Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Σηματων Ελεγχου και Ρομποτικης

Συναισθηματική Ανάλυση προτάσεων με χρήση τεχνικών Θεματικής Μοντελοποίησης

Διπλωματικη Εργασια

της

Ευσταθίας Χριστοπούλου

Επιβλέπων: Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εργαστηριο Ορασης Υπολογιστων, Επεξεργασιας Λογου και Επεξεργασιας Σηματων
Αθήνα, Ιούλιος 2016

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων Ελέγχου και Ρομποτικής

Συναισθηματική Ανάλυση προτάσεων με χρήση τεχνικών Θεματικής Μοντελοποίησης

Διπλωματικη Εργασια

της

Ευσταθίας Χριστοπούλου

Επιβλέπων: Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την επιτροπή στις 13η Ιουλίου 2016.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Αλέξανδρος Ποταμιάνος	Πέτρος Μαραγκός	Αναστασία Κριθαρά
Αναπληρωτής Καθηγητής	Καθηγητής	Επιστ. Συνεργάτης
Ε.Μ.Π.	Ε.Μ.Π.	Ε.Κ.Ε.Φ.Ε. 'Δημόκριτος'

Αθήνα, Ιούλιος 2016

.....

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

Copyright ©–All rights reserved Ευσταθία Χριστοπούλου, 2016.
Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ΄ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν το συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Ευχαριστίες

Η Διπλωματική αυτή εργασία, σηματοδοτεί το τέλος της φοιτητικής μου ζωής στη σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών στο Εθνικό Μετσόβιο Πολυτεχνείο.

Πρώτα από όλα, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου Αλέξανδρο Ποταμιάνο που με εισήγαγε στο πεδίο της Μηχανικής Μάθησης και της Επεξεργασίας Φυσικής Γλώσσας μέσα από το μάθημα της Αναγνώρισης Προτύπων. Η προσέγγισή του έπαιξε σημαντικό ρόλο τόσο στην απόφασή μου να στραφώ σε εκείνον, κατά την αναζήτησή μου για θέμα διπλωματικής εργασίας, όσο και στο να συνεχίσω τις σπουδές μου σε αυτόν τον τομέα. Δεν μπορώ να μην τον ευχαριστήσω για την εμπιστοσύνη που έδειξε, τη γνώση, την καθοδήγηση και τις νέες ευκαιρίες που μου προσέφερε για ακαδημαϊκή μου καριέρα, σε συνδυασμό με την ενθάρρυνση που ποτέ δεν έπαψε να μου δείνει για να στοχεύω υψηλότερα μέσα σε όλη αυτή τη χρονιά.

Επιπλέον, θα ήθελα να ευχαριστήσω τον Δρ. Γιώργο Παλιούρα και την Δρ. Αναστασία Κριθαρά από το Ε.Κ.Ε.Φ.Ε. 'Δημόκριτος' για την υποστήριξη και τη συμβολή τους σε ένα μεγάλο μέρος της παρούσας διπλωματικής εργασίας, καθώς και για τις εποικοδομητικές συζητήσεις και ιδέες που μοιραστήκαμε. Δεν μπορώ να παραβλέψω τη βοήθεια και την υποστήριξη όλων των συναδέλφων μου, μέλη της ομάδας του Ε.Μ.Π. και του Πολυτεχνείου Κρήτης, ιδιαίτερα τον Δρ. Ηλία Ιωσήφ και Ελισάβετ Παλογιαννίδη με τους οποίους άρχισα να δουλεύω από την αρχή, συμπεριλαμβανομένων των Σάσα Κολοβού και Φίλιππο Κόκκινο που ήταν μέλη της ομάδας 'Tweester' που συμμετείχε στο διαγωνισμό SemEval-2016.

Τέλος, θα ήθελα να εκφράσω την αγάπη και την ευγνωμοσύνη μου σε όλους τους φίλους και την οικογένειά μου, που έμειναν στο πλευρό μου κατά τη διάρκεια αυτών των ετών, πίστευαν σε μένα, με βοήθησαν να ξεπεράσω τις δυσκολίες που ήρθαν σε και να γίνω το άτομο που είμαι σήμερα. Θα ήθελα να ευχαριστήσω τον Χρήστο για την υπομονή και την υποστήριξή του όλο αυτό το διάστημα.

Η Διπλωματική αυτή εργασία αφιερώνεται στον παππού μου Σταύρο και τη γιαγιά μου Περσεφόνη.

Ευσταθία Χριστοπούλου
13 Ιουλίου, 2016

Περίληψη

Η συναισθηματική ανάλυση κειμένου, είναι ένας κλάδος της Μηχανικής Μάθησης που αποσκοπεί στην αυτόματη αναγνώριση υποκειμενικής πληροφορίας σε γραπτές πηγές. Λόγω της ανάπτυξης του διαδικτύου και συνεπώς του όγκου πληροφοριών που διανέμονται στα κοινωνικά δίκτυα, πολλαπλές εφαρμογές χρησιμοποιούν αυτή την τεχνολογία για να βελτιόσουν την λήψη των αποφάσεών τους.

