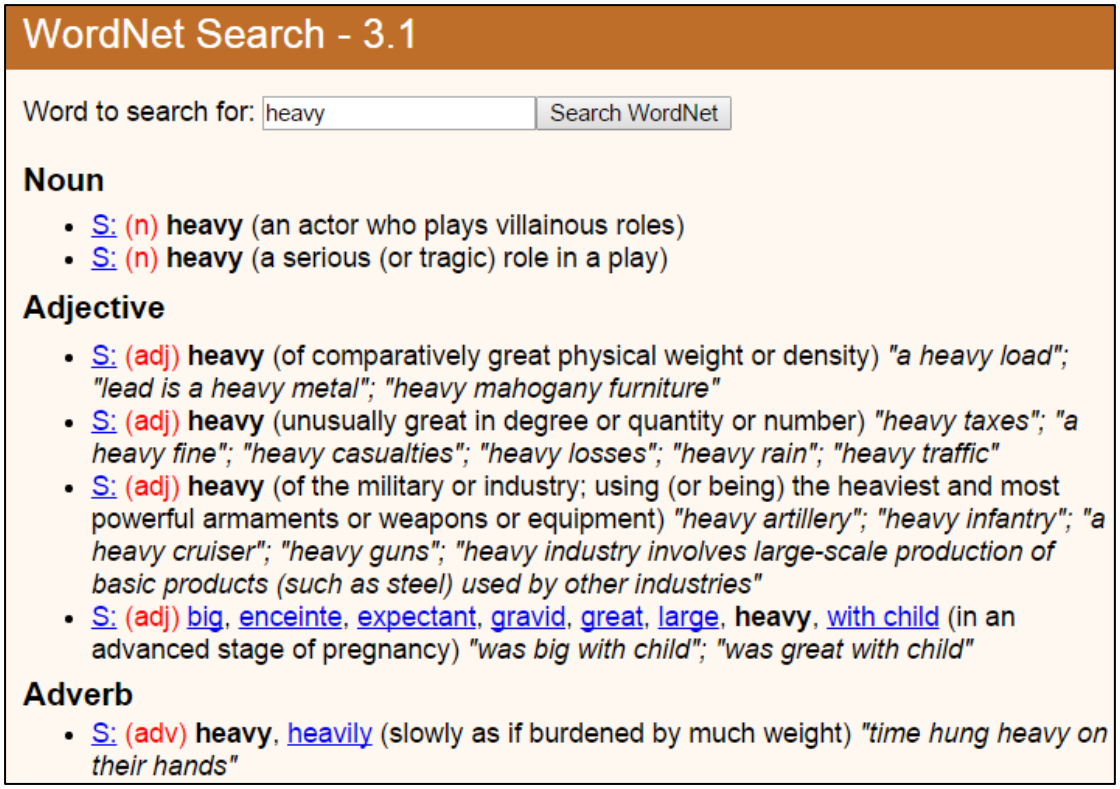
Το WordNet είναι ένας λεξικολογικός πόρος, που αποτέλεσε καθοριστικό εργαλείο σε μία πλειάδα μεθόδων Ανάλυσης Συναισθήματος. Αναπτύχτηκε στο Εργαστήριο Γνωσιακής Επιστήμης του Πανεπιστήμιου του Πρίνστον, ως μία μεγάλη βάση δεδομένων λέξεων της αγγλικής γλώσσας, συνιστώντας ένα ανοιχτού λογισμικού, ηλεκτρονικό σημασιολογικό λεξικό, το οποίο συνδυάζει τις ιδιότητες μιας γνωσιακής βάσης και μιας ιεραρχικής δομής εννοιών. Πιο συγκεκριμένα, για κάθε λέξη που περιέχεται στο λεξικό, παρέχεται η έννοια στην οποία αναφέρεται αλλά και οι σημασιολογικές σχέσεις που έχει με τις υπόλοιπες λέξεις.

Δομικές μονάδες του WordNet είναι ουσιαστικά, ρήματα, επίθετα και επιρρήματα, ο συνδυασμός των οποίων σχηματίζει σύνολα συνωνύμων (synsets), εάν η ερμηνεία τους ταυτίζεται. Κάθε ερμηνεία λέξης συνοδεύεται από τη συχνότητα με την οποία παρουσιάζεται η συγκεκριμένη έννοια στις πρωτότυπες αρχειακές πηγές των ερευνητών του Πρίνστον, φράσεις που λειτουργούν ως υπόδειγμα (gloss) για την ορθή χρήση της λέξης, με βάση την συγκεκριμένη ερμηνεία και τα σύνολα συνωνύμων τα οποία συνιστούν παρεμφερείς ερμηνείες. Κάθε synset είναι μία έκφραση εννοιολογικής μοναδικότητας, και παρόλο που μία πολύσημη λέξη μπορεί να ανήκει αντίστοιχα σε πολλά synsets, σε κάθε διαφορετικό περιβάλλον αποκτά μία εντελώς διαφορετική ερμηνευτική αξία.

Κάθε λέξη ανήκει τουλάχιστον σε έναν από τους βασικούς τύπους σχέσεων, στην ιεραρχία των εννοιών που υποστηρίζει το λεξικό:

1. **Συνώνυμο (Synonym):** Η έννοια της λέξης ταυτίζεται με την έννοια κάποιας άλλης, και μπορεί να την αντικαταστήσει χωρίς σημασιολογική παρέκκλιση.
2. **Υπερώνυμο (Hypernym):** Η έννοια της λέξης είναι γενικότερη σε σχέση με λέξεις σε κατώτερο επίπεδο της ιεραρχίας.
3. **Υπώνυμο (Hyponym):** Η έννοια της λέξης είναι ειδικότερη σε σχέση με λέξεις σε ανώτερο επίπεδο της ιεραρχίας.
4. **Ολώνυμο (Holonym):** Η έννοια της λέξης είναι κατηγορία των λέξεων στο άμεσα κατώτερο επίπεδο της ιεραρχίας.
5. **Μερώνυμο (Meronym):** Η έννοια της λέξης είναι είδος της λέξης στο άμεσα ανώτερο επίπεδο της ιεραρχίας.
6. **Πεδίο (Domain):** Η ιεραρχική θεματική ενότητα στην οποία υπάγεται η έννοια της λέξης.
7. **Συντεταγμένοι όροι (Coordinate terms):** Λέξεις που υπάγονται σε κοινό Υπερώνυμο.

Η συνεχής ανάπτυξη του και η διεύρυνση των λεξικολογικών του πόρων, καθιστούν το WordNet, ένα από τα κορυφαία ηλεκτρονικά ερμηνευτικά λεξικά, το οποίο στην τρίτη έκδοσή του (WordNet 3.1, 2012) περιλαμβάνει πλέον των 150.000 αυτοτελών λέξεων, 117.000 ομάδες συνωνύμων, με παραπάνω από 200.000 σχέσεις μεταξύ τους. Η βάση των δεδομένων του διατίθεται ελεύθερα στο διαδίκτυο και τα σχετιζόμενα με αυτήν εργαλεία έχουν κατοχυρωμένα δικαιώματα χρήσης κάτω από την προστασία μίας BSD άδειας. Η μεγάλη απήχηση του οδήγησε σύντομα στην δημιουργία άλλων ταυτόσημων λεξικογραφικών βάσεων, όπως το WordNets και το BalkaNet, με όμοια χαρακτηριστικά και διαφοροποιήσεις στην αριθμητική ευρύτητα των αναλυόμενων όρων.



Σχήμα 3.1: Παράδειγμα αποτελέσματος αναζήτησης στο WordNet[7]

Ο όγκος των δεδομένων που παράγονται σε κάθε αναζήτηση λεκτικού όρου από το χρήστη, αυξάνονται εκθετικά, γεγονός που δικαιολογείται από τη δενδρική δομή που αποκτά η πληροφορία και την εξάρτησή της, από το πλήθος των δεσμών και των σχέσεων μεταξύ των λέξεων και των synsets. Η ποικιλομορφία των εννοιών που καταγράφει το WordNet, καθώς και οι λεπτές ειδοποιές διαφορές που τις χαρακτηρίζουν, αποτελούν το κυριότερο αίτιο της απόκλισης του τελικού

[7] http://wordnetweb.princeton.edu/perl/webwn

αποτελέσματος, σε σύγκριση με την δυνητικά αναζητούμενη έννοια, που οφείλεται στην αδυναμία ακριβής διατύπωσης της, από το χρήστη.

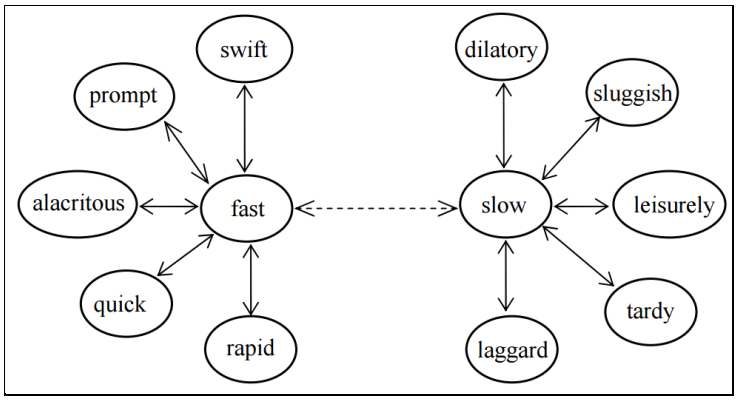
3.2 Μέθοδοι Κατασκευής Λεξικών Συναισθήματος

Οι λίστες λέξεων που εκφράζουν συναίσθημα, είναι μείζονος σημασίας για την κατασκευή των Λεξικών Συναισθήματος. Με τον όρο λέξεις συναισθήματος (sentiment word) ονομάζονται οι λέξεις που εκφράζουν άποψη (opinion words) στις οποίες περιλαμβάνονται λέξεις με δύο βασικές ιδιότητες, την υποκειμενικότητα και την πολικότητα. Η υποκειμενικότητα αναφέρεται στην υποκειμενική ή αντικειμενική σημασία κάθε όρου, και η πολικότητα στη θετική ή αρνητική χροιά κάθε υποκειμενικής λέξης. Οι λέξεις θετικού συναισθήματος χρησιμοποιούνται για να εκφράσουν κάποιες επιθυμητές καταστάσεις ή ιδιότητες, ενώ οι λέξεις αρνητικού συναισθήματος χρησιμοποιούνται για να εκφράσουν κάποιες ανεπιθύμητες καταστάσεις ή ιδιότητες. Παραδείγματα λέξεις θετικής πολικότητας είναι οι «beautiful, wonderful» και «amazing». Παραδείγματα λέξεων αρνητικής πολικότητας είναι οι «bad, awful» και «poor». Εκτός από μεμονωμένες λέξεις, υπάρχουν και φράσεις συναισθημάτων και ιδιωματισμοί.

Οι λέξεις συναισθήματος μπορούν να χωριστούν σε δύο κατηγορίες, σε βασικού τύπου και συγκριτικού τύπου. Όλα τα παραπάνω παραδείγματα είναι βασικού τύπου. Οι λέξεις συναισθήματος συγκριτικού τύπου (οι οποίες περιλαμβάνουν και τον υπερθετικό τύπο) χρησιμοποιούνται για να εκφράσουν απόψεις συγκριτικά. Παραδείγματα τέτοιων λέξεων είναι οι «better, worse, best, worst», οι οποίες είναι ο συγκριτικός ή ο υπερθετικό βαθμός των επιθέτων ή επιρρημάτων, που ανήκουν στις βασικού τύπου. Σε αντίθεση με τις λέξεις συναισθήματος βασικού τύπου, οι λέξεις συναισθήματος συγκριτικού τύπου δεν εκφράζουν μια συνηθισμένη γνώμη σχετικά με μια οντότητα, αλλά μια συγκριτική γνώμη για περισσότερες από μία οντότητες. Αυτή η φράση δεν εκφέρει γνώμη, δηλαδή ότι κάποιο από τα δύο στοιχεία είναι καλό ή κακό. Η πλειονότητα των λεξικών συναισθήματος που χρησιμοποιούνται στα αυτοματοποιημένα συστήματα ανάλυσης συναισθήματος και εξόρυξης γνώμης από κειμενικές πηγές, συνίστανται σε εκτενείς λίστες των παραπάνω δύο τύπων, πρωτεύοντα ρόλο στις οποίες έχουν οι λέξεις συναισθήματος βασικού τύπου.

Οι περισσότερες μέθοδοι για τη δημιουργία λεξικών συναισθήματος, έχουν ως κοινό άξονα την δειγματοληπτική επιλογή λεξικολογικών μονάδων, από γλωσσικές βάσεις δεδομένων, όπως το WordNet, την διαλογή τους με κριτήριο την υποκειμενικότητά τους, το διαχωρισμό τους σε υποομάδες σύμφωνα με την πολικότητά τους και την αξιοποίηση των ιεραρχικών δομών για την επέκταση του αρχικού συνόλου δειγματοληψίας σε ένα ευρύτερο σύνολο με σαφείς σημασιολογικές ιδιότητες.

Πιο συγκεκριμένα, οι πρώτες κατευθύνσεις σε αυτή τη μεθοδολογία δόθηκαν από την τεχνική που εφάρμοσαν οι Hu and Liu (2004) [40], για τη δημιουργία ενός Opinion Lexicon. Διαπιστώνοντας ότι τα επίθετα που εκφράζουν υποκειμενικότητα έχουν κοινό θετικό ή αρνητικό χαρακτήρα με τα συνώνυμά τους, αλλά και ότι για κάθε ένα τέτοιο επίθετο είναι δυνατόν να εντοπιστεί ένα άλλο, αντώνυμο με ένα αντίστοιχο σύνολο συνωνύμων, επέλεξαν ένα αρχέτυπο σύνολο επιθέτων-σπόρων (seed set). Χρησιμοποιώντας αυτά τα επίθετα και τις σχέσεις σημασιολογικής συνάφειας που περικλείονται στα synsets του WordNet, που ανήκουν, προέβησαν στην αξιολόγηση ενός πολύ μεγαλύτερου συνόλου λέξεων που άγγιξε τους 6500 όρους.



Σχήμα 3.2: Διαδικασία κατασκευής Λεξικού Συναισθήματος

Στο ίδιο μήκος κύματος κινούνται οι προσπάθειες για την περαιτέρω ανάλυση της πολικότητας της κάθε λέξης, σύμφωνα με τις οποίες κάθε όρος δεν χαρακτηρίζεται μονάχα από την θετική ή αρνητική χροιά, αλλά και από τα συναισθήματα που προκαλεί στο χρήστη (αηδία, φόβος, λύπη, έκπληξη, χαρά) αλλά και την ένταση η οποία τα διακρίνει [73]. Συνεπώς, οι λέξεις-σπόροι του WordNet, διαχωρίζονται σε ένα μεγαλύτερο αριθμό υποομάδων και συνοδεύονται από μία τιμή που εκφράζει το θετικό ή αρνητικό σημασιολογικό μέγεθος που κρύβει (PosScore/NegScore), με βαθμολογίες τριών ή πέντε αστέρων [30]. Η ομαδοποίηση και ο χαρακτηρισμός επεκτείνεται ενίοτε όχι μόνο στις λεκτικές μονάδες αλλά και σε ολόκληρα synsets, με την αντίστοιχη διαβάθμιση στην ποιότητα και την ένταση του εξαχθέντος συναισθήματος.

Ο συνδυασμός των παραπάνω μεθόδων με ήδη υπάρχοντα λεξικά συναισθήματος οδήγησε στη δημιουργία ακόμη πιο ισχυρών λεξικών, τα οποία ως σπόρους είχαν ολόκληρα σημασιολογικά προσημασμένα σύνολα λέξεων. Χαρακτηριστικό παράδειγμα αποτελεί η εργασία των Cerini et al. (2007) [18], όπου ένα δείγμα 100 όρων από ήδη υπάρχοντος λεξικού συναισθήματος, επεκτάθηκε σε ένα σύνολο 1105 synset του WordNet, τριών επιμέρους υποομάδων πολικότητας (Common, Group1,Group2). Κάθε λέξη του παραγόμενου λεξικού (Micro-WNOp) φέρει χαρακτηριστικά που πηγάζουν από το ερμηνευτικό και το συναισθηματικό λεξικό από το οποίο προέρχεται, όπως το μέρος του λόγου στο οποίο ανήκει, η ερμηνεία και η κατεύθυνση του synset της.

	Adjectives	Nouns	Verbs	Adverbs	Total
WordNet	18563 (16%)	79689 (69%)	3664 (3%)	13508 (12%)	115424
Whole Micro-WNOp	284 (26%)	500 (45%)	32 (3%)	289 (26%)	1105
Micro-WNOp *Common*	28 (25%)	51 (46%)	2 (2%)	29 (26%)	110
Micro-WNOp *Group1*	138 (28%)	214 (43%)	9 (2%)	135 (27%)	496
Micro-WNOp *Group2*	118 (24%)	235 (47%)	21 (4%)	125 (25%)	499

Πίνακας 3.1: Διαμέριση synset του WORDNET στις ομάδες του Micro-WNOp[8]

Η εκτεταμένη χρήση αυτού του είδους των λεξικών στην ανάλυση συναισθημάτων απέδειξε ότι καθοριστικός παράγοντας για την ορθή απόφανση περί του σημασιολογικού προσανατολισμού ενός κειμένου δεν αποτελεί μόνο η εξακρίβωση την πολικότητας κάθε όρου, αλλά και οι ενδογενείς αποχρώσεις που λαμβάνει και εξαρτώνται από το νοηματικό περιβάλλον στο οποίο βρίσκεται. Κατά αυτόν τον τρόπο, πολλές μέθοδοι επικεντρώνονται στην ανάπτυξη λεξικών αξιοποιώντας λεξιλογικούς πόρους που αναφέρονται σε συγκεκριμένα γνωσιακά πεδία, επενδύοντας στην αποδοτικότητα της συμφραστικής πολικότητας (contextual polarity), δηλαδή της ιδιαίτερης σημασιολογικής κλίσης που αποκτά μία λέξη μέσα στον τομέα που συναντάται, η οποία ενίοτε διαφέρει από την a priori πολικότητά της [83]. Τα λεξικά αυτά επιτυγχάνουν υψηλά ποσοστά ακρίβειας στην απόδοση εννοιών μεμονωμένων πεδίων, όπως η οικονομία, η τέχνη ή η πολιτική.

3.3 Λεξικά Συναισθήματος με Διαβαθμίσεις

3.3.1 SentiWordNet

Το SentiWordNet είναι ένας λεξικολογικός πόρος που έχει σχεδιαστεί για χρήση σε εφαρμογές ανάλυσης συναισθήματος πολυμεσικού περιεχομένου. Συνδέεται με ένα web-based GUI και έχει κατά κόρον αξιοποιηθεί σε ποικίλα ερευνητικά προγράμματα παγκοσμίως. Ακρογωνιαίος λίθος στη δημιουργία του, είναι το WordNet και η αυτοματοποιημένη διαδικασία σχολιασμού των synset που περιέχει.
Ομάδες συνωνύμων, δεδομένης πόλωσης, αρχικά συνδέονται με άλλες αντίστοιχες ομάδες και στη συνέχεια αξιοποιούνται οι ορισμοί τους (glosses) στην εκμάθηση ενός ταξινομητή κλάσης (classifier) και κάθε synset χαρακτηρίζεται με μία σημασιολογική ετικέτα (Pos, Neg,Obj). Η επαναληπτική εφαρμογή της διαδικασίας με αντίστοιχη μεταβολή της ακτίνας, δηλαδή του αριθμού των συσχετιζόμενων synset σε κάθε επανάληψη, σε συνδυασμό με ημιαυτόματες μεθόδους μάθησης παράγουν οκτώ τριμερείς ταξινομητές, των οποίων η ακρίβεια είναι παρεμφερής αλλά η συμπεριφορά τους κατά την ταξινόμηση διαφέρει. Πέραν από την μάθηση χαμηλής εποπτείας, εφαρμόζεται ένας αλγόριθμος τυχαίου περιπάτου, που μετουσιώνει τη δενδρική δομή των synset σε γράφους, συντάσσοντας δύο διαφορετικές τιμές πολικότητας σε κάθε ένα από αυτά, που εν τέλει κανονικοποιούνται στο διάστημα [0,1].

[8] http://www-3.unipv.it/wnop

Η παραπάνω διαδικασία μόχλευσης των λεξικογραφικών πόρων του WordNet συντελεί στην δημιουργία ενός διαβαθμισμένου λεξικού συναισθήματος, όπου κάθε όρος χαρακτηρίζεται όχι μόνο από την υποκειμενικότητά του αλλά και την έντασή της. Η τιμή 1 αναφέρεται σε έναν απόλυτα υποκειμενικό όρο, ενώ η τιμή 0 σε έναν ουδέτερο, σύμφωνα με τη γενική αρχή: Objective = 1 - (Positive + Negative). Αναλύοντας ποσοτικά τους ορισμούς των synset και όχι τις ίδιες τις λέξεις που τα αποτελούν, το SentiWordNet καταφέρνει να εξετάσει πλουραλιστικά μία μεγάλη γκάμα ερμηνειών που ενδεχομένως αποδίδονται στην ίδια λέξη, σε διαφορετικά εκφραστικά χωρία [30].

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	00001740	0.125	0	able#1	(usually followed by `to') having the necessary means or [...]
a	00002098	0	0.75	unable#1	(usually followed by `to') not having the necessary means or [...]
a	00002312	0	0	dorsal#2 abaxial#1	facing away from the axis of an organ or organism; [...]
a	00002527	0	0	ventral#2 adaxial#1	nearest to or facing toward the axis of an organ or organism; [...]
a	00002730	0	0	acroscopic#1	facing or on the side toward the apex
a	00002843	0	0	basiscopic#1	facing or on the side toward the base
a	00002956	0	0	abducting#1 abducent#1	especially of muscles; [...]
a	00003131	0	0	adductive#1 adducting#1 adducent#1	especially of muscles; [...]
a	00003356	0	0	nascent#1	being born or beginning; [...]
a	00003553	0	0	emerging#2 emergent#2	coming into existence; [...]

Σχήμα 3.3: Παράδειγμα δομής SentiWordNet[9]

3.3.2 Bing Liu's Opinion Lexicon

Το λεξικό του Bing Liu αναπτύχθηκε στο πλαίσιο της ευρύτερης εργασίας του για την εξόρυξη γνώμης και τον εντοπισμό της πολικότητάς της, σε κειμενικές πηγές που προέρχονταν από κριτικές πελατών για προϊόντα και υπηρεσίες. Χρησιμοποιεί μία καινοτόμο μέθοδο αλλά και συνάμα απλή ως προς τη σύλληψη της, για τη δημιουργία μίας λίστας θετικών και αρνητικών λέξεων, της Αγγλικής γλώσσας [41].
Ξεκινώντας από μία μικρή ομάδα επιθέτων-σπόρων, αξιοποίησε τις διπολικές ιδιότητες των στοιχείων του WordNet που διαχωρίζονται σε αμφίσημες συστάδες όρων (πχ. fast/slow) για τον εντοπισμό υποσυστάδων και μέσω αυτών νέων επιθέτων. Κάθε ημισυστάδα (half cluster) έχει ως κεφαλή ένα synset και ένα σύνολο περιφερειακών αυτής ομάδων συνώνυμων. Αντίστοιχα η υπολειπόμενη ημισυστάδα δημιουργεί ένα σύνολο synset αντωνύμων. Η στρατηγική του περιελάμβανε 30 επίθετα που παρουσιάζονται με μεγάλη συχνότητα, των οποίων ο σημασιολογικός προσανατολισμός καθορίστηκε χειροκίνητα (πχ great, fantastic, nice, cool, bad, dull), και μία επαναληπτική διαδικασία για την εισαγωγή στο λεξικό νέων όρων. Ως αποτέλεσμα αυτής της μεθόδου είχε δύο λίστες θετικών και αρνητικών επιθέτων, με 2000 και 4800 επίθετα αντίστοιχα, στα οποία περιλαμβάνονται όχι μόνο δόκιμοι όροι

[9] http://sentiment.christopherpotts.net/lexicons.html#resources

αλλά και αντικανονικές μορφές που συναντούνται συχνά στα κείμενα χρηστών του διαδικτύου (αργκό, ορθογραφικά λάθη, λεκτικοί εκφυλισμοί)[10].

3.3.3 ANEW

Το λεξικό ANEW (Affective Norms for English Words) αναπτύχθηκε για να παρέχει ένα σύνολο κανονιστικών συναισθηματικών βαθμολογιών για ένα μεγάλο αριθμό λέξεων στην αγγλική γλώσσα, που έχει αξιολογηθεί από την άποψη της ευχαρίστησης, της διέγερσης, και της κυριαρχίας. Συμπλήρωσε το Διεθνές Σύστημα Συναισθηματικής Εικόνας (IAP), και το Διεθνές Σύστημα Ψηφιοποιημένων Συναισθηματικών Ήχων (IADS), οι οποίες είναι συλλογές ερεθισμάτων εικόνας και ήχου, που αναπτύχθηκαν και διανέμονταν από το Κέντρο Συναισθήματος και Προσοχής (CSEA), προκειμένου να παρέχουν τυποποιημένο υλικό για περαιτέρω ερευνητικές μελέτες του συναισθήματος και της προσοχής. Για την αξιολόγηση των τριών παραπάνω συναισθηματικών διαστάσεων εφαρμόζεται μία συναισθηματική βαθμολογία, σύμφωνα με το σύστημα Self-Assessment Manikin, με διπολικές κλίμακες 9 σημείων, που απεικονίζουν διαφορετικές τιμές κατά μήκος κάθε συναισθηματικής διάστασης. Η μέση βαθμολογία χαράς κάθε λέξης κυμαίνεται από το πολύ δυσάρεστο μέχρι το πολύ ευχάριστο, και κατανέμονται ομοιόμορφα σε όλο το δείγμα, και στις ουδέτερες λέξεις αποδόθηκαν χαμηλές τιμές στον τομέα της διέγερσης. Η βαθμολόγηση κάθε όρου πραγματοποιήθηκε από φοιτητές Ψυχολογίας, σε ένα σύνολο 1040 λέξεων.

Affective Norms for English Words. **Female Subjects**

Description	Word No.	Valence Mean(SD)	Arousal Mean(SD)	Dominance Mean (SD)	Word Frequency
excellence	151	8.41 (1.02)	5.83 (2.78)	7.24 (2.40)	15
excitement	152	7.83 (2.09)	8.07 (1.46)	6.48 (2.29)	32
excuse	153	4.19 (1.02)	4.20 (2.02)	4.32 (1.95)	27
execution	154	1.78 (1.54)	5.83 (2.77)	4.00 (2.95)	15
exercise	155	7.26 (1.36)	6.65 (1.99)	5.43 (2.25)	58
fabric	742	5.38 (0.97)	4.32 (2.15)	5.24 (1.76)	15
face	556	6.65 (1.41)	5.31 (2.00)	6.04 (1.59)	371
failure	156	1.36 (0.81)	5.52 (2.86)	2.24 (2.33)	89
fall	743	4.32 (2.23)	4.81 (2.32)	3.71 (2.17)	147
FALSE	744	3.30 (1.22)	3.60 (1.96)	3.80 (1.44)	29
fame	157	7.93 (1.31)	7.14 (1.96)	7.00 (2.11)	18
family	158	7.77 (1.37)	4.64 (2.81)	6.00 (1.92)	331
famous	745	6.69 (2.49)	5.68 (2.66)	6.12 (2.27)	89
fantasy	746	7.42 (2.16)	4.54 (2.94)	6.23 (2.14)	14
farm	557	5.65 (1.74)	4.19 (1.96)	5.50 (1.94)	125
fascinate	159	7.44 (1.39)	5.72 (2.82)	6.38 (1.98)	3
fat	160	1.75 (1.62)	6.50 (2.54)	3.60 (3.07)	60
father	161	7.26 (2.24)	6.22 (2.35)	6.83 (2.21)	383

Σχήμα 3.4: Παράδειγμα δομής ANEW

3.3.4 AFFIN

Ο Arup Finn Nielsen είναι ο δημιουργός του λεξικού συναισθήματος AFINN, για την ανάλυση συναισθήματος σε κειμενικές πηγές του διαδικτύου. Το λεξικό AFFIN είναι

[10] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

μία λίστα αγγλικών λέξεων, βαθμονομημένων με μία κλίμακα μεταξύ -5 (αρνητικό μέγιστο) και 5 (θετικό μέγιστο), βαθμολογώντας με 0 οποιαδήποτε λέξη εκτός της βάσης δεδομένων. Οι λέξεις βαθμολογήθηκαν από τον ίδιο το συγγραφέα στο διάστημα 2009-2011. Η αρχική έκδοση του λεξικού (AFINN-96) περιελάμβανε 1468 μοναδικούς όρους, χωρίς αλφαβητική διάταξη. Η δεύτερη επαυξημένη έκδοση (AFINN-111) περιλαμβάνει συνολικά 2477 λέξεις και φράσεις, και βρίσκεται κάτω από άδεια χρήσης Open Database (ODbL)[11].

abandon	-2	accepts	1	admiring	3	abhorred	-3
abandoned	-2	accident	-2	admit	-1	abhorrent	-3
abandons	-2	accidental	-2	admits	-1	abhors	-3
abducted	-2	accidentally	-2	admitted	-1	abilities	2
abduction	-2	accidents	-2	admonish	-2	ability	2
abductions	-2	accomplish	2	admonished	-2	aboard	1
abhor	-3	accomplished	2	adopt	1	absentee	-1

Σχήμα 3.5: Παράδειγμα διαβάθμισης πολικότητας στο AFFIN

3.4 Λεξικά Συναισθήματος χωρίς Διαβαθμίσεις

3.4.1 MPQA

Το Multi-perspective Question Answering (MPQA) Λεξικό Υποκειμενικότητας αποτελεί τον πυρήνα της εφαρμογής ανάλυσης συναισθήματος Opinion Finder. Περιλαμβάνει μεθόδους επέκτασης μίας υπάρχουσας λίστας περισσότερων από 8000 λέξεων, που εκφράζουν υποκειμενικότητα (subjectivity clues) με χρήση πόρων με γλωσσολογικά χαρακτηριστικά, οι οποίες αρχικά ομαδοποιούνται σύμφωνα με την αξιοπιστία τους και ταξινομούνται σημασιολογικά. Κάθε λέξη περιέχει στοιχεία για το μήκος την συναισθηματική της κλίση, τα παραγλωσσικά μοτίβα της (μέρος του λόγου, μορφή), και την αρχέτυπη πολικότητά της (prior polarity), ανεξάρτητα από το γνωσιακό αντικείμενο στο οποίο εντοπίζεται.

[11] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

<table>
<tr><td>Word Features
word token
word part-of-speech
word context
prior polarity: positive, negative, both, neutral
reliability class: strongsubj or weaksubj</td><td rowspan="2">Sentence Features
strongsubj clues in current sentence: count
strongsubj clues in previous sentence: count
strongsubj clues in next sentence: count
weaksubj clues in current sentence: count
weaksubj clues in previous sentence: count
weaksubj clues in next sentence: count
adjectives in sentence: count
adverbs in sentence (other than not): count
cardinal number in sentence: binary
pronoun in sentence: binary
modal in sentence (other than will): binary</td></tr>
<tr><td>Modification Features
preceeded by adjective: binary
preceeded by adverb (other than not): binary
preceeded by intensifier: binary
is intensifier: binary
modifies strongsubj: binary
modifies weaksubj: binary
modified by strongsubj: binary
modified by weaksubj: binary</td></tr>
<tr><td>Document Feature
document topic</td><td>Structure Features
in subject: binary
in copular: binary
in passive: binary</td></tr>
</table>

Σχήμα 3.6: Χαρακτηριστικά ουδέτερης-πολικής κατάταξης στο MPQA

Οι πειραματικές μελέτες των Riloff και Wiebe (2005) [83] αποτέλεσαν τη λυδία λίθο για την δημιουργία μίας εφαρμογής αυτόματης διάκρισης της πολικότητας μίας λέξης, αξιοποιώντας τόσο την πρότερη της εγνωσμένη κλίση, όσο και την αξιολόγηση της με βάση τα συμφραζόμενα της. Ξεκινώντας με ένα μεγάλο σύνολο από ρίζες δεδομένου σημασιολογικού προσανατολισμού, προσδιορίζεται η συμφραστική πολικότητα που περιέχουν οι περιπτώσεις των εν λόγω στοιχείων, στο εξεταζόμενο σώμα κειμένου. Χρησιμοποιείται μια διαδικασία δύο σταδίων που χρησιμοποιεί μηχανική μάθηση και μια ποικιλία άλλων χαρακτηριστικών. Το πρώτο βήμα ταξινομεί κάθε φράση που περιέχει μια τέτοια ένδειξη, ως ουδέτερη ή πολική. Το δεύτερο βήμα λαμβάνει όλες τις φράσεις που σημειώνονται στο πρώτο στάδιο και αποσαφηνίζει την πολικότητα τους (θετική, αρνητική, και τα δύο, ή ουδέτερη). Με την προσέγγιση αυτή, το σύστημα είναι σε θέση να προσδιορίσει αυτόματα την συμφραστική πολικότητα για ένα μεγάλο υποσύνολο εκφράσεων συναισθήματος, επιτυγχάνοντας αποτελέσματα που είναι σαφώς καλύτερα από την αρχική της τιμή.

3.4.2 Linguistic Inquiry and Word Count

Η ανάγκη για μελέτη της ανθρώπινης προσωπικότητας όπως αυτή εκφράζεται μέσα από τα διάφορα λεκτικά στοιχεία, οδήγησε στη δημιουργία μιας εφαρμογής ανάλυσης κειμένου, στον πυρήνα της οποίας βρίσκεται το λεξικό συναισθημάτων Γλωσσικής Έρευνας και Λεκτικής Καταμέτρησης, ή όπως αρχικά αναφέρεται στην ερευνητική μελέτη της γλώσσας των Francis and Pennebaker (1993) ως Linguistic Inquiry and Word Count. Η πιο εξελιγμένη έκδοση της εφαρμογής (LIWC2007) περιλαμβάνει πάνω από 4000 λέξεις, οι οποίες συνοδεύονται από μία εντυπωσιακή πλειάδα χαρακτηριστικών που αφορούν είτε τα καθαρά σημασιολογικά τους στοιχεία είτε τις συντακτικές τους ιδιότητες, συντελώντας στη δημιουργία υπολεξικών ψυχολογικών καταστάσεων και γραμματολογικών δομών, σε 82 διαφορετικές γλώσσες.
Σύμφωνα με την μέθοδο ανάλυσης που εφαρμόζει, κάθε λέξη εξετάζεται μονομερώς και κατατάσσεται σε μία από τις επιμέρους κατηγορίες ανάλογα με το σημασιολογικό

της περιεχόμενο αυξάνοντας ταυτόχρονα τον αντίστοιχο μετρητή του ποσοστού της κατηγορίας ή των κατηγοριών στις οποίες ανήκει. Αυτές οι κατηγορίες αναφέρονται και σε στοιχεία της προσωπικότητας (δραστήριος, οκνηρός) και σε ψυχολογικά χαρακτηριστικά (λυπημένος, ευδιάθετος). Μετά την ανάλυση κάθε λέξης αυξάνονται και οι μετρητές των συνολικών γραμματικών και συντακτικών φαινομένων του κειμένου (μετρητές σημείων στίξης, άρθρων, ρημάτων σε παρελθοντικό χρόνο). Ως τελικό αποτέλεσμα στον χρήστη αποδίδεται μία περιγραφή το αρχικού κειμενικού υλικού από 80 μεταβλητές που αφορούν εκτός των άλλων, τον συνολικό αριθμό λέξεων, το μήκος των λέξεων, την λεκτική περιεκτικότητα κάθε πρότασης, τα μέρη του λόγου, τους ψυχολογικούς συντελεστές, τα στοιχεία προσωπικών δραστηριοτήτων και τους όρους που αφορούν τον κοινωνικό χώρο.

LIWC Results

Details of Writer: 28 year old Female
Date/Time: 24 February 2015, 10:26 pm

LIWC Dimension	*Your Data*	*Personal Texts*	*Formal Texts*
Self-references (I, me, my)	7.17	11.4	4.2
Social words	8.63	9.5	8.0
Positive emotions	3.84	2.7	2.6
Negative emotions	1.34	2.6	1.6
Overall cognitive words	7.84	7.8	5.4
Articles (a, an, the)	3.81	5.0	7.2
Big words (> 6 letters)	12.64	13.1	19.6

The text you submitted was 20580 words in length.

Σχήμα 3.7: Παράδειγμα αποτελέσματος ανάλυσης κειμένου με το LIWC[12]

Όπως είναι αναμενόμενο μία τόσο σύνθετη διάσταση αξιολόγησης των λεκτικών όρων, είναι αποτέλεσμα διαδοχικών σταδίων εξέλιξης και τεκμηρίωσης κάθε ενός όρου ξεχωριστά και επαναληπτικά. Η συνολική διαδικασία κατασκευής του λεξικού περιελάμβανε τα στάδια της συλλογής των λεκτικών πόρων, των ριζών και των λημμάτων τους, της αξιολόγησης από ανεξάρτητους κριτές τόσο κατά την επιλογή όσο και κατά την διαγραφή των πόρων, την ψυχομετρική κατάταξη των ιδιοτήτων τους και την συνεχή ενημέρωση των λεξικολογικών βάσεων δεδομένων του, με πηγές γραπτού αλλά και προφορικού λόγου [64].