Σκοπός αυτής της Διπλωματικής εργασίας είναι η διερεύνηση μιας μεθόδου προσαμοργής για την βελτίωση της αυτόματης, δυαδικής αναγνώρισης συναισθήματος σε προτάσεις (θετική/αρνητική), με τη βοήθεια τεχνικών Θεματικής Μοντελοποίησης. Σε σύγκριση με τις κλασσικές μεθόδους που χρησιμοποιούν μια ποκιλια λεξιλογικών χαρακτηριστικών και λεξικών σαν είσοδο σε ταξινομητές, τα θεματικά μοντέλα έχουν την ικανότητα να συλλαμβάνουν το νοηματικό πλαίσιο στο οποίο ανήκει μια πρόταση και έτσι να εκτιμούν τη σημασία της με μεγαλύτερη λεπτομέρεια.

Το πρώτο μέρος της Διπλωματικής εργασίας, ασχολείται με τις τεχνικές Θεματικής μοντελοποίησης οι οποίες διευκολύνουν την εξαγωγή πληροφορίας από αδόμητες συλλογές κειμένων. Περιγράφεται η Latent Dirichlet Allocation (LDA) τεχνική, που προσπαθεί να ανακατασκευάσει την διαδικασία συγγραφής ενός κειμένου και συνεπώς να παράγει τις θεματικές ενότητες από τις οποίες αποτελείται. Στο δεύτερο μέρος, τα μοντέλα διανυσματικού χώρου διερευνόνται, τα οποία στοχεύουν στο να εκφράσουν τη σημασιολογική ομοιότητα μεταξύ λέξεων, κωδικοποιώντας τη γλώσσα ως μια μαθηματική κατανομή και συνεπώς σχηματίζοντας ένα σημασιολογικό χώρο. Επιπλέον, παρουσιάζονται μοντέλα στο χώρο των συναισθημάτων, τα οποία κωδικοποιούν το συναίσθημα των λέξεων, και συγκεκριμένα πώς είναι εφικτό το πέρασμα από το σημασιολογικό στον συναισθηματικό χώρο μέσω ενός μοντέλου αντιστοίχισης, χρησιμοποιώντας ένα λεξικό, ώστε να εκτιμηθεί το συναίσθημα νέων λέξεων.

Ο αλγόριθμος που παρουσιάζεται επικεντρώνεται κυρίως στην προσαρμογή του Σημασιολογικού χώρου, και ονομάζεται Σημασιολογικό Μοντέλο Προσαρμογής (ΣΜΠ). Στην προσέγγιση αυτή, ο σημασιολογικός χώρος μιας πρότασης αναπαρίσταται ως ένα ζυγισμένο μείγμα από διαφορετικά θεματικά-σημασιολογικά μοντέλα με βάση την εκτίμηση ενός εκπαιδευμένου πιθανοτικού θεματικού μοντέλου. Η τελική εκτίμηση του συναισθήματος μια πρότασης είναι το αποτέλεσμα ενός σημασιολογικού-συναισθηματικού μοντέλου που ενώνει τον σημασιολογικό και συναισθηματικό χώρο μέσω μιας αντιστοίχισης και ενός υπάρχοντος συναισθηματικού λεξικού.

Η απόδοση του ΣΜΠ μοντέλου μπορεί να εκτιμηθεί τόσο σε εφαρμογές σημασιολογικής

ομοιότητας λέξεων αλλά και σε αναγνώριση συναισθήματος προτάσεων, δείχνοτας μια βελτίωση στη μέτρηση της συσχέτισης με τιμές ομοιότητας δοσμένες από ανθρώπους, παράλληλα με μια αύξηση της ακρίβειας ταξινόμησης για προτάσεις, σε δεδομένα γενικού περιεχομένου αλλά και από το Twitter σε σύγκριση με ένα σύστημα που δεν χρησιμοποιεί θεματικές ενότητες.

Η αντίστοιχη εργασία μπορεί να βρεθεί στα αγγλικά στο σύνδεσμο https://bitbucket.org/fenchri/diploma-thesis/raw/22ac4e37fb6d918be1c766a1e37c824dbdf30769/Sentence-level%20Sentiment%20Analysis%20using%20Topic%20Modeling.pdf.

Λέξεις Κλειδιά

συναισθηματική αναλύση προτάσεων, θεματική μοντελοποίηση, LDA, σημασιολογικά μοντέλα, word2vec, συναισθηματικά μοντέλα, προσαρμογή

Abstract

Affective Text Analysis, is a field of Machine Learning that aims to automatically detect subjective information in textual data. Due to the development of the World Wide Web and consequently the amount of opinions distributed in social media, numerous applications use this technology to improve their decision making.

The scope of this Diploma Thesis is to explore an adaptation algorithm for improving the automatic binary sentiment classification of sentences in text (positive/negative), with the help of Topic modeling techniques. Compared to classic approaches that use a variety of lexical features and existing annotated lexicons as input to classifiers, topic models are able to capture the context of a sentence and thus estimate its meaning in greater detail.

The first part of the thesis, is devoted to Topic Modeling techniques that facilitate the extraction of meaningful information from unstructured collections of documents. The Latent Dirichlet Allocation (LDA) method is described which tries to reconstruct the process of writing a document and generate the topics that it consists of. In the second part, different Vector Space Models are investigated, which aim to express how similar a word is to another, by representing the language as a distribution, forming a semantic space. Additionally, Affective Space Models, that measure the sentiment of words are explored, and in particular, how one can create a mapping from the semantic to the affective space with the help of a small lexicon in order to estimate the sentiment of new words.

The presented algorithm focuses on the adaptation of the Semantic space, named Semantic Model Adaptation (SMA). In the SMA approach, the semantic space of each sentence is represented as a weighted mixture of different topic-semantic models according to a trained probabilistic topic model. The final estimation of a sentence's score is the result of a semantic-affective model, that connects the semantic with the affective space through a mapping and an existing annotated affective lexicon.

The proposed model can be evaluated on both pair-wise semantic similarity and sentence affective estimation tasks, showing a significant improvement of correlation with human annotations, as well as an important increase in the classification accuracy for sentence-level binary sentiment detection in generic and Twitter data, compared to a baseline method that does not use topic models.

Keywords

affective analysis, sentence-level sentiment analysis, topic modeling, LDA, distributional semantic models, word2vec, adaptation

Περιεχόμενα

Ευχαριστίες **7**

Περίληψη **9**

Abstract **11**

Περιεχόμενα **14**

Κατάλογος Σχημάτων **15**

Κατάλογος Πινάκων **17**

1 Εισαγωγή **21**
1.1 Κίνητρο 21
1.2 Συναισθήματική Ανάλυση Κειμένου 22
1.2.1 Αναπαράσταση Συναισθήματος 22
1.2.2 Θεματικά Μοντέλα 24
1.2.3 Διανεμητική Σημασιολογία 25
1.3 Σκοπός & Συνεισφορά της Εργασίας 25
1.4 Διάρθρωση της Εργασίας 26

2 Προϋπάρχουσα Εργασία **29**
2.1 Συναισθηματική Ανάλυση Προτάσεων 29
2.1.1 Τεχνικές Μηχανικής Μάθησης 29
2.1.2 Χώρος Χαρακτηριστικών 31
2.2 Τεχνικές Σημασιολογικής Κατεύθυνσης 32
2.2.1 Lexicon-based approaches 33
2.3 Διανεμητικά Σημασιολογικά Μοντέλα 34
2.4 Τεχνικές Θεματικής Μοντελοποίησης 35
2.5 Συναισθηματική Ανάλυση χρησιμοποιώντας ΘΜ 39

3 Θεματική Μοντελοποίηση **41**
3.1 Θεωρητική Εξήγηση 41

3.2 Στατιστική Εξήγηση . . . 43
3.2.1 Συμβολισμός και Ορολογία . . . 44
3.2.2 Αλγόριθμος . . . 45
3.2.3 Τεχνικές Εξαγωγής Συμπεράσματος . . . 46

4 Σημασιολογικά & Συναισθηματικά Μοντέλα 47
4.1 Διανεμητικά Σημασιολογικά Μοντέλα . . . 47
4.1.1 Ορισμός . . . 47
4.1.2 Λειτουργικότητα . . . 48
4.1.3 Word2Vec . . . 50
4.2 Συναισθηματική Μοντέλο . . . 52
4.2.1 Από το Σημασιολογικό στο Συναισθηματικό Χώρο . . . 52
4.2.2 Επισημείωση Λέξεων . . . 53
4.2.3 Επισημείωση προτάσεων . . . 56

5 Προσαρμοστικό Μοντέλο 59
5.1 Κίνητρο . . . 59
5.2 Topic Modeling . . . 60
5.2.1 Εκπαίδευση LDA . . . 60
5.2.2 Κατασκευή Topics . . . 61
5.3 Προσαρμογή Σημασιολογικών Μοντέλων . . . 62
5.3.1 Εκτίμηση Συναισθήματος . . . 63

6 Πειραματική Διαδικασία & Αποτελέσματα 65
6.1 Περιγραφή Δεδομένων . . . 65
6.2 Ανάλυση Topics . . . 67
6.3 Σημασιολογική Ομοιότητα . . . 71
6.4 Εκτίμηση Συναισθήματος . . . 75
6.4.1 Baseline Training . . . 75
6.4.2 Semantic Model Adaptation . . . 75