3.4.3 General Inquirer

Το λεξικό συναισθημάτων General Inquirer, έχει καταρτιστεί από το Πανεπιστήμιο του Harvard, ως ένα σύνολο λιστών κατάταξης λέξεων στην Αγγλική γλώσσα. Η πρωτόλεια ιδέα για τη σύνθεσή του βασίστηκε στο μοτίβο συναισθηματικών καταστάσεων Osgood. [59] και αποτελεί δομικό συστατικό μιας ευρύτερης εφάρμογής, που το περιεχόμενό της συνεχώς εμπλουτίζεται. Η σημερινή δομή του

[12] https://itp.nyu.edu/classes/roy-spring2015/files/2015/02/DE_words_LIWC.jpg

λεξικού διαχωρίζεται σε τέσσερεις ξεχωριστές λεξικολογικές κατηγορίες, η κάθε μία από τις οποίες ειδικεύεται σε λεξιλόγιο διαφορετικών γλωσσολογικών προσεγγίσεων [72].
Τα κύρια επιμέρους λεξικά που περιλαμβάνει είναι το Harvard IV-4 Dictionary και το Lasswell Value Dictionary. Το Harvard IV-4 συνίσταται στις τρεις συναισθηματικές διαστάσεις του "Osgood". Οι κατηγορίες αυτές αντανακλούν ειδοποιούς διαφορές σχετικά με τη βασική, καθολική γλώσσα, με τρία διαφορετικά επίπεδα «έντασης» για κάθε κατηγορία, που συνδυάζονται και κάθε λέξη μπορεί να έχει περισσότερες από μία διάσταση, ανάλογα με την περίπτωση. Οι κατηγορίες που περιλαμβάνει αντανακλούν μια κοινωνιολογική προοπτική και αναφέρονται σε ευχαρίστηση, πόνο, την αρετή, υπερβολή, συγκράτηση, ρόλους, συλλογικότητες, τελετουργίες και μορφές των διαπροσωπικών σχέσεων, παρουσία ή η απουσία συναισθηματικής εκφραστικότητας. Οι καταχωρήσεις του Lasswell αναπτύχθηκαν από τους Namenwirth και Weber (1987) [56]. Καταρχήν, χωρίζει τους όρους του σε τέσσερις τομείς σεβασμού: δύναμη, εντιμότητα, σεβασμός, ασφάλεια, καθώς και τέσσερις τομείς κοινωνικής ευημερίας: πλούτος, ευεξία, διαφωτισμός και ικανότητα. Σε κάθε τέτοιο τομέα, υπάρχουν υποκατηγορίες, όπως τα κέρδη, οι απώλειες, οι συμμετέχοντες, τα όρια, και τα πεδία δράσης.

```
HAPPINESS H4Lvd Pos Pstv Pleasure EMOT WLBPSYC WLBTOT Noun  | noun: The quality or state of being happy
HAPPY#1 H4Lvd Pos Pstv Pleasure EMOT WLBPSYC WLBTOT Modif  | 78% adjective: Joyous, pleased.
HAPPY#2 H4Lvd Pos Pstv Pleasure POSAFF Modif | 13% adverb: ""Happily""--in a joyous manner
HAPPY#3 H4Lvd Pos Pstv Pleasure EMOT WLBPSYC WLBTOT Modif  | 7% adjective: ""Happier,"" comparative of sense 1
HAPPY#4 H4Lvd Pos Pstv Pleasure EMOT WLBPSYC WLBTOT Modif  | 2% adjective: ""Happiest,"" superlative of sense 1
HARBOR#1 H4Lvd Econ* ECON PLACE Aquatic Noun |
HARBOR#2 H4Lvd Strng Stay IAV SUPV |
HARBOUR#1 H4 |
HARD#1 H4Lvd Neg Ngtv Strng Vice EVAL WLBPSYC WLBTOT Modif  | 47% adj: Difficult, trying, severe (especially of people)
```

Σχήμα 3.8: Παράδειγμα δομής General Inquirer[13]

Συνολικά το Harvard General Inquirer Lexicon περιλαμβάνει πάνω από 180 κατηγορίες στις οποίες εντάσσονται περίπου 4000 λέξεις. Η άνιση κατανομή των λέξεων σε κάθε λίστα του λεξικού (πχ συγκριτικά περισσότερες λέξεις αντιπάθειας από ότι χαράς) καθιστά αναγκαία την κανονικοποίηση των μετρήσεων πριν την επιλογή του τελικού σημασιολογικού προσανατολισμού ενός κειμένου.

[13] http://www.wjh.harvard.edu/~inquirer/inqdict.txt

3.5 Λεξικά Συναισθήματος της Ελληνικής Γλώσσας

Η μετάφραση των όρων ενός υπάρχοντος λεξικού συναισθήματος της Αγγλικής Γλώσσας σε κάποια άλλη γλώσσα και η διατήρηση του ίδιου σημασιολογικού προσανατολισμού τους, αποτέλεσε μία απλή, αλλά άμεση, αυτοματοποιημένη τεχνική, για τη δημιουργία πολύγλωσσων λεξικών συναισθήματος. Σε αυτές τις μεθόδους αρχικά επιλέγονται οι λέξεις με μεγάλη ένταση συναισθήματος, γεγονός που αυξάνει την πιθανότητα διατήρησης του και στο λεξιλόγιο στης προκύπτουσας γλώσσας. Στη συνέχεια κατηγοριοποιούνται σύμφωνα με τα διάφορα συντακτικά χαρακτηριστικά τους και μεταφράζονται από on-line εφαρμογές μετάφρασης (πχ. Google translation) [23].

Παρόλαυτα, μια τέτοια προσέγγιση εμπίπτει σε πολλούς περιορισμούς από την ίδια της τη φύση, αφού κάθε λέξη ενδέχεται να μην μεταφέρει το ίδιο ή και κανένα συναισθηματικό φορτίο στην αποδιδόμενη γλώσσα, συγκριτικά με την αγγλική, χωλαίνοντας σημαντικά στον τομέα της συμφραστικής πολικότητας. Η συνέπεια αυτή παρατηρείται και στα λεξικά με διαβαθμίσεις, όπου δεν υπάρχει πάντοτε ταύτιση της έντασης του συναισθήματος της αρχικής με την αποδιδόμενη λέξη. Για τη μεγιστοποίηση της απόδοσης τέτοιων λεξικών κρίνεται απαραίτητη η συνεχής ανανέωση του λεξιλογίου τους και ο συνδυασμός τους με ερμηνευτικά λεξικά της γλώσσας στην οποία αποδίδονται, που θα εστιάζουν στα ιδιαίτερα πολιτιστικά χαρακτηριστικά των λέξεων της.

Στον αντίποδα των παραπάνω τεχνικών βρίσκονται οι έρευνες που έχουν ως στόχο τη δημιουργία ενός αμιγώς ελληνικού λεξικού συναισθήματος, που θα αντλεί τους όρους του από ελληνικούς λεξικολογικούς πόρους, είτε αυτοί θα είναι λήμματα ενός ελληνικού ερμηνευτικού λεξικού, είτε κατάλληλα επεξεργασμένα στοιχεία από κειμενικές πηγές ελληνικής βιβλιογραφίας ή ελληνικών σελίδων του διαδικτύου. Σε αυτές τις περιπτώσεις παρατηρούμε την εφαρμογή μεθόδων παρόμοιων με αυτές που συναντήσαμε στην κατασκευή αντίστοιχων αγγλόγλωσσων λεξικών συναισθήματος.

Ενδεικτικό παράδειγμα αυτής της προσέγγισης αποτελεί η ελληνική εταιρία Qualia (http://www.qualia.gr/), που σε συνεργασία με το Ινστιτούτο Γλώσσας και Επεξεργασίας Λόγου, ήδη από το 2004 αναπτύσσει ένα εξελιγμένο σύστημα εξαγωγής πληροφορίας που ανιχνεύει web σελίδες και τις μετατρέπει σε δομημένα δεδομένα που είναι προσαρμοσμένα για την ανάκτηση πληροφοριών. Χρησιμοποιώντας λογισμικό κάθετης εξαγωγής πληροφορίας (vertical information extraction), με σκοπό την εξαγωγή δεδομένων από συγκεκριμένες θέσεις μέσα από τις σελίδες, ανέλυσε και μετέτρεψε πολύτιμες πληροφορίες σε δομημένη μορφή και τις ενσωμάτωσε σε ένα σύστημα βάσεων δεδομένων για περαιτέρω επεξεργασία. Για την κατηγοριοποίηση του κειμένου, εφάρμοσαν πλειάδα αλγορίθμων, όπως οι bootstrapping αλγόριθμοι για την μάθηση μερικώς σχολιασμένων όρων, η πολυδιάστατη γραμμική κατηγοριοποίηση, οι μπευσιανοί αλγόριθμοι και οι μέθοδοι kernel, δημιουργώντας ένα λεξικό συναισθημάτων που αποτελείται από μία λίστα 1363 ελληνικών λέξεων. Κάθε μία από αυτές τις λέξεις συνοδεύεται από έναν ακέραιο αριθμό που ανήκει στο διάστημα [-5,5] και εκφράζει τον σημασιολογικό προσανατολισμό της.

Μία καινοτόμα προσπάθεια στον ίδιο τομέα, αποτελεί το Greek Sentiment Lexicon, η πρώτη δημόσια και ελεύθερα διαθέσιμη έκδοση λεξικού συναισθήματος στα Ελληνικά. Η ανάγκη για ένα τέτοιο λεξικό προήλθε από έρευνες των Tsakalidis και Papadopoulos (2014) [76], σε μεθόδους επιβλεπόμενης και μη επιβλεπόμενης μάθησης, που βασίζονταν σε λεξικά και είχαν ως συμπέρασμα ότι ο συνδυασμός τους μπορούν να βελτιώσουν σημαντικά την ακρίβεια εντοπισμού συναισθήματος.

Σε αυτά τα πλαίσια, επινοήθηκε μια ημι-αυτόματη προσέγγιση για την κατασκευή του λεξικού. Εν ολίγοις, πραγματοποιήθηκε αναζήτηση για πρώτη φορά στην ηλεκτρονική έκδοση του "Λεξικού της κοινής νεοελληνικής" του Μ. Τριανταφυλλίδη και συλλέχτηκαν όροι που μπορούν να κατηγοριοποιηθούν σε έναν από τους παρακάτω σημασιολογικούς τόνους: ειρωνικός, μειωτικός, υβριστικός, χυδαίος και σκωπτικός. Προστέθηκαν επίσης όροι που περιέχουν συναισθηματικές λέξεις στην περιγραφή τους (π.χ. αίσθηση, αγάπη, κτλ). Στο τελικό στάδιο σε κάθε όρο αποδόθηκε σχολιασμός τεσσάρων ερευνητών (δύο από την επιστήμη των υπολογιστών, και δύο της υπολογιστικής γλωσσολογίας) μήκους οκτώ διαστάσεις: υποκειμενικότητα, πολικότητα και έξι συναισθήματα (ευτυχία, λύπη, θυμός, φόβος, αηδία, έκπληξη), με μία κλίματα έντασης συναισθήματος 5 βαθμίδων, δημιουργώντας μία λίστα 2317 λέξεων[14].

Term	POS1	POS2	POS3	POS4	Subjectivity1	Subjectivity2	Subjectivity3	Subjectivity4	Polarity1	Polarity2	Polarit
αβάφτιστος	ADJ	ADJ	ADJ	ADJ	SUBJ-	OBJ	SUBJ-	OBJ	BOTH	N/A	BOTH
αβάπτιστος	ADJ	ADJ	ADJ	ADJ	SUBJ-	OBJ	SUBJ-	OBJ	BOTH	N/A	BOTH
αγανάκτηση	NOUN	NOUN	NOUN	NOUN	SUBJ+	SUBJ+	SUBJ+	SUBJ+	NEG	NEG	NEG
αγανακτώ	VERB	VERB	VERB	VERB	SUBJ+	SUBJ+	SUBJ+	SUBJ+	NEG	NEG	NEG
αγάπη	NOUN	NOUN	NOUN	NOUN	SUBJ+	SUBJ+	SUBJ+	SUBJ+	POS	POS	POS
αγαπημένος	PART	PART	PART	PART	SUBJ+	SUBJ+	SUBJ+	SUBJ+	POS	POS	POS
αγαπητός	ADJ	ADJ	ADJ	ADJ	SUBJ+	SUBJ+	SUBJ+	SUBJ+	POS	POS	POS
αγαπώ	VERB	VERB	VERB	VERB	SUBJ+	SUBJ+	SUBJ+	SUBJ+	POS	POS	POS
άγγελος	NOUN	NOUN	NOUN	NOUN	SUBJ+	SUBJ-	SUBJ+	OBJ	POS	BOTH	POS
αγέλη	NOUN	NOUN	NOUN	NOUN	SUBJ-	SUBJ-	SUBJ-	OBJ	NEG	NEG	NEG
αγευσία	NOUN	NOUN	NOUN	NOUN	OBJ	OBJ	SUBJ+	OBJ	N/A	N/A	NEG
αγκουσεύω	VERB	VERB	VERB	VERB	SUBJ+	SUBJ+	SUBJ+	SUBJ+	NEG	NEG	NEG
αγνός	ADJ	ADJ	ADJ	ADJ	SUBJ+	SUBJ-	SUBJ+	SUBJ+	BOTH	POS	POS

Σχήμα 3.9: Παράδειγμα δομής Greek Sentiment Lexicon[15]

[14] http://www.socialsensor.eu/
[15] https://github.com/MKLab-ITI/greek-sentiment-lexicon/blob/master/greek_sentiment_lexicon.tsv

4. Μέθοδοι βασιζόμενες σε Λεξικά Συναισθήματος

Είναι σαφές ότι οι λίστες με λέξεις και φράσεις που μεταφέρουν θετικά ή αρνητικά συναισθήματα, είναι καθοριστικός παράγοντας για την Ανάλυση Συναισθήματος. Έχουν προταθεί πολλές προσεγγίσεις για την επιλογή των κατάλληλων λέξεων ενός κειμένου για την ανάλυση του. Οι τρεις κύριες προσεγγίσεις είναι η χειροκίνητη προσέγγιση, η προσέγγιση βασιζόμενη σε ερμηνευτικό λεξικό και η προσέγγιση βασιζόμενη σε ηλεκτρονικά σώματα κειμένων.

Η συλλογή λέξεων σύμφωνα με ένα ερμηνευτικό λεξικό αναφέρεται στην κατάρτιση λιστών λέξεων και φράσεων, τον χαρακτηρισμό τους ως θετικές ή αρνητικές, και μέσω αυτών τη σημασιολογική κατάταξη των προτάσεων ενός κειμένου, αξιοποιώντας τις ιδιότητες νοηματικής σύνδεσης (συνωνυμία, αντωνυμία, κοινή ρίζα) των λέξεων αυτών των λεξικών. Η χρήση του λεξιλογίου των ίδιων των ηλεκτρονικών σωμάτων κειμένου, για την εξόρυξη γνώμης, βασίζεται στα κοινά χαρακτηριστικά που παρουσιάζουν οι λέξεις σε κείμενα συγκεκριμένων τομέων, ως προς τη συχνότητα συνεμφάνισης, τα μέρη του λόγου στα οποία ανήκουν και οι σύνδεσμοι που τις συνδέουν και η δυνατότητα ομαδοποίηση τους. Η χειροκίνητη προσέγγιση είναι χρονοβόρα και απαιτεί εντατική εργασία, ως εκ τούτου συνήθως δεν χρησιμοποιείται μόνη της αλλά ως ένας τελικός έλεγχος, σε συνδυασμό με αυτοματοποιημένες προσεγγίσεις, επειδή οι τελευταίες εμπεριέχουν αρκετά σφάλματα.

4.1 Προσεγγίσεις βασιζόμενες σε Ερμηνευτικό Λεξικό

4.1.1 Μέθοδοι μη Επιβλεπόμενης Μάθησης

Η χρησιμοποίηση ενός λεξικού για τη συγκέντρωση λέξεων συναισθημάτων είναι μια προφανής προσέγγιση, επειδή τα περισσότερα λεξικά (π.χ. το WordNet) καταγράφουν συνώνυμα και αντώνυμα για κάθε λέξη. Έτσι, μια απλή τεχνική σε αυτή την προσέγγιση είναι να χρησιμοποιηθούν κατά την εκκίνηση ως λέξεις συναισθημάτων, λίγα λήμματα ενός ερμηνευτικού λεξικού, αξιοποιώντας το νοητικό σχήμα «συνώνυμο – αντώνυμο» που εμπεριέχεται στη δομή του. Πιο συγκεκριμένα, σε αυτή τη μέθοδο, αρχικά ένα μικρό σύνολο λέξεων συναισθήματος, με γνωστό θετικό ή αρνητικό προσανατολισμό, επιλέγεται χειρονακτικά. Ο αλγόριθμος έπειτα αυξάνει αυτό το σύνολο κάνοντας αναζήτηση στο WordNet (ή σε κάποιο άλλο on-

line ερμηνευτικό λεξικό) για τα συνώνυμα και αντώνυμα τους. Οι εκ νέου ευρεθείσες λέξεις, προστίθενται στον κατάλογο των λημμάτων και αρχίζει η επόμενη επανάληψη. Η επαναληπτική διαδικασία τελειώνει όταν δεν μπορούν να βρεθούν περισσότερες νέες λέξεις. Οι Hu και Liu (2004) [41], ακολουθώντας την ίδια μέθοδο, μετά την ολοκλήρωση της διαδικασίας, πραγματοποίησαν χειροκίνητο έλεγχο για τον καθαρισμό της λίστας. Μια παρόμοια μέθοδος χρησιμοποιήθηκε επίσης από τους Valitutti, Strapparava και Stock (2004) [73]. Οι Kim και Hovy (2004) [46] προσπάθησαν να διακρίνουν στις προκύπτουσες λέξεις τα σφάλματα και να εκχωρήσουν μια δύναμη συναισθήματος (sentiment strength) σε κάθε λέξη, χρησιμοποιώντας μια Πιθανοτική Μέθοδο. Οι Mohammad και Dunne (2009) [54] επιπλέον αξιοποιούν τις ιδιότητες των προθεμάτων άρνησης της μορφής X και disX (πχ honest–dishonest), για να αυξήσουν την κάλυψη του συνόλου λημμάτων.

Μια πιο σύνθετη προσέγγιση προτάθηκε από τους Kamps et al., (2004) [44], η οποία χρησιμοποίησε μια μέθοδο αποστάσεων (distance method), για να καθορίσει τον συναισθηματικό προσανατολισμό ενός δοθέντος επίθετου. Η απόσταση $d(t_1, t_2)$ μεταξύ των όρων t_1 και t_2 , είναι το μήκος της συντομότερης διαδρομής που συνδέει τους δύο όρους στο WordNet. Ο προσανατολισμός ενός επιθέτου t καθορίζεται από τη σχετική απόστασή του, από δύο λήμματα αναφοράς, τους όρους «good» και «bad», δηλαδή, $SO = \frac{d(t,bad)-d(t\ good)}{d(good,bad)}$, όπου ο όρος t είναι θετικός αν και μόνο αν $SO(t) > 0$, και είναι αρνητικός διαφορετικά. Η δύναμη του συναισθήματος ορίζεται ως η απόλυτη τιμή του $SO(t)$. Συνδυάζοντας τις δύο τελευταίες μεθόδους, οι Williams και Anand (2009) μελέτησαν το πρόβλημα της ανάθεσης της δύναμης συναισθήματος σε κάθε λέξη ξεχωριστά.

Οι Rao και Ravichandran, (2009) [67], με τρεις μεθόδους ημι-επιβλεπόμενης μάθησης σε γράφους, προσπάθησαν να διαχωρίσουν τις θετικές και αρνητικές λέξεις, χρησιμοποιώντας ένα σύνολο θετικών λημμάτων, ένα σύνολο αρνητικών λημμάτων, και ένα γράφο συνωνύμων που εξήγαγαν από το WordNet. Οι τρεις αλγόριθμοι ήταν: ο Αλγόριθμος Ελάχιστης Τομής (Mincut), ο Τυχαιοκρατικός Αλγόριθμος Ελάχιστης Τομής (Randomized Min-Cut) και ο Αλγόριθμος Διάδοσης Ετικέτας (Label Propagation) Zhu και Ghahramani, (2002) [85]. Αποδείχθηκε ότι ο Mincut και ο Randomized Min-Cut παράγουν καλύτερες βαθμολογίες, αλλά ο Διάδοσης Ετικέτας έδωσε σημαντικά υψηλότερη ακρίβεια, με χαμηλή ανάκληση.

Στην εργασία των Turney και Littman, (2003) [79], χρησιμοποιείται η ίδια μέθοδος που βασίζεται στον PMI Turney, (2002) [78], για τον υπολογισμό του συναισθηματικού προσανατολισμού μιας δεδομένης λέξης. Συγκεκριμένα, υπολογίζει τον προσανατολισμό της λέξης από τη δύναμη της συσχέτιση της με μια σειρά από λέξεις θετικού προσανατολισμού (good, nice, excellent, positive,fortunate, correct, and superior), μείον την δύναμη της συσχέτιση της με μια σειρά από λέξεις αρνητικού προσανατολισμού (bad, nasty, poor, negative, unfortunate, wrong, andinferior). Η δύναμη συσχέτισης μετριέται με τη χρήση του PMI.

4.1.2 Μέθοδοι Επιβλεπόμενης Μάθησης

Οι Esuli και Sebastiani (2005) [29] χρησιμοποιούν επιβλεπόμενη μάθηση για να ταξινομήσουν τις λέξεις σε θετικά και αρνητικά σύνολα. Δεδομένου ενός συνόλου θετικών λημμάτων P και και ενός αρνητικών N, τα δύο σύνολα αρχικά επεκτείνονται με τη χρήση σχέσεων συνωνύμων και αντωνύμων στο WordNet, για να δημιουργήσουν ένα διευρυμένο σετ P 'και N', τα οποία αποτελούν το σύνολο εκπαίδευσης. Ο αλγόριθμος στη συνέχεια, χρησιμοποιεί όλες τις ερμηνείες στο λεξικό για κάθε όρο στο $P' \cup N'$, για να δημιουργήσει ένα διάνυσμα χαρακτηριστικών. Στη συνέχεια κατασκευάστηκε ένας δυαδικός ταξινομητής με τη χρήση διαφορετικών αλγορίθμων μάθησης. Η διαδικασία μπορεί επίσης να εκτελεστεί επαναληπτικά. Δηλαδή, οι άρτι εντοπισθέντες θετικοί και τα αρνητικοί όροι και τα συνώνυμά και αντώνυμα τους, προστίθενται στο σύνολο εκπαίδευσης, και ένας επικαιροποιημένος ταξινομητής μπορεί πλέον να κατασκευαστεί. Στην εργασία των Esuli και Sebastiani (2006) [29], περιλαμβάνεται επίσης, η ουδέτερη κατηγορία. Για να επεκταθεί το σύνολο ουδέτερων λημμάτων, χρησιμοποιήθηκαν υπερώνυμα, παράλληλα με συνώνυμα και αντώνυμα. Στη συνέχεια χρησιμοποιήθηκαν διαφορετικές στρατηγικές για να κάνουν την κατάταξη στις τρεις σημασιολογικές κατηγορίες. Με βάση την παραπάνω μέθοδο, χρησιμοποίησαν ένα σύνολο ταξινομητών για την κατασκευή του SentiWordNet, μια λεξιλογική πηγή στην οποία, κάθε εννοιολογικό σύνολο (synset) του WordNet συνδέεται με τρεις αριθμητικές βαθμολογίες *Obj(s)*, *Pos(s)* και *Neg(s)*, περιγράφοντας πόσο ουδέτεροι, θετικοί ή αρνητικοί είναι οι όροι στο εννοιολογικό σύνολο. Η μέθοδος των Kim και Hovy (2006) [48], αρχίζει επίσης με τρία σύνολα λημμάτων (θετικά, αρνητικά και ουδέτερα). Στη συνέχεια βρίσκει τα συνώνυμά τους στο WordNet. Τα διευρυμένα σύνολα, ωστόσο, έχουν πολλά σφάλματα. Η μέθοδος στη συνέχεια χρησιμοποιεί ένα Μπεϋζιανό τύπο για να υπολογίσει την εγγύτητα της κάθε λέξης σε κάθε κατηγορία για να προσδιοριστεί η πλέον πιθανή κατηγορία της.

Οι Velikovich et al. (2010) [81] πρότειναν επίσης μια μέθοδο για την κατασκευή ενός λεξικού συναισθημάτων χρησιμοποιώντας ιστοσελίδες. Εφάρμοσαν έναν Αλγόριθμο Διάδοσης Γράφων (Graph Propagation Algorithm), πάνω σε ένα γράφο με όμοιες φράσεις. Και εδώ ως είσοδος θεωρείται ένα σύνολο από φράσεις θετικών λημμάτων και ένα σύνολο από φράσεις αρνητικών λημμάτων. Οι κόμβοι στο γράφο ήταν οι υποψήφιες φράσεις που επιλέχθηκαν από όλα τα N-γράμματα μήκους έως 10. Μόνο 20 εκατομμύρια υποψήφιες φράσεις επιλέχθηκαν χρησιμοποιώντας διάφορες ευριστικές τεχνικές, όπως η συχνότητα και η αμοιβαία πληροφορία. Στη συνέχεια κατασκευάστηκε ένα διάνυσμα πληροφορίας πλαισίου (contextual information) για κάθε υποψήφια φράση βασισμένο σε ένα παράθυρο λέξης μεγέθους έξι, που συγκεντρώθηκε από όλες τις αναφορές για την κάθε φράση, στα 4 δισεκατομμύρια έγγραφα. Το σύνολο των ακμών κατασκευάστηκε με τον υπολογισμό της ομοιότητας συνημίτονου των διανυσμάτων πλαισίου, των υποψήφιων φράσεων. Όλες οι ακμές (*vi*, *vj*) απορρίφθηκαν αν δεν ήταν από τις 25 υψηλότερες σταθμισμένες ακμές των κόμβων vi ή vj. Το βάρος ακμής ορίστηκε στην αντίστοιχη τιμή του συνημίτονου ομοιότητας. Στο τέλος χρησιμοποιείται μια μέθοδος διάδοσης σε γράφους για τον

υπολογισμό του συναισθήματος κάθε φράσης ως το άθροισμα όλων των καλύτερων διαδρομών προς τα λήμματα.

4.1.3 Bootstrapping Μέθοδοι

Ως bootstrapping καλείται κάθε μέθοδος για την εξαγωγή ισχυρών εκτιμήσεων για τυπικά σφάλματα και για διαστήματα εμπιστοσύνης, όπως η μέση τιμή, η διάμεσος, η απλή αναλογία, η αναλογία πιθανοτήτων, ο συντελεστής συσχέτισης και ο συντελεστής παλινδρόμησης. Μπορεί επίσης να χρησιμοποιηθεί για την κατασκευή δοκιμαστικών υποθέσεων. Οι bootstrapping μέθοδοι είναι πιο χρήσιμες ως εναλλακτική λύση για παραμετρικές εκτιμήσεις, όταν οι παραδοχές των μεθόδων αυτών είναι υπό αμφισβήτηση (όπως στην περίπτωση των μοντέλων παλινδρόμησης με ετεροσκεδαστικά υπολείμματα σε μικρά δείγματα), ή όπου τα παραμετρικά συμπεράσματα είναι αδύνατα ή απαιτούν πολύ πολύπλοκους τύπους για τον υπολογισμό του σφάλματος τυπικής απόκλισης.

Οι Andreevskaia και Bergler (2006) [5] πρότειναν μια εξελιγμένη bootstrapping μέθοδο με διάφορες τεχνικές για να επεκτείνουν τα αρχικά σύνολα θετικών και αρνητικών λημμάτων και να καθαρίσουν τα επεκταμένα σύνολα (απομάκρυνση μη επιθέτων και λέξεις που ανήκουν και στα δυο σύνολα). Επιπλέον, ο αλγόριθμος τους εκτελεί πολλαπλώς μια bootstrapping διαδικασία, χρησιμοποιώντας μη-επικαλυπτόμενα υποσύνολα λημμάτων. Κάθε εκτέλεση βρίσκει συνήθως ένα ελαφρώς διαφορετικό σύνολο συναισθηματικών λέξεων. Στη συνέχεια υπολογίζεται για κάθε λέξη, μια βαθμολογία επικάλυψης (overlapping score), με βάση το πόσες φορές η λέξη εντοπίστηκε σε κάποια εκτέλεση, ως θετική ή ως αρνητική. Η βαθμολογία κανονικοποιείται στο διάστημα [0, 1], σύμφωνα με τη Θεωρία Ασαφών Συνόλων.

Οι Blair-Goldensohn et al., (2008) [10] πρότειναν μια διαφορετική bootstrapping μέθοδο, η οποία χρησιμοποιεί ένα σύνολο θετικών λημμάτων, ένα σύνολο αρνητικών λημμάτων και ένα ουδέτερων. Η προσέγγιση έργων που βασίζονται σε ένα κατευθυνόμενο, σημασιολογικό γράφο με βάρη, όπου οι γειτονικοί κόμβοι είναι συνώνυμα ή αντώνυμα των λέξεων στο WordNet και δεν αποτελούν μέρος του σύνολο ουδέτερων λημμάτων. Το ουδέτερο σύνολο χρησιμοποιείται για να σταματήσει η διάδοση των συναισθημάτων μέσω των ουδέτερων λέξεων. Τα βάρη των ακμών ορίζονται προκαταβολικώς με βάση μια παράμετρο κλιμάκωσης για τα διάφορα είδη των άκρων, δηλαδή συνώνυμο ή αντώνυμο άκρο. Σε κάθε λέξη έπειτα δίνεται μία τιμή συναισθήματος χρησιμοποιώντας μία τροποποιημένη εκδοχή ενός αλγορίθμου Διάδοσης Ετικέτας των Zhu και Ghahramani, (2002) [85]. Στην αρχή, σε κάθε θετικό λήμμα δίνεται η βαθμολογία +1, σε κάθε αρνητικό η βαθμολογία -1, και σε όλες τις άλλες λέξεις 0. Οι βαθμολογίες αναθεωρούνται κατά τη διάρκεια της διαδικασίας διάδοσης. Όταν η διάδοση σταματάει μετά από έναν αριθμό επαναλήψεων, οι τελικές βαθμολογίες μετά από μια λογαριθμική κλιμάκωση

ανατίθενται στις λέξεις ανάλογα με τους βαθμό θετικότητας ή αρνητικότητας που αποκόμισαν.

Στο έργο των Dragut et al., (2010) [25], προτάθηκε μια πολύ διαφορετική bootstrapping μέθοδος χρησιμοποιώντας το WordNet. Λαμβάνοντας υπόψη ένα σύνολο λημμάτων, αντί για την απλή ακολουθία του λεξικού, οι συγγραφείς προτείνουν μια σειρά από σύνθετους κανόνες συμπερασμού για να καθορίσουν άλλες λέξεις συναισθήματος προσανατολισμούς μέσω μιας αφαιρετικής διαδικασίας. Ο αλγόριθμος δηλαδή παίρνει τα λήμματα με τους γνωστούς προσανατολισμούς ως είσοδο και παράγει σύνολα συνωνύμων (synsets) με κατευθύνσεις. Τα synsets με τις υπολογισθείσες κατευθύνσεις μπορούν στη συνέχεια να χρησιμοποιηθούν για να συναχθούν οι περαιτέρω πολικότητες των άλλων λέξεων.

4.1.4 Στατιστικές Μέθοδοι

Οι Hassan and Radev (2010) [38] παρουσίασαν ένα Μαρκοβιανό μοντέλο Τυχαίου Περιπάτου (Random Walk), πάνω σε ένα γράφο λεκτικής συσχέτισης για να παράγουν μια εκτίμηση συναισθήματος για κάθε δοθείσα λέξη. Καταρχάς, χρησιμοποιούν συνώνυμα και υπερώνυμα του WordNet, για να χτίσουν ένα γράφο λεκτικής συσχέτισης. Στη συνέχεια ορίζουν μία μονάδα μέτρησης, η οποία ονομάζεται Μέσος Χρόνος Ευστοχίας (Mean Hitting Time) $h(i|S)$, και χρησιμοποιείται για να μετρηθεί η απόσταση από τον κόμβο i σε ένα σύνολο κόμβων S, η οποία ισούται με το μέσο αριθμό βημάτων που θα χρειαστεί ένας Τυχαίος Περίπατος, ξεκινώντας από την κατάσταση (i, S), για να φτάσει στην κατάσταση (k, S) για πρώτη φορά. Δεδομένου ενός συνόλου θετικών λημμάτων $S+$ και ενός συνόλου αρνητικών λημμάτων $S-$, για να εκτιμηθεί ο συναισθηματικός προσανατολισμός μιας δεδομένης λέξης w, υπολογίζονται οι δείκτες $h(w|S+)$ και $h(w|S-)$. Αν ο $h(w|S+)$ είναι μεγαλύτερο από τον $h(w|S-)$, η λέξη ταξινομείται ως αρνητική, αλλιώς ως θετική.

Στην εργασία των Kaji και Kitsuregawa, (2006) [43], χρησιμοποιήθηκαν πολλοί Ευριστικοί Κανόνες για την κατασκευή ενός λεξικού συναισθήματος από έγγραφα HTML, με βάση τις δομές διάταξης των δικτυακών σελίδων. Για παράδειγμα, ένας πίνακας σε μια ιστοσελίδα μπορεί να έχει μια στήλη που δείχνει σαφώς θετικές ή αρνητικές κατευθύνσεις (π.χ. Πλεονεκτήματα και Μειονεκτήματα) . Αυτά τα στοιχεία μπορούν να αξιοποιηθούν για να εξαχθεί ένας μεγάλος αριθμός υποψηφίων θετικών και αρνητικών προτάσεων που εκφράζουν άποψη από ένα μεγάλο σύνολο ιστοσελίδων. Οι φράσεις που περιέχουν επίθετα εξάγονται και τους αποδίδεται συναισθηματικός προσανατολισμός που βασίζεται σε διαφορετικές στατιστικές των εμφανίσεων τους στο θετικό και αρνητικό σύνολο προτάσεων, αντίστοιχα.

4.2 Προσεγγίσεις βασιζόμενες σε Ηλεκτρονικά Σώματα Κειμένων

Η προσέγγιση που βασίζεται σε ηλεκτρονικά σώματα κειμένων έχει εφαρμοστεί σε δύο βασικές περιπτώσεις εξόρυξης γνώμης. Καταρχάς, σε μια λίστα με λήμματα από λέξεις συναισθήματος, για την ανακάλυψη άλλων λέξεων συναισθήματος και τον προσανατολισμό τους, σε ένα ηλεκτρονικό σώμα κειμένου, και δευτερευόντως γα την προσαρμογή ενός λεξικού συναισθημάτων σε ένα νέο, γενικού σκοπού, χρησιμοποιώντας ένα σώμα κειμένου, για εφαρμογές ανάλυσης συναισθήματος. Ωστόσο, το ζήτημα είναι πιο περίπλοκο από ότι η οικοδόμηση λεξικού συναισθήματος ενός συγκεκριμένου τομέα, γιατί στον ίδιο τομέα η ίδια λέξη μπορεί να είναι θετική σε ένα πλαίσιο, αλλά αρνητική σε ένα άλλο. Αν και η προσέγγιση που βασίζεται σε ηλεκτρονικά σώματα κειμένων μπορεί επίσης να χρησιμοποιηθεί για την κατασκευή ενός λεξικό συναισθημάτων γενικού σκοπού εάν είναι διαθέσιμο ένα πολύ μεγάλο και διαφορετικό σώμα, η προσέγγιση που βασίζεται σε ερμηνευτικό λεξικό συνήθως είναι πιο αποτελεσματικό επειδή ένα λεξικό έχει όλες τις λέξεις.

4.2.1 Μέθοδοι Κοινού Γνωσιακού Τομέα

Μια από τις βασικές μεθόδους αυτής της προσέγγισης προτάθηκε από τους Hazivassiloglou και McKeown (1997) [39]. Οι συγγραφείς χρησιμοποίησαν ένα σώμα κειμένου και μερικά λήμματα επίθετων ως λέξεις συναισθήματος, για να βρουν πρόσθετα επίθετα συναισθήματος στο σώμα. Η τεχνική τους εκμεταλλευόταν ένα σύνολο γλωσσικών κανόνων στον τομέα των συνδέσμων για να εντοπίσουν περισσότερα επίθετα συναισθήματος και τους προσανατολισμούς τους από το σώμα. Ένα τέτοιο είδος γλωσσικών κανόνων είναι η συμπλεκτική σύνδεση δύο όρων («and»), που συνήθως δηλώνει ότι τα συνδεόμενα μέρη, εν προκειμένω τα επίθετα, παρουσιάζουν τον ίδιο σημασιολογικό προσανατολισμό. Αυτό συμβαίνει επειδή οι άνθρωποι συνήθως εκφράζουν το ίδιο συναίσθημα και στις δύο πλευρές του συνδέσμου. Έχουν επίσης σχεδιαστεί κανόνες για άλλους συνδέσμους, («or», “but”, “either-or”, και «neither-or»). Η τεχνική αυτή, ονομάζεται Συνεκτικότητα Συναισθήματος (sentiment consistency) . Στην πράξη, δεν είναι πάντοτε συνεπής. Έτσι, εφαρμόστηκε και ένα βήμα εκμάθησης για τον καθορισμό εάν τα δύο συμπλεγμένα επίθετα έχουν ίδιες ή διαφορετικές κατευθύνσεις.

Κατ 'αρχάς, σχηματίστηκε ένας γράφος με ίδιου και διαφορετικού προσανατολισμού δεσμούς (links) μεταξύ των επίθετων. Για να διαμεριστούν οι κόμβοι του, σε υποσύνολα, ίδιου προσανατολισμού, χρησιμοποιήθηκε μια επαναληπτική διαδικασία βελτιστοποίησης, για κάθε συνδεδεμένο, με βάση έναν μη ιεραρχικό αλγόριθμο, τη «μέθοδο της ανταλλαγής». Ορίστηκε μια Αντικειμενική Συνάρτηση Φ, σημειώνοντας κάθε δυνατή διαμέριση P από τα επίθετα, σε δύο υποομάδες C1 και C2: $\Phi(P) =$

$\sum_{i=1}^{2}\left(\frac{1}{|Ci|}\sum_{\substack{x,y\in Ci\\ x\neq y}} d(x,y)\right)$ όπου ως Ci συμβολίζεται το πλήθος των συστάδων κόμβων i, και d(x, y) είναι η ανομοιότητα μεταξύ των επίθετων x και y. Ο αλγόριθμος δημιουργίας συστάδων, διαχώρισε κάθε συνιστώσα του γράφου, ανάμεσα σε δύο ομάδες επιθέτων, αλλά δεν επισήμανε, εάν τα επίθετα είναι θετικά ή αρνητικά. Για αυτό, χρησιμοποιήθηκε ένα απλό κριτήριο που ισχύει μόνο για ζευγάρια ή ομάδες λέξεων, αντίθετου προσανατολισμού. Στα ζεύγη αντίθετων διαβαθμίσιμων επιθέτων, εάν το ένα μέλος είναι μη σημασμένο, το άλλο σημασμένο μέλος είναι το πιο συχνό, με ποσοστό εμφάνισης 81%. Η σημασιολογική σήμανση παρουσίαζε ισχυρή συσχέτιση με τον προσανατολισμό, αφού το σημασμένο μέλος σχεδόν πάντα έχει θετικό προσανατολισμό. Στη συνέχεια διεξήχθη η ομαδοποίηση τους στο γράφο, για να παραχθούν δύο σύνολα θετικών και αρνητικών λέξεων.

Οι Kanayama και Nasukawa (2006) [45] επέκτειναν αυτή την αρχική ιδέα εισάγοντας τις έννοιες της ενδο-προτασιακής και της δια-προτασιακής συνεκτικότητας (intra-sentential και inter-sentential sentiment consistency) αναζητώντας αναλογίες σημασιολογικού προσανατολισμού και σε επίπεδο ολόκληρης περιόδου λόγου και όχι μεμονωμένων προτάσεων. Η ενδο-προτασιακή συνεκτικότητα εφαρμόζει αυτή την ιδέα σε γειτονικές προτάσεις. Δηλαδή, ο ίδιος σημασιολογικός προσανατολισμός εκφράζεται συνήθως σε διαδοχικές προτάσεις. Οι αλλαγές συναισθήματος υποδεικνύονται από αντιθετικές εκφράσεις όπως «but» και «however». Προτείνουν επίσης ορισμένα κριτήρια για τον καθορισμό της ένταξης της λέξης σε ένα θετικό ή αρνητικό λεξικό. Η μελέτη βασίστηκε σε ιαπωνικό κείμενο και χρησιμοποιήθηκε για να βρει λέξεις συναισθήματος με τους προσανατολισμούς τους, που εξαρτώνται από ένα συγκεκριμένο αντικείμενο αναφοράς.

Παρόλο που ο εντοπισμός του προσανατολισμού λέξεων συναισθήματος για ένα συγκεκριμένο αντικείμενο αναφοράς είναι χρήσιμο, στην πράξη είναι ανεπαρκές. Οι Ding, Liu και Yu (2008) [24] έδειξαν ότι πολλές λέξεις, στο ίδιο αντικείμενο αναφοράς, μπορεί να έχουν διαφορετικούς προσανατολισμούς σε διαφορετικά συμφραζόμενα ακόμα και κατά τη χρήση της στο ίδιο πεδίο. Διαπιστώνεται ότι τόσο η οπτική όσο και το συναίσθημα που εκφράζουν τα λόγια ήταν το ίδιο σημαντικό. Στη συνέχεια, πρότειναν να χρησιμοποιηθεί το ζεύγος (χαρακτηριστικό, λέξη συναισθήματος) ως πλαίσιο γνώμης (opinion context), π.χ., («διάρκεια ζωής της μπαταρίας", "μεγάλη"). Η μέθοδος τους προσδιορίζει έτσι τις λέξεις συναισθήματος και τους προσανατολισμούς τους μαζί με τα χαρακτηριστικά που τροποποιούν. Για να προσδιοριστεί αν ένα ζευγάρι είναι θετικό ή αρνητικό, εξακολουθούν να εφαρμόζονται οι παραπάνω ενδοπροτασιακοί και δια-προτασιακοί κανόνες συνεκτικότητας Το έργο των Ganapathibhotla και Liu, (2008) [35] υιοθέτησε τον ίδιο ορισμό αλλά τον χρησιμοποίησαν για την ανάλυση των συγκριτικών προτάσεων. Οι Wu και Wen (2010) [84] ασχολήθηκαν με ένα παρόμοιο πρόβλημα στην κινέζικη γλώσσα. Ωστόσο, επικεντρώθηκαν μόνο σε ζευγάρια στα οποία οι επίθετα είναι ποσοδείκτες, όπως «big, small, low and high» και η μέθοδός τους βασίζεται σε συντακτικά μοτίβα [78].

Οι Lu et al. (2011) [51] χρησιμοποίησαν το ίδιο αντικείμενο αναφοράς, όπως και οι Ding, Liu και Yu, (2008) [24],και υπέθεσαν ότι το σύνολο των χαρακτηριστικών ήταν δεδομένο. Οι τελευταίοι διετύπωσαν το πρόβλημα της ανάθεσης σε κάθε ζεύγος θετικό ή αρνητικό συναίσθημα ως ένα πρόβλημα βελτιστοποίησης με μια σειρά από περιορισμούς. Η αντικειμενική συνάρτηση και οι περιορισμοί έχουν σχεδιαστεί με βάση στοιχεία όπως ένα λεξικό συναισθημάτων γενικού σκοπού, η συνολική βαθμολογία συναισθήματος κάθε αναφοράς, τα συνώνυμα και αντώνυμα, καθώς και οι συμπλεκτικοί και αντιθετικοί κανόνες που πηγάζουν από την παρουσία των συνδεσμων «and« και «but» αντίστοιχα. Σε κάποιο βαθμό, η μέθοδος του (Turney, 2002) [78] μπορεί επίσης να θεωρηθεί ως έμμεση μέθοδος για την εύρεση απόψεων σε συγκεκριμένα συμφραζόμενα, αλλά δεν χρησιμοποίησαν την ιδέα της Συναισθηματικής Συνεκτικότητας, αλλά χρησιμοποίησαν το Διαδίκτυο για να βρουν τον προσανατολισμό τους.

Στην ίδια γραμμή κινούνται και οι Wilson, Wiebe, και Hoffmann (2005) [83], που μελέτησαν τις συμφραζόμενες υποκειμενικότητες και τα συναισθήματα σε επίπεδο φράσης ή έκφρασης. Ως Συμφραζόμενο Συναίσθημα (Contextual sentiment) ορίζουν το σημασιολογικό προσανατολισμό των συμφραζομένων της φράσης στην οποία ανήκει μία θετική ή αρνητική λέξη του λεξικού, ανεξάρτητα από το δικό της προσανατολισμό. Σε αυτό το έργο, αρχικά επισημαίνονται οι υποκειμενικές εκφράσεις στο σώμα, δηλαδή αυτές οι εκφράσεις που περιέχουν υποκειμενικές λέξεις ή φράσεις με βάση ένα συγκεκριμένο λεξικό υποκειμενικότητας. Σημειώνεται ότι ένα Λεξικό Υποκειμενικότητας (subjectivity lexicon) είναι ελαφρώς διαφορετικό από ένα Λεξικό Συναισθημάτων, επειδή περιέχει λέξεις που δηλώνουν μόνο την υποκειμενικότητα, αλλά όχι το συναίσθημα.

Ο στόχος του έργου ήταν να χαρακτηρίσει το συμφραζόμενο συναίσθημα στις δεδομένες εκφράσεις που περιέχουν εκφάνσεις υποκειμενικών στοιχείων, στο λεξικό υποκειμενικότητας. Χρησιμοποιήθηκε μία μέθοδος επιβλεπόμενης μάθησης με δύο στάδια. Στο πρώτο, προσδιορίζει αν η έκφραση είναι υποκειμενική ή αντικειμενική και στο δεύτερο, αν η υποκειμενική έκφραση είναι θετική, αρνητική, και τα δύο, ή ουδέτερη. Στην περίπτωση που μία έκφραση είναι και θετική και αρνητική, συνυπάρχουν και τα δύο συναισθήματα. Η περίπτωση της ουδέτερης εξακολουθεί να περιλαμβάνεται, επειδή το πρώτο στάδιο μπορεί να κάνει λάθη και να άφησε κάποιες ουδέτερες εκφράσεις αταυτοποίητες. Για την κατάταξη υποκειμενικότητας, χρησιμοποιήθηκε ένα μεγάλο και πλούσιο σύνολο χαρακτηριστικών (λεκτικών, προτασιακών, δομικών, διαμόρφωσης και εγγράφου). Για το δεύτερο στάδιο της ταξινόμησης συναισθήματος, χρησιμοποιηθήκαν χαρακτηριστικά, όπως «word tokens», «word prior sentiments», «negations», «modified by polarity» και «conj polarity».

Μια ταξινόμηση συναισθήματος σε επίπεδο έκφρασης πραγματοποιήθηκε και από τους Choi και Cardie, (2008) [20], που ταξινομούν τις εκφράσεις στο ηλεκτρονικό σώμα κειμένου Multi-Perspective Question Answering (MPQA), πειραματιζόμενοι τόσο στην ταξινόμηση με βάση κάποιο λεξικό όσο και με επιβλεπόμενη μάθηση. Οι

Breck, Choi και Cardie, (2007) [14], μελέτησαν το πρόβλημα της εξαγωγής εκφράσεων συναισθήματος με οποιοδήποτε αριθμό των λέξεων χρησιμοποιώντας υπό συνθήκη μαρκοβιανά πεδία (Conditional Random Fields - CRF).

4.2.2 Μέθοδοι Διαφορετικού Γνωσιακού Τομέα

Το πρόβλημα της προσαρμογής ενός γενικού λεξικού σε ένα νέο, για την ταξινόμηση συναισθήματος ενός αντικείμενου αναφοράς σε επίπεδο συγκεκριμένης φράσης, μελετήθηκε από τους Choi και Cardie (2009) [21]. Η τεχνική τους προσάρμοσε τις πολικότητες σε επίπεδο λέξης, ενός λεξικού συναισθημάτων γενικού σκοπού, σε ένα συγκεκριμένο τομέα με την αξιοποίηση της πολικότητας σε επίπεδο φράσης, και αντιστρόφως. Οι προσαρμοσμένες πολικότητες σε επίπεδο λέξης χρησιμοποιήθηκαν για να βελτιώσουν τις πολικότητες σε επίπεδο φράσης. Η συσχέτιση του επιπέδου λέξης και φράσης μοντελοποιήθηκε ως ένα σύνολο περιορισμών και το πρόβλημα λύθηκε με τη χρήση ακέραιου γραμμικού προγραμματισμού. Συγκεκριμένα, θεώρησαν ότι υπήρχε λεξικό πολικότητας γενικής χρήσης L, και ένας αλγόριθμος ταξινόμησης πολικότητας $f(e_l, L)$, που μπορεί να προσδιορίσει την πολικότητα της γνώμης μιας φράσης e_l, με βάση τις λέξεις στην e_l και L .

Οι Du et al. (2010) [26] μελέτησαν το πρόβλημα της προσαρμογής του λεξικού συναισθημάτων ενός αντικειμένου αναφοράς (όχι γενικής χρήσης) σε ένα άλλο. Ως είσοδο, ο αλγόριθμος υποθέτει την ύπαρξη ενός συνόλου εγγράφων ενός αντικειμένου αναφοράς, που φέρουν ένα σύνολο λέξεων συναισθήματος από τα έγγραφα, καθώς και μια σειρά από έγγραφα εκτός αντικειμένου. Ο στόχος ήταν να κάνουν το λεξικό εντός αντικειμένου αναφοράς κατάλληλο για έγγραφα εκτός αντικειμένου. Η μέθοδος αυτή βασίζεται σε δύο ιδέες. Πρώτον, ένα έγγραφο θα πρέπει να είναι θετικό (ή αρνητικό), εάν περιέχει πολλές θετικές (ή αρνητικές) λέξεις, και μια λέξη πρέπει να είναι θετική (ή αρνητική) εάν εμφανιστεί σε πολλά θετικά (ή αρνητικά) έγγραφα. Αυτή είναι η σχέση αμοιβαίας ενίσχυσης (mutual reinforcement relationships). Δεύτερον, ακόμη και αν τα δύο αντικείμενα αναφοράς μπορούν να είναι υπό διαφορετικές κατανομές, είναι δυνατόν να προσδιοριστεί ένα κοινό αντικείμενο μεταξύ αυτών (π.χ. η ίδια λέξη έχει τον ίδιο προσανατολισμό). Η προσαρμογή ενός λεξικού συναισθημάτων λύθηκε χρησιμοποιώντας το πλαίσιο πληροφοριών συμφόρησης (information bottleneck framework).

Σε ένα ελαφρώς διαφορετικό θέμα, Wiebe και Mihalcea (2006) [82] διερεύνησαν τη δυνατότητα ανάθεσης ετικετών υποκειμενικότητας στις λεκτικές έννοιες βάσει σε ένα ηλεκτρονικό σώμα κειμένου. Διεξήχθησαν δύο μελέτες. Η πρώτη μελέτη εξέτασε τη συμφωνία μεταξύ σχολιαστών που χειροκίνητα ανέθεσαν υποκειμενικές, αντικειμενικές, ή και τις δύο, ετικέτες σε έννοιες του WordNet. Η δεύτερη μελέτη αξιολόγησε μια μέθοδο για την αυτόματη εκχώρηση των ετικετών υποκειμενικότητας στην έννοια κάθε λέξης. Η μέθοδος αυτή βασίζεται στην ομοιότητα κατανομής (distributional similarity). Απέδειξαν ότι η υποκειμενικότητα είναι μια ιδιότητα που

μπορεί να σχετίζεται με τις έννοιες μιας λέξης και η αποσαφήνιση της έννοιας μιας λέξης μπορεί να επωφεληθεί άμεσα από το χαρακτηρισμό της υποκειμενικότητας. Μια μεταγενέστερη εργασία των Akkaya, Wiebe και Mihalcea, (2009) [1], μελέτησε επίσης το πρόβλημα και πραγματοποίησε μια περιπτωσιολογική μελέτη για την αναγνώριση της υποκειμενικότητας.

Οι Brody και Diakopoulos (2011) [15] μελέτησαν την επιμήκυνση των λέξεων (π.χ., slooooow) σε microblogs. Έδειξαν ότι η επιμήκυνση συνδέεται στενά με την υποκειμενικότητα και το συναίσθημα, και παρουσίασαν έναν αυτόματο τρόπο για να αξιοποιήσουν αυτή τη συσχέτιση για τον εντοπισμό του αντικειμένου αναφοράς συναισθήματος και τη συγκίνηση στις λέξεις.

Τέλος, οι Feng, Bose και Choi (2011) [32] μελέτησαν το πρόβλημα της παραγωγής ενός λεξικού χροιάς (connotation lexicon). Ένα λεξικό χροιάς διαφέρει από ένα λεξικό συναισθήματος από το γεγονός ότι το δεύτερο αφορά λέξεις που εκφράζουν το συναίσθημα, είτε ρητά είτε σιωπηρά, ενώ το πρώτο αφορά τις λέξεις που συχνά συνδέονται με μια συγκεκριμένη πολικότητα συναισθήματος, π.χ., οι λέξεις «award» και «promotion» έχουν θετική χροιά και οι «cancer» και «war» έχουν αρνητική χροιά. Για να λυθεί το παραπάνω πρόβλημα προτάθηκε μία μέθοδος, που βασίζεται σε γράφους και αντλεί πληροφορίες από την αμοιβαία ενίσχυση.

4.3 Μικτές Προσεγγίσεις

Οι Peng και Park (2011) [65] παρουσίασαν μια μέθοδο δημιουργίας λεξικού συναισθημάτων χρησιμοποιώντας παραγοντοποίηση μη αρνητικής μήτρας με περιορισμούς συμμετρίας (CSNMF). Αρχικά χρησιμοποιείται μια bootstrapping μέθοδος για να βρεθεί μια σειρά από υποψήφιες λέξεις συναισθήματος σε ένα λεξικό και στη συνέχεια, χρησιμοποιείται ένα μεγάλο σώμα κειμένου που αποδίδει βαθμολογίες πολικότητας σε κάθε λέξη. Αυτή η μέθοδος χρησιμοποιεί έτσι και το λεξικό, αλλά και το σώμα κειμένου, όπως και αρκετές άλλες ολοκληρωμένες μέθοδοι που συνδυάζουν τις δύο παραπάνω προσεγγίσεις, που βασίζονται στη διάδοση ετικέτας σε γράφους ομοιότητας [85], και αναλύονται εκτενώς στα επόμενα κεφάλαια.

5. Διάδοση Ετικέτας

5.1 Προσημασμένα Δεδομένα

Οι εργασίες μέχρι τις αρχές του 1990 σχετικά με τομείς συναφείς της Ανάλυσης Συναισθήματος, όπως ο καθορισμός της γνώμης ενός κειμένου και της αναγνώρισης σύνθετων προβλημάτων, σε γενικές γραμμές θεωρούσαν απαραίτητη την ύπαρξη υποσυστημάτων για μερικές εξεζητημένες εργασίες της Επεξεργασίας Φυσικής Γλώσσας (NLP). Με δεδομένη την έλλειψη επαρκών ποσοτήτων προσημασμένων δεδομένων (labeled data), η έρευνα που περιγράφεται σε αυτές τις πρώτες εργασίες, αναγκαστικά λάμβανε υπόψη μόνο τις προτάσεις για αυτά τα πρωτότυπα συστήματα, χωρίς μεγάλης κλίμακας εμπειρική αξιολόγηση. Κατά κανόνα, δεν αναμιγνυόταν καμία συνιστώσα μάθησης σε αυτά. Τα λειτουργικά συστήματα επικεντρώνονταν σε απλούστερα θέματα κατάταξης, όπως χειροποίητα λεξικά διακεκριμένων λέξεων [3], δεδομένου ότι τέτοια λεξικά, μπορούσαν να χαρακτηρίσουν μία μονάδα κειμένου με την εξέταση ποιοι όροι ή φράσεις από το λεξικό εμφανίζονταν στο δεδομένο κείμενο.

Η άνοδος των ευρύτερα, ερευνητικώς διαθέσιμων συλλογών, προσανατολισμένων εγγράφων (ιστοσελίδες συζήτησης οικονομικών νέων και συλλογής κριτικών, όπως πχ η Epinions) και άλλων σωμάτων κειμένου πιο γενικού περιεχομένου (π.χ. ειδήσεων) και άλλων πόρων (π.χ. το WordNet) αποτέλεσαν το εφαλτήριο μιας μεγάλης στροφής, προς τις προσεγγίσεις που βασίζονται σε προσημασμένα δεδομένα. Κατ 'αρχάς, η διαθεσιμότητα των πρώτων μη προσημασμένων κειμένων έδωσε τη δυνατότητα στην δημιουργία λεξικών με μη επιβλεπόμενο τρόπο αντί για χειροκίνητα.

Τα δεδομένα τα οποία έχουν προηγουμένως χαρακτηρισθεί με ετικέτες, είναι απαραίτητα για την επιβλεπόμενη μάθηση. Ωστόσο συχνά είναι διαθέσιμα μόνο σε μικρές ποσότητες, ενώ τα μη προσημασμένα δεδομένα μπορεί να είναι άφθονα. Η χρήση μη προσημασμένων δεδομένων μαζί με δεδομένα με ετικέτες, προκαλούν τόσο θεωρητικό όσο και πρακτικό ενδιαφέρον.

5.2 Κ-Πλησιέστεροι Γείτονες

Πολλές προσεγγίσεις έχουν προταθεί για το συνδυασμό μη επισημασμένων και προσημασμένων δεδομένων [70]. Ανάμεσά τους υπάρχει μια πολλά υποσχόμενη οικογένεια των μεθόδων που υποθέτουν ότι τα κοντινότερα σημεία δεδομένων τείνουν να έχουν παρόμοιες ετικέτες τάξεων, κατά τρόπο ανάλογο με τον Κ-

Πλησιέστερο Γείτονα (K-Nearest Neighbors – K-NN), στην παραδοσιακή επιβλεπόμενη μάθηση.

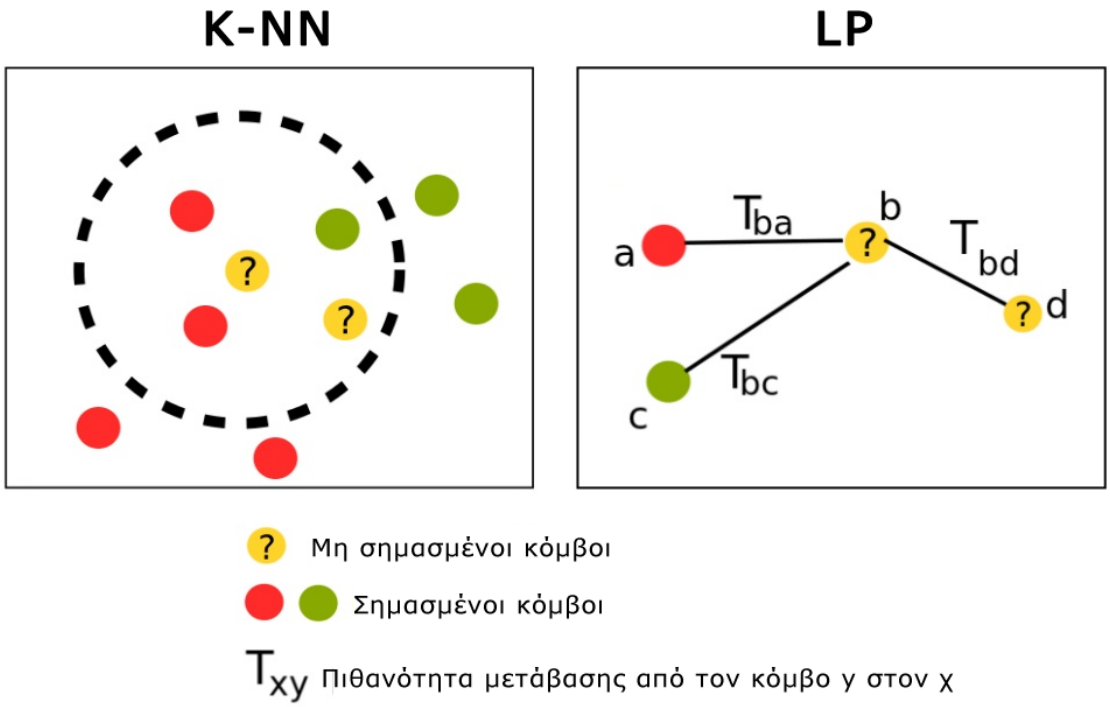
Στην Αναγνώριση Προτύπων, ο αλγόριθμος Κ-Πλησιέστερων Γειτόνων (K-NN) είναι μια μη-παραμετρική μέθοδος που χρησιμοποιείται για την ταξινόμηση και την παλινδρόμηση. Σε αμφότερες τις περιπτώσεις, η είσοδος αποτελείται από τα Κ πλησιέστερα δείγματα στο χώρο των χαρακτηριστικών. Η έξοδος εξαρτάται από το αν αλγόριθμος k-NN χρησιμοποιείται για την ταξινόμηση ή την παλινδρόμηση. Στην K-NN ταξινόμηση, η έξοδος ανήκει σε κάποια κατηγορία. Ένα αντικείμενο χαρακτηρίζεται από την πλειοψηφία των γειτόνων του, και υπάγεται στην πιο κοινή κλάση μεταξύ των Κ πλησιέστερων γειτόνων του (όπου Κ ένας θετικός ακέραιος). Αν $k = 1$ τότε το αντικείμενο ανατίθεται στην κατηγορία του εν λόγω πλησιέστερου γείτονα. Αντιθέτως, στην k-NN παλινδρόμηση, η έξοδος είναι η τιμή του μέσου όρου των τιμών των Κ πλησιέστερων γειτόνων της [4].

Τόσο για την ταξινόμηση όσο και την παλινδρόμηση μπορεί να είναι χρήσιμο να ορισθεί το βάρος σε κάθε συνεισφορά ενός γείτονα, έτσι ώστε οι πιο κοντινοί από τους γείτονες να συμβάλλουν περισσότερο στο μέσο όρο από τους πιο μακρινούς. Για παράδειγμα, ένα συνηθισμένο σχήμα στάθμισης συνίσταται στην παροχή σε κάθε γείτονα ενός βάρους $\frac{1}{d}$, όπου d είναι η απόσταση από το γείτονα. Οι γείτονες επιλέγονται από ένα σύνολο αντικειμένων για τα οποία είναι γνωστή η κατηγορία (για την ταξινόμηση K-NN) ή η αξία της τιμής τους (για την K-NN παλινδρόμηση). Αυτό μπορεί να θεωρηθεί ως το σύνολο εκπαίδευσης του αλγόριθμου, αν και δεν απαιτείται ρητή βήμα εκπαίδευση.

Ο αλγόριθμος K-NN είναι ένας τύπος μάθησης που βασίζεται σε δείγματα (lazy learning), όπου η συνάρτηση προσεγγίζεται μόνο σε τοπικό επίπεδο και όλος ο υπολογισμός αναβάλλεται μέχρι την τελική ταξινόμηση. Ένα μειονέκτημα του αλγορίθμου K-NN είναι ότι είναι ευαίσθητος στην τοπική δομή των δεδομένων. Ο K-NN είναι μεταξύ των πιο απλών αλγορίθμων μηχανικής μάθησης. Ως αποτέλεσμα, οι μέθοδοι αυτές διαδίδουν ετικέτες σε πυκνές περιοχές μη προσημασμένων δεδομένων.

5.3 Αλγόριθμος Διάδοσης Ετικέτας

Εμπνεόμενοι από τον K-NN, οι Zhu, X., & Ghahramani, Z. (2002) [85], προτείνουν ένα νέο αλγόριθμο Διάδοσης Ετικέτας (Label Propagation - LP) και την αξιοποίηση της μάθησης από προσημασμένα ή μη, δεδομένα. Διατυπώνουν το πρόβλημα ως μια ιδιαίτερη μορφή της διάδοσης ετικέτας, όπου οι ετικέτες ενός κόμβου διαδίδονται σε όλους τους άλλους κόμβους, σύμφωνα με την εγγύτητά τους, διορθώνοντας παράλληλα τα ήδη προσημασμένα δεδομένα και πολλαπλασιάζοντας με έναν συνδυασμό Τυχαίου Περιπάτου (random walk) και Σύσφιξης (clamping). Έτσι, τα δεδομένα στα οποία έχουν αποδοθεί ετικέτες, λειτουργούν σαν πηγές που ωθούν αυτές τις ετικέτες στα μη προσημασμένα δεδομένα.



Σχήμα 5.1: Σύγκριση μεθόδου σήμανσης K-NN και LP

Αποδεικνύουν έπειτα τη σύγκλιση του αλγορίθμου τους, βρίσκοντας μια λύση κλειστού τύπου για ένα σταθερό σημείο και αναλύοντας τη συμπεριφορά του σε διάφορα σύνολα δεδομένων. Προτείνουν επίσης μια ευριστική μέθοδο που βασίζεται σε ένα ελάχιστο συνδετικό δέντρο, και το κριτήριο ελαχιστοποίησης της εντροπίας με δυνατότητα μάθησης παραμέτρων, και ανίχνευσης μη σχετικών χαρακτηριστικών. Όπως και με διάφορους αλγόριθμους ημι-επιβλεπόμενης μάθησης αυτού του είδους, ο LP λειτουργεί αποδοτικά μόνο αν η δομή της διανομής των δεδομένων, όπως αποκαλύπτεται από τα άφθονα μη προσημασμένα δεδομένα, ταιριάζει στην κατηγοριοποίηση που στοχεύουμε.

5.3.1 Διατύπωση του Προβλήματος

Έστω $(x_1, y_1) \dots (x_l, y_l)$ τα προσημασμένα δεδομένα, όπου $Y_L = \{y_1 \dots y_l\} \in \{1 \dots C\}$ οι ετικέτες κάθε κατηγορίας. Θεωρούμε ότι ο αριθμός των κατηγοριών C είναι δεδομένος και όλες οι κατηγορίες εμφανίζονται στα προσημασμένα δεδομένα. Έστω $(x_{1+1}, y_{1+1}) \dots (x_{l+u}, y_{l+u})$ τα μη προσημασμένα δεδομένα όπου $Y_U = \{y_{l+1} \dots y_{l+u}\}$ δεν έχουν παρατηρηθεί, με δεδομένο ότι συνήθως $l \ll u$. Έστω $X = \{x_1 \dots x_{l+u}\} \in R^D$. Το πρόβλημα εδράζεται στην εύρεση του Y_U με δεδομένα τα X και Y_L.

Διαισθητικά θα θέλαμε σημεία δεδομένων τα οποία είναι κοντά σε ό,τι αφορά τη συνάφεια των ετικετών τους. Δημιουργείτε λοιπόν ένας πλήρης γράφος που περιέχει ακμές για κάθε ζεύγος κόμβων του, όπου οι κόμβοι είναι σημεία δεδομένων, είτε επισημασμένα, είτε όχι. Η ακμή ανάμεσα σε δύο τέτοιους κόμβους i, j έχει τέτοιο βάρος ώστε όσο πιο κοντά είναι οι κόμβοι, σύμφωνα με την ευκλείδεια απόσταση d_{ij} , τόσο πιο μεγάλο να είναι το βάρος w_{ij}. Τα βάρη ελέγχονται από μία παράμετρο σ:

$$w_{ij} = exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) = exp\left(-\frac{\sum_{d=1}^{D}\left(x_i^d - x_j^d\right)^2}{\sigma^2}\right)$$

Και άλλες επιλογές εκτός από την ευκλείδεια απόσταση είναι δυνατές, και ίσως ακόμη πιο κατάλληλες, εάν το x είναι θετικό ή διακριτό. Δίνεται παρόλα αυτά έμφαση στην ευκλείδεια απόσταση με τη δυνατότητα χρήσης διαφορετικών σ για κάθε διάσταση ανάλογα με τις κλίμακες μήκους προς διάδοση.

Όλοι οι κόμβοι έχουν ελαφριές ετικέτες οι οποίες μπορούν να ερμηνευτούν ως κατανομές στις υπόλοιπες ετικέτες. Οι ετικέτες κάθε κόμβου διαδίδονται, μέσω των ακμών του σε όλους τους άλλους κόμβους. Οι ακμές με μεγαλύτερα βάρη επιτρέπουν την διάδοση μίας ετικέτας ευκολότερα.

Ορίζεται μία πιθανολογική μήτρα μετάβασης T , διαστάσεων $(l+u) \times (l+u)$:

$$T_{ij} = P(j \to i) = \frac{w_i}{\sum_{k=1}^{l+u} w_{kj}}$$

όπου T_{ij} είναι η πιθανότητα μετάβασης από τον κόμβο j στον i . Στα μαθηματικά, μία πιθανολογική μήτρα μετάβασης (που ονομάζεται επίσης στοχαστική μήτρα ή μήτρα Markov) είναι ένας πίνακας που χρησιμοποιείται για να περιγράψει τις μεταβάσεις σε μια αλυσίδα Markov, δηλαδή σε μια τυχαία διαδικασία, που δε διατηρεί μνήμη για τις προηγούμενες μεταβολές και κάθε επόμενη κατάσταση εξαρτάται μόνο από την τωρινή κατάσταση και σε καμιά περίπτωση από αυτές που προηγήθηκαν [6]. Οι καταχωρήσεις της είναι ένας μη αρνητικός πραγματικός αριθμός που αντιπροσωπεύει μια πιθανότητα.

Επίσης ορίζεται ένας πίνακας ετικετών Y, διαστάσεων $(l+u) \times C$, του οποίου η σειρά i, αντιπροσωπεύει τις πιθανότητες των ετικετών του κόμβου y_i. Η αρχικοποίηση των σειρών του Y εξαρτάται από τα μη προσημασμένα δεδομένα και δεν είναι καθοριστικής σημασίας.

5.3.2 Δομή Αλγορίθμου

Ο αλγόριθμος διάδοσης ετικέτας ακολουθεί τα εξής βήματα:

1. Όλοι οι κόμβοι διαδίδουν τις ετικέτες τους για ένα βήμα $Y \leftarrow TY$
2. Πραγματοποιείται κανονικοποίηση σε κάθε σειρά της μήτρας Y, για να διατηρηθεί η πιθανότητα κάθε κατηγορίας, σύμφωνα με την ορισθείσα ερμηνεία της
3. Πραγματοποιείται σύσφιξη των προσημασμένων δεδομένων και επανάληψη της διαδικασίας από το βήμα 2 μέχρι η μήτρα Y να συγκλίνει.

Το βήμα 3 είναι καθοριστικής σημασίας. Τα προσημασμένα σημεία δεδομένων δεν αφήνονται να εξασθενίσουν. Αν θεωρήσουμε ότι η X είναι μια διακριτή τυχαία μεταβλητή που εκφράζει την ετικέτα κάθε σημείου δεδομένων, της οποίας το σύνολο των τιμών είναι πεπερασμένο, τότε η συνάρτηση $f_X(x) = P_r\{X = x\}$ ονομάζεται Συνάρτηση Μάζας Πιθανότητας (Probability Mass Function) της X και έχει την ιδιότητα $\sum f_X(x) = 1$ [50]. Για να αποφευχθεί η εξασθένιση των προσημασμένων δεδομένων συσφίγγονται οι κατηγορίες $Y_{ic} = \delta(y_i, c)$, ώστε η μάζα πιθανότητας να συγκεντρώνεται στη συγκεκριμένη κατηγορία. Με τον όρο σύσφιξη νοείται η διαδικασία του περιορισμού μίας θέσης σε μια περιοχή, με τη μετακίνηση ενός σημείου δεδομένων στην πλησιέστερη διαθέσιμη τιμή ετικέτας.

Ακόμη και διαισθητικά, μπορεί να διαπιστωθεί, ότι με μία σταθερή ώθηση από τους κόμβους προσημασμένων δεδομένων, τα όρια των κατηγοριών θα εξωθηθούν μέσα από περιοχές δεδομένων υψηλής πυκνότητας και θα εγκατασταθούν σε κενά χαμηλής πυκνότητας. Αν η δομή των δεδομένων ταιριάζει με το στόχο της ταξινόμησης, ο αλγόριθμος αυτός μπορεί να χρησιμοποιήσει μη προσημασμένα δεδομένα, για να υποβοηθήσει τις τεχνικές μάθησης.

Ο αλγόριθμος συγκλίνει σε μία απλή λύση. Αρχικά, τα βήματα 1 και 2 μπορούν να συνδυαστούν, ώστε να πραγματοποιούν το σχηματισμό της μήτρας $Y \leftarrow \bar{T}Y$, με το $\bar{T}$ να είναι ένας, ανά σειρά, κανονικοποιημένος πίνακας του T. Η κανονικοποιημένη μορφή του πίνακα προκύπτει από τη σχέση $\bar{T}_{ij} = \frac{T_{ij}}{\sum_k T_{ik}}$. Έστω Y_L οι κορυφαίες l σειρές, του Y, που εκφράζει τον πίνακα με τα προσημασμένα δεδομένα και Y_U οι εναπομείνασες u σειρές. Παρατηρούμε ότι ο πίνακας Y_L δεν μεταβάλλεται ποτέ στην πραγματικότητα, από τη στιγμή που συσφίγγεται στο βήμα 3, και το επίκεντρο του ενδιαφέροντος παραμένει ο πίνακας Y_U.

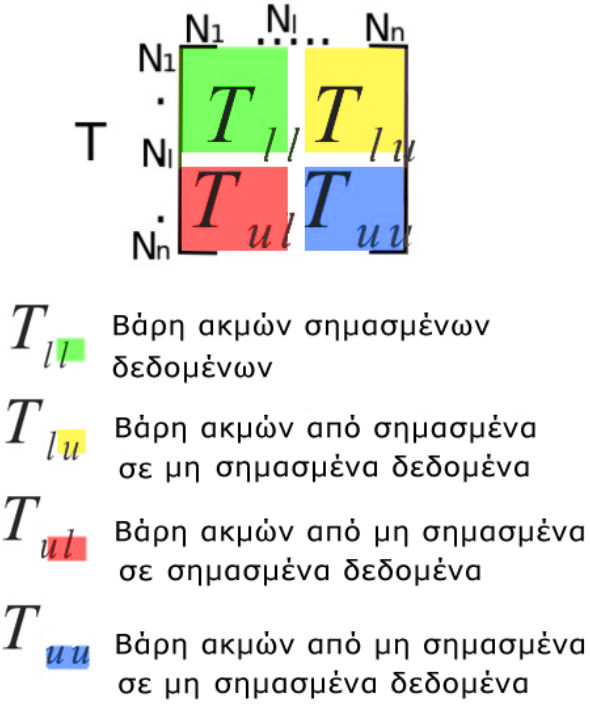
Η παραπάνω επαναληπτική μέθοδος εξαρτάται από την άμεση λύση των εξισώσεων μήτρας, που περιλαμβάνουν μήτρες πιο γενικές από τριδιαγώνιες μήτρες, δηλαδή πίνακες στους οποίους είναι δεν μη μηδενικά, μόνο τα στοιχεία της κύριας διαγωνίου και της διαγωνίου πάνω και κάτω από την κύρια. Αυτές οι εξισώσεις μητρών μπορεί συχνά να λυθούν άμεσα και αποτελεσματικότερα όταν γράφονται σαν διάσπαση της υπάρχουσας μήτρας. Στην αριθμητική γραμμική άλγεβρα, μια διάσπαση της μήτρας

είναι μια έκφραση που αντιπροσωπεύει μια δεδομένη μήτρα, ως άθροισμα ή διαφορά, άλλων μητρών [80]. Για αυτό το λόγο διασπάται η $\bar{T}$ μετά τη σειρά l και τη στήλη l, σε τέσσερεις υποπίνακες, όπως φαίνεται ακολούθως:

$$\bar{T} = \begin{bmatrix} \bar{T}_u & \bar{T}_{lu} \\ \bar{T}_{ul} & \bar{T}_{uu} \end{bmatrix}$$

Αποδεικνύεται ότι ο αλγόριθμος εξάγει το αποτέλεσμα της πράξης $Y_U \leftarrow \bar{T}_{uu} Y_U + \bar{T}_{ul} Y_L$, η οποία οδηγεί στο συμπέρασμα ότι $\lim_{n\to\infty} \bar{T}_{uu}^n Y^0 + \left[\sum_{i=1}^n \bar{T}_{uu}^{(i-1)}\right] \bar{T}_{ul} Y_L$, όπου ως Y^0 ορίζεται ο αρχικοποιημένος πίνακας Y.