7 Επίλογος 83
7.1 Ανακεφαλαίωση & Συμπεράσματα . . . 83
7.2 Εν εξελίξει Εργασία . . . 84
7.2.1 Semantic-affective model adaptation (SAMA) . . . 84
7.3 Μελλοντικές Επεκτάσεις . . . 86

A′ Gensim Toolbox 89
A′.1 Topic Modeling . . . 89
A′.2 Word2Vec . . . 91

Bibliography 93

Κατάλογος Σχημάτων

1.1 Διαφορετικές Αναπαραστάσεις Συναισθήματος. 23
1.2 Βασικές περιοχές ενασχόλησης της εργασίας. 26

2.1 Sentiment Analysis categories. 30
2.2 Polarity classification via subjectivity detection Pang and Lee (2004). . . . 31
2.3 Singular Value Decomposition technique used in LSI for dimensionality reduction. 36
2.4 Graphical model representation of the CTM (from Blei and Lafferty (2007)). 38

3.1 Example of Topics produced by LDA from Steyvers and Griffiths (2007). . 42
3.2 Explanation of the intuition behind LDA, Blei (2012). 44
3.3 Generative process along with the problem of inference (source: Steyvers and Griffiths (2007)) 46

4.1 Example of term-term BOW matrix, using term co-occurrence frequency and window = 1. Image from Clark (2012). 49
4.2 CBOW and Skipgram architectures of Word2Vec. 51
4.3 Semantic-Affective model as proposed in Malandrakis et al. (2014). 52
4.4 Semantic-Affective Mapping concept. 53
4.5 word affective score estimation procedure using SAM. 55

5.1 Topic construction diagram. 62
5.2 Mixture of Semantic models diagram. 63

6.1 Intra-topic coherence measured after LDA training. 68
6.2 Example of visualization for 60 topics (*60 Topics Example Link*). 69
6.3 Example of visualization for 90 topics (*90 Topics Example Link*). 70
6.4 Number of sentences classified per number of topics. 72
6.5 Posteriors histogram for MEN dataset. 75
6.6 10-fold Cross Validation on ANEW dataset using Ridge Regression and $\lambda = 0.35$. 77
6.7 Posteriors histogram for 1000 news headlines sentences. 77
6.8 Accuracy and Spearman Correlation using linear fusion. 78

6.9 Accuracy and Spearman Correlation using weighted fusion. 78
6.10 Accuracy and Spearman Correlation using non-linear max fusion. 79
6.11 Histograms for seed words valence scores obtain from ANEW. 79
6.12 Classification accuracy using Linear Logistic Regression classifier. 80
6.13 Classification accuracy using Naive Bayes classifier. 81

Κατάλογος Πινάκων

2.1 Basic classifiers tested in Pang et al. (2002). 30
2.2 List of most known existing affective dictionaries 34

3.1 Distributions used in LDA table. 45

4.1 List of similarity metrics for word-vectors comparison. 50
4.2 Εξαμπλε οφ τραινινγ ωιτη 10 σεεδ ωορδς φρομ ΑΝΕΩ, υσινγ ΛΣΕ. 55

5.1 Example of Topic Modeling output, distribution words in a topic, using the top 10 most probable words. 61

6.1 ANEW affective lexicon. 66
6.2 Word-pairs similarity datasets. 66
6.3 Sentences datasets for polarity classification. 67
6.4 Statistics for number of sentences collected per topic using a posterior probability threshold = 0.1 and the max posterior probability. 70
6.5 Number of sentences not classified. 71
6.6 Number of pairs for different similarity ranges for each dataset. 72
6.7 Spearman correlation for MEN dataset, using different similarity ranges. . . 73
6.8 Spearman correlation for WordSim353 dataset, using different similarity ranges. 73
6.9 Number of pairs for MEN dataset, that are classified to more than one topic along with the standard deviation of their posterior probabilities for the topics they belong. 74
6.10 Different number of iterations when trained the semantic model for 30, 60 and 90 topics for WS-353 dataset. 76
6.11 Different number of iterations when trained the semantic model for 30, 60 and 90 topics for MEN dataset. 76
6.12 Results summary (CA: Classification Accuracy, CC: Classification Correlation (Spearman)) for sentence polarity detection using SMA. 80
6.13 Classification Results for SemEval-2016 Twitter Task B dataset. 81
6.14 Number of positive and negative tweets in the SemEval-2016 Task 4, Subtask B dataset. 81