Η ολοκλήρωση της παραπάνω απόδειξης καταλήγει στην αναγκαιότητα της απόδειξης ότι $\bar{T}_{uu}^n Y^0 \to 0$. Από κατασκευαστικής απόψεως όλα τα στοιχεία στο πίνακα $\bar{T}$ είναι μεγαλύτερα του μηδενός. Από τη στιγμή που ο $\bar{T}$ είναι ένας κανονικοποιημένος πίνακας ανά σειρά, και ο $\bar{T}_{uu}$ είναι υποπίνακας του $\bar{T}$, συνεπάγεται ότι $\exists \gamma < 1$, τέτοιο ώστε $\sum_{j=1}^u \bar{T}_{uuij} \le \gamma \,, \forall i = 1 \ldots u$. Ισχύει ακόμα ότι $\sum_{j=1}^u \bar{T}_{uuij} = \sum_j \sum_k \bar{T}_{uuik}^{(n-1)} \bar{T}_{uukj} = \sum_k \bar{T}_{uuik}^{(n-1)} \sum_j \bar{T}_{uukj} \le \sum_k \bar{T}_{uuik}^{(n-1)} \gamma \le \gamma^n$. Συνεπώς το άθροισμα των σειρών του $\bar{T}_{uu}^n$ συγκλίνει στο μηδέν, το οποίο με τη σειρά του συνεπάγεται ότι $\bar{T}_{uu}^n Y^0 \to 0$. Άρα το αρχικό σημείο Y^0 είναι επουσιώδες. Προφανώς η ισότητα $Y_U = (I - \bar{T}_{uu})^{-1} \bar{T}_{ul} Y_L$ είναι ένα σταθερό σημείο. Ως εκ τούτου, είναι το μοναδικό σταθερό σημείο και η λύση στον παραπάνω επαναληπτικό αλγόριθμο.



Σχήμα 5.2: Δομή Πιθανολογικής Μήτρας Μετάβασης

5.3.3 Αλγόριθμος LP/1-NN

Μία μέθοδος που συνδυάζει τον LP και τον K-NN (Label Propagating 1NN – LP/1NN) εντοπίζει τα σημεία δεδομένων x_u, μεταξύ των μη προσημασμένων δεδομένων, τα οποία είναι πιο κοντά στα προσημασμένα δεδομένα (έστω x_1). Στη συνέχεια αποδίδεται στο στοιχείο x_u η ετικέτα του x_1 και το στοιχείο x_u εντάσσεται πλέον στο σύνολο των επισημασμένων στοιχείων και η διαδικασία επαναλαμβάνεται. Ο αλγόριθμος LP/1-NN αποτελεί σαφώς, μία ακατέργαστη έκδοση, του LP.

Σχήμα 5.3: Στάδια διάδοσης ετικέτας στον LP/1-NN

5.4 Μέθοδοι Παραμετροποίησης

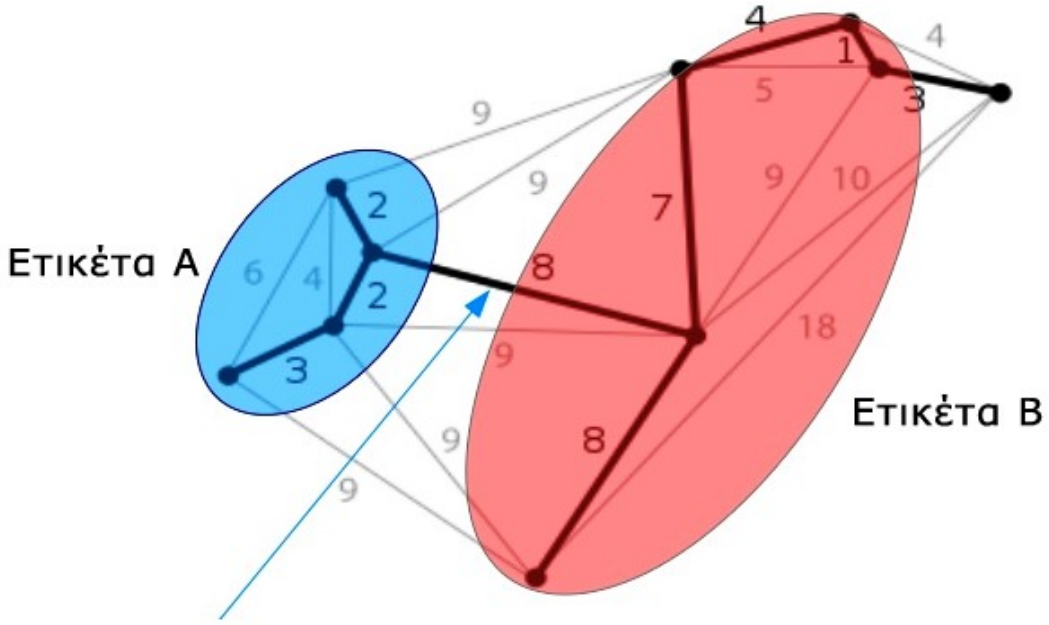
5.4.1 Ευριστική Μέθοδος

Με τον όρο Ευριστική (Heuristic) εννοούμε οποιαδήποτε μέθοδο ή προσέγγιση για την μάθηση, την ανακάλυψη και την επίλυση του προβλήματος, που χρησιμοποιεί μια πρακτική μέθοδο, η οποία δεν είναι εγγυημένη, βέλτιστη ή τέλεια, αλλά επαρκής για τους άμεσους στόχους. Συνεπώς, είναι οι στρατηγικές που προέρχονται από την εμπειρία με παρόμοια προβλήματα, με τη χρήση, εύκολα προσβάσιμων, πληροφοριών, που η εφαρμογή τους όμως ενδέχεται να είναι μικρή [63].

Η ευριστική στην συγκεκριμένη περίπτωση προϋποθέτει τον υπολογισμό του Ελάχιστα Συνδετικού Δένδρου (Minimum Spanning Tree - MST) πάνω από όλα τα σημεία δεδομένων, σύμφωνα με την Ευκλείδεια απόσταση d_{ij}, με χρήση του Αλγόριθμου Kruskal [49]. Στο ξεκίνημα κανένας κόμβος δεν είναι συνδεδεμένος με το δένδρο. Κατά τη διάρκεια της ανάπτυξης του δένδρου, οι άκρες εξετάζονται μία

προς μία από την μικρότερη στη μεγαλύτερη. Μια ακμή προστίθεται στο MST, εάν συνδέει δύο ξεχωριστά στοιχεία. Η διαδικασία επαναλαμβάνεται μέχρι να συνδεθεί ο συνολικός γράφος. Αυτό σημαίνει ότι ο αλγόριθμος βρίσκει ένα υποσύνολο των ακμών που σχηματίζουν ένα δέντρο, το οποίο περιλαμβάνει κάθε κορυφή, και ελαχιστοποιεί το συνολικό βάρος όλων των ακμών σε αυτό.

Καταρχάς, εντοπίζεται η πρώτη ακμή δέντρου που συνδέει δύο κόμβους ή υποδέντρα του γράφου με διαφορετικά προσημασμένα σημεία δεδομένων. Θεωρούμε το μήκος της ακμής αυτής d^0, ως την ευρεστική της ελάχιστης απόστασης μεταξύ των κατηγοριών. Με την ελπίδα ότι η τοπική διάδοση της ετικέτας είναι ως επί το πλείστον μέσα στις επιθυμητές κατηγορίες, τίθεται η μεταβλητή $\sigma = \frac{d^0}{3}$, έτσι ώστε το βάρος αυτής της μέγιστης, έως τώρα, ακμής να είναι κοντά στο 0.



Σχήμα 5.4: Παράδειγμα ελάχιστης ακμής υπογράφων διαφορετικής ετικέτας

5.4.2 Μέθοδος Εντροπίας

Η επιλογή της τιμής του σ, εναλλακτικά, θα μπορούσε να προκύψει ως αποτέλεσμα μίας διαδικασίας μάθησης βασιζόμενη στο κριτήριο της εντροπίας. Όταν το $\sigma \to 0$, το αποτέλεσμα του LP προσεγγίζει εκείνο του P1NN, διότι υπό την επιβάρυνση των εκθετικών βαρών κάθε ακμής, η επίδραση του κοντινότερου σημείου δεδομένων επικρατεί. Όταν το $\sigma \to \infty$, όλο το σύνολο δεδομένων συρρικνώνεται δραματικά, σε ένα μόνο σημείο. Όλα τα μη προσημασμένα δεδομένα δέχονται την ίδια επίδραση από όλα τα προσημασμένα σημεία, γεγονός που δημιουργεί ίδιες πιθανότητες κατηγοριοποίησης. Το κατάλληλο σ προφανώς και βρίσκεται ανάμεσα σε αυτές τις ακραίες τιμές.

Ως Εντροπία Γράφου Η(G, P) στη Θεωρίας Πληροφορίας, θεωρούμε μία λειτουργικότητα, ενός γράφου G, με μία κατανομή πιθανότητας P στους κόμβους του. Παράλληλα, η Εντροπία μιας διακριτής τυχαίας μεταβλητής Χ είναι ένα μέτρο της ποσότητας της αβεβαιότητας που σχετίζεται με τη τιμή του Χ. Η εντροπία πληροφοριών ποσοτικοποιεί την αβεβαιότητα που εμπλέκεται στην πρόβλεψη της τιμής μίας τυχαίας μεταβλητής. Στόχος της μεθόδου, συνεπώς, είναι η ελαχιστοποίηση της εντροπίας $H = -\sum_{ij} Y_{ij} log Y_{ij}$, η οποία είναι το άθροισμα της εντροπίας των μεμονωμένων σημείων. Αυτό αποτυπώνει την διαίσθηση ότι ένα κατάλληλο σ θα πρέπει να επισημαίνει όλα τα σημεία με βεβαιότητα.

Υπάρχουν πολλές αυθαίρετες αποδόσεις ετικέτας σε μη προσημασμένα δεδομένα που έχουν αρκετά χαμηλή εντροπία, η οποία θα μπορούσε να υποδηλώνει ότι το κριτήριο αυτό δεν είναι ιδιαίτερα λειτουργικό. Ωστόσο, είναι σημαντικό να επισημανθεί ότι οι περισσότερες από αυτές τις αυθαίρετες επισημάνσεις χαμηλής εντροπίας δεν παράγονται από τον εξεταζόμενο αλγόριθμο διάδοσης ετικέτας. Στην πραγματικότητα, διαπιστώνεται ότι ο χώρος των ετικετών χαμηλής εντροπίας, είναι μικρός και αποτελεί πρόσφορο μέσο για ρύθμιση της παραμέτρου σ .

Η μόνη επιπλοκή που παραμένει είναι η ελαχιστοποίηση του H και ο μηδενισμός του, όταν $\sigma \to 0$. Για τον παραπάνω λόγο απαιτείται η εξομάλυνση της μήτρας T με μία ομοιόμορφη μήτρα μετάβασης U , όπου $U_{ij} = \frac{1}{(l+u)}$, $\forall\, i, j$ [57]. Μετά την εξομάλυνση η μήτρα T αποκτά την ακόλουθη μορφή: $\tilde{T} = \epsilon U + (1-\epsilon)T$ και η μήτρα $\tilde{T}$ αντικαθιστά πλήρως την T, κατά την εκτέλεση του αλγορίθμου, με τον καθορισμό των παραμέτρων του, με τη μέθοδο της εντροπίας. Παρά το γεγονός ότι είναι αναγκαία η εισαγωγή μίας ακόμη παραμέτρου ϵ για τον καθορισμό της σ , το πλεονέκτημα είναι εμφανές όταν χρησιμοποιούνται πολλαπλές παράμετροι $\sigma_1 \ldots \sigma_D$ για κάθε μία διάσταση. Τα βάρη των ακμών πλέον παίρνουν την εξής μορφή:

$$w_{ij} = exp\left(-\frac{d_{ij}^2}{\sigma_d^2}\right) = exp\left(-\frac{\sum_{d=1}^{D}\left(x_i^d - x_j^d\right)^2}{\sigma_d^2}\right)$$

Οι παράμετροι σ_d είναι ανάλογες με τις κλίμακες μήκους μίας γκαουσιανής στοχαστικής διαδικασίας, δηλαδή μιας στατιστικής κατανομής όπου συμβαίνουν παρατηρήσεις σε μια συνεχή περιοχή και δηλώνουν πόσο κοντά πρέπει να είναι δύο σημεία ώστε να αλληλοεπηρεάζονται σημαντικά [68]. Σε μία γκαουσιανή διαδικασία, κάθε σημείο, στον συνεχή χώρο εισόδου, συνδέεται με μια κανονικά κατανεμημένη τυχαία μεταβλητή. Για τον υπολογισμό των $\sigma_1 \ldots \sigma_D$ που ελαχιστοποιούν την H χρησιμοποιείται η στοχαστική επικλινής κάθοδος (stochastic gradient descent), η οποία αποτελεί μία μέθοδο ελαχιστοποίησης μιας αντικειμενικής συνάρτησης, που γράφεται ως ένα άθροισμα διαφορίσιμων συναρτήσεων [13]. Με τις πολλαπλές παραμέτρους σ_d, ο αλγόριθμος μπορεί να ανιχνεύσει μη σχετικές διαστάσεις. Οι μεγάλες τιμές τους επιτρέπουν στις ετικέτες να διαδίδονται ελεύθερα κατά μήκος μίας διάστασης ανεξάρτητης από την κατηγοριοποίηση, εάν δεν είναι σχετική με αυτή,

καθώς το σύνολο δεδομένων της αυξάνει κατά τη διάρκεια της διαδικασίας μάθησης, παρόλο που τα δεδομένα θυσανώνονται προς αυτές τις διαστάσεις.

5.4.3 Σύγκριση Μεθόδων Παραμετροποίησης

Η χρήση της ευριστικής μεθόδου για τον υπολογισμό του σ μπορεί να λειτουργεί πολύ καλά σε κάποιες ειδικές εφαρμογές αλλά δεν προτιμάται πάντα, αφού το κριτήριο της εντροπίας μπορεί να εφαρμοστεί ακόμη και σε πιο γενικευμένες περιπτώσεις. Ούτως ή άλλως, η πιθανότητα ύπαρξης μίας ετικέτας σε ένα σημείο δεδομένων δεν μπορεί να λειτουργήσει ως κριτήριο, αφού με αυτή τη μέθοδο η ποιότητα της λύσης εξαρτάται από το πώς τα μη προσημασμένα δεδομένα δέχονται τις ετικέτες των προσημασμένων, και τα τελευταία ενίοτε είναι πολύ λίγα. Στην εφαρμογή του αλγόριθμου διάδοσης ετικέτας και οι δύο παραπάνω μέθοδοι μπορούν να αξιοποιηθούν για τον καθορισμό των παραμέτρων του, σύμφωνα με τη μορφή των δεδομένων.

5.5 Εξισορρόπηση κατανομών στις κατηγορίες

5.5.1 ML-Μέθοδος

Για τους σκοπούς της ταξινόμησης, μόλις υπολογισθεί το Y_U επιλέγεται ως ετικέτα για κάθε μη προσημασμένο σημείο δεδομένων η πιο πιθανή (most likely - ML) κατηγορία. Προφανώς μία προσέγγιση σύμφωνα με τον αλγόριθμο K-NN θα αποτύγχανε να ακολουθήσει τη δομή των δεδομένων, και η χρήση της ML-Μεθόδου αποδεικνύεται καταλυτική. Ωστόσο, η διαδικασία αυτή δεν παρέχει κανένα έλεγχο επί των τελικών κατανομών των κατηγοριών, οι οποίες εμμέσως καθορίζονται από την κατανομή των δεδομένων. Αν οι κατηγορίες δεν είναι καλά διαχωρισμένες και τα προσημασμένα στοιχεία δεδομένων είναι λιγοστά, απαιτείται η ενσωμάτωση περιορισμών στις κατανομές κατηγοριών για τη βελτίωση της τελικής κατάταξης. Οι κατανομές των κατηγοριών $P_1 \ldots P_C$ $(\sum_c P_c = 1)$ υπολογίζονται από τα προσημασμένα δεδομένα ή θεωρούνται ότι είναι δεδομένα a priori από μία εξωτερική αδιαμφισβήτητη πηγή αλήθειας (Oracle).

5.5.2 Μέθοδοι Μετεπεξεργασίας

Για την ανάθεση μίας κατηγορίας σε κάποιο δεδομένο, εκτός από την ML-Μέθοδο, μπορούν να εφαρμοστούν εναλλακτικά δύο μέθοδοι μετεπεξεργασίας (post-processing) των δεδομένων.

- CN-Μέθοδος: Αφορά την κανονικοποίηση της μάζας κάθε κατηγορίας (Class mass Normalization - CN). Πιο συγκεκριμένα υπολογίζονται οι συντελεστές λ_c , για την κλιμάκωση των στηλών Y_U, ώστε να ισχύει η ισότητα $\lambda_1 \sum Y_{U_1} : \ldots : \lambda_c \sum Y_{U_c} = P_1 : \ldots : P_C$. Η διαδικασία αυτή παρόλα αυτά δεν εγγυάται ότι μόλις ληφθούν οι αποφάσεις για κάθε σημείο δεδομένων, η κατανομή των κατηγοριών θα είναι στην πραγματικότητα ακριβώς ίση με $P_1 : \ldots : P_C$.

- LB-Μέθοδος: Ονομάζεται «Προσφορά Ετικέτας» (Label Biding - LB), θεωρώντας ότι υπάρχουν $u\, P_C$ στον αριθμό ετικέτες, για την κατηγορία , για διάθεση σε «πελάτες». Κάθε σημείο δεδομένων i , προσφέρει $Y_{U_{ic}}$ μονάδες αξίας, για την κατηγορία c. Οι προσφορές υποβάλλονται σε επεξεργασία από τις υψηλότερες μονάδες αξίας προς τις χαμηλότερες. Ως $Y_{U_{ic}}$ θεωρείται η στιγμιαία μέγιστη προσφορά. Εάν οι ετικέτες της κατηγορίας c παραμείνουν αμετάβλητες, μια ετικέτα της κατηγορίας c , «πωλείται» στο σημείο δεδομένων i , το οποίο στη συνέχεια κλείνει την προσφορά. Σε αντίθετη περίπτωση, η προσφορά αγνοείται και η δεύτερη υψηλότερη προσφορά τίθεται προς επεξεργασία, και ούτω καθεξής. Η μέθοδος «Προσφοράς Ετικέτας» σε αντίθεση με τη μέθοδο κανονικοποίησης της μάζας κάθε κατηγορίας, εγγυάται ότι θα υπάρξει αυστηρή ταύτιση των κατανομών των κατηγοριών και αποδίδει τα μέγιστα εάν οι ακριβείς κατανομές των κατηγοριών είναι γνωστές.

5.6 Σύγκριση με Αλγορίθμους Γράφων

5.6.1 Αλγόριθμος Τυχαίων Περιπάτων

Ο LP είναι στενά συνδεδεμένος με το μαρκοβιανό αλγόριθμο Τυχαίων Περιπάτων (Random Walks-RW). Και οι δύο χρησιμοποιούν την πολλαπλή δομή των δεδομένων, που ορίζεται από τη μεγάλη ποσότητα μη προσημασμένων στοιχείων τους, θεωρώντας ότι η δομή τους συσχετίζεται με το στόχο της ταξινόμησης. Και οι δύο επίσης, ορίζουν μια πιθανολογική διαδικασία για την διέλευση των ετικετών μεταξύ των κόμβων.

Ο RW προσεγγίζει το ίδιο πρόβλημα όμως από μια διαφορετική σκοπιά. Χρησιμοποιεί τη διαδικασία μετάβασης για να υπολογίσει τον πρόγονο t βημάτων βάθους συγγένειας κάθε κόμβου i, ο οποίος δεδομένου ότι ο τυχαίος περίπατος βρίσκεται στον κόμβο i, ισοδυναμεί με την πιθανότητα να βρισκόταν σε κάποιο κόμβο j πριν t βήματα. Κάθε κόμβος έχει δύο ξεχωριστά σύνολα ετικετών, ένα

παρατηρήσιμο και ένα μη. Μία παρατηρήσιμη ετικέτα του κόμβου i είναι ο μέσος όρος όλων των κρυφών ετικετών του κόμβου βεβαρυμμένες από τους προγόνους τους. Ο αλγόριθμος είναι ευαίσθητος στην χρονική κλίμακα t, αφού όταν $t \to \infty$, κάθε κόμβος φαίνεται εξίσου όμοιος σαν πρόγονος, και όλες οι παρατηρήσιμες ετικέτες είναι ίδιες. Με παρόμοιο τρόπο στον LP, τα προσημασμένα δεδομένα αποτελούν σταθερές πηγές οι οποίες ωθούν τις ετικέτες και το σύστημα επιτυγχάνει ισορροπία όταν $t \to \infty$.

5.6.2 Παλινδρόμηση Kernel

Η Παλινδρόμηση Kernel (Kernel Regression - KR) είναι μια μη-παραμετρική τεχνική, που εφαρμόζεται σε στατιστικά στοιχεία, για την εκτίμηση της, υπό όρους, προσδοκίας της τιμής, μιας τυχαίας μεταβλητής. Στόχος της είναι να βρεθεί μια μη γραμμική σχέση μεταξύ ενός ζεύγους τυχαίων μεταβλητών X και Y. Σε κάθε μη παραμετρική παλινδρόμηση, η υπό όρους προσδοκία μιας μεταβλητής Y σε σχέση με μία μεταβλητή X μπορεί να γραφεί: $E(Y|X) = m(X)$, όπου m είναι μία άγνωστη συνάρτηση [55].

Ο αλγόριθμος KR, όταν εφαρμόζεται σε γράφους δεδομένων, βασίζεται στην ιδέα της μοντελοποίησης της κατανομής των ετικετών κάθε κόμβου τους, για κάθε πεπερασμένο σύνολο δεδομένων, από μια κανονική κατανομή πολλών μεταβλητών. Εάν τα σημεία δεδομένων ενός γράφου G υπακούουν στην υπόθεση Markov και η διαδικασία είναι χωρίς μνήμη, οι ανεξάρτητοι περιορισμοί ανακλώνται στις μηδενικές εγγραφές στον Αντίστροφο Πίνακα Συνδιασποράς. Αυτό περιορίζει σαφώς την επιλογή των πιθανών συναρτήσεων του πυρήνα, σε τέτοια δεδομένα [37]. Μία συνάρτηση kernel που βασίζεται σε τυχαίους περιπάτους, μπορεί να εκμεταλλευτεί το μήκος όλων των διαδρομών, ανάμεσα σε όλα τα ζεύγη κόμβων, μιας δεδομένης ετικέτας.

Η παραπάνω διαδικασία ταυτίζεται σε αρκετά σημεία τόσο με τον αλγόριθμο τυχαίων περιπάτων, όσο και με τον LP. Και σε αυτόν χρησιμοποιείται η διαδικασία μετάβασης, για να υπολογίσει τον πρόγονο t βημάτων βάθους συγγένειας, κάθε κόμβου i, που ισοδυναμεί με την πιθανότητα να βρισκόταν σε κάποιο κόμβο j πριν t βήματα. Στη συγκεκριμένη περίπτωση θεωρούμε ότι ο πυρήνας (kernel) είναι πρόγονος βάθους συγγένειας t βημάτων και οι μη παρατηρίσημες ετικέτες διέρχονται της διαδικασίας μάθησης, έτσι ώστε η πιθανότητα των παρατηρούμενων ετικετών να βελτιστοποιείται.

5.6.3 Προσέγγιση Μέσου Πεδίου

Μεγάλη ομοιότητα παρουσιάζει το αποτέλεσμα της σύγκλισης της λύσης, μεταξύ LP και Προσέγγισης Μέσου Πεδίου (Mean Field Approximation - MFA) [42]. Στον LP,

μετά από τη σύγκλιση έχουμε τις ακόλουθες εξισώσεις για τα μη προσημασμένα δεδομένα: $Y_{ic} = \frac{\sum_j T_{ij} Y_{jc}}{\sum_{c'} \sum_j T_{ij} Y_{jc'}}$. Ο γράφος προσημασμένων δεδομένων μπορεί να θεωρηθεί ως ένα τυχαίο πεδίο F, που πληρεί τις υποθέσεις Markov με ζεύγη αλληλεπίδρασης w_{ij}, μεταξύ των κόμβων i, j, και με τους κόμβους προσημασμένων δεδομένων συνεσφιγμένους. Κάθε μη συνεσφιγμένος κόμβος i στο F, μπορεί να είναι σε μία από τις καταστάσεις C , που μπορούν να συμβολιστούν ως $Y_{ic} = \big(\delta(y_i, 1) \dots \delta(y_i, C)\big)$. Η πιθανότητα μίας συγκεκριμένης διαμόρφωσης του Y στο F είναι $P_F(Y) = \frac{1}{z} exp\big[log\big(\sum_{ij} w_{ij} Y_i Y_j^T\big)\big]$. Ο LP προσεγγίζει την παραπάνω λύση, σε ένα πεδίο F' , που προσεγγίζει το F. Το F' υπολογίζεται από τη σχέση $P_{F'}(Y) = \frac{1}{z} exp\big[log\big(\sum_{ij} w_{ij} Y_i Y_j^T\big)\big] \approx \frac{1}{z} exp\big[log\big(\sum_{ij} w_{ij} Y_i Y_j^T - 1\big)\big] = P_F(Y)$.

Εξάγεται λοιπόν ότι η λύση της προσέγγισης μέσου πεδίου F' είναι: $\langle Y_{ic} \rangle = \frac{\sum_j w_{ij} \langle Y_{jc} \rangle}{\sum_{c'} \sum_j w_{ij} \langle Y_{jc} \rangle}$ όπου με $\langle\ \rangle$ συμβολίζεται το μέσο. Συνεπώς η Y_{ic} προσεγγίζει την $\langle Y_{ic} \rangle$ υπό την έννοια ότι εάν τα αθροίσματα $\sum_k w_{ik}$ είναι ίδια για όλα τα i, μπορεί να αντικατασταθεί η μήτρα T_{ij} με την w_{ij} στην εξίσωση της Y_{ic} , το οποίο συνεπάγεται ότι ο η λύση του LP προσεγγίζει του MFA, στο F.

5.6.4 Αλγόριθμος Ελάχιστης Τομής

Στη θεωρία γραφημάτων μια ελάχιστη τομή ενός γράφου είναι η ελάχιστη διαμέριση των κορυφών του σε δύο ξένα μεταξύ τους υποσύνολα, που ενώνονται με ένα τουλάχιστον άκρο. Στην περίπτωση που ο γράφος είναι βεβαρυμμένων ακμών και μη κατευθυνόμενος, η τομή χωρίζει ένα συγκεκριμένο ζεύγος κόμβων και έχει το ελάχιστο βάρος [12].

Ο Αλγόριθμος Ελάχιστης Τομής (Mincut Algorithm - MA) βρίσκει την πιο πιθανή διαμόρφωση της κατάστασης του ίδιου τυχαίου μαρκοβιανού πεδίου F, δεδομένου ότι η ελάχιστη τομή αντιστοιχεί στην ελάχιστη ενέργεια. Σύμφωνα με τη φιλοσοφία του MA, τα σημεία των μη προσημασμένων δεδομένων, τα οποία δέχονται επιρροή από δύο διαφορετικές επισημασμένες περιοχές, θα λάβουν εν τέλει ετικέτες μόνο της μίας, η οποία και θα κυριαρχήσει στο σύνολο των ετικετών, των κόμβων. Σε αντίθεση με αυτήν την τεχνική, στον LP τα ίδια μη προσημασμένα σημεία, θα επηρεασθούν και από τις δύο εκατέρωθεν περιοχές. Επιπλέον, ο LP δεν περιορίζεται μονάχα στη δυαδική ταξινόμηση όπως ο MA.

6. Διάδοση Ετικέτας στην Ανάλυση Συναισθήματος

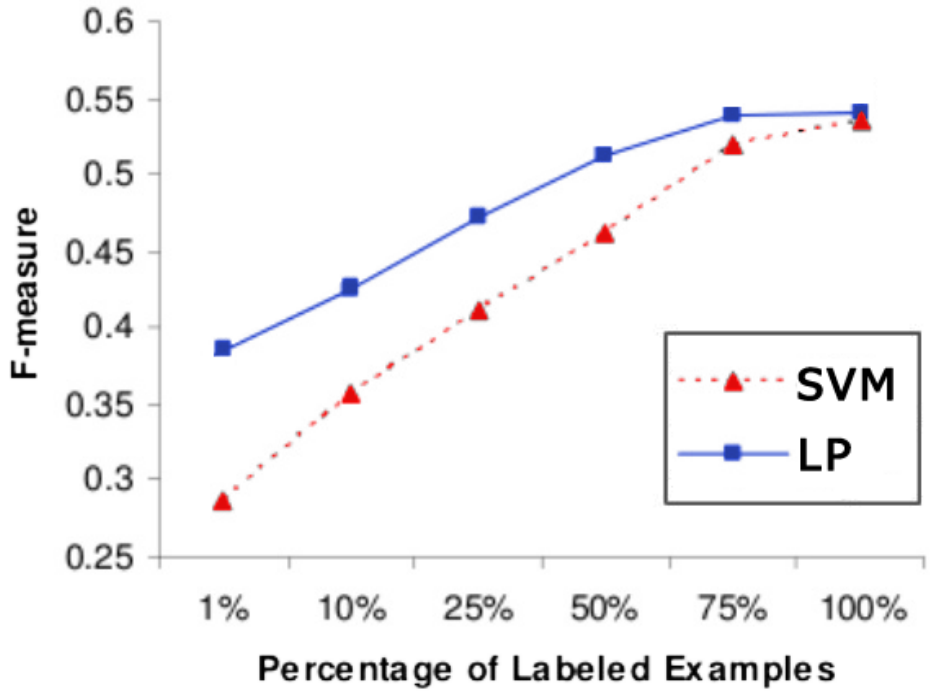
Οι επιβλεπόμενες μέθοδοι μηχανικής μάθησης έχουν επιτύχει σημαντικά σε μια ευρεία ποικιλία των τομέων που κυμαίνονται από την Επεξεργασίας Φυσικής Γλώσσας μέχρι την Αναγνώριση Ομιλίας. Οι πρώτες μέθοδοι επιβλεπόμενης ταξινόμησης [60] ως αντιστάθμισμα απαιτούσαν κείμενα προσημασμένης πολικότητας, ως είσοδο και δεν είχαν τη δυνατότητα προσαρμογής στις αλλαγές στη χρήση της γλώσσας. Δυστυχώς, η προετοιμασία των προσημασμένων δεδομένων, για τις μεθόδους αυτές ήταν συχνά δαπανηρή και χρονοβόρα, ενώ ήταν ευρέως διαθέσιμα σε πολλές περιπτώσεις τα μη προσημασμένα δεδομένα. Ένας τρόπος αντιμετώπισης αυτού του φαινομένου ήταν να χρησιμοποιηθούν ετικέτες, που να δίνουν την δυνατότητα εποπτείας, εντός του περιεχομένου του κειμένου [69].

Οι απλούστερες προσεγγίσεις τους είχαν εξαρχής ως βάση, την παρουσία λέξεων ή emoticons, ως δείκτες θετικής ή αρνητικής πολικότητας [58] ή τον υπολογισμό της αναλογίας θετικών και αρνητικών όρων στα ηλεκτρονικά σώματα κειμένων [21]. Αν και αυτές αποτέλεσαν ένα χρήσιμο πρώτο στάδιο, οι γλωσσικές αποχρώσεις και η ποικιλία τους συχνά τις υπερέβαιναν οδηγώντας σε αστοχία.

Οι αδυναμίες αυτών των τεχνικών ήταν ένα μείζων κίνητρο που οδήγησε στην ανάπτυξη των ημι-επιβλεπόμενων αλγορίθμων, οι οποίοι βασίζονται σε γράφους και μαθαίνουν από περιορισμένες ποσότητες προσημασμένων δεδομένων και τεράστιες ποσότητες ελεύθερα διαθέσιμων δεδομένων χωρίς καμία σήμανση. Οι ημι–επιβλεπόμενες μέθοδοι που περιλαμβάνουν την χρήση λεξικών πολικότητας και ετικετών, μπόρεσαν επίσης να μειώσουν την εξάρτηση από τα προσημασμένα σώματα κειμένων. Τέτοιες μέθοδοι συμβολίζουν τα δεδομένα ως κορυφές ενός γράφου με ακμές μεταξύ των κορυφών του, που κωδικοποιούν τις ομοιότητες μεταξύ τους. Οι παραπάνω αλγόριθμοι συχνά διαδίδουν τις πληροφορίες των ετικετών, από τις λίγες προσημασμένες κορυφές, σε ολόκληρο το γράφο.

Παρόλα αυτά, η πλήρης ανάλυση συναισθήματος σε ένα συγκεκριμένο ερώτημα ή θέμα απαιτεί πολλά στάδια, όπως το φιλτράρισμα των δεδομένων με βάση ένα αρχικό ερώτημα, την απόρριψη των spam και άσχετων αντικείμενων από αυτά, τον προσδιορισμό των αντικειμενικών και υποκειμενικών δεδομένων, και την αναγνώριση της πολικότητας αυτών των τελευταίων. Οι εφαρμογές του LP συνήθως επικεντρώνεται στην τελευταία φάση, την ταξινόμηση της πολικότητας και είναι αρκετοί αυτοί οι οποία τα τελευταία χρόνια συνδύασαν τις παραπάνω μεθόδους με τον LP, στις τεχνικές απόδοσης ετικετών πολικότητας ξεκινώντας από ένα μικρό δείγμα χειροκίνητα προσημασμένων δεδομένων συνεισφέροντας στην μερική αποδέσμευση από την παρουσία εκτεταμένων προσημασμένων συνόλων.

Κάθε μορφή κειμένου που καλούμαστε να αναλύσουμε ως προς το συναίσθημά του, στα κοινωνικά δίκτυα, δεν δημιουργείται σε απομόνωση, αλλά συνδέεται με άλλα από τον ίδιο συγγραφέα, και κάθε συγγραφέας επηρεάζεται από μηνύματα ή κείμενα αυτών που ακολουθεί. Κοινό λεξιλόγιο και θέματα συζήτησης επίσης συνδέουν αυτά τα μηνύματα μεταξύ τους. Μέθοδοι βασιζόμενοι σε γράφους , όπως ο LP, παρέχουν ένα φυσικό μέσο για να απεικονιστούν και να αξιοποιηθούν τέτοιες σχέσεις, προκειμένου να βελτιωθεί η ταξινόμηση, συχνά με μικρότερες απαιτήσεις επίβλεψης, σε σχέση με τη συμβατική προσέγγιση της ταξινόμηση άλλων μεθόδων.



Σχήμα 6.1: Σύγκριση LP και SVM σε διαφορετικά μεγέθη προσημασμένων δεδομένων[16]

Εν τούτοις, οι ιδιαιτερότητες των δεδομένων στα μέσα κοινωνικής δικτύωσης οδήγησαν σε προσαρμογές του πρωτότυπου αλγορίθμου του Zhu και Ghahramani (2002) [85], οι οποίες τον καθιστούν σαφώς αποτελεσματικότερο σε μία τέτοια ανάλυση. Η ανάδειξη συγκεκριμένων χαρακτηριστικών (aspects) του κειμένου και η απεξάρτηση από τη γλώσσα, το γνωσιακό τομέα και τις λεξικολογικές πηγές αποτελούν τις κύριες στοχεύσεις των παραπάνω προσπαθειών, οι οποίες κρίνονται ως κομβικής σημασίας λόγω της ποικιλομορφίας των δεδομένων των κοινωνικών δικτύων, όπως φανερώνεται εκτενέστερα στη συνέχεια.

[16] http://www.slideshare.net/dav009/label-propagation-semisupervised-learning-with-applications-to-nlp

6.1 Αλγόριθμος Διάδοσης Ετικέτας Συνδυαζόμενων Σχέσεων

Τα κείμενα που εκφέρουν άποψη χαρακτηρίζονται από λέξεις ή φράσεις που μεταδίδουν θετικά ή αρνητικά συναισθήματα. Αυτοί οι όροι, και η προϋπάρχουσα γνώση της πολικότητάς τους, θα μπορούσαν να χρησιμοποιηθούν ως χαρακτηριστικά γνωρίσματα σε ένα πλαίσιο επιβλεπόμενης κατηγοριοποίησης, ώστε να προσδιοριστεί το συναίσθημα του κειμένου άποψης. Έτσι, λεξικά που δείχνουν την πολικότητα τέτοιων λέξεων αποτελούν απαραίτητους πόρους όχι μόνο για την αυτόματη ανάλυση συναισθημάτων, αλλά και για άλλους σκοπούς κατανόησης φυσικών γλωσσών, όπως την εξαγωγή συμπερασμάτων από κείμενο. Ενώ είναι εφικτή η δόμηση με το χέρι τέτοιων πόρων για μια γλώσσα, η προσπάθεια που απαιτείται είναι μεγάλη. Αυτό οδηγεί στην ανάγκη για αυτόματες γλωσσο–αγνωστικές μεθόδους για τη δόμηση λεξικών συναισθημάτων.