Συντομογραφίες

BOW	Bag-of-words
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
pLSA	Probabilistic Latent Semantic Analysis
pLSI	Probabilistic Latent Semantic Indexing
CTM	Correlated Topic Model
PMI	Point-wise Mutual Information
TF-IDF	Term frequency-Inverse Topic Frequency
SAM	Semantic-Affective Model
POS	Part-of-Speech
SVM	Support Vector Machine
NB	Naive Bayes
LSE	Least Squares Estimation
RR	Ridge Regression
ANEW	Affective Norms for English Words
CBOW	Continuous Bag-of-words
W2V	Word2Vec
SVD	Singular Value Decomposition
VSM	Vector Space Model
DSM	Distributional Semantic Model
ME	Maximum Entropy
SVR	Support Vector Regression
OVA	One-versus-all
EM	Expectation Maximization
HDP	Hierarchical Dirichlet Process
HMM	Hidden Markov Model
JST	Joint Sentiment Topic
PMF	Probability Mass Function
MCMC	Monte Carlo Markov Chain
KL	Kullback-Leibler

Κεφάλαιο 1

Εισαγωγή

1.1 Κίνητρο

Οι γνωμοδοτήσεις διαδραματίζουν σημαντικό ρόλο στις ανθρώπινες δραστηριότητες και επηρεάζουν τις συμπεριφορές των ανθρώπων. Η ανάπτυξη του Διαδικτύου και η μετατροπή των πληροφοριών σε ψηφιακή μορφή, επέτρεψε στους χρήστες να έχουν πρόσβαση σε άφθονες ποσότητες δεδομένων, καθώς και να μοιράζονται τις απόψεις τους και τα συναισθήματά τους για το περιβάλλον τους άμεσα. Πολλές ενέργειες δείχνουν αυτού του είδος την επιρροή στις ζωές των ανθρώπων, από την αναζήτηση των χρηστών για κριτικές ταινιών μέχρι την αναζήτηση φαρμάκων, μέσω του διαδικτύου.

Ωστόσο, ο χρόνος και η προσπάθεια για την ανάλυση αυτών των δεδομένων είναι τυπικά ανάλογα προς το μέγεθος τους, κάνοντας την επεξεργασία τους εξαιρετικά δύσκολη για ένα μέσο άνθρωπο. Για πολλά χρόνια, έρευνες, δημοσκοπήσεις και ομάδες εστίασης είχαν ως στόχο να συγκεντρώσουν όσο το δυνατόν περισσότερες πληροφορίες για να συλλάβουν την γενική προτίμηση ή δυσαρέσκεια του κοινού σχετικά με προϊόντα, υπηρεσίες ή άλλα σύγχρονα θέματα. Ως εκ τούτου, το πεδίο της επιστήμης των υπολογιστών ασχολείται με την συναισθηματική ανάλυση κειμένου (Affective Text Analysis), η οποία σημαίνει την αυτόματη ανίχνευση υποκειμενικών πληροφοριών, εναλλακτικά γνωστό ως αναγνώριση συναισθημάτων, σε γραπτές πηγές, με στόχο να λύσει αυτό το πρόβλημα. Ωστόσο, η άνοδος των κοινωνικών μέσων μαζικής ενημέρωσης, όπως το Facebook και το Twitter έχει στρέψει την προσοχή της έρευνας σε μικρά κομμάτια του κειμένου συνήθως προτάσεις, ως ο κύριος τρόπος έκφρασης σε κοινωνικές ιστοσελίδες.

Υπάρχουν πολλοί τρόποι για να εκτιμηθεί η γνώμη που εκφράζει μια πρόταση, όπως αναλύεται στο κεφάλαιο 2. Ωστόσο, ένα γνωστό πρόβλημα σε αυτό το έργο είναι η μεγάλη ποικιλία θεμάτων που ανιχνεύονται σε φράσεις και πώς αυτά επηρεάζουν την ερμηνεία τους. Το νόημα των λέξεων αλλάζει όχι μόνο σε διαφορετικούς τομείς, αλλά και όταν χρησιμοποιείται διαφορετικό στυλ γραφής (π.χ. τυπική, σαρκαστική) με αποτέλεσμα πολλαπλές έννοιες και νοήματα για λέξεις. Η πλειονότητα των διαθέσιμων δεδομένων αφορούν σε συγκεκριμένους τομείς και συνεπώς δεν παρέχουν μια λύση σε προβλήματα πάνω σε διαφορετικούς τομείς. Ένας δημοφιλής τρόπος για την αντιμετώπιση αυτού του προβλήματος είναι η Θεματική Μο-

ντελοποίηση δεδομένων γενικού σκοπού. Αυτές οι μέθοδοι προσπαθούν να αναπαραστήσουν τις προτάσεις ως ένα πιθανολογικό μείγμα διαφορετικών συγκεκριμένων θεματικών μοντέλων και μπορούν να οδηγήσουν σε μια πιο ακριβή εκτίμηση της πολικότητας της πρότασης (θετική / αρνητική στις απλούστερες περιπτώσεις). Ωστόσο, προκειμένου να καθοριστεί τελικά αν μια πρόταση είναι θετική ή αρνητική, διάφοροι επιστημονικοί τομείς πρέπει να συνδυαστούν όπως Αναδιανεμητικά Σημασιολογικά Μοντέλα και Συναισθηματική Μοντέλα.

1.2 Συναισθήματική Ανάλυση Κειμένου

Η ανάλυση του συναισθηματικού περιεχομένου των κειμένων είναι γνωστό ως συναισθηματική ανάλυση κειμένου. Ο αγγλικός όρος 'Affect' είναι ένας όρος ομπρέλα που περιγράφει τα συναισθήματα, τις διαθέσεις και τις ιδιοσυγκρασίες και χρησιμοποιείται τόσο συχνά όσο τον όρος 'Emotion¨. Μια εργασία στη συναισθηματική ανάλυση μπορεί να καθορίζεται από το είδος των συναισθημάτων προς ανίχνευση (π.χ. θυμός, ευτυχία) ή το επίπεδο της ανάλυσης που θα εκτελέσει (λέξεις, προτάσεις, παράγράφοι, εγγράφα, συλλογές). Ο κύριος στόχος αυτού του πεδίου είναι να προσδιορίσει τη γνώμη του συγγραφέα ή του αναγνώστη, σε μικρότερα ή μεγαλύτερα τμήματα κειμένου. Πρόκειται για μια μεγάλη περιοχή που είναι άρρηκτα συνδεδεμένη με τον τομέα της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing (NLP)) και των Αναδιανεμητικών Σημασιολογικών Μοντέλων.

Σε γενικές γραμμές, ο τομέας που ασχολείται με την ανάπτυξη των ευφυών συστημάτων που είναι σε θέση να αντιληφθούν και να κωδικοποιήσουν το συναίσθημα είναι γνωστός ως συναισθηματική υπολογισιμότητα (Affective Computing). Για να οικοδομήσουμε συστήματα ως εκ τούτου, θα πρέπει να αναπαραστήσουμε τα ανθρώπινα συναισθήματα σε μια κλίμακα κατανοητή από τους υπολογιστές.

1.2.1 Αναπαράσταση Συναισθήματος

Τα συναισθήματα μπορεί να γίνουν αντιληπτά είτε σε δισδιάστατο ή σε τρισδιάστατο χώρο. Ο Paul Ekman καθόρισε τα έξι βασικά συναισθήματα που είναι κοινά μεταξύ των ανθρώπων σε όλους τους πολιτισμούς: τον θυμό, την ευτυχία, την έκπληξη, την αηδία, τη θλίψη και το φόβο. [1].

Για θεωρητικούς και πρακτικούς λόγους οι ερευνητές καθόρισαν τα συναισθήματα σύμφωνα με μία ή περισσότερες διαστάσεις. Ο Wilhelm Max Wundt, ο πατέρας της σύγχρονης ψυχολογίας, πρότεινε ότι τα συναισθήματα μπορεί να περιγραφούν με τρεις διαστάσεις: 'Ευχαρίστηση έναντι Δυσαρέσκεια¨, 'Διέργεση ή Υποταγή' και 'Ένταση ή Χαλάρωση' [2]. Επιπλέον, ο Harold Schlosberg ονόμασε τρεις διαστάσεις του συναισθήματος: 'ευχαρίστηση-δυσαρέσκεια¨, 'προσοχή-απόρριψη» και 'επίπεδο ενεργοποίησης' (Schlosberg (1954)). Οι βασικές συναισθηματικές διαστάσεις περιγράφονται παρακάτω:

Valence είναι ένα υποκειμενικό αίσθημα ευχαρίστησης ή δυσαρέσκειας. Κυμαίνεται από πολύ

[1] ηττπς://ωωω.παυλεκμαν.ςομ/ωπ-ςοντεντ/υπλοαδς/2013/07/Βασις-Εμοτιονς.πδφ
[2] ηττπ://πσψςηςλασσιςς.ψορκυ.ςα/Ωυνδτ/Ουτλινες/