Η σημαντικότητα αυτού του προβλήματος παρακίνησε τους Rao και Ravichandran (2009) [67] να μελετήσουν το πρόβλημα της ανίχνευσης της δυαδικής πολικότητας των λέξεων για την επαγωγική δημιουργία λεξικών πολικότητας. Στην εργασία τους χειρίζονται την ανίχνευση πολικότητας ως πρόβλημα ημι-επιβλεπόμενης διάδοσης ετικέτας σε γράφο. Στον γράφο κάθε κόμβος αντιπροσωπεύει μια λέξη της οποίας η πολικότητα πρέπει να προσδιοριστεί, κάθε βεβαρυμμένη ακμή κωδικοποιεί μια σχέση μεταξύ δύο λέξεων και κάθε κόμβος μπορεί να έχει είτε θετική είτε αρνητική ετικέτα. Οι ακμές μεταξύ των κόμβων κωδικοποιούν κάποια έννοια ομοιότητας. Μερικοί από αυτούς τους κόμβους, σε μορφή μεταγωγής, επισημαίνονται χρησιμοποιώντας παραδείγματα-πυρήνες (seed examples) και οι ετικέτες για τους υπόλοιπους κόμβους παράγονται χρησιμοποιώντας αυτούς τους πυρήνες.

Ο Αλγόριθμος Διάδοσης Ετικέτας Συνδυαζόμενων Σχέσεων (Combined Relationships LP – CR LP) που κατασκευάζουν, βασίζεται στον κλασικό LP, συνδυάζοντας όμως τις σχέσεις συνωνυμίας ή υπερωνυμίας των ερμηνευτικών λεξικών. Σε αυτό το πλαίσιο, εξετάζουν δύο διαφορετικά σενάρια διαθέσιμων πόρων χρησιμοποιώντας το WordNet, και όπου αυτό δεν είναι διαθέσιμο, τον θησαυρό του OpenOffice, έναν ελεύθερα διαθέσιμο πολύγλωσσο πόρο, που ελάχιστα χρησιμοποιείται στην NLP βιβλιογραφία. Τα αποτελέσματά τους αναφέρονται σε τρεις διαφορετικές γλώσσες : Αγγλικά, Γαλλικά και Χίντι, αποδεικνύοντας ότι η Διάδοση Ετικέτας μπορεί να βελτιωθεί σημαντικά σε σχέση με άλλες ημι-επιβλεπόμενες μεθόδους μάθησης, όπως ο αλγόριθμος Ελάχιστης Τομής (Mincuts) και ο Τυχαιοκρατικός Αλγόριθμος Ελάχιστης Τομής (Randomized Mincuts), για τον ίδιο σκοπό.

6.1.1 Δομή Αλγορίθμου CR LP

Οι αλγόριθμοι διάδοσης ετικετών είναι ένα πλαίσιο μεταβατικής μάθησης (transductive learning) που χρησιμοποιεί μερικά παραδείγματα, ή πυρήνες λημμάτων (seeds), για να επισημάνει με ετικέτες ένα μεγάλο αριθμό μη προσημασμένων δεδομένων. Στην εκδοχή των Rao και Ravichandran (2009) [67], πέραν του πυρήνα λημμάτων, ο CR LP χρησιμοποιεί μια σχέση μεταξύ των παραδειγμάτων. Η σχέση μεταξύ των δειγμάτων, πρέπει να πληροί δύο απαιτήσεις, δηλαδή να είναι μεταβατική και να κωδικοποιεί κάποια έννοια της συγγένειας μεταξύ των παραδειγμάτων.

Τα περισσότερα δεδομένα μιας φυσικής γλώσσας έχουν κάποια δομή που μπορεί να αξιοποιηθεί, ακόμη και ελλείψει πλήρως προσημασμένων δεδομένων. Για παράδειγμα, τα έγγραφα είναι παρόμοια ως προς τους όρους που περιέχουν οι λέξεις τους, μπορεί να είναι συνώνυμες η μία της άλλης και ούτω καθεξής. Τέτοιες πληροφορίες μπορούν να κωδικοποιηθούν εύκολα σε ένα γράφο, όπου η παρουσία μιας ακμής μεταξύ δύο κόμβων θα δείχνει μια σχέση μεταξύ των δύο κόμβων και προαιρετικά, το βάρος της ακμής θα μπορούσε να κωδικοποιεί τη δύναμη της σχέσης. Αυτές οι πρόσθετες πληροφορίες βοηθούν τη μάθηση όταν υπάρχουν πολύ λίγα σχολιασμένα παραδείγματα.

Μερικά παραδείγματα τέτοιων σχέσεων είναι μεταξύ άλλων, η συνωνυμία, η υπερωνυμία και η ομοιότητα μέσα κάποιο μετρικό χώρο. Αυτή η σχέση μεταξύ των παραδειγμάτων μπορεί να κωδικοποιηθεί εύκολα ως γράφος. Έτσι, κάθε κόμβος του γράφου είναι ένα παράδειγμα και η ακμή αντιπροσωπεύει τη σχέση. Επίσης συνδεμένη με κάθε κόμβο είναι μια κατανομή πιθανότητας επί των ετικετών του κόμβου. Για τους κόμβους του αρχικού πυρήνα, αυτή η κατανομή είναι γνωστή και διατηρείται σταθερή. Ο στόχος είναι να εξαχθούν οι κατανομές για τους υπόλοιπους κόμβους.

Στον αλγόριθμο θεωρούμε ένα γράφος $G(V, E, W)$, με κορυφές V, ακμές E, και μια μήτρα του βάρους των παραπάνω ακμών $W = [w_{ij}]$, μεγέθους $n \times n$, όπου $n = |V|$. Ο LP ελαχιστοποιεί την τετραγωνική ενεργειακή συνάρτηση: $\varepsilon = \frac{1}{2}\sum_{(i,j)\in E} w_{ij}(y_i - y_j)^2$ όπου y_i και y_j είναι οι ετικέτες των κόμβων i και j αντίστοιχα. Στην ειδική περίπτωση που πραγματοποιείται δυαδική κατηγοριοποίηση, όπως εδώ, ισχύει ότι $y_k \in \{-1, +1\}$. Έτσι, για την εξαγωγή των ετικετών στο y_i, θέτουν $\frac{\partial}{\partial y_i}\varepsilon = 0$, φτάνοντας στην ακόλουθη προσαρμοσμένη εξίσωση: $y_i = \frac{\sum_{(i,j)\in E} w_{ij} y_i}{\sum_{(i,j)\in E} w_{ij}}$.

Όπως και στον αυθεντικό αλγόριθμο των Zhu και Ghahramani (2002) [85], χρησιμοποιείται μία $n \times n$ στοχαστική μήτρα μετάβασης T, που παράγεται κανονικοποιώντας τον W ως προς τις γραμμές, ως εξής: $T_{ij} = P(j \to i) = \frac{w_{kj}}{\sum_{k=1}^{n} w_{kj}}$, όπου ως T_{ij} θεωρείται η πιθανότητα μετάβασης από τον κόμβο j στον κόμβο i.

Ο αλγόριθμος εξελίσσεται στα ακόλουθα βήματα, ως εξής:

1. Δώσε σε μια, $n \times C$, μήτρα Y, τις αρχικές ετικέτες, όπου C είναι ο αριθμός των κατηγοριών.
2. Πρόσθεσε μια ειδική ετικέτα " DEFAULT " στο υπάρχον σύνολο ετικετών.
3. Θέσε $P(\ DEFAULT\ |unlabeled\ node\ =\ u\ ) = 1$
4. Θέσε $P(\ L\ |\ seed\ node\ =s\ ) = 1$
5. Διάδωσε τις ετικέτες για όλους τους κόμβους υπολογίζοντας την $Y = TY$
6. Κανονικοποίησε ως προς τις σειρές, την Y , έτσι ώστε κάθε σειρά να αθροίζεται στη μονάδα.
7. Σύσφιξε τους πυρήνες λημμάτων (παραδείγματα), της μήτρας Y, στις αρχικές τους τιμές.
8. Επανάλαβε τη διαδικασία από το βήμα 5, μέχρι η μήτρα Y να συγκλίνει.

Ως $P(\ DEFAULT\ |unlabeled\ node\ =\ u\ ) = 1$ ορίζεται η πιθανότητα όλων των μη προσημασμένων κόμβων u, να έχουν ως μοναδική ετικέτα την " DEFAULT " και ως $P(\ L\ |\ seed\ node\ =s\ ) = 1$, η πιθανότητα όλων των κόμβων-πυρήνων s, να έχουν ετικέτα κατηγορίας L. Αυτό εξασφαλίζει ότι οι κόμβοι που δεν μπορούν να επισημανθούν καθόλου θα διατηρήσουν $P(\ DEFAULT\ ) = 1$ οδηγώντας έτσι σε μια γρήγορη σύγκλιση.

Ο αλγόριθμος παράγει μια κατανομή πιθανότητας πάνω στις ετικέτες για όλα τα μη προσημασμένα σημεία δεδομένων. Αυτό καθιστά τη μέθοδο αυτή ιδιαίτερα κατάλληλη για προσεγγίσεις που συνδυάζουν διαφορετικούς ταξινομητές συναισθήματος. Στην παρούσα εργασία, απλά επιλέγεται η πιο πιθανή ετικέτα ως η προβλεπόμενη ετικέτα για κάθε σημείο και ο αλγόριθμος καταλήγει πάντοτε σε σύγκλιση. Σημειώνεται ότι κατά την εφαρμογή του αλγορίθμου παραλείπεται το στάδιο του καθορισμού των παραμέτρων, καθιστώντας έτσι περιττή την ανάγκη για μια σειρά ξεχωριστών δοκιμών ανάπτυξης.

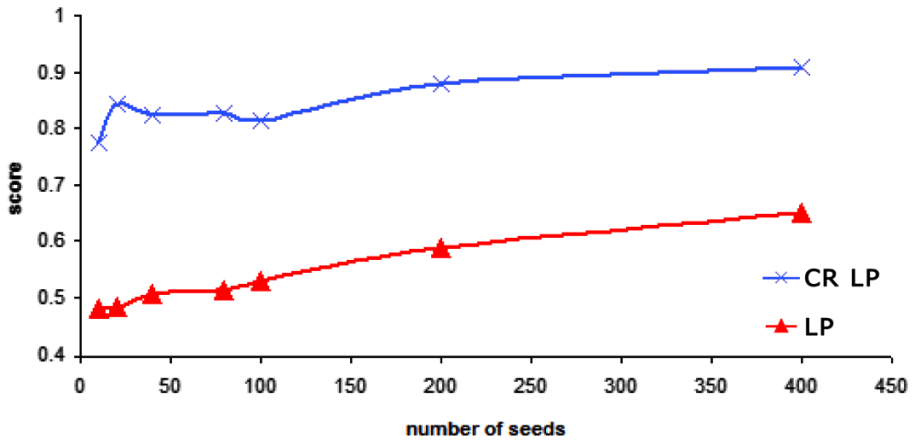
6.1.2 Χαρακτηριστικά Μεθόδου CR LP

Η μέθοδος των Rao και Ravichandran (2009) [67] λειτουργεί ως βελτίωση των μεθόδων των Kim και Hovy (2006) [47] και Kamps et al (2004) [44]. Στην μεν πρώτη, εμπλουτίζεται ένα λεξικό συναισθημάτων από το WordNet, με τα συνώνυμα μιας θετικής λέξης να είναι θετικά, και τα αντώνυμα αρνητικά. Αυτή η προσέγγιση πάσχει από πολύ πτωχή ανάκληση στα επίθετα. Στη δεύτερη, για τη μέτρηση του σημειολογικού προσανατολισμού με το WordNet χρησιμοποιούνται πρότυπα παραδείγματα λέξεων, όπως το "καλό" και "κακό" (Μέθοδος Πρωτοτύπων).

Κατά την προσθήκη του LP στις δύο μεθόδους, ο γράφος συνωνύμων εξάγεται από το WordNet με μία ακμή μεταξύ δύο κόμβων να ορίζεται μόνο όταν ο ένας είναι συνώνυμο του άλλου. Ο αλγόριθμος έτσι αποδίδει αισθητά καλύτερα σε ακρίβεια

συγκριτικά με τις προηγούμενες μεθόδους σε γράφους. Παρόλα αυτά, όταν ο LP εκτελείται, η επαναληπτική μέθοδος των Kim-Hovy συνεχίζει να πάσχει από πτωχή ανάκληση. Το αίτιο της χαμηλότερης ανάκλησης οφείλεται στην έλλειψη συνδεσιμότητας μεταξύ των πιθανά σχετικών κόμβων, κάτι που δεν διευκολύνει την «εξάπλωση» των ετικετών από τους προσημασμένους κόμβους-πυρήνες προς τους μη προσημασμένους.

Η συνολική βελτίωση στους Mincuts και Randomized Mincuts, ήταν αρκετά σημαντική, αλλά όχι ανάλογη της εφαρμογής του LP, ακόμα και όταν τα προσημασμένα δεδομένα ήταν εξαιρετικά λίγα, γεγονός που αποτέλεσε και το αρχικό κίνητρο για τη χρήση της ημι-επιβλεπόμενης μάθησης. Για την αντιμετώπιση αυτού του προβλήματος επιλέχθηκε η προσθήκη επιπλέον ακμών, που εκφράζουν τα υπερώνυμα, στο γράφο συνωνύμων. Ο κύριος λόγος για τη χαμηλή ανάκληση στη διάδοση ετικέτας είναι ότι ο γράφος συνωνύμων του WordNet είναι εξαιρετικά ασύνδετος. Ακόμα και σε κόμβους που λογικά είναι σχετικοί λείπουν οι μεταξύ τους διαδρομές. Έτσι, ενσωματώνοντας τα υπερώνυμα κάθε κόμβου βελτιώνεται η συνδεσιμότητα. Η εκτέλεση του LP σε αυτό το συνδυασμένο γράφο δίνει πολύ καλύτερα αποτελέσματα με πολύ υψηλότερη ανάκληση, ακόμα και με ελαφρώς καλύτερη ακρίβεια.



Σχήμα 6.2: Σύγκριση απόδοσης LP και CR LP σε ουσιαστικά

Ένα φυσικό ερώτημα που προκύπτει είναι αν μπορούν να χρησιμοποιηθούν και άλλες σχέσεις του WordNet, πέραν των συνωνύμων και υπερωνύμων, στον LP. Μια σχέση που θα ήταν ενδιαφέρουσα και χρήσιμη είναι η σχέση των αντώνυμων. Οι ακμές αντώνυμων όμως δεν μπορούν να προστεθούν με έναν άμεσο τρόπο στον γράφο για τη διάδοση ετικέτας, καθώς η σχέση αντωνύμων κωδικοποιεί αρνητική ομοιότητα (ανομοιότητα) και η σχέση ανομοιότητας διάδοσης δεν είναι μεταβατική, πράγμα που αντιβαίνει στα αρχικά κριτήρια του CR LP.

6.2 Αλγόριθμος Διάδοσης Ετικέτας βασιζόμενος σε Χαρακτηριστικά

Τα συστήματα Ανάλυσης Συναισθήματος Βασισμένης σε Χαρακτηριστικά (Aspect Based Sentiment Analysis - ABSA) δέχονται ως είσοδο ένα σύνολο κειμένων, όπως οι κριτικές προϊόντων ή μηνύματα κοινωνικών δικτύων, και εντοπίζουν τα κυριότερα και πιο συχνά σχολιαζόμενα χαρακτηριστικά (aspects) τους, καθώς και να εκτιμούν το μέσο συναίσθημα (θετικό ή αρνητικό) που εκφράζουν τα κείμενα, για κάθε χαρακτηριστικό της οντότητας [62]. Τα πειράματα των Brody και Elhadad (2010) [16], επιβεβαιώνουν την αξία μιας πλήρως μη-επιβλεπόμενης προσέγγισης, που στοχεύει στην ανίχνευση των χαρακτηριστικών και την ανάλυσης συναισθήματος. Η εργασία τους εστιάζει στον εντοπισμό των δύο πρωταρχικών στοιχείων σε κάθε κείμενο, δηλαδή τα χαρακτηριστικά και το συναίσθημα, τα οποία μέχρι τότε είχαν αντιμετωπιστεί επί το πλείστον ως δύο ξεχωριστοί στόχοι.

Οι μη επιβλεπόμενες μέθοδοι είναι επιθυμητές για το παραπάνω εγχείρημα, κυρίως για δύο λόγους. Πρώτον, λόγω της ευρύτητας και ποικιλίας των κοινωνικών δικτύων, το πλαίσιο πρέπει να είναι εύρωστο και εύκολα μεταβιβάσιμο από τον ένα τομέα γνώσης στον άλλο. Ο δεύτερος λόγος είναι η φύση των δεδομένων. Τα μηνύματα στα κοινωνικά δίκτυα είναι συχνά σύντομα και χωρίς δομή, και μπορεί να περιέχουν πολλά ορθογραφικά και γραμματικά σφάλματα, καθώς και αργκό ή εξειδικευμένη ορολογία. Οι παράγοντες αυτοί δημιουργούν συχνά ένα πρόβλημα στις μεθόδους που στηρίζονται αποκλειστικά σε λεξικά και σε χειροποίητους πόρους γνώσης, καθώς αυτά μπορεί να παραλείπουν κάποιο σημαντικό χαρακτηριστικό ή μια ένδειξη συναισθήματος. Οι μη επιβλεπόμενες μέθοδοι, από την άλλη πλευρά, δεν επηρεάζονται από τη λεξικολογική μορφή, και μπορούν να χειριστούν άγνωστες λέξεις ή λεξο-μορφές, αρκεί αυτές να εμφανίζονται αρκετά συχνά. Αυτό εξασφαλίζει ότι κάθε θέμα, που έχει εμφανή παρουσία στα δεδομένα, θα τύχει επεξεργασίας από το σύστημα.

Για τον προσδιορισμό της πολικότητας συνθέτουν έναν Αλγόριθμο Διάδοσης Ετικέτας βασιζόμενο στα Χαρακτηριστικά (Aspect Based LP – AB LP). Τα χαρακτηριστικά προσδιορίζονται μέσω μιας τοπικής έκδοσης της Αφανούς Κατανομής Dirichlet. Στην επεξεργασία φυσικής γλώσσας, η Αφανής Κατανομή Dirichlet (Latent Dirichlet Allocation - LDA) είναι ένα παραγωγικό στατιστικό μοντέλο, που επιτρέπει σε σύνολα παρατηρήσεων, να ερμηνεύονται από απαρατήρητες ομάδες, εξηγώντας γιατί ορισμένα μέρη των δεδομένων είναι παρόμοια [11]. Για παράδειγμα, εάν οι παρατηρήσεις είναι λέξεις που συλλέγονται από έγγραφα, αυτό προϋποθέτει ότι κάθε έγγραφο είναι ένα μίγμα από ένα μικρό αριθμό θεμάτων, και ότι η δημιουργία κάθε λέξης οφείλεται σε ένα από τα θέματα του εγγράφου.

6.2.1 Δομή Αλγορίθμου AB LP

Όπως και στον αυθεντικό αλγόριθμο των Zhu και Ghahramani (2002) [85] χρησιμοποιείται μία στοχαστική μήτρα μετάβασης T, οι τιμές της οποίας σε αυτή την περίπτωση καθορίζονται από την αναδρομική επαναληπτική σχέση:

$$T_{x,y} = P^t(x \to y) = \frac{\sum_{y \in N(x)} w(y,x) \cdot p^{t-1}(y)}{\sum_{y \in N(x)} w(y,x)}$$

όπου ως $T_{x,y}$ θεωρείται η πιθανότητα μετάβασης από τον κόμβο x στον κόμβο y, ως t η επανάληψη αναθεώρησης, $N(x)$ το σύνολο των γειτονικών κόμβων του x,ως $p^t(x)$ η πολικότητα του επιθέτου x στο βήμα t και ως $w(y,x)$ το βάρος της ακμής μεταξύ των κόμβων x και y. Το βάρος κάθε ακμής καθορίζεται από τη σχέση $w(y,x) = 1 + log\big(\#mod(y,x)\big)$, όπου ως $\#mod(y,x)$ ορίζεται ο αριθμός των περιπτώσεων, που τα y και x τροποποιούν το ίδιο ουσιαστικό.

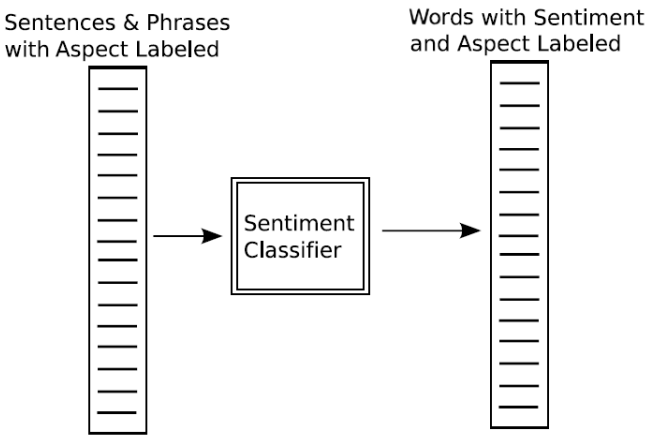
Ο αλγόριθμος εξελίσσεται στα ακόλουθα βήματα, ως εξής:

1. Θεώρησε μοναδιαίες τις πολικότητες των θετικών πυρήνων και μηδενικές των αρνητικών πυρήνων.
2. Θεώρησε ίσες με 0.5 τις πολικότητες όλων των μη προσημασμένων πυρήνων.
3. Όλοι οι κόμβοι διαδίδουν τις ετικέτες τους για ένα βήμα $Y \leftarrow TY$, προς τους μη προσημασμένους κόμβους.
4. Κανονικοποίησε ως προς τις σειρές, την Y , έτσι ώστε κάθε σειρά να αθροίζεται στη μονάδα.
5. Σύσφιξε τους πυρήνες λημμάτων, της μήτρας Y, στις αρχικές τους τιμές.
6. Επανάλαβε τη διαδικασία από το βήμα 3, μέχρι η μήτρα Y να συγκλίνει.

6.2.2 Χαρακτηριστικά Μεθόδου AB LP

Ως ένα βήμα προ-επεξεργασίας αναλύονται τα δεδομένα για την ανίχνευση της άρνησης και των συνδέσμων. Αν ένα επίθετο A συμμετέχει σε μια άρνηση στην πρόταση, αντικαθίσταται από ένα νέο επίθετο μη-A. Στη συνέχεια εξάγονται όλες οι περιπτώσεις, όπου ένα επίθετο τροποποιεί ένα ουσιαστικό. Αυτή η μέθοδός για τον προσδιορισμό πολικότητας συναισθήματος βασίζεται σε μια προσαρμογή των Hatzivassiloglou και McKeown (1997) [39], όπου αγνοούνται τα επίθετα χωρίς προσανατολισμό. Επιπλέον, επίθετα των οποίων ο προσανατολισμός εξαρτάται από τα συμφραζόμενα αγνοήθηκαν επίσης. Τα επίθετα αυτά έχουν ιδιαίτερο ενδιαφέρον και είναι πιθανό να λείπουν ή να έχουν λανθασμένη επισήμανση στα συνήθη λεξικά συναισθήματος. Δεδομένου ότι πρέπει να χειριστούν επίθετα που εκφράζουν διάφορες αποχρώσεις συναισθήματος (και όχι μόνο έντονα θετικές ή αρνητικές) πραγματοποιείται μια μέθοδος βαθμολόγησης και όχι μια δυαδική επισήμανση.

Για την οικοδόμηση του γράφου πολικότητας χρησιμοποιούνται διαζευκτικές λέξεις (π.χ. «αλλά») ως δείκτες της αντίθετης πόλωσης. Αντί να χρησιμοποιηθούν συνήθεις εκφράσεις για τη συλλογή ρητών γραμματικών συνδέσεων, ανακτώνται όλες οι περιπτώσεις, κατά τις οποίες δύο επίθετα τροποποιούν το ίδιο ουσιαστικό, στην ίδια πρόταση. Για να εξασφαλιστεί ότι τα εξειδικευμένα σε χαρακτηριστικά (aspect-specific) επίθετα τυγχάνουν σωστού χειρισμού, δομείται ένας ξεχωριστός γράφος για κάθε χαρακτηριστικό, επιλέγοντας τις περιπτώσεις όπου το τροποποιημένο ουσιαστικό είναι μία από τις αντιπροσωπευτικές λέξεις για αυτό το χαρακτηριστικό.



Σχήμα 6.3: Παράδειγμα Ταξινομητή Πολικότητας

Για τη δόμηση του συνόλου πυρήνων κάθε τέτοιου γράφου χρησιμοποιούνται μορφολογικές πληροφορίες και εμφανείς αρνήσεις, για τον εντοπισμό ζευγών με αντίθετη πολικότητα. Συγκεκριμένα, από το σύνολο πυρήνων επιλέγονται ζεύγη επιθέτων που διακρίνονταν μόνο από τα αρνητικά προθέματα «un», «in», «dis», «non» ή από το δείκτη άρνησης «not». Αρχίζοντας από το πλέον συχνό ζεύγος δίνεται μια θετική πολικότητα στο συχνότερο μέλος του ζεύγους. Στη συνέχεια, κατά φθίνουσα σειρά συχνότητας, αποδίδεται πολικότητα στα άλλα ζεύγη πυρήνων, με βάση τη συντομότερη διαδρομή που είχε ένα από τα μέλη μέχρι ένα προηγουμένως επισημασμένο επίθετο. Αυτό το μέλος έλαβε την πολικότητα του γείτονά του, και το άλλο μέλος του ζεύγους έλαβε την αντίθετη πολικότητα. Όταν όλα τα ζεύγη έλαβαν ετικέτες, διορθώνονται οι εσφαλμένες ταξινομήσεις, επαναλαμβάνοντας μέσω των ζευγών και αντιστρέφοντας την πολικότητα, εάν αυτό βελτιώνει τη συνοχή, δηλαδή αν κάνει τα μέλη του ζεύγους να ταιριάζουν σε πολικότητα με αυτή των περισσοτέρων από τους γείτονές τους. Τέλος, αντιστρέφεται η πολικότητα των ομάδων-πυρήνων, αν η αρνητική ομάδα έχει υψηλότερη συνολική συχνότητα.

Για την αξιολόγηση της απόδοσης του τμήματος συναισθήματος του συστήματος, για κάθε ένα από τα πρώτα οκτώ αυτομάτως τεκμαρθέντα χαρακτηριστικά,

κατασκευάστηκε ένας γράφος πολικότητας, ανακτώντας έναν κατάλογο όλων των επιθέτων, που μετείχαν σε πέντε ή περισσότερες τροποποιήσεις ουσιαστικών από εκείνο το συγκεκριμένο χαρακτηριστικό. Τα δεδομένα χωρίστηκαν σε δέκα τμήματα και για κάθε τμήμα βαθμολογήθηκαν από δύο εθελοντές ανά επίθετο, ανάλογα με την πολικότητα συναισθήματος, που εκφράζει στα συμφραζόμενα του συγκεκριμένου χαρακτηριστικού. Η κλίμακα είχε τις ακόλουθες διαβαθμίσεις: Έντονα αρνητικό, ασθενώς αρνητικό, ουδέτερο, ασθενώς θετικό, έντονα θετικό, και μη εφαρμόσιμο. Μπορεί μεν η ακριβής συμφωνία μεταξύ των σχολιαστών να ήταν χαμηλή, αλλά θεωρώντας τις παρακείμενες εκτιμήσεις ισοδύναμες, η συμφωνία ήταν μεγαλύτερη του 90%.

Οι σχολιασμένες βαθμολογίες μεταφράστηκαν σε μια αριθμητική κλίμακα, από −2 (έντονα Αρνητικό) έως +2 (έντονα Θετικό), με διαστήματα ακεραίων μονάδων. Αφού απορρίφθηκαν τα επίθετα στα οποία ένας ή περισσότεροι σχολιαστές έδωσαν την ετικέτα «μη εφαρμόσιμο», υπολογίστηκε ο μέσος όρος της βαθμολογίας των δύο σχολιαστών και αυτά τα δεδομένα χρησιμοποιήθηκαν σαν κεντρικός κανόνας της αξιολόγησής. Ο AB LP επιτυγχάνει καλά αποτελέσματα, ενώ το πιο σημαντικό είναι ότι ο συσχετισμός με αυτόν τον κανόνα δεν δείχνει μεγαλύτερη καταλληλότητα, μόνο στην ανίχνευση συναισθήματος σε κριτικές. Τα χαρακτηριστικά που έχουν το υψηλότερο ποσοστό επιθέτων είναι αυτά που βαθμολογήθηκαν ως ουδέτερα από τους σχολιαστές. Εντούτοις, σε πολλές περιπτώσεις, αυτά τα επίθετα στην πραγματικότητα φέρουν κάποιο συναίσθημα στα συμφραζόμενα τους.

Εφαρμόζοντας τον AB LP σε δεδομένα προερχόμενα αποκλειστικά από το Twitter, οι Brody και Diakopoulos (2012) [15] διαπίστωσαν ότι λόγω της φύσης του τομέα, ο οποίος είναι πολύ άτυπος και χωρίς δομή, η ακριβής ανάλυση είναι δύσκολη και απαιτείται ιδιαίτερη προσαρμογή του αλγορίθμου. Πιο συγκεκριμένα, εξετάζονται όλες οι υποψήφιες λέξεις ως κόμβοι, μαζί με τις λέξεις στα θετικά και αρνητικά σύνολα σπόρων. Σαν διαμεσολαβητής (proxy) της συντακτικής σχέσης, οι ακμές σταθμίζονται ως συνάρτηση του αριθμού των φορών, που δύο λέξεις εμφανίζονται μέσα σε ένα παράθυρο τριών λέξεων, από κάθε άλλη λέξη, στο σύνολο δεδομένων και αφαιρούνται οι κόμβοι των οποίων οι γειτονικές ακμές έχουν ένα συνδυασμένο βάρος μικρότερο του 20, που σημαίνει ότι συμμετέχουν σε σχετικά λίγες σχέσεις συνεμφάνισης, με άλλες λέξεις στο γράφο. Δίνονται, τέλος, τιμές 1 (έντονα αρνητική) έως 5 (έντονα θετική) στις αξιολογήσεις, και υπολόγισαν το μέσο όρο, μεταξύ δύο αξιολογήσεων, για κάθε μία λέξη. Λέξεις με μέση βαθμολογία 3 θεωρήθηκαν ουδέτερες, και εκείνες με χαμηλότερες και υψηλότερες θεωρήθηκαν αρνητικές και θετικές, αντίστοιχα. Οι μεταβολές αυτές των παραμέτρων του AB LP δίνουν καλύτερα αποτελέσματα σε σύγκριση με την αρχική του μορφή, καθιστώντας τον αποδοτικότερο κατά την εφαρμογή του στα κοινωνικά δίκτυα.

6.3 Αλγόριθμος Διάδοσης Ετικέτας προσανατολισμένος σε Χαρακτηριστικά

Η μελέτη των Blair-Goldensohn et al. (2008) [10] εξετάζει το πρόβλημα της σύνοψης συναισθήματος βασιζομένου σε χαρακτηριστικά. Ένα σύστημα σύνοψης βασιζόμενου σε χαρακτηριστικά λαμβάνει ως εισαγωγή ένα σύνολο κριτικών χρήστη για ένα συγκεκριμένο προϊόν ή υπηρεσία και παράγει ένα σύνολο σχετικών γνωρισμάτων, μια συνολική βαθμολογία για κάθε χαρακτηριστικό, και υποστηρικτικές ενδείξεις του κειμένου. Η μέθοδος τους έχει ως στόχο να δημιουργήσει ένα γενικό σύστημα, υιοθετώντας μια τυποποιημένη αρχιτεκτονική κατάλληλη για σύνοψη βασισμένη σε χαρακτηριστικά και αποτελείται από τρία βήματα:

1. Προσδιορισμός όλων των τμημάτων κειμένων των κριτικών που είναι φορτωμένα με συναίσθημα.

2. Προσδιορισμός των σχετικών χαρακτηριστικών που αναφέρονται σ' αυτά τα τμήματα.

3. Συγκέντρωση του συναισθήματος σε κάθε χαρακτηριστικό, σύμφωνα με το εκάστοτε συναίσθημα των αναφορών.

Στο σύστημά του, χρησιμοποιήθηκε ένα υβρίδιο επειδή σε τέτοια δεδομένα απαιτείται ένας γενικός ταξινομητής συναισθήματος λεξικού, ανεξάρτητος μεν από το γνωσιακό τομέα, αλλά με την ισχύ ενός ταξινομητή μηχανικής εκμάθησης, που να μπορεί να βελτιστοποιήσει τις παραμέτρους του συστήματος, σε ένα μεγάλο σύνολο δεδομένων.

Η ταξινόμηση συναισθήματος στο επίπεδο πρότασης δεν είναι ένας σχεδιασμένος στόχος, δεδομένου ότι οι χρήστες συνήθως έχουν δώσει μια αριθμητική βαθμολόγηση συναισθήματος για ολόκληρη την κριτική. Ακόμη και ισχυρά θετικές κριτικές μπορεί να περιέχουν αρνητικές γνώμες και αντίστροφα. Έτσι, για την αυτόματη ταξινόμηση των προτάσεων κατασκευάζουν τον αλγόριθμο Διάδοσης Ετικέτας, που είναι προσανατολισμένος στην μετέπειτα εξαγωγή χαρακτηριστικών (Aspect Oriented LP - AO LP), αλλά και πάλι τα μοντέλα θα πρέπει να λάβουν υπόψη οποιεσδήποτε τις αριθμητικές εκτιμήσεις χρήστη, όπου υπάρχουν τέτοιες.

6.3.1 Δομή Αλγορίθμου AO LP

Το πρώτο βήμα στο παραπάνω υβριδικό μοντέλο είναι η κατασκευή ενός γενικού λεξικού συναισθήματος. Η παρούσα μέθοδος δεν δημιουργεί μόνο αυτά τα σύνολα,

αλλά και τη βαρύτητα σε κάθε μέλος του συνόλου, με ένα μέτρο εμπιστοσύνης που αντιπροσωπεύει το πόσο πιθανό είναι μια δεδομένη λέξη να έχει το οριζόμενο θετικό ή αρνητικό συναίσθημα. Κατά συνέπεια, χρησιμοποιείται μια τροποποιημένη έκδοση του LP πάνω σε γράφους, προσαρμοσμένη στη δημιουργία λεξικών συναισθήματος, στην οποία επισυνάπτονται απλουστευμένες ετικέτες του μέρους του λόγου κάθε λέξης (επίθετο, επίρρημα, ουσιαστικό ή ρήμα) στο σύνολο-πυρήνα, προκειμένου να βοηθηθεί η διάκριση, μεταξύ προτάσεων πολλαπλών λέξεων.

Οι είσοδοι στον αλγόριθμο είναι τα τρία χειρονακτικώς δομημένα σύνολα-πυρήνες που συμβολίζονται με P (θετικό), N (αρνητικό), και M (ουδέτερο). Επίσης, ως είσοδος δίδονται τα σύνολα συνωνύμων και αντωνύμων, που εξάγονται από το WordNet, για την αυθαίρετη λέξη w, που συμβολίζεται με $syn(w)$ και $ant(w)$ αντίστοιχα. Ο αλγόριθμος αρχίζει με τον καθορισμό ενός διανύσματος βαθμολογίας s^m, που κωδικοποιεί τις βαθμολογίες των λέξεων συναισθήματος, για κάθε λέξη του WordNet. Αυτό το διάνυσμα ενημερώνεται επαναληπτικά, με κάθε επανάληψη να υποδεικνύεται από τον εκθέτη m. Η τιμή του s^0 αρχικοποιείται ως εξής:

$$s_i{}^0 = \begin{cases} +1 & \text{εάν } w_i \in P \\ -1 & \text{εάν } w_i \in N \\ 0 \ \forall w_i \in \text{WordNet} - P \cup N \end{cases}$$

Το s^0 αρχικοποιείται έτσι ώστε όλες οι θετικές λέξεις πυρήνες να πάρουν την τιμή $+1$, όλες οι αρνητικές λέξεις πυρήνες παίρνουν τιμή -1, και όλες οι άλλες λέξεις τιμή 0. Όλα τα παραπάνω διανύσματα των λέξεων w_i, συνθέτουν την μήτρα S. Έπειτα, επιλέγεται ένας παράγοντας κλιμάκωσης $\lambda < 1$, για τον ορισμό μίας μήτρας γειτνίασης, για το σύνολο όλων των λέξεων w_i στο λεξικό WordNet $A = (a_{ij})$:

$$\boldsymbol{A} = a_{ij} = \begin{cases} 1+\lambda & \text{εάν } i == \text{j} \\ +\lambda \text{ εάν } w_i \in syn(w) \text{ και } w_i \notin M \\ -\lambda \text{ εάν } w_i \in ant(w) \text{και } w_i \notin M \\ 0 \qquad \sigma\varepsilon \text{ κάθε άλλη } \pi\varepsilon\rho\acute{\iota}\pi\tau\omega\sigma\eta \end{cases}$$

Ο $\boldsymbol{A}$ είναι απλά μια μήτρα που αντιπροσωπεύει έναν κατευθυνόμενο, βεβαρυμμένο σημασιολογικό γράφο, όπου οι γειτονικοί κόμβοι είναι συνώνυμα ή αντώνυμα και δεν είναι μέλη του προκαθορισμένου ουδέτερου συνόλου, ώστε να σταματά η διάδοση συναισθήματος, μέσω των ουδέτερων λέξεων.

Η παράμετρος αποσύνθεσης λ χρησιμοποιείται για να περιοριστεί το μέγεθος των βαθμολογιών που είναι μακριά από τους πυρήνες στο γράφο. Μεγαλύτερα λάμδα οδηγούν σε πολύ ασύμμετρη κατανομή της βαθμολογίας, δηλαδή τα υψηλά σκορ

λέξης υπερτερούν κατά πολύ όλων των άλλων σκορ. Αντίθετα, ένα πάρα πολύ μικρό λάμδα δίνει στις λέξεις πυρήνες υπερβολική σημαντικότητα.

Η βαθμολογία συναισθήματος διαδίδεται πάνω στον γράφο, μέσω επαναλαμβανόμενων πολλαπλασιασμών του $\boldsymbol{A}$, επί τα διανύσματα βαθμολογίας s^m, που αυξάνεται με μια συνάρτηση διόρθωσης πρόσημου των λέξεων-πυρήνων, ώστε να αντισταθμίσει τις σχέσεις που έχουν λιγότερο νόημα, στα συμφραζόμενα των κριτικών. Αν για παράδειγμα, μια λέξη με συνήθως καλή έννοια στις κριτικές φανεί σε κάποια κριτική ως αρνητική, αντί να επισημανθεί τεχνητά ως ουδέτερη, επιλέγεται ως θετική λέξη πυρήνα, και διατηρείται το πρόσημό της σε κάθε μια από τις m επαναλήψεις:

$$for\ m := 1\ to\ M$$

$$s^m := \text{sign} - \text{correct}(\mathbf{A}\ s^m\ - 1)$$

Εδώ, η συνάρτηση $t = \text{sign} - \text{correct(s)}$ διατηρεί την ισότητα $|t_i| = |s_i| \forall i$ γεγονός που εξασφαλίζει ότι η συνάρτηση πρόσημου $\text{sign}(t_i) = {s_i}^0$, για όλες τις w_i λέξεις πυρήνες, και παράλληλα συντηρείται το πρόσημο όλων των άλλων λέξεων. Το τελικό διάνυσμα αποτελέσματος s παράγεται με λογαριθμική κλιμάκωση του s^m στην σχέση:

$$s_i := \begin{cases} log(|{s_i}^M|) * sign({s_i}^M) & εάν\ |{s_i}^M| > 1 \\ 0 & σε\ κάθε\ άλλη\ περίπτωση \end{cases}$$

Η παραπάνω κλιμάκωση των βαθμολογιών, περιορίζει την επίδραση που ασκούν οι όροι με υψηλή βαθμολογία, πάνω στις τελικές αποφάσεις ταξινόμησης, δεδομένου ότι αυτά τα σκορ μπορεί συχνά να είναι αρκετά υψηλά.

Ο αλγόριθμος εξελίσσεται στα ακόλουθα βήματα ως εξής:

1. Θεώρησε ως 1 την τιμή της πολικότητας των θετικών πυρήνων του συνόλου P και ως -1 των αρνητικών πυρήνων του συνόλου N.
2. Θεώρησε μηδενικές τις πολικότητες όλων των υπολοίπων όρων του Wordnet.
3. Όλοι οι κόμβοι διαδίδουν τις ετικέτες τους για ένα βήμα $S \leftarrow \boldsymbol{A}S$, προς τους μη προσημασμένους κόμβους.
4. Εφάρμοσε τη συνάρτηση διόρθωσης πρόσημου $t = \text{sign} - \text{correct(s)}$.
5. Κανονικοποίησε ως προς τις σειρές, την S, έτσι ώστε κάθε σειρά να αθροίζεται στη μονάδα.
6. Κλιμάκωσε λογαριθμικά τους πυρήνες λημμάτων, της μήτρας S.
7. Επανάλαβε τη διαδικασία από το βήμα 3, μέχρι η μήτρα S να συγκλίνει.

Σε κάθε επανάληψη του αλγορίθμου στο γράφο οι λέξεις οι οποίες είναι θετικά συνδεδεμένες σε έναν μεγάλο αριθμό γειτόνων, με παρόμοιο συναίσθημα, θα πάρουν μια αυξημένη βαθμολογία. Κατά συνέπεια, μια λέξη που δεν είναι λέξη πυρήνας,

αλλά είναι γείτονας τουλάχιστον μιας λέξης πυρήνα, θα λάβει μια βαθμολογία συναισθήματος παρόμοια με αυτή των συνδεδεμένων λέξεων πυρήνων. Αυτό διαδίδεται έπειτα σε άλλες λέξεις, και ούτω καθεξής. Ταυτόχρονα αξιοποιείται η αποσαφήνιση, που προσφέρεται με τις ετικέτες, που δηλώνουν το μέρους του λόγου, στο οποίο ανήκουν οι λέξεις του WordNet, και διατρέχει την ιεραρχία του.

Η χρήση του παραχθέντος λεξικού μέσω του AO LP, συμβάλει στην ταξινόμηση του συναισθήματος των προτάσεων ή άλλων τεμαχίων κειμένου. Λαμβάνοντας μια τεμαχισμένη σειρά λέξεων $x = (w_1, w_2, \dots, w_n)$, το συναίσθημά της πρότασης ταξινομείται, χρησιμοποιώντας την συνάρτηση: $raw - score\ (x) = \sum_{i=1}^{n} s_i$. Η βαθμολογία s_i για οποιονδήποτε όρο δίνεται από το παραχθέν λεξικό, αντιστρέφοντας το πρόσημο του s_i, σε περιπτώσεις όπου προηγείται σαν όρος άρνησης, όπως το «όχι» ή το «μη». Όταν το $|raw - score\ (x)|$ είναι κάτω από ένα όριο, το x ταξινομείται ως ουδέτερο, διαφορετικά ως θετικό ή αρνητικό, ανάλογα με το πρόσημό του. Επιπλέον, ταξινομούνται προτάσεις με βάση το μέγεθος. Ένα σύνηθες πρόσθετο μέτρο ενδιαφέροντος, που χρησιμοποιείται στους LPαλγόριθμους είναι η καθαρότητα ενός τεμαχίου:

$$\text{purity}(x) = \frac{raw - score\ (x)}{\sum_{i=1}^{n} |\text{s}_\text{i}|}$$

Αυτό το σκορ είναι πάντα στο διάστημα $[-1, 1]$ και έχει συσχετισμό με το σταθμισμένο κλάσμα λέξεων στο x, που ταιριάζουν με το γενικό συναίσθημα της ακατέργαστης βαθμολογίας και δίνει ένα πρόσθετο μέτρο τάσης του x. Για παράδειγμα, εάν δύο τεμάχια, x_i και x_j, με σκορ 2, όπου όμως το x_i , το έλαβε μέσω δύο λέξεων με σκορ 1, ενώ το x_j, το έλαβε μέσω 2 λέξεων με σκορ 3 και -1, τότε το x_i θεωρείται καθαρότερο, υπό τη θετική έννοια, λόγω της έλλειψης οποιασδήποτε αρνητικής ένδειξης.

Αν και ο ταξινομητής που βασίζονται στο λεξικό του AO LP, μπορεί να είναι ισχυρός προγνώστης, δεν εκμεταλλεύεται κάποιο τοπικό ή σφαιρικό συμφραζόμενο. Επιπλέον, οι βαθμολογίες θέτονται έτσι ώστε να χρησιμοποιούν ad-hoc συναρτήσεις αποσύνθεσης, αντί συναρτήσεις βελτιστοποίησης δεδομένων. Προκειμένου να ξεπεραστούν και τα δύο αυτά προβλήματα, μπορεί να χρησιμοποιηθεί ένας ταξινομητής μέγιστης εντροπίας για να προβλέψει αυτές τις βαθμολογήσεις, βασισμένος σε έναν μικρό αριθμό τοπικών και σφαιρικών χαρακτηριστικών των συμφραζομένων, για μια πρόταση x_i που εμφανίζεται στην κριτική $r = (x_1, x_2, \dots, x_m)$, δηλαδή:

1. $raw - score\ (x_\text{i})$ και $\text{purity}(x_\text{i})$

2. $raw - score\ (x_{\text{i}-1})$ και $\text{purity}(x_{\text{i}-1})$

3. $raw - score\ (x_{\text{i}+1})$ και $\text{purity}(x_{\text{i}+1})$

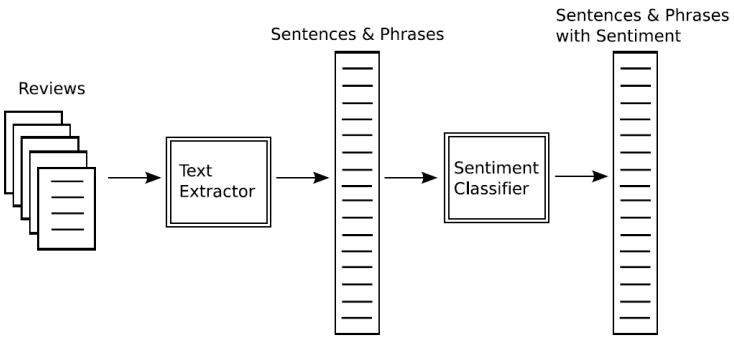
4. $raw - score\ (r)$ και purity(r)

Ένα κοινό θέμα στο σύστημά είναι η χρησιμοποίηση όσο το δυνατόν περισσότερης a priori πληροφορίας. Συνεπώς, στα δεδομένα της κριτικής, μπορούν να αξιοποιηθούν βαθμολογίες με αστέρια, παρεχόμενες από χρήστες, οι οποίες ουσιαστικά περιγράφουν το γενικό συναίσθημα. Αυτό το συναίσθημα δεν προκαθορίζει το συναίσθημα μεμονωμένων προτάσεων, αλλά μόνο το συναίσθημα που μεταβιβάζεται συνολικά από την κριτική. Συμβαίνει συχνά μια κριτική να έχει ένα καλό ή ένα κακό συνολικό συναίσθημα, αλλά να έχει ταυτόχρονα μερικές προτάσεις, με αντίθετη πολικότητα. Αυτό είναι ιδιαίτερα συχνό στις κριτικές με το συναίσθημα στη μεσαία κλίμακα. Κατά συνέπεια, αυτή η πληροφορία πρέπει να χρησιμοποιηθεί μόνο ως πρόσθετο σήμα κατά την ταξινόμηση και όχι ως άκαμπτος κανόνας κατά τον καθορισμό του συναισθήματος των προτάσεων.

6.3.2 Χαρακτηριστικά Μεθόδου AO LP

Η είσοδος στο σύστημα είναι ένα σύνολο κριτικών. Ο εξαγωγέας κειμένων τεμαχίζει αυτά τα κείμενα κριτικών σε ένα σύνολο τεμαχίων κειμένων που μπορεί να είναι χρήσιμα σε μια σύνοψη. Αυτό μπορεί να περιλαμβάνει προτάσεις και μικρές ή μεγάλες φράσεις. Αυτά τα τεμάχια κειμένων θα χρησιμοποιηθούν για να συγκεντρωθούν βαθμολογίες για οποιοδήποτε χαρακτηριστικό αναφέρεται μέσα σε αυτά, αλλά και ως υποψήφια για την τελική σύνοψη, όπου θα περιληφθούν ενδείξεις για κάθε βαθμολογία χαρακτηριστικού. Το σύστημά χρησιμοποιεί τεμάχια κειμένου τόσο επιπέδου πρότασης όσο και φράσης κατά την παραγωγή μιας σύνοψης.

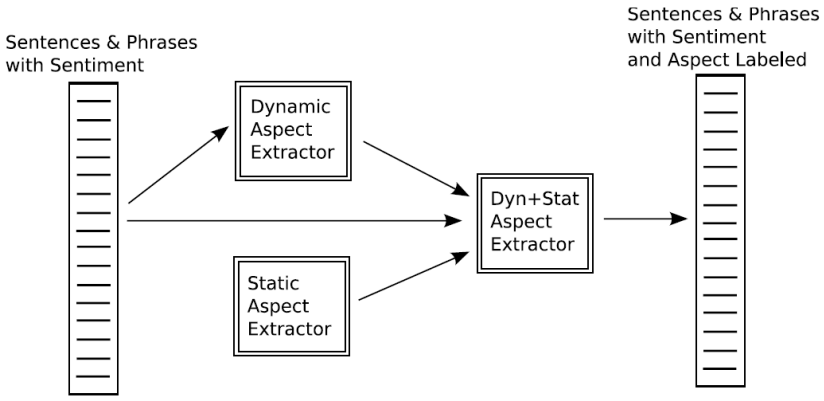
Το πρώτο στάδιο περιλαμβάνει την ταξινόμηση των εξαχθεισών προτάσεων, ως θετικές, αρνητικές ή ουδέτερες, ως προς την άποψη. Το μοντέλο που υιοθετείται για την ταξινόμηση συναισθήματος χρησιμοποιεί αλγορίθμους βασισμένους σε λεξικά. Η χρήση ημι-επιβλεπόμενης μηχανικής μάθησης, με τη μοντελοποίηση των συμφραζομένων μιας πρότασης, καθώς και των σφαιρικών πληροφοριών που παρέχονται από το χρήστη, όπως μια γενική βαθμολογία με αστέρια, βελτιώνουν την ταξινόμηση συναισθήματος στο επίπεδο πρότασης.



Σχήμα 6.4: Παράδειγμα διαδικασίας ταξινόμησης σε επίπεδο πρότασης

Το αρχικό σύνολο πυρήνων του AO LP, περιέχει διψήφιο πλήθος αρνητικών και θετικών λέξεων, που επιλέγονται με το χέρι, για να μεγιστοποιήσουν την κάλυψη γνωστικών περιοχών, καθώς επίσης και τριψήφιο πλήθος ουδέτερων λέξεων, που αποτελούνται κατά ένα μεγάλο μέρος, από stopwords, δηλαδή λέξεις μεγάλης συχνότητας, αλλά μειωμένης αξίας, στην ανάκτηση πληροφορίας σε ένα κείμενο. Αυτές οι ουδέτερες λέξεις χρησιμεύουν ως ένα είδος ελέγχου λογικότητας, διότι δεν επιτρέπουν τη διάδοση προσημασμένων βαθμολογιών, μέσω μιας ουδέτερης λέξης πυρήνα. Το τρέξιμο του αλγορίθμου οδηγεί σε ένα επεκταμένο λεξικό συναισθήματος. Τα επίθετα αποτελούν το μεγαλύτερο μέρος του παραχθέντος λεξιλογίου και ακολουθούν αριθμητικά τα ρήματα, τα ουσιαστικά και τελικά τα επιρρήματα. Οι περισσότερες από τις πολικότητες αποτελέσματος συμφωνούν με την ανθρώπινη διαίσθηση, αν και όχι σε όλες τις περιπτώσεις. Οι βαθμολογίες είναι συνήθως σωστές, παρόλο που μερικά βάρη που συμβάλλουν, έχουν λανθασμένη πολικότητα ή βασισμένη σε μια σπάνια έννοια της λέξης.

Στο δεύτερο στάδιο εξάγονται τα χαρακτηριστικά, που καθορίζονται μόνο από το κείμενο της κριτικής με ένα δυναμικό εξαγωγέα χαρακτηριστικών, και έναν στατικό εξαγωγέα, για τα χαρακτηριστικά που είναι προκαθορισμένα. Οι ταξινομητές εξαγωγής είναι εκπαιδευμένοι σε ένα σύνολο επισημασμένων δεδομένων. Οι στατικοί εξαγωγείς αξιοποιούν τα ήδη γνωστά χαρακτηριστικά κάποιου γνωσιακού τομέα, βελτιώνοντας τη γενική ακρίβεια του συστήματος.



Σχήμα 6.5: Παράδειγμα διαδικασίας εξαγωγής χαρακτηριστικών

Η έξοδος του ταξινομητή συναισθήματος και του εξαγωγέα χαρακτηριστικών είναι ένα σύνολο προτάσεων που έχουν επισημανθεί με συναίσθημα και τα αντίστοιχα χαρακτηριστικά τους. Αυτές οι προτάσεις εισάγονται έπειτα στον τελικό συνοψιστή (summarizer) ο οποίος υπολογίζει το μέσο συναίσθημα πάνω από κάθε γνώρισμα και επιλέγει τις κατάλληλες κειμενικές ενδείξεις, που θα περιληφθούν στην σύνοψη.

6.4 Αλγόριθμος Διάδοσης Γράφου

Οι Velikovich et al. (2010) [81] ερευνούν η βιωσιμότητα ιστογενών (web-derived) λεξικών πολικότητας, που προέρχονται αποκλειστικά από μη προσημασμένα έγγραφα του διαδικτύου. Για την εκμετάλλευση της αφθονίας της πληροφορίας συμφραζομένων, που υπάρχει στα πλήρη έγγραφα, προτείνουν μια μέθοδο βασισμένη στον Αλγόριθμο Διάδοση Γράφου (Graph Propagation - GP), εμπνευσμένη από τις προηγούμενες εργασίες των Blair-Goldensohn et al. (2008) [10] και Rao και Ravichandran (2009) [67], πάνω στην κατασκευή λεξικών πολικότητας, από λεξικολογικούς γράφους με χρήση του LP.

Για το σκοπό αυτό διεξάγουν ποιοτική και ποσοτική ανάλυση, ενός αγγλικού λεξικού δικτυακής προέλευσης, σε σχέση, με δύο προηγουμένως δημοσιευμένα λεξικά, όπως αυτό που χρησιμοποίησαν οι Wilson et al. (2005) [83] και αυτό που χρησιμοποίησαν οι Blair-Goldensohn et al. (2008) [10]. Τα πειράματά τους δείχνουν ότι ένα διαδικτυακής προέλευσης λεξικό είναι, όχι μόνο σημαντικά ευρύτερο, αλλά έχει και βελτιωμένη ακρίβεια στην ταξινόμηση πολικότητας φράσεων, κάτι που αποτελεί σημαντικό πρόβλημα σε πολλές εφαρμογές ανάλυσης συναισθήματος, περιλαμβανομένης της ομαδοποίησης (sentiment aggregation) και της σύνοψης συναισθήματος.

6.4.1 Δομή Αλγορίθμου GP

Έστω ένας μη κατευθυνόμενος γράφος βεβαρυμμένων ακμών $G(V, E)$, όπου $w_{ij} \in [0,1]$ το βάρος της ακμής $(v_i, v_j) \in E$. Το σύνολο κόμβων V είναι το σύνολο φράσεων, οι οποίες είναι υποψήφιες να συμπεριληφθούν στο λεξικό συναισθήματος. Ο γράφος G κωδικοποιεί σημασιολογικές ομοιότητες μεταξύ δύο κόμβων. Κατά την εκτέλεση του αλγορίθμου, θεωρούμε ότι $w_{ij} > w_{ik}$, εάν v_i = *«καλό»*, v_j = *«υπέροχο» και* v_k = *«κακό»*. Υποθέτουμε επίσης, σαν είσοδο του GP, δύο σύνολα φράσεων-πυρήνων, το σύνολο P για τους θετικούς πυρήνες και το σύνολο N για τους αρνητικούς.

Ως έξοδος λαμβάνουμε ένα διάνυσμα πολικότητας $\text{pol}_i \in \mathbb{R}^{|V|}$ τέτοιο ώστε το pol_i είναι ο βαθμός πολικότητας για την i^{th}υποψήφια φράση (ή ο κόμβος υπ' αριθ. i^{th}στο G). Το διάνυσμα pol έχει την ακόλουθη σημασιολογία:

$$\text{pol}_i = \begin{cases} > 0 & i^{th} \text{ φράση έχει θετική πολικότητα} \\ < 0 & i^{th} \text{ φράση έχει αρνητική πολικότητα} \\ = 0 & i^{th} \text{ φράση δεν έχει κανένα συναίσθημα} \end{cases}$$

Διαισθητικά, ο αλγόριθμος δουλεύει υπολογίζοντας ένα θετικό και ένα αρνητικό μέγεθος πολικότητας για κάθε κόμβο του γράφου, αποκαλούμενα pol_i^+ και pol_i^-. Αυτές οι τιμές είναι ίσες με το άθροισμα της μέγιστης, βεβαρυμμένης διαδρομής, από κάθε λέξη-πυρήνα (θετική ή αρνητική), έως τον κόμβο v_i. Φράσεις που συνδέονται με πολλαπλές θετικές λέξεις-πυρήνες μέσω σύντομων αλλά υψηλά βεβαρυμμένων διαδρομών, θα λάβουν υψηλές θετικές τιμές. Τότε η τελική πολικότητα μιας φράσης εξάγεται από τη σχέση $\mathrm{pol}_i = \mathrm{pol}_i^+ - \beta\mathrm{pol}_i^-$, όπου β μια σταθερά, που παριστά τη διαφορά της ολικής μάζας θετικής και αρνητικής ροής, στο γράφο. Έτσι, αφού τρέξει ο αλγόριθμος, εάν μια φράση έχει υψηλότερο θετικό, από αρνητικό βαθμό πολικότητας, η τελική πολικότητά της θα είναι θετική, διαφορετικά θα είναι αρνητική.

Ο αλγόριθμος έχει γραφτεί μέσα σε ένα επαναληπτικό πλαίσιο, όπου σε κάθε επανάληψη λαμβάνονται υπόψη διαδρομές με αυξανόμενο μήκος. Η μεταβλητή εισόδου T ελέγχει το μέγιστο μήκος διαδρομής που λαμβάνει υπόψη ο αλγόριθμος. Σ' αυτή μπορεί να τεθεί μια μικρή τιμή στην πράξη, δεδομένου ότι τα πολλαπλασιαστικά βάρη των διαδρομών έχουν ως αποτέλεσμα μακριές διαδρομές οι οποίες σπάνια συμβάλλουν στο βαθμό πολικότητας. Η παράμετρος γ λειτουργεί ως ένα κατώφλι, που καθορίζει το ελάχιστο μέγεθος πολικότητας που πρέπει να έχει μια φράση, ώστε να περιληφθεί στο λεξικό.

Ο αλγόριθμος εξελίσσεται στα ακόλουθα βήματα, ως εξής:

1. Θεώρησε μηδενική κάθε είδους πολικότητα, και μοναδιαίες τις πολικότητες, θετικές ή αρνητικές, των κόμβων των συνόλων P και N, αντίστοιχα
2. Θεώρησε μηδενικές τις βεβαρυμμένες διαδρομές μεταξύ όλων των κόμβων
3. Θεώρησε ως F, το εκάστοτε σύνολο κόμβων που έχουν σημανθεί ως θετικοί
4. Για κάθε $v_i \in P$ και για όλα τα $t : 1 \ldots T$ και για όλα τα ζεύγη ακμών (v_k, v_j) για τα οποία ισχύει ότι $v_k \in F$, υπολόγισε τη μέγιστη βεβαρυμμένη διαδρομή $a_{ij} = max\{a_{ij}, a_{ik} \cdot w_{kj}\}$ και το νέο σύνολο θετικών κόμβων $F = F \cup \{v_j\}$
5. Για κάθε κόμβο υπολόγισε το διάνυσμα θετικής πολικότητας pol_j^+
6. Επανάλαβε τα βήματα 3-4 για το σύνολο N, των αρνητικά προσημασμένων κόμβων
7. Για κάθε κόμβο, υπολόγισε το διάνυσμα αρνητικής πολικότητας pol_j^-
8. Υπολόγισε τη διαφορά της ολικής μάζας θετικής και αρνητικής ροής, β, στο γράφο
9. Για κάθε φράση, υπολόγισε την τελική πολικότητα, pol_i
10. Εάν $|\mathrm{pol}_i| < \gamma$ θεώρησε μηδενική την τελική πολικότητα μίας φράσης

Το T και το γ ρυθμίζονται πάνω σε προσημασμένα δεδομένα, τα οποία χωρίστηκαν εξ ημισείας, σε ομάδες πυρήνων και δοκιμών (hold-out validation). Για να δομηθεί το τελικό λεξικό, εξάγονται οι εναπομείναντες κόμβοι, με βαθμό πολικότητας μεγαλύτερο του γ, και τους αποδίδεται η αντίστοιχη πολικότητα.

Αλγόριθμος Διάδοσης Ετικέτας	Αλγόριθμος Διάδοσης Γράφου
Είσοδος: *G = (V,E),* $w_{ij} \in [0,1]$ *P, N* Έξοδος: pol $\in \mathbb{R}^{\lvert V \rvert}$	Είσοδος: *G = (V,E),* $w_{ij} \in [0,1]$ *P, N,* $\gamma \in \mathbb{R}, T \in \mathbb{N}$ Έξοδος: pol $\in \mathbb{R}^{\lvert V \rvert}$
Αρχικοποίηση: $\text{pol}_i = 1.0$ για όλα τα $v_i \in P$ $\text{pol}_i = -1.0$ για όλα τα $v_i \in N$ $\text{pol}_i = 0.0\ \forall v_i \notin P \cup N$	Αρχικοποίηση: $\text{pol}_i, \text{pol}_i^+, \text{pol}_i^- = 0$, για όλα τα i $\text{pol}_i^+ = 1.0$ για όλα τα $v_i \in P$ $\text{pol}_i^- = 1.0$ για όλα τα $v_i \in N$
1. Για: t .. T 2. $\text{pol}_i = \frac{\sum_{(v_i,v_j)\in E} w_{ij} \times pol_j}{\sum_{(v_i,v_j)} w_{ij}}, \forall v_i \in V$ 3. Επαναφορά $\text{pol}_i = 1.0 \forall \in P$ 4. Επαναφορά $\text{pol}_i = -1.0 \forall \in N$	1. Θέσε $a_{ij} = 0$ για όλα τα i, j 2. Για $v_i \in P$ 3. $F = \{v_i\}$ 4. Για $t : 1 \ldots T$ 5. Για $(v_k, v_j) \in E$ ώστε $v_k \in F$ 6. $a_{ij} = max\{a_{ij}, a_{ik} \cdot w_{kj}\}$ $F = F \cup \{v_j\}$ 7. Για $v_j \in V$ 8. $\text{pol}_i^+ = \sum v_i \in P^{a_{ij}}$ 9. Επανάλαβε τα βήματα 1-8 χρησιμοποιώντας το *N* για να υπολογίσεις το pol^- 10. $\beta = \sum_i \text{pol}_i^+ / \sum_i \text{pol}_i^-$ 11. $\text{pol}_i = \text{pol}_i^+ - \beta \text{pol}_i^-$, για όλα τα i 12. Αν $\lvert \text{pol}_i \rvert < \gamma$ τότε $\text{pol}_i = 0.0$, για όλα τα i

Πίνακας 6.1: Σύγκριση βημάτων διάδοσης ετικέτας LP και GP

6.4.2 Χαρακτηριστικά Μεθόδου GP

Οι Velikovich et al. (2010) [81] χρησιμοποίησαν μια διαφορετική μέθοδο διάδοσης ετικέτας από το συμβατικό LP. Αντί να βασίζονται σε διάχυση κατά μήκος ολόκληρου του γράφου, η μέθοδος τους λαμβάνει υπόψη μόνο τη μία ισχυρότερη διαδρομή μεταξύ κάθε υποψήφιας και κάθε λέξης σπόρου. Ο αλγόριθμος τους είναι καταλληλότερος από εκείνον των Zhu και Ghahramani (2002) [85] για ένα σύνολο δεδομένων βασισμένων στο διαδίκτυο, ο οποίος περιέχει πολλούς πυκνούς υπογράφους και αναξιόπιστες συσχετίσεις βασισμένες μόνο σε στατιστικές συνεμφανίσεων.

Ο GP εκτελέστηκε πάνω σε ένα γράφο δομημένο από το διαδίκτυο, χρησιμοποιώντας χειροκινήτως κατασκευασμένα σύνολα, θετικών και αρνητικών πυρήνων, μεγέθους 187 και 192 λέξεων αντίστοιχα, πολλές από τις οποίες ήταν μορφολογικές παραλλαγές της ίδιας ρίζας. Ο αλγόριθμος παρήγαγε ένα λεξικό που περιείχε 178 χιλιάδες λήμματα. Ανάλογα με το κατώφλι γ, αυτό το λεξικό μπορούσε να είναι μεγαλύτερο ή μικρότερο.

Για να προσδιοριστεί η πρακτική χρησιμότητα ενός λεξικού πολικότητας που προέρχεται από το διαδίκτυο, μετρήθηκε η απόδοση του, σε μια εργασία κατηγοριοποίησης/κατάταξης προτάσεων. Ως είσοδος δόθηκε ένα σύνολο από προτάσεις και ως έξοδος μια συναισθηματική ταξινόμηση των προτάσεων, ως θετικών, αρνητικών ή ουδέτερων. Επιπλέον, το σύστημα εξήγαγε δύο σειρές κατάταξης: Η πρώτη είναι μια κατάταξη της πρότασης ως προς την θετική πολικότητα και η δεύτερη μια κατάταξη, ως προς την αρνητική πολικότητα.

Για την ταξινόμηση των προτάσεων ως θετικές, αρνητικές ή ουδέτερες, χρησιμοποιήθηκε ο ενισχυμένος αλγόριθμος αντιστροφής ψήφου (augmented vote-flip algorithm) [21] σύμφωνα με τον οποίο μετράται ο αριθμός των αντιστοιχισμένων θετικών και αρνητικών φράσεων από το λεξικό, και όποια κατηγορία έχει τις περισσότερες ψήφους κερδίζει. Ο αλγόριθμος αντιστρέφει την απόφαση, εάν ο αριθμός των αρνήσεων είναι μονός. Όπως και σε άλλες μεθόδους αλγορίθμων LP, για την κατάταξη των προτάσεων ορίστηκε η καθαρότητα X μιας πρότασης, ως το κανονικοποιημένο άθροισμα των βαθμολογιών συναισθήματος για κάθε φράση x στην πρόταση, η οποία είναι μια κανονικοποιημένη βαθμολογία στο διάστημα $[-1, 1]$, δηλαδή $\text{purity(X)} = \frac{\sum_{x \in X} \text{pol}_x}{\delta + \sum_{x \in X} |\text{pol}_x|}$.

Τα δεδομένα που χρησιμοποιήθηκαν ήταν ένα σύνολο από κριτικές καταναλωτών στην αγγλική γλώσσα, όπως περιγράφονται από τους McDonald et al. (2007) [52]. Τα χαρακτηριστικά που περιλαμβάνονταν στον ταξινομητή ήταν ο βαθμός καθαρότητας, ο αριθμός θετικών και αρνητικών αντιστοιχήσεων του λεξικού, ο αριθμός των αρνήσεων στην πρόταση, καθώς επίσης και αλληλουχίες αυτών των χαρακτηριστικών

στο εσωτερικό της πρότασης και με τα ίδια χαρακτηριστικά παραγόμενα από τις προτάσεις σε ένα παράθυρο μοναδιαίου μεγέθους. Επιπλέον, όλες οι προτάσεις τοποθετήθηκαν στις θετικές και αρνητικές ιεραρχήσεις με βάση την πιθανότητα που έδωσε ο ταξινομητής στις θετικές και αρνητικές κατηγορίες, αντίστοιχα.

Το πιο ενδιαφέρον χαρακτηριστικό του παραγόμενου ιστογενούς λεξικού είναι ότι το συχνότερο μήκος φράσης είναι το 2 και όχι 1. Ο κύριος λόγος γι' αυτό είναι η αφθονία φράσεων με επίθετα, που αποτελούνται από ένα επίρρημα και ένα επίθετο. Σχεδόν κάθε επίθετο μοναδιαίου μήκους, συνδυάζεται με τέτοιο τρόπο στον ιστό, έτσι ώστε δεν προκαλεί έκπληξη το ότι βλέπουμε πολλές από αυτές τις φράσεις στο λεξικό και αυτό το επίθετο τους έχει την υψηλότερη πολικότητα. Αυτές οι φράσεις είναι αναγκαστικά πιο κοινές και έτσι έχουν περισσότερες ακμές με μεγαλύτερα βάρη στο γράφο και συνεπώς μεγαλύτερη πιθανότητα συσσώρευσης υψηλής βαθμολογίας συναισθήματος.

Οι Brody και Diakopoulos (2012) [15] εφαρμόζοντας τον GP αποκλειστικά σε δεδομένα κοινωνικών δικτύων (Twitter), όπου οι γράφοι είναι συγκριτικά μικρότεροι, διαπιστώνουν ότι δεν χρειάζεται να περιοριστεί το μήκος της διαδρομής T, ούτε να χρησιμοποιηθεί το κατώφλι γ, αλλά μόνο μια απλή αποκοπή των n κορυφαίων λέξεων. Η μέθοδός, επίσης, είναι ικανή να ανιχνεύσει όρους οι οποίοι σχετίζονται με συναίσθημα σε διαφορετικά χρονικά σημεία, κάτι που δεν είναι εφικτό με ένα σταθερό λεξικό.

Τόσο θετικές όσο και αρνητικές φράσεις, συνήθως επίθετα, που αναμένεται να υπάρχουν σε ένα λεξικό συναισθήματος δεν περιλαμβάνονται στα προσημασμένα σύνολα πυρήνων. Παράλληλα, οι παραλλαγές ορθογραφίας και τα λάθη είναι πολύ περισσότερα στις θετικές φράσεις, απ' ότι στις αρνητικές. Πολλές από αυτές αντιστοιχούν στο κείμενο των κοινωνικών μέσων δικτύωσης, όπου κανείς εκφράζει ένα αυξημένο επίπεδο συναισθήματος με την επανάληψη χαρακτήρων και αντίστοιχα με τη χυδαιότητα στις αρνητικές φράσεις, να είναι πολύ μεγαλύτερη από ό, τι στις θετικές φράσεις. Υπάρχει επίσης ένας αριθμός υποτιμητικών και ρατσιστικών όρων στο λεξικό, οι περισσότεροι εκ των οποίων έλαβαν αρνητικό συναίσθημα λόγω της τυπικά δυσφημιστικής χρήσης τους.

6.5 Αλγόριθμος Διάδοσης Ετικέτας Τροποποιημένης Προσρόφησης

Μία μεγάλη συζήτηση έχει πραγματοποιηθεί για τα αυτοματοποιημένα εργαλεία, που θα μπορούσαν να αποδώσουν πολικότητα σε κείμενα από μικρο-blog, όπως οι αναρτήσεις του Facebook και οι δημοσιεύσεις (tweet) στο Twitter. Αυτό όμως είναι δύσκολο λόγω της συντομίας και της ανεπισημότητας τους, πέρα από την μεγάλη

ποικιλία και ταχεία εξέλιξη της γλώσσας στα μέσα κοινωνικής δικτύωσης. Έτσι, δεν είναι πρακτική η χρήση τυποποιημένων τεχνικών επιβλεπόμενης μηχανικής μάθησης, που βασίζονται μόνο σε προσημασμένα δείγματα εκπαίδευσης.

Οι Speriosu et al. (2011) [71] εργάστηκαν πάνω σε μία μέθοδο, χωρίς έναν τέτοιο σχολιασμό, με τη χρήση του LP για να ενσωματώσουν ετικέτες από έναν ταξινομητή μέγιστης εντροπίας εκπαιδευμένο σε θορυβώδεις ετικέτες και σε προϋπάρχουσες γνώσεις σχετικά με τους τύπους λέξεων, που είναι κωδικοποιημένους σε ένα λεξικό, σε συνδυασμό με τον follower graph της Twitter. Η μέθοδος τους βασίζεται στον Αλγόριθμο Διάδοσης Ετικέτας Τροποποιημένης Προσρόφησης (Modified Adsorption LP – MAD LP) [75]. Τα αποτελέσματα της ταξινόμησης πολικότητας, για αρκετά σύνολα δεδομένων, δείχνουν ότι η δική τους προσέγγιση στη χρήση του LP, συναγωνίζεται επάξια σε απόδοση αντίστοιχα επιβλεπόμενα μοντέλα, με σχολιασμένα tweet κάποιου συγκεκριμένου γνωσιακού τομέα, και ξεπερνά σε απόδοση το θορυβωδώς επιβλεπόμενο ταξινομητή τον οποίο εκμεταλλεύεται, καθώς και έναν ταξινομητή αναλογίας πολικότητας (polarity ratio) βασισμένο σε λεξικό.

6.5.1 Δομή Αλγορίθμου MAD LP

Όπως και στους αλγόριθμους διάδοσης ετικέτας που εξετάστηκαν μέχρι στιγμής, ο MAD LP εξαπλώνει τις ετικέτες, από ένα μικρό σύνολο κόμβων, με πυρήνα κάποια αρχική πληροφορία ετικέτας, σε όλο το γράφο. Οι κατανομές ετικέτας εξαπλώνονται σε ένα γράφο $G(V, E, W)$, όπου V είναι το σύνολο n κόμβων, E είναι ένα σύνολο m ακμών και W είναι ένας $n \times n$ πίνακας βαρών, με W_{ij} ως το βάρος της ακμής (i, j). Στην Διάδοση Ετικέτας Τροποποιημένης Προσρόφησης η διάδοση πραγματοποιείται πάνω σε ένα γράφο, με κόμβους, που αντιπροσωπεύουν tweet, συγγραφείς και χαρακτηριστικά, ενώ μεταβάλλονται οι πληροφορίες του πυρήνα και η δομή των συνόλων ακμών.