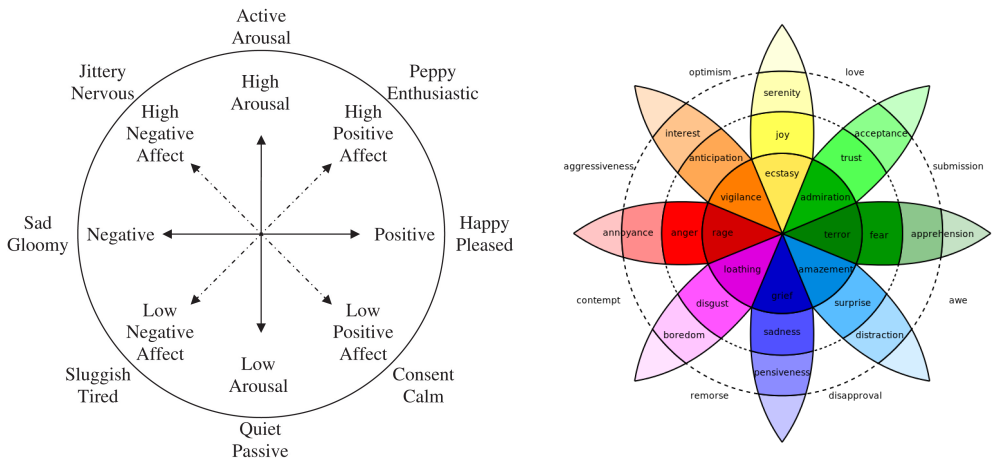
θετικές έως πολύ αρνητικές τιμές. Για παράδειγμα, τόσο ο 'θυμός' όσο και ο 'φόβος' είναι δυσάρεστα συναισθήματα, και βρίσκονται ψηλά στην κλίμακα δυσαρέσκειας. Ωστόσο η 'χαρά' είναι ένα ευχάριστο συναίσθημα.

Arousal είναι μια υποκειμενική κατάσταση του αισθήματος ενεργοποίησης ή απενεργοποίησης. Εναλλακτικά, μπορεί να βρεθεί ως διέγερση-ενεργοποίηση, η οποία είναι η διάσταση των συναισθημάτων που εκφράζει σε ποιο βαθμό ένα άτομο είναι έτοιμο να πράξει ή να χαλαρώσει. Η ενεργοποίηση/διέγερση μετρά την ένταση του συναισθήματος. Για παράδειγμα, ενώ και 'θυμός' και 'οργή' είναι δυσάρεστα συναισθήματα, η 'οργή' έχει υψηλότερη ένταση ή υψηλότερη διεγερτική κατάσταση. Ωστόσο η 'πλήξη', η οποία είναι επίσης μια δυσάρεστη κατάσταση, έχει χαμηλή τιμή διέγερσης.

Dominance αντιπροσωπεύει τον ελεγχόμενο και κυρίαρχο χαρακτήρα του συναισθήματος. Για παράδειγμα, ενώ και 'φόβος' και 'θυμός' είναι δυσάρεστα συναισθήματα, ο 'θυμός' είναι ένα κυρίαρχο συναίσθημα, ενώ ο 'φόβος' είναι ένα υποτακτικό συναίσθημα.

Κατά συνέπεια, τα ανθρώπινα συναισθήματα μπορούν να καθοριστούν σε δύο ή τρεις διαστάσεις. Τα περισσότερα μοντέλα χρησιμοποιούν το valence και το arousal, καθώς το arousal συσχετίζεται έντονα με το dominance. Τα δύο πιο σημαντικά μοντέλα είναι α) το circumplex μοντέλο και β) το μοντέλο του Plutchik.

Το circumplex μοντέλο αναπτύχθηκε από τον James Russell (Russell (1980)). Αυτό το μοντέλο προτείνει ότι τα συναισθήματα διανέμονται σε ένα δισδιάστατο κυκλικό χώρο, που περιέχει διαστάσεις διέγερσης και σθένους. Η Διέγερση αντιπροσωπεύει τον κάθετο άξονα και το σθένους αντιπροσωπεύει τον οριζόντιο άξονα, ενώ το κέντρο του κύκλου αντιπροσωπεύει ένα ουδέτερο σθένος και ένα μέσο επίπεδο διέγερσης. Σε αυτό το μοντέλο, συναισθηματικές καταστάσεις μπορεί να εκπροσωπούνται σε οποιοδήποτε επίπεδο του σθένους και της διέγερσης, ή σε ουδέτερο επίπεδο του ενός ή και των δύο από αυτούς τους παράγοντες.



(α') Αναπαράσταση Circumplex μοντέλου.

(β') Plutchik's wheel of emotions.

Σχήμα 1.1: Διαφορετικές Αναπαραστάσεις Συναισθήματος.

Ο Robert Plutchik προσφέρει ένα τρισδιάστατο μοντέλο που είναι ένα υβρίδιο των δύο βασικών κατηγοριών (Plutchik (2001)). Οργανώνει τα συναισθήματα σε ομόκεντρους κύκλους, όπου εσωτερικοί κύκλοι είναι πιο απλοί και εξωτερικούς κύκλους πιο περίπλοκοι. Αξίζει να σημειωθεί ότι, εξωτερικοί κύκλοι σχηματίζονται επίσης με ανάμιξη των εσωτερικών συναισθηματικών κύκλων. Το μοντέλο Plutchik, όπως του Russel, προέρχεται από μια circumplex αναπαράσταση, όπου οι συναισθηματικές λέξεις απεικονίζονται με βάση την ομοιότητα.