Κάθε στιγμιότυπο του κατασκευαζόμενου γράφου, κατά τη διάρκεια εκτέλεσης του αλγορίθμου, σχετίζεται με δύο γραμμικά διανύσματα $Y_v, \hat{Y}_v \in \mathbb{R}_+^m$. Το l στοιχείο του Y_v κωδικοποιεί την προυπάρχουσα γνώση για τον κόμβο v. Όσο μεγαλύτερη είναι η τιμή του Y_{vl} , τόσο ισχυρότερη είναι η πεποίθηση ότι η ετικέτα του v, θα πρέπει να είναι η $l \in L = \{1 \ldots m\}$, και μία μηδενική τιμή, $Y_{vl} = 0$, δηλώνει άγνοια για την ετικέτα του. Το $\hat{Y}_v$ είναι η έξοδος του αλγορίθμου, με χαρακτηριστικά αντίστοιχα με το Y_v.

Η εξάπλωση των κατανομών ετικέτας μπορεί να θεωρηθεί ως ένας ελεγχόμενος Τυχαίος Περίπατος, με τρεις πιθανές δράσεις: (i) τον εμβολιασμό ενός «σπαρμένου κόμβου», δηλαδή ενός κόμβου του πυρήνα, με την αντίστοιχη ετικέτα «σπόρο» (inject), (ii) τη συνέχιση της πορείας, από τον τρέχοντα κόμβο προς γειτονικό κόμβο (continue), και (iii) την εγκατάλειψη της πορείας (abandon), με τις αντίστοιχες

πιθανότητες να αθροίζονται στη μονάδα, $p_v^{in\,j} + p_v^{cont} + p_v^{abnd} = 1$. Για την απόδοση ετικέτας σε ένα κόμβο, προσημασμένο ή μη, εκκινείται ένας τυχαίος περίπατος ξεκινώντας από αυτόν. Στην πρώτη περίπτωση, που πραγματοποιείται η πιθανότητα $p_v^{in\,j}$ ο περίπατος σταματάει και επιστρέφει το διάνυσμα πληροφορίας προσήμανσης Y_v, θεωρώντας ως $p_v^{in\,j} = 0$ για τους μη σημασμένους κόμβους. Στη δεύτερη περίπτωση, που πραγματοποιείται η πιθανότητα p_v^{cont}, ο περίπατος συνεχίζεται σε κάποιον από τους γειτονικούς κόμβους, με πιθανότητα ανάλογη του βάρους της ακμής, $W_{v'v} \geq 0$, που τους ενώνει. Στην τρίτη περίπτωση, που πραγματοποιείται η πιθανότητα p_v^{abnd} , ο περίπατος εγκαταλείπει την διαδικασία απόδοσης ετικετών και επιστρέφει μηδενικής τιμής διάνυσμα r, όπου οι απορριφθείσες ετικέτες κωδικοποιούνται ως $\nu \notin L$, $r_l = 0$ για $l \neq \nu$ και $r_\nu = 1$.

Σε κάθε περίπτωση, οι πιθανότητες μετάβασης δίνονται από τη σχέση:

$$Pr[v'|u] = \begin{cases} \dfrac{W_{v'u}}{\sum_{v:(u,v)\in E} W_{uv}} & (v', v) \in E \\ 0 \end{cases}$$

Ο MAD LP παίρνει, επίσης, τρεις παραμέτρους, μ_1, μ_2 και μ_3, οι οποίες ελέγχουν τη σχετική σπουδαιότητα, κάθε μιας από τις τρεις παραπάνω ενέργειες, αντίστοιχα. Η τροποποιημένη προσρόφηση απαιτεί από κάποιους κόμβους στον γράφο, να έχουν κατανομές κόμβων πυρήνα (seed nodes), οι οποίες μπορεί να προέρχονται από μια ποικιλία γνωσιακών τομέων.

Θεωρώντας την επίλυση της κατασκευής του γράφου, ως ένα γραμμικό σύστημα $Mx = b \Rightarrow (\mu_1 S + \mu_2 L + \mu_3 I)\hat{Y}_l = \mu_1 S Y_l + \mu_3 R_l$, όπου $L = D + \overline{D} - (p_v^{cont} W_{vu}) - (p_v^{cont} W_{vu})^\top$, $S = p_v^{in\,j}$ και R η μήτρα με μηδενικές τις πρώτες m στήλες και τις υπόλοιπες ίσες με τα στοιχεία του διανύσματος $p_v^{abnd} \times r$, εφαρμόζεται η ιακωβιανή επαναληπτική μέθοδο, και η αναμενόμενη βαθμολογία $\hat{Y}_v$ για ένα κόμβο $v \in V$ παίρνει την ακόλουθη μορφή:

$$\hat{Y}_v = \frac{1}{M_{vv}} \left(\mu_1 \times p_v^{in\,j} \times Y_v + \mu_2 \times D_v + \mu_3 \times p_v^{abnd} \times r \right)$$

Όπου $M_{vv} = \mu_1 \times p_v^{in\,j} + \mu_2 \sum_{u \neq v} (p_v^{cont} W_{vu} + p_u^{cont} W_{uv}) + \mu_3$ και $D_v = \sum_u (p_v^{cont} W_{vu} + p_u^{cont} W_{vu})\, \hat{Y}_v$.

Αλγόριθμος Διάδοσης Ετικέτας	Αλγόριθμος Διάδοσης Ετικέτας Τροποποιημένης Προσρόφησης
Είσοδος: $G = (V,E)$, $w_{ij} \in [0,1]$ P, N Έξοδος: pol $\in \mathbb{R}^{\lvert V \rvert}$	Είσοδος: $G = (V,E,W)$, $Y_v \in \mathbb{R}^{m+1}$, $p_v^{in\ j}$, p_v^{cont}, p_v^{abnd} Έξοδος: $\hat{Y}_v$ για όλα τα $v \in V$
Αρχικοποίηση: $\text{pol}_i = 1.0$ για όλα τα $v_i \in P$ $\text{pol}_i = -1.0$ για όλα τα $v_i \in N$ $\text{pol}_i = 0.0\ \forall v_i \notin P \cup N$	Αρχικοποίηση: $\hat{Y}_v = Y_v$ για όλα τα $v \in V$
1. Για: t .. T 2. $\text{pol}_i = \frac{\sum_{(v_i,v_j)\in E} w_{ij} \times pol_j}{\sum_{(v_i,v_j)} w_{ij}}, \forall v_i \in V$ 3. Επαναφορά $\text{pol}_i = 1.0 \forall \in P$ 4. Επαναφορά $\text{pol}_i = -1.0 \forall \in N$	1. $M_{vv} = \mu_1 \times p_v^{in\ j} + \mu_2 \sum_{v \neq u} (p_v^{cont} W_{vu} + p_u^{cont} W_{vu}) + \mu_3$ 2. $D_v = \sum_u (p_v^{cont} W_{vu} + p_u^{cont} W_{vu})\, \hat{Y}_v$ 3. Για όλα τα $v \in V$: 4. $\hat{Y}_v = \frac{1}{M_{vv}} \left(\mu_1 \times p_v^{in\ j} \times Y_v + \mu_2 \times D_v + \mu_3 \times p_v^{abnd} \times r\right)$ 5. Επανάλαβε τα βήματα 1-5 μέχρι να επέλθει σύγκλιση

Πίνακας 6.2: Σύγκριση βημάτων διάδοσης ετικέτας LP και MAD LP

Ο αλγόριθμος εξελίσσεται στα ακόλουθα βήματα ως εξής:

1. Θεώρησε μηδενικές τις πολικότητες όλων των μη προσημασμένων κόμβων.
2. Αρχικοποίησε την μήτρα μετάβασης $\hat{Y}_v$, σύμφωνα με την προϋπάρχουσα γνώση για τις ετικέτες των προσημασμένων κόμβων, των διανυσμάτων στη μήτρα Y_v.
3. Όλοι οι κόμβοι διαδίδουν τις ετικέτες τους, μέσω τυχαίων περιπάτων, προς τους μη προσημασμένους κόμβους, σύμφωνα με τις τρεις δυνατές περιπτώσεις.
4. Υπολόγισε τη μήτρα M_{vv} , σύμφωνα με τις πιθανότητες κάθε περίπτωσης για κάθε κόμβο.
5. Υπολόγισε τη μήτρα D_v , όπως αυτή προκύπτει από τις πιθανότητες κάθε περίπτωσης και την αρχική μήτρα $\hat{Y}_v$.
6. Για κάθε κόμβο υπολόγισε τη νέα μήτρα $\hat{Y}_v$.
7. Επανάλαβε τη διαδικασία από το βήμα 5, μέχρι η μήτρα $\hat{Y}_v$ να συγκλίνει.

- Παραλλαγές κατασκευής γράφου:

    1. **Maxent-seed:** Κάθε tweet αναπαριστάται με ένα κόμβο στο γράφο και «σπέρνεται» με τις προγνώσεις πολικότητας για τα tweet, όπως αυτές προήλθαν από τον εκπαιδευμένο, σε emoticon, ταξινομητή.

    2. **Lexicon- seed:** Δημιουργούνται κόμβοι για κάθε λέξη στο λεξικό. Οι θετικές λέξεις «σπείρονται» ως 90% θετικές, εάν είναι έντονα υποκειμενικές και ως 80% θετικές, εάν είναι ασθενώς υποκειμενικές. Το ίδιο και αντίστροφα γίνεται για τις αρνητικές λέξεις. Κάθε tweet συνδέεται μέσω μιας ακμής με κάθε λέξη στο λεξικό πολικότητας που περιέχει.

    3. **Emoticon-seed:** Δημιουργούνται κόμβοι για emoticons και «σπείρονται», ως 90% θετικοί ή αρνητικοί, ανάλογα με την πολικότητα τους.

    4. **Annotated-seed:** Χρησιμοποιούνται οι σχολιασμοί, ενός εκπαιδευμένου ταξινομητή σε συγκεκριμένο γνωσιακό τομέα, για να «σπαρούν» τα tweet από αυτό το σύνολο δεδομένων, ως 100% θετικά ή αρνητικά, σύμφωνα με την προσήμανσή τους.

- Παραλλαγές απόδοσης βαρών:

    1. **Follower-edges:** Όταν ένας χρήστης A διαδέχεται έναν άλλο χρήστη B,

προστίθεται μια ακμή από τον A στον B, με βάρος 1,0, που είναι συγκρίσιμο με εκείνο μιας μέτρια συχνής λέξης στο Feature-edges κατωτέρω.

2. **Feature-edges:** Προστίθενται κόμβοι για τα hashtag και τα χαρακτηριστικά των μονογραμμάτων και διγραμμάτων και συνδέονται με τα tweet, που τα περιέχουν. Μια ακμή, που συνδέει ένα tweet t με ένα γνώρισμα f, έχει βάρος w_{tf} χρησιμοποιώντας αναλογίες σχετικής συχνότητας του χαρακτηριστικού μεταξύ του εξεταζόμενου συνόλου δεδομένων d και του συνόλου δεδομένων των emoticon, ως σώμα κειμένου αναφοράς r :

$$w_{tf} = \begin{cases} \log \frac{P_d(f)}{P_r(f)} & if\ P_d(f) > P_r(f) \\ 0 & o.w. \end{cases}$$

Οι τρείς πρώτες τεχνικές κατασκευής γράφου περιέχουν μεγάλο ποσοστό θορύβου (noisy seeding) αφού περιλαμβάνουν πλήθος κόμβων με όρους, εκτός του εκάστοτε εξεταζόμενου γνωσιακού τομέα. Κύριο θέμα της κατασκευής γράφου είναι ο προσδιορισμός των ακμών και των βαρών τους. Κάθε χρήστης u_n είναι συνδεδεμένος με οποιονδήποτε τον ακολουθεί ή με οποιονδήποτε αυτός ακολουθεί. Κάθε χρήστης συνδέεται επίσης με τα tweet που έγραψε. Οι λέξεις, από το λεξικό του MAD LP συνδέονται με tweet, που τις περιέχουν και ομοίως τα hashtag, τα emoticon, τα μονογράμματα και τα διγράμματα. Τα emoticon και οι λέξεις από το λεξικό, «σπείρονται» και στο τέλος, σε όλες τις ακμές, εκτός από τις ακμές-χαρακτηριστικά (feature-edges) δίνεται μοναδιαίο βάρος.

6.5.2 Χαρακτηριστικά Μεθόδου MAD LP

Η πλήρης ανάλυση συναισθήματος για ένα συγκεκριμένο ερώτημα ή θέμα απαιτεί πολλά στάδια. Οι Speriosu et al. (2011) [71] όπως και οι περισσότερες εργασίες στην ανάλυση συναισθήματος, εστιάζουν στο τελευταίο στάδιο, την ταξινόμηση πολικότητας. Στην εργασία τους, συγκεντρώνουν αρκετές από τις παραπάνω προσεγγίσεις, στις οποίες αρχίζοντας με ένα μικρό αριθμό χειρονακτικά προσημασμένων λέξεων, παράγουν ένα λεξικό, μέσω της διάδοσης ετικέτας, για χρήση στην ταξινόμηση πολικότητας.

Η προσέγγιση του MAD LP, συγκρίνεται, έπειτα, την, με τον ίδιο τον υπό συνθήκες θορύβου επιβλεπόμενο ταξινομητή, και με μια μέθοδο βασιζόμενη σε λεξικό που παράγει ο αλγόριθμος, χρησιμοποιώντας αναλογίες θετικών / αρνητικών όρων. Πέρα από τη μέτρηση της ακρίβειας του προτύπου ανά tweet, μετράται και η ακρίβεια, ανά στόχο και μια συνολική μετρική σφάλματος, πάνω σε όλους τους χρήστες, στο σύνολο της δοκιμής, που καταγράφει πόσο κοντά είναι η προβλεφθείσα θετικότητα,

του κάθε χρήστη, με την πραγματική του θετικότητα.

Για την Τροποποιημένη Προσρόφηση χρησιμοποιούνται 100 επαναλήψεις και μια παράμετρος εισαγωγής πυρήνα μ_1, με τιμή 0.005, δίνει την καλύτερη ισορροπία, επιτρέποντας στις κατανομές πυρήνων, να επηρεάσουν άλλους κόμβους, χωρίς όμως να τους κατακλύσουν. Για τα μ_2 και μ_3 χρησιμοποιείται η προκαθορισμένη τιμή 0,01. Για όλα τα σύνολα δεδομένων ο MAD LP με ακμές-χαρακτηριστικά και noisy-seed τεχνικές κατασκευής γράφων ξεπερνά ή είναι ισάξιο όλων των άλλων μεθόδων.

Σε ότι αφορά τις τεχνικές δόμησης του γράφου και απόδοσης βαρών, ένας γράφος με ακμές-γνωρίσματα, «σπαρμένος» με ετικέτες με την τεχνική Annotated-seed, επιτυγχάνει μόνο 64,6% ακρίβεια ανά tweet και ένας ταξινομητής μέγιστης εντροπίας 66,7%. Η καλύτερή προσέγγιση διάδοσης ετικέτας ξεπερνά και τα δύο, με 71,2%. Η τεχνική Follower-edges υστερεί εν γένει, απαιτώντας περαιτέρω εκλέπτυνση της μοντελοποίησης του κοινωνικού γράφου, ώστε να αυξήσει την ακρίβεια της, επιτρέποντας στους χρήστες να επηρεάζουν τους ακολούθους τους, περισσότερο από ό, τι το αντίστροφο.

Ταξινομητής	Μέσο Τετραγωνικό Σφάλμα
MAD LP (Follower-edges, Maxent-seed)	0,233
MAD LP (All-edges, Lexicon-seed)	0,187
MAD LP (Feature-edges, Noisy-seed	0,148
MAD LP (All-edges, Noisy-seed)	0,148

Πίνακας 6.3: Εξάρτηση Σφάλματος από τον Ταξινομητή του MAD LP

6.6 Σύγκριση Αλγορίθμων Διάδοσης Ετικέτας

6.6.1 Γλωσσική Ανεξαρτησία

Ο LP αλγόριθμος με κατεξοχήν δυνατότητες εφαρμογής σε διαφορετικές γλώσσες είναι αυτός των Rao και Ravichandran (2009) [67]. Ο CR LP επιτρέπει με φυσικό τρόπο τον συνδυασμό αρκετών σχέσεων από το WordNet. Η προσέγγισή αυτή λειτουργεί για όλες τις κατηγορίες λέξεων, και όχι μόνο για τα επίθετα, όπως σε αρκετές περιπτώσεις πριν από αυτή. Αν και δημοσιεύουν αποτελέσματα για τα αγγλικά, χίντι, και γαλλικά, οι μέθοδός τους μπορεί εύκολα να αναπαραχθεί και σε άλλες γλώσσες, στις οποίες είναι διαθέσιμο το WordNet. Προκειμένου να αποδείξουν την ευκολία προσαρμογής της μεθόδου και σε άλλες γλώσσες,

χρησιμοποιήθηκε το Hindi WordNet για να παραχθεί ο αντίστοιχος γράφος συνώνυμων επιθέτων. Τα Χίντι είναι μια ινδοαρειανή γλώσσα που μιλιέται στην Ινδία από 185 εκατομμύρια άτομα ως μητρική, γεγονός που την κάνει την 5η πιο διαδεδομένη γλώσσα στον κόσμο σε αριθμό ομιλητών ως μητρική, αλλά παρόλα αυτά έχει μικρό αριθμό χρηστών στο διαδίκτυο[17]. Ο CR LP εκτελέστηκε επιτυχώς, χρησιμοποιώντας μία λίστα πυρήνα 489 επιθέτων, προσημασμένη από δύο φυσικούς ομιλητές της γλώσσας.

Επειδή το WordNet μπορεί να μην είναι ελεύθερα διαθέσιμο για όλες τις γλώσσες ή μπορεί να μην υπάρχει καν σε κάποιες, μπορεί να αρκεί η δόμηση του γράφου από ένα διαθέσιμο θησαυρό. Για παράδειγμα, αν και το French WordNet είναι διαθέσιμο, το κόστος αγοράς από πολλούς μελετητές κρίνεται ως απαγορευτικό. Παρατηρείται ότι αν χρησιμοποιηθεί μόνο η σχέση συνωνυμίας στο WordNet, τότε μπορεί να χρησιμοποιηθεί αντί αυτού, κάθε θησαυρός. Για το λόγο αυτό, επιλέχθηκε ο θησαυρός του γαλλικού OpenOffice που είναι ελεύθερα διαθέσιμο. Αν και τα αποτελέσματα δεν ήταν εξίσου καλά όπως στα χίντι, αυτό αποδόθηκε αφενός στην υψηλότερη συμφωνία μεταξύ των φυσικών σχολιαστών (70% στη Χίντι σε σύγκριση με 55% στη γαλλική γλώσσα), αφετέρου στο ότι το πείραμα με τη Χίντι, όπως και με τα αγγλικά, χρησιμοποιήθηκε το WordNet, ενώ στο πείραμα στα γαλλικά πραγματοποιήθηκε σε γράφους, που προήλθαν από τον θησαυρό του OpenOffice, λόγω έλλειψης ελεύθερα διαθέσιμου γαλλικού WordNet.

Εν γένει, σε καταστάσεις έλλειψης πόρων, όπου δεν είναι διαθέσιμο το WordNet ή άλλοι θησαυροί, αλλά μόνο μονόγλωσσο ακατέργαστο κείμενο, μπορεί ομοίως να εκτελεστεί ο CR LP, στους πλησιέστερους γειτονικούς γράφους, που προέρχονται απ' ευθείας από ακατέργαστο κείμενο, χρησιμοποιώντας μεθόδους ομοιότητας κατανομής. Γίνεται αντιληπτό ότι ο βαθμός εξάρτησης από μία δομημένη λεξικολογική πηγή, όπως το WordNet, καθορίζει και τη δυνατότητα μεταφερσιμότητας μία έκδοσης του LP σε κάποια άλλη γλώσσα.

Για τους παραπάνω λόγους και ο αλγόριθμος των Velikovich et al. (2010) [81] έχει αρκετά χαρακτηριστικά γλωσσικής ανεξαρτησίας. Ενώ οι περισσότερες προσπάθειες χρησιμοποιούν το WordNet για να δομήσουν το λεξικολογικό γράφο, πάνω στον οποίο τρέχει ο LP, τα δικά τους λεξικά δομούνται από τον GP, με τη χρήση γράφου, από στατιστικές συνεμφάνισης, από ολόκληρο τον Ιστό. Αν και ο αλγόριθμός τους εφαρμόστηκε μόνο σε κείμενα της αγγλικής, η μέθοδος που ερευνούν μπορεί να θεωρηθεί ως συνδυασμός μεθόδων για τη διάδοση συναισθήματος μέσω λεξικολογικών γράφων και μεθόδων δόμησης λεξικών συναισθήματος, βασισμένων σε χαρακτηριστικά της κατανομής φράσεων σε ακατέργαστα δεδομένα, ανεξάρτητα γλωσσικής προέλευσης [78].

Στο ίδιο μήκος κύματος κινείται και ο αλγόριθμος των Brody και Elhadad (2010) [16] ο οποίος όμως από την κατασκευή του εστιάζει περισσότερο από τον GP σε αυτόν

[17] https://en.wikipedia.org/wiki/Hindi

τον τομέα. Η προσέγγισή τους έχει σχεδιαστεί, ώστε να είναι όσο το δυνατό λιγότερο επιβλεπόμενη και με λιγότερη απαίτηση γνώσης, για να μπορεί να μεταφερθεί σε διάφορα είδη προϊόντων και υπηρεσιών, καθώς και σε πολλές γλώσσες. Η τοπική έκδοση της LDA που εφαρμόζουν, λειτουργεί πρωτίστως πάνω σε προτάσεις και δευτερεύοντος σε έγγραφα, χρησιμοποιώντας ένα μικρό αριθμό θεμάτων που αντιστοιχούν άμεσα σε κάποια χαρακτηριστικά. Χρησιμοποιούν δείκτες μορφολογικής άρνησης, για να δημιουργήσει αυτόματα ένα σύνολο πυρήνων, με πολύ σχετικά θετικά και αρνητικά επίθετα, εγγυημένης συνάφειας με το εκάστοτε χαρακτηριστικό. Αυτά τα αυτομάτως παραγόμενα σύνολα-πυρήνες, με την εφαρμογή του AB LP, πετυχαίνουν συγκρίσιμα αποτελέσματα με τα παραγόμενα χειρωνακτικώς, και η χρήση της άρνησης μπορεί εύκολα να μεταφερθεί σε άλλες γλώσσες.

Από την άλλη πλευρά, αλγόριθμοι οι οποίοι εξαρτώνται πλήρως από το Wordnet ή άλλες αντίστοιχες συμπαγείς λεξιλογικές πηγές, κρίνονται ακατάλληλοι για διαγλωσσική ανάλυση συναισθήματος. Ενδεικτικό παράδειγμα αποτελεί η μέθοδος των Blair -Goldensohn et al. (2008) [10], στην οποία, μετά τον καθορισμό ενός μικρού, αρχικού, λεξικού-πυρήνα γνωστών θετικών και αρνητικών όρων συναισθήματος, το λεξικό επεκτείνεται κατά την εκτέλεση του AO LP μέσω δεσμών συνωνύμων ή και αντωνύμων, του WordNet. Παρόμοια προσέγγιση, ως προς το γλωσσικό περιορισμό, έχει και η μέθοδος των Speriosu et al. (2011) [71], οι οποίοι δημιούργησαν ένα σύνολο εκπαίδευσης από ένα δείγμα μίας ροής ενημερώσεων του (Twitter feed), λαμβάνοντας μια ισορροπημένη αναλογία θετικών/αρνητικών ετικετών και αποκλείοντας τα μη-αγγλικά tweet, που δεν ανήκαν στο CMU Pronouncing Dictionary. Κατά την εφαρμογή του MAD LP, το σύνολο δεδομένων τους περιείχε ελάχιστα μη αγγλόφωνα tweet, τα οποία διήλθαν από το παραπάνω φιλτράρισμα.

6.6.2 Ανεξαρτησία από Λεξικολογικές Πηγές

Οι Velikovich et al. (2010) [81] ασχολούνται κατ' αποκλειστικότητα με την πλήρη απεξάρτηση από προϋπάρχοντες λεξικολογικούς πόρους. Αν και ένα ιστογενές λεξικό κατασκευάζεται από ένα λεξικολογικό γράφο, που περιέχει θόρυβο, ο GP εμφανίζεται αρκετά ανθεκτικός σε αυτό το θόρυβο και ικανός να παράγει μεγάλα και ακριβή λεξικά πολικότητας. Ένα από τα πλεονέκτημα της απεξάρτησης από το WordNet είναι επίσης ότι επιτρέπει στα λεξικά να συμπεριλάβουν μη-συμβατικά λήμματα, ιδίως ορθογραφικά λάθη, παραλλαγές, αργκό, και εκφράσεις πολλαπλών λέξεων.

Το ιστογενές λεξικό, που παράγει ο GP, είναι περισσότερο από μια τάξη μεγέθους, μεγαλύτερο από τα αντίστοιχα λεξικά που κατασκευάζονται από την πλειονότητα των άλλων αλγορίθμων. Αυτό από μόνο του δεν θα ήταν σημαντικό επίτευγμα, αν οι επιπρόσθετες φράσεις ήταν κακής ποιότητας. Στην περίπτωσή του GP όμως, κάθε ένα

φρασίδιο (token) ορίζεται απλώς με διαστήματα και σημεία στίξης, με το σημείο στίξης να μετρά κι αυτό ως φρασίδιο. Επί το πλείστον, βλέπουμε ότι καθώς ο αριθμός των φρασιδίων αυξάνεται, ο αριθμός των αντίστοιχων φράσεων στο λεξικό μειώνεται. Οι μεγαλύτερες φράσεις είναι λιγότερο συχνές και συνεπώς, έχουμε λιγότερες και χαμηλότερου βάρους ακμές, στους συνδεδεμένους κόμβους του γράφου.

Οι Brody και Elhadad (2010) [16], αν και δεν έχουν ως πρωταρχικό μέλημα την απεξάρτηση από τα υπάρχοντα λεξικά συναισθήματος όπως ο GP, επειδή χρησιμοποιούν μια μέθοδο για την αυτόματη παραγωγή ενός μη επιβλεπόμενου συνόλου πυρήνων, από θετικά και αρνητικά επίθετα, για την ανίχνευση συναισθήματος, η μέθοδος τους αντικαθιστά επιτυχώς, τις χειρονακτικά δομημένες μεθόδους, που χρησιμοποιούνται συνήθως. Για αυτό το λόγο, το πρωταρχικό σύνολο δεδομένων στο οποίο εκτελείται ο AB LP, είναι το δημοσίως διαθέσιμο σώμα κειμένων, που χρησιμοποιήθηκε από τους Ganu et al. (2009) [36], το οποίο περιέχει πάνω από 50.000 κριτικές.

Η υψηλή επίδοση του AB LP επιβεβαίωσε ότι τα χαρακτηριστικά, που τεκμαίρονται από δεδομένα, είναι περισσότερο αντιπροσωπευτικά από αυτά που παράγονται χειρονακτικά. Τα επίθετα μπορούν να μεταφέρουν διαφορετικά συναισθήματα ανάλογα με το χαρακτηριστικό που εξετάζεται, όμως υπάρχουν και άλλα μέρη του λόγου που μπορούν να φανούν πολύ χρήσιμα για αυτόν τον σκοπό. Αφού οι online κριτικές ανήκουν σε ένα ανεπίσημο είδος γραφής, με πολλές φορές ευρηματική ορθογραφία και εξειδικευμένη ορολογία, είναι ανεπαρκές να στηριχθεί κανείς μόνο σε λεξικά, τόσο για το χαρακτηριστικό όσο και για το συναίσθημα, συνεπώς η χρήση μίας μεθόδου ανεξάρτητης, από δομημένες λεξιλογικές πηγές, κρίνεται καθοριστικής σημασίας.

Οι Speriosu et al. (2011) [71] δημιουργούν τον γράφο πυρήνων, χρησιμοποιώντας τις τιμές πολικότητας του λεξικού OpinionFinder, τη γνωστή πολικότητα των emoticon, και έναν ταξινομητή μέγιστης εντροπίας, εκπαιδευμένο πάνω σε 1,8 εκατομμύρια tweet με αυτομάτως εκχωρούμενες ετικέτες, βάσει της παρουσίας θετικών και αρνητικών emoticon. Συνεπώς δεν υπάρχει πλήρης εξάρτηση τόσο από το λεξικό, όσο από τις δύο άλλες πηγές πληροφορίας συναισθήματος. Παράλληλα, για την εκπαίδευση των ταξινομητών πολικότητας χρησιμοποίησαν τρία προϋπάρχοντα σχολιασμένα σύνολα δεδομένων, το Stanford Twitter Sentiment, το Obama-McCain Debate και το Health Care Reform. Ο MAD LP, όπως εκτελέστηκε σε μία καθαρά ημι-επιβλεπόμενη μέθοδο, ήταν αναμενόμενο να έχει μία σημαντική εξάρτηση από άλλες προγενέστερες συλλογές, όχι όμως σε βαθμό που η εφαρμογή του σε ένα περιβάλλον χωρίς αυτές να είναι αδύνατη. Σαφώς όμως χωρίς την ευχέρεια αλλαγή συνόλων δεδομένων των GP και AB LP.

Οι Rao και Ravichandran (2009) [67] από την άλλη, προτείνουν την εφαρμογή του CR LP, στα πλαίσια μιας, επίσης ημι-επιβλεπόμενης μεθόδου μάθησης για την επαγωγική δημιουργία λεξικών συναισθήματος, αποκλειστικά από WordNet ή άλλους

παρεμφερείς θησαυρούς, όπως του OpenOffice. Για την αξιολόγηση των επιδόσεών του αλγορίθμου χρησιμοποιήθηκαν τα δεδομένα του General Inquirer, και πιο συγκεκριμένα μία μόνο από τις συναισθηματικές του διαστάσεις, που δηλώνει σθένος. Συνεπώς, η εξάρτηση από αυτές τις πηγές, και μάλιστα από κάποιες ιδιαίτερα εξεζητημένες γλωσσολογικές ιδιότητες των λέξεων, είναι ζωτικής σημασίας για την απόδοση του CR LP, υστερώντας εμφανώς σε αυτόν τον τομέα, σε σύγκριση με τους προηγούμενους αλγορίθμους.

Η μέθοδος των Blair-Goldensohn et al.(2008) [10], παρόλο που ανήκει σε εκείνες της ανάλυσης συναισθήματος βασιζόμενης σε χαρακτηριστικά, όπως και ο AB LP, εκτελεί τον AO LP σε ένα σύνολο δεδομένων, αξιοποιώντας κατά κόρον στις ιδιότητες του WordNet. Η ίδια η μήτρα μετάβασης, που χρησιμοποιεί στην επαναληπτική του διαδικασία, κατασκευάζεται σύμφωνα με το αν ανήκει ένας κόμβος στο σύνολο των συνωνύμων ή αντωνύμων, καθιστώντας αδύνατη την εκτέλεσή του όχι μόνο σε ένα περιβάλλον ανεξάρτητο από λεξικολογικές πηγές αλλά και συγκεκριμένα από κάποια έκδοση του WordNet. Ο AO LP κρίνεται λοιπόν ως ο πλέον ακατάλληλος, για μεθόδους ανεξάρτητες από προϋπάρχοντες πόρους.

6.6.3 Ανεξαρτησία από το Γνωσιακό Τομέα

Οι παρόμοιες μελέτες, με αυτές των AB LP και AO LP, κάνουν ιδιαίτερα περιοριστικές υποθέσεις, θεωρώντας ότι δεν υπάρχει καμία εκ των προτέρων γνώση του γνωσιακού τομέα που συνοψίζεται, και ότι κάθε κριτική αποτελείται μόνο από το κείμενο της κριτικής. Στην πραγματικότητα, τα περισσότερα μηνύματα του Ιστού συνοδεύονται με τουλάχιστον κάποια επισήμανση, αφού συνήθως υπάρχει ένδειξη για το γενικό συναίσθημα του κειμένου, και συχνά υπάρχει επίσης κάποια πρότερη γνώση του συγκεκριμένου γνωσιακού τομέα.

Το μη επιβλεπόμενο σύστημα των Brody και Elhadad (2010) [16], για την εξαγωγή χαρακτηριστικών και τον προσδιορισμό συναισθήματος σε ένα κείμενο κριτικής, είναι το πιο ευπροσάρμοστο ως προς το γνωσιακό τομέα, και λαμβάνει υπόψη την επιρροή των χαρακτηριστικών, στην πολικότητα του συναισθήματος, το οποίο αποτελεί ένα ζήτημα που έχει αγνοηθεί σε μεγάλο βαθμό, στους άλλους αλγορίθμους. Η επιτυχημένη εφαρμογή του AB LP δείχνει την αποτελεσματικότητα του συστήματος και στους δύο ανωτέρω στόχους, όπου επιτυγχάνει παρόμοια αποτελέσματα με τις υπόλοιπες πολύπλοκες ημι-επιβλεπόμενες μεθόδους, οι οποίες έχουν τον περιορισμό ότι βασίζονται στον χειρονακτικό σχολιασμό και σε πηγές εκτεταμένης γνώσης.

Είναι ενδιαφέρον ακόμη ότι κατά την αξιολόγηση της μεθόδου του AB LP, οι βαθμολογίες των όρων αλλάζουν ανάλογα με το χαρακτηριστικό και συνδέονται στενά με συγκεκριμένα χαρακτηριστικά, και όχι με όλα. Αυτή η τάση μπορεί να εντοπισθεί αυτόματα με την παραπάνω μέθοδό και σε πολύ πιο λεπτομερές επίπεδο,

από αυτό που παρατηρεί ένας άνθρωπος, που αναλύει τα δεδομένα. Αυτό δείχνει ότι υπάρχει κάποια δυσκολία διάκρισης μεταξύ των λεπτομερών κατηγοριών κάθε γνωσιακού τομέα, αλλά υπάρχει υψηλή συμφωνία σε πιο χονδροειδές επίπεδο σε ποσοστό μεγαλύτερο του 90% των περιπτώσεων, ξεχωρίζοντας έτσι από τις άλλες μεθόδους σε αυτό τον τομέα.

Με παρόμοιο τρόπο, κεντρική θέση στο σύστημα των Blair-Goldensohn et al. (2008) [10] έχει η δυνατότητα αξιοποίησης διαφορετικών πηγών πληροφορίας, όπου αυτές είναι διαθέσιμες. Αφού το σύστημα εξάγει όλες τις προτάσεις ενός γνωσιακού τομέα, το επόμενο στάδιο είναι να ταξινομηθεί κάθε πρόταση ως θετική, αρνητική ή ουδέτερη σε κάποια αριθμητική κλίμακα, με την εκτέλεση του AO LP. Ειδικότερα, αποδεικνύεται ότι το συναίσθημα που παρέχεται σε επίπεδο εγγράφου από τον χρήστη, μπορεί να βοηθήσει στην πρόβλεψη του συναισθήματος σε επίπεδο φράσης ή πρότασης, μέσω μιας ποικιλίας μοντέλων. Επιπλέον, ο εξεταζόμενος γνωσιακός τομέας έχει συγκεκριμένα χαρακτηριστικά, που μπορούν να αξιοποιηθούν προκειμένου να βελτιωθούν τόσο η ποιότητα, όσο και το εύρος των παραγόμενων συνόψεων.

Οι Velikovich et al. (2010) [81] διαπιστώνοντας την αδυναμία του GP, να αξιοποιήσει στο έπακρο τη συναισθηματική πληροφορία που περιέχει κάθε γνωσιακός τομέας, και για τη βελτίωση της απόδοσης του αλγορίθμου, χρησιμοποίησαν έναν ταξινομητή συμφραζομένων, ο οποίος προβλέπει την πολικότητα μιας πρότασης, χρησιμοποιώντας χαρακτηριστικά γνωρίσματα της και των συμφραζομένων της. Στην πειραματική μέθοδο αυτός ήταν ένας ταξινομητής μέγιστης εντροπίας, εκπαιδευμένος και αξιολογημένος με τη χρήση διασταυρούμενης επικύρωσης 10 δειγμάτων (10-fold cross validation) των δεδομένων αξιολόγησης. Για κάθε πρόταση, ο ταξινομητής συμφραζομένων προέβλεπε μια θετική, αρνητική ή ουδέτερη ταξινόμηση με βάση την ετικέτα με την υψηλότερη πιθανότητα, αναπληρώνοντας, μερικώς, το υστέρημα του GP, σε σύγκριση με τους AO LP και AB LP.

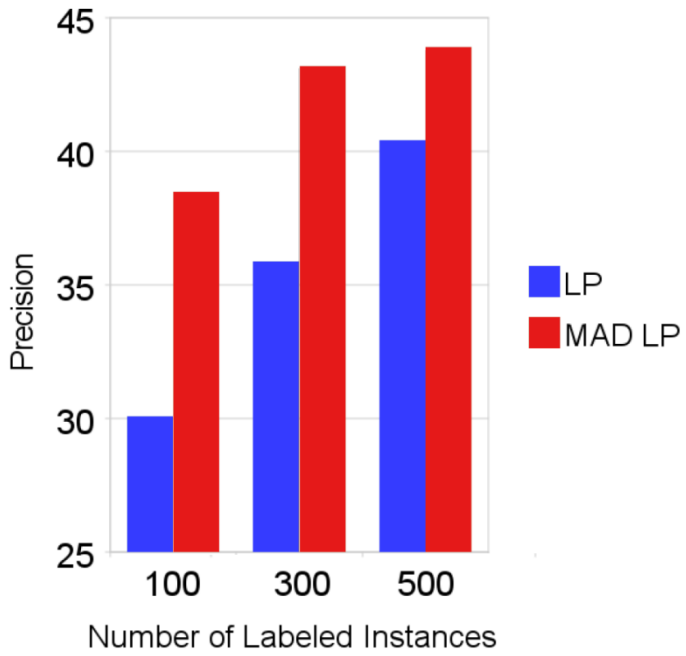
Ο CR LP ερευνά τις φυσικές πηγές των λέξεων του γράφου, όπως το WordNet, και εκμεταλλεύεται διάφορες σχέσεις μέσα σε αυτό, όπως η συνωνυμία και η υπερωνυμία. Η μέθοδός στην οποία εκτελείται δεν περιορίζεται μόνο στο WordNet, αλλά μπορεί να χρησιμοποιήσει οποιαδήποτε πηγή, που παραθέτει συνώνυμα. Συμπεραίνουμε λοιπόν, ότι σε μία τέτοια περίπτωση, απαιτείται είτε ένα εξειδικευμένο λεξικό που να μπορεί να αποδώσει τις ορθές νοηματικές συσχετίσεις, όπως αυτές εμφανίζονται σε κάθε συγκεκριμένο γνωσιακό τομέα, είτε ένα λεξικό γενικής χρήσης, με πολλαπλά χαρακτηριστικά για κάθε όρο. Οι παραπάνω απαιτήσεις καταδεικνύουν την αδυναμία του CR LP να ανταπεξέλθει σε ένα πολυτομεϊκό περιβάλλον, χωρίς πρότερη επεξεργασία του λεξιλογίου του.

Η αξιολόγηση της μεθόδου των Speriosu et al. (2011) [71] γίνεται σε διάφορα σύνολα δεδομένων, από tweet τα οποία έχουν σχολιασμένη την πολικότητα, όπως το Σύνολο Αισθήματος Twitter του Stanford, tweet, από το ντιμπέιτ του 2008 μεταξύ Ομπάμα

και McCain, καθώς και ένα νέο σύνολο δεδομένων από tweet που δημιούργησαν οι ίδιοι, σε σχέση με τη μεταρρύθμιση της υγειονομικής περίθαλψης, δηλαδή με μεγάλη δέσμευση από τον εκάστοτε γνωσιακό τομέα (πχ. πολιτική, υγεια). Η εκτέλεση του MAD LP σε ανομοιογενές νοηματικά περιβάλλον θα ήταν όντως προβληματική, αφού η μέθοδος του βασίζεται σε κάποιο βαθμό σε επιβλεπόμενα εκπαιδευμένους ταξινομητές, σε σύνολα δεδομένων προκαθορισμένου γνωσιακού τομέα.

6.6.4 Ανάλυση Κοινωνικών Δικτύων

Ο πλέον εξειδικευμένος από όλους τους παραπάνω αλγορίθμους πάνω σε κοινωνικά δίκτυα όπως το Twitter, παραμένει ο MAD LP. Η μέθοδος των Speriosu et al. (2011) [71] αξιοποιεί πλήρως όλα τα χαρακτηριστικά του Twitter για τη βελτίωση της ταξινόμηση της πολικότητας, με την παραδοχή ότι οι άνθρωποι επηρεάζουν ο ένας τον άλλο ή έχουν κοινές κλίσεις προς τα διάφορα θέματα. Ο γράφος που κατασκευάζεται, ως κόμβους έχει χρήστες, tweet, μονογράμματα και διγράμματα λέξεων, hashtag και emoticon. Με βάση τον παραπάνω follower graph, οι χρήστες συνδέονται με τα tweet που δημιούργησαν, και τα tweet συνδέονται με μονογράμματα, διγράμματα, hashtag και emoticon που περιέχουν.



Σχήμα 6.6: Σύγκριση ακρίβειας LP και MAD LP σε δεδομένα κοινωνικών δικτύων

Μια ακόμη ελκυστική ιδιότητα του MAD LP είναι ότι μπορεί να επιτύχει διανομές ετικέτας σε άλλους κόμβους, πέρα από τα tweet, και το σημαντικότερο, σε κόμβους

που δεν έχουν πυρήνα. Για παράδειγμα, όλοι οι κόμβοι χαρακτηριστικών (μονογραμμάτων, διγραμμάτων και hashtag) έχουν ένα φορτίο για τις θετικές και αρνητικές ετικέτες. Αυτές μπορούν να χρησιμοποιηθούν για διάφορες απεικονίσεις των αποτελεσμάτων της ταξινόμησης πολικότητας, συμπεριλαμβανομένων των όρων, που είναι οι πιο θετικοί και αρνητικοί, καθώς και για τον τονισμό τέτοιων όρων, όταν εμφανίζονται τα ατομικά tweet ενός χρήστη.

Σε όλα τα σύνολα δεδομένων και μέτρα, ο MAD LP ήταν σταθερά καλύτερος, ακόμη και σε συνθήκες έντονου θορύβου, από τις μεθόδους που βασίζονται στο συμβατικό LP. Η ημι-επιβλεπόμενη μέθοδος στην οποία εκτελείται ο αλγόριθμος συγκρίνεται ευνοϊκά με τις πλήρως επιβλεπόμενες προσεγγίσεις. Η τεχνική της προσρόφησης έχει επιτυχώς εφαρμοστεί και στο παρελθόν σε δεδομένα μεγάλης κλίμακας αλλά και για ανάλυση συναισθήματος στο YouTube, αποδίδοντας στην πολυεπίπεδη κατηγοριοποίηση, στον παραλληλισμό και την κλιμάκωση των ογκωδών δεδομένων, χαρακτηριστικό πολύ σημαντικό για μία ημι-επιβλεπόμενη μέθοδο [8].

Εκτός από τα στοιχεία των μέσων κοινωνικής δικτύωσης, που διαχειρίζεται κάθε αλγόριθμος, σημαντικό ρόλο παίζει και ο τρόπος με τον οποίο δομεί με αυτά, τον αντίστοιχο γράφο της πολικότητας τους. Ο CR LP, όπως και ο AO LP, ανήκουν σε μία οικογένεια επαναληπτικών αλγορίθμων LP, όπου κάθε κόμβος παίρνει το βεβαρυμμένο μέσο όρο των τιμών των γειτόνων του, από την προηγούμενη επανάληψη. Το αποτέλεσμα είναι ότι οι κόμβοι με πολλά μονοπάτια προς πυρήνες παίρνουν υψηλές πολικότητες λόγω της επιρροής από τους γείτονές τους.

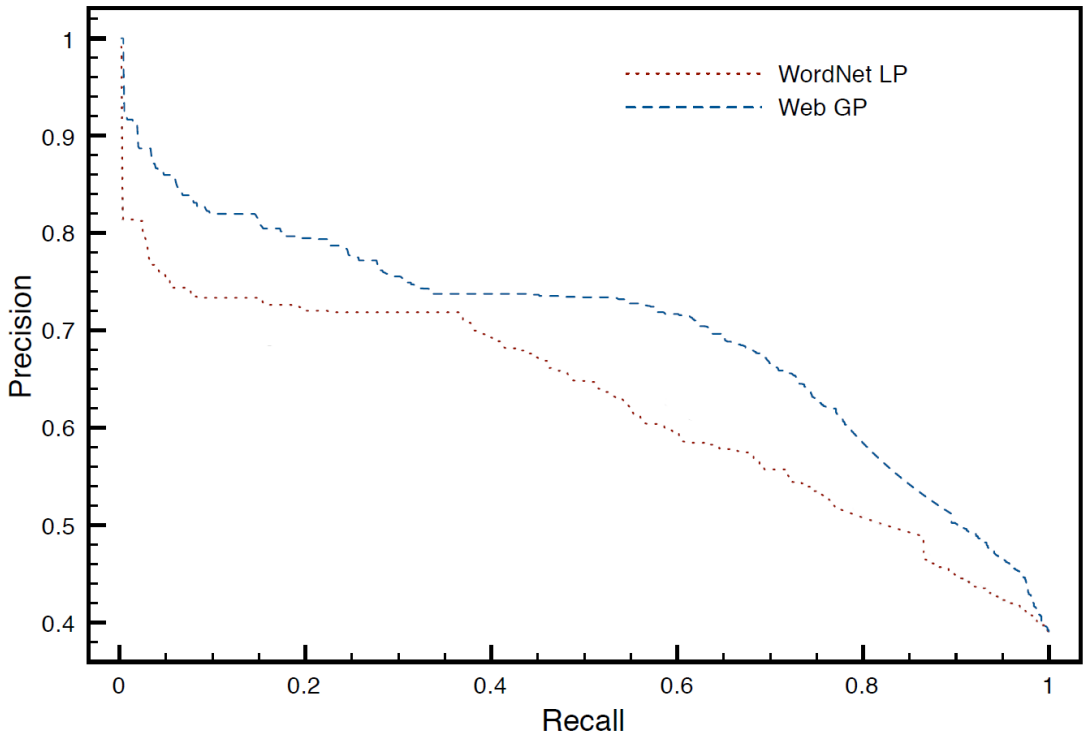
Η κύρια διαφορά μεταξύ των παραπάνω εκδοχών του LP και του GP είναι ότι ένας κόμβος με πολλαπλές διαδρομές προς ένα πυρήνα, θα επηρεάζεται από όλα αυτά τα μονοπάτια στον LP, ενώ μόνο η μία διαδρομή από έναν πυρήνα θα επηρεάζει την πολικότητα ενός κόμβου στον GP, δηλαδή η διαδρομή με το μεγαλύτερο βάρος. Αν ένας κόμβος έχει πολλαπλές διαδρομές για έναν πυρήνα, αυτό θα πρέπει να αντικατοπτρίζεται σε υψηλότερη βαθμολογία. Αυτό ισχύει ασφαλώς όταν ο γράφος είναι υψηλής ποιότητας και όλες οι διαδρομές αξιόπιστες. Ωστόσο, σε ένα γράφο κατασκευασμένο από στατιστικά συνεμφάνισης από κοινωνικά δίκτυα, αυτό συμβαίνει σπάνια.

Συν τοις άλλοις, ένας γράφος αποτελείται από πολλούς πυκνούς υπογράφους, με τον καθένα να αντιπροσωπεύει κάποια σημασιολογική κατηγορία σε σχέση με κάποιο θέμα που συζητείται στα κοινωνικά δίκτυα. Προβλήματα προκύπτουν, στον LP, όταν η πολικότητα εισρέει μέσα σ' αυτούς τους πυκνούς υπογράφους. Σε αυτή την περίπτωση, αυτή η ροή θα μεγεθυνόταν, αφού ο πυκνός υπογράφος θα παρείχε εκθετικά πολλές διαδρομές, από κάθε κόμβο προς την πηγή της ροής, πράγμα το οποίο θα προκαλούσε ένα φαινόμενο επανενίσχυσης. Ως αποτέλεσμα το λεξικό θα αποτελείτο από κατηγορίες μεγάλου πλήθους. Αυτό θα οδηγούσε επίσης σε προβλήματα σύγκλισης, αφού η πολικότητα θα διαιρείτο κατ' αναλογία με το μέγεθος του πυκνού υπογράφου. Επιπλέον, οι αρνητικές φράσεις, στο γράφο, βρίσκονται σε πυκνότερα συνδεδεμένες περιοχές, το οποίο έχει ως αποτέλεσμα τα τελικά λεξικά να

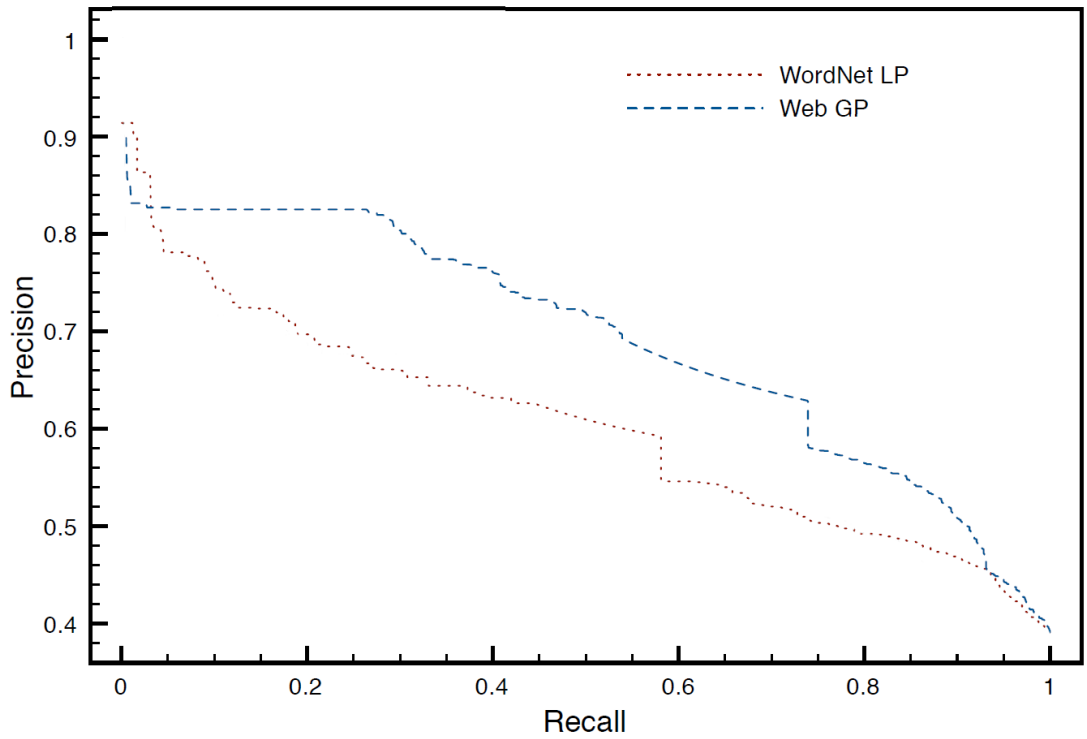
έχουν υψηλή κλίση προς τα αρνητικά λήμματα, λόγω της επιρροής των πολλαπλών διαδρομών προς τις λέξεις-πυρήνες.

Στον GP αυτά τα προβλήματα είναι λιγότερο έντονα, καθώς κάθε κόμβος σε ένα πυκνό υπογράφο παίρνει την πολικότητα μόνο μία φορά από κάθε πυρήνα, η οποία υποβαθμίζεται από το γεγονός ότι τα βάρη των ακμών είναι μικρότερα της μονάδας. Αυτή η παράμετρος έχει ως αποτελέσματα, οι περισσότερες επιμήκεις διαδρομές να έχουν βάρη κοντά στο μηδέν, το οποίο με τη σειρά του οδηγεί σε ταχεία σύγκλιση.

Εξετάζοντας την καταλληλότητα των LP αλγορίθμων στα κοινωνικά δίκτυα, και ειδικότερα στο Twitter, οι Brody και Diakopoulos (2012) [15] διαπίστωσαν ότι οι αλγόριθμοι AB LP και GP, αποδίδουν συγκριτικά καλύτερα, στις θετικές λέξεις και ο GP είναι εν τέλει, πιο ακριβής από τη μέθοδο των Brody και Elhadad (2010) [16]. Η διαφορά στην επίδοση μπορεί να εξηγηθεί από τις συσχετίσεις που χρησιμοποιούν οι αλγόριθμοι. Ο αλγόριθμος GP λαμβάνει υπόψη την ισχυρότερη διαδρομή για κάθε λέξη σπόρο, ενώ ο AB LP διαδίδεται από τον καθένα σπόρο, προς τους γείτονές του και περαιτέρω, γεγονός που τον καθιστά ευαίσθητο στις ισχυρές συσχετίσεις, ανάμεσα σε μία λέξη και έναν μεμονωμένο σπόρο. Επειδή ο γράφος στον AB LP κατασκευάζεται με ακμές συνεμφάνισης μεταξύ λέξεων, παρά με συντακτικές σχέσεις μεταξύ επιθέτων, εισάγονται θορυβώδεις ακμές, προκαλώντας εσφαλμένες συσχετίσεις. Ο αλγόριθμος GP, από την άλλη, βρίσκει λέξεις που έχουν ισχυρή σχέση με το θετικό ή αρνητικό σύνολο πυρήνων, σαν ολότητα, κάτι που τον καθιστά πιο εύρωστο.



Σχήμα 6.7: Σύγκριση Ακρίβειας/Ανάκλησης AO LP και GP σε θετικές κατηγορίες



Σχήμα 6.8: Σύγκριση Ακρίβειας/Ανάκλησης AO LP και GP σε αρνητικές κατηγορίες

6.6.5 Συνολική Σύγκριση Αλγορίθμων

Σύμφωνα με την παραπάνω συγκριτική μελέτη των επιμέρους χαρακτηριστικών των αλγορίθμων είναι πλέον εφικτή η συνολική τους σύγκριση και αποτίμηση της συνεισφοράς τους, στην ανάλυση συναισθήματος.

Ο AO LP παρουσιάζει έντονη εξάρτηση από το γλωσσικό περιβάλλον στο οποίο εκτελείται, μεγάλη εξάρτηση από προϋπάρχουσες λεξικολογικές πηγές και μειωμένη ικανότητα ανάλυσης δεδομένων που προέρχονται από τα κοινωνικά δίκτυα, γεγονός αναμενόμενο, αν ληφθεί υπόψη η γλωσσική και λεκτική ποικιλομορφία τους. Μεγάλη ανεξαρτησία, παρόλα αυτά, παρουσιάζει από το γνωσιακό τομέα, καθιστώντας τον, το δεύτερο καλύτερο σε απόδοση ανάλυσης δεδομένων διαφορετικών τομέων.

Με μικρές αποκλίσεις στη συνολική απόδοση, ακολουθεί ο CR LP, του οποίου η πολύ υψηλή γλωσσική ανεξαρτησία δεν μπορεί να αντισταθμίσει την αδυναμία του να λειτουργήσει αυτόνομα από λεξικολογικούς πόρους ή να προσαρμοστεί σε ποικίλα γνωσιακά πεδία. Πέραν τούτο, έκδηλη είναι και η αδυναμία ανάλυσης των δεδομένων των κοινωνικών δικτύων.

Σε ένα αξιόλογο επίπεδο απόδοσης κινείται ο AB LP, ο οποίος εκμεταλλευόμενος την μέθοδο ανάλυσης χαρακτηριστικών, καταφέρνει να επιβεβαιώσει την δυνατότητά του να ελίσσεται μεταξύ των γνωσιακών περιβαλλόντων, χωρίς να έχει απώλειες στην απόδοση του. Η μέθοδός του προσφέρει συνεπώς, γλωσσική απεξάρτηση και αποδεικνύεται αρκούντως αξιόπιστη στην διαχείριση δεδομένων από κοινωνικά δίκτυα.

Σε εκείνους τους αλγόριθμους, μεγάλων δυνατοτήτων, οι οποίοι παρουσιάζουν ανοχή στις μεταβολές των γλωσσικών και λεξικολογικών χαρακτηριστικών, δίκαια εντάσσεται ο GP. Τα ιδιαίτερα χαρακτηριστικά του και η φιλοσοφία με την οποία αναπαριστά τα δεδομένα του σε γράφους, του επιτρέπουν να αναλύει αποτελεσματικά τα δεδομένα κοινωνικών δικτύων, διαφορετικών γνωσιακών τομέων, μέσα σε γραφικούς χώρους πυκνού θορύβου. Συνολικά, αποτελεί τον ιδανικότερο αλγόριθμο διάδοσης ετικέτας, για πλήρως μη επιβλεπόμενες μεθόδους ανάλυσης συναισθήματος.

Παρόλο που ο MAD LP εμφανώς χωλαίνει σε θέματα γλωσσικής και γνωσιακής ανεξαρτησίας, η μεγάλη του ικανότητα ανάλυσης της πολικότητας στα δεδομένα κοινωνικών δικτύων μπορεί και εξισορροπεί αυτές του τις ελλείψεις. Ξεχωρίζει από κάθε άλλο αλγόριθμο διάδοσης ετικέτας, αφού κάθε κόμβος του γράφου του μπορεί να αντιπροσωπεύσει δομικές μονάδες των κοινωνικών δικτύων, και όχι απλά μεμονωμένες λέξεις. Σε συνδυασμό με την ανεξαρτησία από προϋπάρχοντα λεξικά συναισθήματος, τον καθιστά τον καταλληλότερο για μεθόδους αυτοματοποιημένης ανάλυσης στα κοινωνικά δίκτυα, ανταγωνιζόμενος επάξια τον GP σε απόδοση.

7. Σύνοψη και Συμπεράσματα

Βασικός στόχος της εργασίας αποτέλεσε η μελέτη των ημι-επιβλεπόμενων και μη επιβλεπόμενων μεθόδων ανάλυσης συναισθήματος στα κοινωνικά δίκτυα, μέσα από την εξαντλητική βιβλιογραφική ανασκόπηση των σημαντικότερων και ευρέως αποδεκτών εργασιών, των οποίων οι μέθοδοι επανειλημμένως έχουν χρησιμοποιηθεί και επιβεβαιωθεί πειραματικά, ως προς την απόδοσή τους στον τομέα. Αρχικά, επιχειρήθηκε μια περιγραφή και πολύπλευρη ταξινόμηση, τόσο των μεθόδων Μηχανικής Μάθησης, όσο και των μεθόδων χρήσης Λεξικών, ή στατιστικών κριτηρίων ανάλυσης.

Η παραπάνω διαδικασία αβίαστα οδήγησε στην εμβάθυνση σε προσεγγίσεις και αλγορίθμους, που είτε συνθέτουν, είτε αντλούν τα πορίσματά τους μέσα από τη χρήση Λεξικών Συναισθήματος. Σε αυτό το πλαίσιο πραγματοποιήθηκε μία πλήρης παράθεση σχεδόν όλων των διαθέσιμων, δομημένων, εμπορικά αξιοποιήσιμων ή μη, λεξικών συναισθήματος, κυρίως της αγγλικής γλώσσας, περιλαμβάνοντας και μία μεγάλη γκάμα χαρακτηριστικών, που συνθέτουν την συνολική ταξινόμηση τους, αλλά και μία αναφορά στις προσπάθειες σύνθεσης λεξικών συναισθήματος της ελληνικής γλώσσας. Τα αγγλόγλωσσα λεξικά χαίρουν εκτεταμένης χρήσης στην διεθνή βιβλιογραφία, στο πεδίο της μη επιβλεπόμενης συναισθηματικής ανάλυσης, γεγονός που συντέλεσε στην ανάδειξη πληθώρας ερευνητικών προσπαθειών στο χώρο της αυτοματοποιημένης κατηγοριοποίησης πολικότητας, με χρήση αλγορίθμων που αναπαριστούν τα δεδομένα τους σε δομές γράφων.

Μείζονα θέση ανάμεσα τους αποδείχθηκε ότι έχει η τεχνική της Διάδοσης Ετικέτας, η οποία συνίσταται στην αξιοποίηση περιορισμένου εύρους προσημασμένων δεδομένων, για την επαναληπτική διεύρυνση του συνόλου όρων εγνωσμένης υποκειμενικότητας. Η παραπάνω ιδιότητα κρίθηκε ως πρωτεύουσας σημασίας, δεδομένου ότι στην ανάλυση των κοινωνικών δικτύων καλούμαστε να διαχειριστούμε δεδομένα μεγάλου όγκου (big data). Ως χαρακτηριστικό μέσο κοινωνικής δικτύωσης, επιλέχθηκε το Twitter, το οποίο βιβλιογραφικά επιβεβαιώθηκε, ως το συνηθέστερο μέσο, στο οποίο πραγματοποιείται πειραματική έρευνα στον αλγόριθμό Διάδοσης Ετικέτας.

Η αναζήτηση των καταλληλότερων παραμέτρων του αλγορίθμου και των τροποποιήσεων του, ώστε να είναι συμβατός στις απαιτήσεις των κοινωνικών δικτύων, ανέδειξε αφενός την αποδοτικότητα εκείνων που εντάσσονταν σε μεθόδους ανάλυσης βασιζόμενες σε χαρακτηριστικά και αφετέρου εκείνων, οι οποίες αποδείχθηκαν ακριβέστερες στην ανάλυσή τους και ανθεκτικότερες στα κριτήρια αξιολόγησης (γλωσσική, γνωσιακή, λεξικολογική ανεξαρτησία), που τέθηκαν στα πλαίσια της έρευνας. Στην πρώτη κατηγορία εντάσσονται οι αλγόριθμοι Διάδοσης Ετικέτας που βασίζονται σε χαρακτηριστικά (AB LP) ή προσανατολίζονται σε αυτά (AO LP), ενώ στη δεύτερη, οι αλγόριθμοι Διάδοσης Ετικέτας Τροποποιημένης Προσρόφησης (MAD LP) και Διάδοσης Γράφων (GP).

Αλγόριθμος Διάδοσης Ετικέτας	Γλωσσική Ανεξαρτησία	Ανεξαρτησία Λεξικολογικών Πηγών	Ανεξαρτησία Γνωσιακού Τομέα	Ικανότητα Ανάλυσης Κοινωνικών Δικτύων
MAD LP	**ΧΑΜΗΛΗ**	**ΜΕΤΡΙΑ**	**ΧΑΜΗΛΗ**	**ΥΨΗΛΗ**
GP	**ΜΕΤΡΙΑ**	**ΥΨΗΛΗ**	**ΜΕΤΡΙΑ**	**ΜΕΤΡΙΑ**
AB LP	**ΜΕΤΡΙΑ**	**ΜΕΤΡΙΑ**	**ΥΨΗΛΗ**	**ΜΕΤΡΙΑ**
CR LP	**ΥΨΗΛΗ**	**ΧΑΜΗΛΗ**	**ΧΑΜΗΛΗ**	**ΧΑΜΗΛΗ**
AO LP	**ΧΑΜΗΛΗ**	**ΧΑΜΗΛΗ**	**ΥΨΗΛΗ**	**ΧΑΜΗΛΗ**

Πίνακας 6.4: Συγκριτική απεικόνιση Αλγορίθμων Διάδοσης Ετικέτας

Από τη συγκριτική τους μελέτη διαπιστώθηκε ότι ασθενέστερος όλων υπήρξε ο AO LP, με τη δυναμική του να οφείλεται καθαρά στη συνολική μέθοδο που εκτελείται, και όχι σε κάποια ιδιαίτερα κατασκευαστικά στοιχεία του. Παράλληλα, στον τομέα της γλωσσικής πολυμορφίας αποδείχθηκε ότι υπερτερεί ο CR LP, στην ανεξαρτησία από λεξικολογικές πηγές ο GP, στην ικανότητα διαχείρισης δεδομένων διαφορετικών γνωσιακών τομέων ο AB LP και στην αποτελεσματικότητα καθορισμού της πολικότητας σε δεδομένα μέσων κοινωνικής δικτύωσης ο MAD LP. Τα παραπάνω συμπεράσματα υποδηλώνουν έντονες προοπτικές για την περαιτέρω πειραματική μελέτη τους, τόσο για την δόμηση λεξικών συναισθήματος, όσο και για την κατηγοριοποίηση πολικότητας, σε δεδομένα κοινωνικών δικτύων της ελληνικής γλώσσας.

Βιβλιογραφία

[1] Akkaya, C., Wiebe, J., & Mihalcea, R. (2009, August). Subjectivity word sense disambiguation. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1 (pp. 190-199). Association for Computational Linguistics.

[2] Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, Tools and Applications for Sentiment Analysis Implementation. International Journal of Computer Applications, 125(3).

[3] Alison Huettner and Pero Subasic. Fuzzy typing for document management. In ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes, pages 26–27, 2000.

[4] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175-185.

[5] Andreevskaia, A., & Bergler, S. (2006, April). Mining WordNet for a Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. In EACL (Vol. 6, pp. 209-216).

[6] Asmussen, S. (2008). Applied probability and queues (Vol. 51). Springer Science & Business Media.

[7] Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 2200-2204).

[8] Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S., ... & Aly, M. (2008, April). Video suggestion and discovery for youtube: taking random walks through the view graph. In Proceedings of the 17th international conference on World Wide Web (pp. 895-904). ACM.

[9] Bishop, C. M. (2006). Pattern recognition. Machine Learning, 128.

[10] Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., & Reynar, J. (2008, April). Building a sentiment summarizer for local service reviews. In WWW workshop on NLP in the information explosion era (Vol. 14, pp. 339-348).

[11] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

[12] Blum, A., & Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts.

[13] Bousquet, O., & Bottou, L. (2008). The tradeoffs of large scale learning. In Advances in neural information processing systems (pp. 161-168).

[14] Breck, E., Choi, Y., & Cardie, C. (2007, January). Identifying Expressions of Opinion in Context. In IJCAI (Vol. 7, pp. 2683-2688).

[15] Brody, S., & Diakopoulos, N. (2011, July). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In Proceedings of the conference on empirical methods in natural language processing (pp. 562-570). Association for Computational Linguistics.

[16] Brody, S., & Elhadad, N. (2010, June). An unsupervised aspect-sentiment model for online reviews. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 804-812). Association for Computational Linguistics.

[17] Cambria, E., Schuller, B., Liu, B., Wang, H., & Havasi, C. (2013). Statistical approaches to concept-level sentiment analysis. IEEE Intelligent Systems, 3(28), 6-9.

[18] Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., & Gandini, G. (2007). Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Language resources and linguistic theory: Typology, second language acquisition, English linguistics, 200-210.

[19] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15.

[20] Choi, Y., & Cardie, C. (2008, October). Learning with compositional semantics as structural inference for subsentential sentiment analysis. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 793-801). Association for Computational Linguistics.

[21] Choi, Y., & Cardie, C. (2009, August). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2 (pp. 590-598). Association for Computational Linguistics.

[22] Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin, 70(4), 213.

[23] Das, A., & Bandyopadhyay, S. (2010, November). Towards the Global SentiWordNet. In PACLIC (pp. 799-808).

[24] Ding, X., Liu, B., & Yu, P. S. (2008, February). A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 international conference on web search and data mining (pp. 231-240). ACM.

[25] Dragut, E. C., Yu, C., Sistla, P., & Meng, W. (2010, October). Construction of a sentimental word dictionary. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 1761-1764). ACM.

[26] Du, W., Tan, S., Cheng, X., & Yun, X. (2010, February). Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. In Proceedings of the third ACM international conference on Web search and data mining (pp. 111-120). ACM.

[27] Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22, 457-479.
[28] Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. ACM SIGKDD explorations newsletter, 4(1), 65-75.
[29] Esuli, A., & Sebastiani, F. (2005, October). Determining the semantic orientation of terms through gloss classification. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 617-624). ACM.
[30] Esuli, A., & Sebastiani, F. (2007). SENTIWORDNET: A high-coverage lexical resource for opinion mining. Technical Report 2007-TR-02, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT.
[31] Fahrni, A., & Klenner, M. (2008, April). Old wine or warm beer: Target-specific sentiment analysis of adjectives. In Proc. of the Symposium on Affective Language in Human and Machine, AISB (pp. 60-63).
[32] Feng, S., Bose, R., & Choi, Y. (2011, July). Learning general connotation of words using graph-based algorithms. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1092-1103). Association for Computational Linguistics.
[33] Freedman, D. A. (2009). Statistical models: theory and practice. cambridge university press.
[34] Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005, September). Pulse: Mining customer opinions from free text. In international symposium on intelligent data analysis (pp. 121-132). Springer Berlin Heidelberg.
[35] Ganapathibhotla, M., & Liu, B. (2008, August). Mining opinions in comparative sentences. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 (pp. 241-248). Association for Computational Linguistics.
[36] Ganu, G., Elhadad, N., & Marian, A. (2009, June). Beyond the Stars: Improving Rating Predictions using Review Text Content. In WebDB (Vol. 9, pp. 1-6).
[37] Gärtner, T., Le, G. Q. V., & Smola, A. J. (2006). 1 A Short Tour of Kernel Methods for Graphs.
[38] Hassan, A., Qazvinian, V., & Radev, D. (2010, October). What's with the attitude?: identifying sentences with attitude in online discussions. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 1245-1255). Association for Computational Linguistics.
[39] Hatzivassiloglou, V., & McKeown, K. R. (1997, July). Predicting the semantic orientation of adjectives. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (pp. 174-181). Association for Computational Linguistics.

[40] Hu, M., & Liu, B. (2004, July). Mining opinion features in customer reviews. In*AAAI* (Vol. 4, No. 4, pp. 755-760).

[41] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

[42] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. Machine learning, 37(2), 183-233.ISO 690.

[43] Kaji, N., & Kitsuregawa, M. (2006, July). Automatic construction of polarity-tagged corpus from HTML documents. In Proceedings of the COLING/ACL on Main conference poster sessions (pp. 452-459). Association for Computational Linguistics.

[44] Kamps, J., Marx, M. J., Mokken, R. J., & Rijke, M. D. (2004). Using wordnet to measure semantic orientations of adjectives.

[45] Kanayama, H., & Nasukawa, T. (2006, July). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 355-363). Association for Computational Linguistics.

[46] Kim, S. M., & Hovy, E. (2004, August). Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics (p. 1367). Association for Computational Linguistics.

[47] Kim, S. M., & Hovy, E. (2006, June). Identifying and analyzing judgment opinions. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (pp. 200-207). Association for Computational Linguistics.

[48] Kim, S. M., & Hovy, E. (2006, July). Automatic identification of pro and con reasons in online reviews. In Proceedings of the COLING/ACL on Main conference poster sessions (pp. 483-490). Association for Computational Linguistics.

[49] Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical society, 7(1), 48-50. Simonyi, G. (1995). Graph entropy: a survey. Combinatorial Optimization, 20, 399-441.

[50] Kumar, U. D., Crocker, J., Chitra, T., & Saranga, H. (2006). Reliability and six sigma. Springer Science & Business Media.

[51] Lu, Y., Castellanos, M., Dayal, U., & Zhai, C. (2011, March). Automatic construction of a context-aware sentiment lexicon: an optimization approach. In Proceedings of the 20th international conference on World wide web (pp. 347-356). ACM.

[52] McDonald, R., Hannan, K., Neylon, T., Wells, M., & Reynar, J. (2007, June). Structured models for fine-to-coarse sentiment analysis. In Annual Meeting-Association For Computational Linguistics (Vol. 45, No. 1, p. 432).

[53] Mejova, Y. (2009). Sentiment analysis: an overview. Comprehensive exam paper, available on http://www. cs. uiowa. edu/~ ymejova/publications/CompsYelenaMejova.pdf [2010-02-03].
[54] Mohammad, S., Dunne, C., & Dorr, B. (2009, August). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2 (pp. 599-608). Association for Computational Linguistics.
[55] Nadaraya, E. A. (1964). On estimating regression. Theory of Probability & Its Applications, 9(1), 141-142.
[56] Namenwirth, J. Z., & Weber, R. P. (1987). Dynamics of culture. Allen & Unwin.
[57] Ng, A. Y., Zheng, A. X., & Jordan, M. I. (2001, August). Link analysis, eigenvectors and stability. In International Joint Conference on Artificial Intelligence (Vol. 17, No. 1, pp. 903-910). LAWRENCE ERLBAUM ASSOCIATES LTD.
[58] O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. ICWSM, 11(122-129), 1-2.
[59] Ortony, A., Clore, G., & Collins, A. (1988). The Cognitive Structure of Emotions: Cambridge Uni. Press, New York.
[60] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.
[61] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.
[62] Pavlopoulos, I.(2014). Aspect based sentiment analysis.
[63] Pearl, J. (1984). Heuristics: intelligent search strategies for computer problem solving.
[64] Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic Inquiry and Word Count (LIWC): A computerized text analysis program. Mahwah (NJ), 7.
[65] Peng, W., & Park, D. H. (2004). Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. Urbana, 51, 61801.
[66] Powers, D. M. (2012, April). The problem with kappa. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 345-355). Association for Computational Linguistics.
[67] Rao, D., & Ravichandran, D. (2009, March). Semi-supervised polarity lexicon induction. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 675-682). Association for Computational Linguistics.

[68] Rasmussen, C. E. (2006). Gaussian processes for machine learning.
[69] Read, J. (2005, June). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL student research workshop (pp. 43-48). Association for Computational Linguistics.
[70] Seeger, M. (2000). Learning with labeled and unlabeled data (No. EPFL-REPORT-161327).Sindhwani, V., & Melville, P. (2008, December). Document-word co-regularization for semi-supervised sentiment analysis. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 1025-1030). IEEE.
[71] Speriosu, M., Sudan, N., Upadhyay, S., & Baldridge, J. (2011, July). Twitter polarity classification with label propagation over lexical links and the follower graph. In Proceedings of the First workshop on Unsupervised Learning in NLP (pp. 53-63). Association for Computational Linguistics
[72] Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The General Inquirer: A Computer Approach to Content Analysis.
[73] Strapparava, C., & Valitutti, A. (2004, May). WordNet Affect: an Affective Extension of WordNet. In LREC (Vol. 4, pp. 1083-1086).
[74] Takala, P., Malo, P., Sinha, A., & Ahlgren, O. (2014). Gold-standard for Topic-specific Sentiment Analysis of Economic Texts. In LREC (Vol. 2014, pp. 2152-2157).
[75] Talukdar, P. P., & Crammer, K. (2009). New regularized algorithms for transductive learning. In Machine Learning and Knowledge Discovery in Databases (pp. 442-457). Springer Berlin Heidelberg.
[76] Tsakalidis, A., Papadopoulos, S., & Kompatsiaris, I. (2014, October). An ensemble model for cross-domain polarity classification on twitter. In International Conference on Web Information Systems Engineering (pp. 168-177). Springer International Publishing.
[77] Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. Data Mining and Knowledge Discovery, 24(3), 478-514.
[78] Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics.
[79] Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS), 21(4), 315-346.
[80] Varga, R. S. (1959). Factorization and normalized iterative methods (No. WAPD-T-950). Westinghouse Electric Corp. Bettis Plant, Pittsburgh.
[81] Velikovich, L., Blair-Goldensohn, S., Hannan, K., & McDonald, R. (2010, June). The viability of web-derived polarity lexicons. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 777-785). Association for Computational Linguistics.

[82] Wiebe, J., & Mihalcea, R. (2006, July). Word sense and subjectivity. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 1065-1072). Association for Computational Linguistics.

[83] Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 347-354). Association for Computational Linguistics.

[84] Wu, Y., & Wen, M. (2010, August). Disambiguating dynamic sentiment ambiguous adjectives. In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 1191-1199). Association for Computational Linguistics.

[85] Zhu, X., & Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University.