Με βάση αυτούς τους συναισθηματικούς χώρους, όλα τα συναισθήματα μπορούν να αναπαρασταθούν. Στην απλούστερη περίπτωση, η εκτίμηση μπορεί να γίνει χρησιμοποιώντας μόνο τη διάσταση σθένους, για δύο κατηγορίες, θετική (υψηλό σθένους) και αρνητικό (χαμηλό σθένους), τρεις κατηγορίες, αν η ουδέτερη τάξη (μηδέν σθένος) περιλαμβάνεται (το οποίο δεν περιέχει κανένα συναίσθημα), ή πέντε κατηγορίες και περισσότερο, ανάλογα με την εφαρμογή και πόσο η συναισθηματική κλίμακα πρέπει να επεκταθεί.

Τέλος, η υποκειμενικότητα ή αντικειμενικότητα ενός τμήματος κειμένου είναι μια άλλη κατηγορία 'συναισθήματος'. Ωστόσο, στην περίπτωση των προτάσεων, καθώς είναι αρκετά συχνό να περιέχουν περισσότερα από ένα συναισθήματα, μερικές φορές μια Aspect-Based ανάλυση συναισθήματος είναι πιο κατάλληλη, η οποία ανιχνεύει την γνώμη προς μια συγκεκριμένη πτυχή ή πρόσωπο που περιέχονται στην πρόταση.

1.2.2 Θεματικά Μοντέλα

Τα θεματικά μοντέλα, που ονομάζονται επίσης στατιστικά θεματικά Μοντέλα, είναι μια κατηγορία μαθηματικών μοντέλων, που ενσαρκώνουν μια σειρά από υποθέσεις που συμβάλλουν στην ανακάλυψη της θεματικής δομής των εγγράφων, χρησιμοποιώντας κατανομές πιθανότητας. Πιο συγκεκριμένα, η βασική υπόθεση των τεχνικών μοντελοποίησης είναι ότι κάθε έγγραφο μπορεί να θεωρηθεί ως ένα μείγμα πολλών θεμάτων που εκφράζονται ως κατανομές πιθανότητας. Ο όρος 'Topic' περιγράφει συνήθως μεγάλες θεματικές περιοχές όπως 'υπολογιστές', 'υγεία', 'ταινίες' κλπ, ή να χαρακτηρίζουν πιο συγκεκριμένους τομείς, όπως η 'τιμή', 'ποιότητα' κ.λπ. Ωστόσο, κατά τη μελέτη των θεματικός μοντέλων, ο ίδιος όρος περιγράφει μια πιθανοτική κατανομή πάνω σε ένα σταθερό λεξικό. Εναλλακτικά, ένα 'topic' αποτελείται από μια ομάδα λέξεων που εμφανίζονται συχνά μαζί σε αυτόν τον τομέα. Η θεματική Μοντελοποίηση είναι ένα συχνά χρησιμοποιούμενο εργαλείο εξόρυξης κειμένου για την ανάλυση μεγάλων συλλογών κειμένου. Το κύριο πλεονέκτημα αυτών των μοντέλων είναι αυτόνομα (χωρίς επιτήρηση) και, κατά συνέπεια, δεν απαιτούν την γνώση της κατηγορίας των δεδομένων, τα οποία είναι δύσκολο να βρεθούν.

Χρησιμοποιώντας το σημασιολογικό πλαίσιο των λέξεων σε κείμενα, τα θεματικά μοντέλα μπορούν να αναγνωρίσουν και να συνδέσουν λέξεις με παρόμοια σημασία, καθώς, γίνεται διάκριση μεταξύ των χρήσεων των λέξεων με πολλαπλές σημασίες (λεξιλογική αποσαφήνιση). Τα θεματικά μοντέλα που συνδυάζονται με συναισθηματική ανάλυση μπορούν να χρησιμοποιηθούν για να βελτιωθεί η εκτίμηση της συναισθηματικής βαθμολογίας των λέξεων. Οι πιο γνωστές προσεγγίσεις περιλαμβάνουν την εκμάθηση θεμάτων που βασίζεται σε εκμάθηση ετικετών συναισθήματος κατά την ταυτόχρονη λήψη συναισθημάτων και θεμάτων σε συλλογές

κειμένου. Επιπλέον, η aspect-based συναισθηματική ανάλυση μπορεί να επωφεληθεί από τις θεματικές τεχνικές μοντελοποίησης που είναι σε θέση να διακρίνουν τις διαφορετικές πτυχές που εξετάζει μια κριτική.

1.2.3 Διανεμητική Σημασιολογία

Η Διανεμητική σημασιολογία (Distributional Semantics) είναι μια περιοχή έρευνας που αναπτύσσει και μελετά τις θεωρίες και τις μεθόδους για τον ποσοτικό προσδιορισμό και την κατηγοριοποίηση των σημασιολογικών ομοιοτήτων μεταξύ των γλωσσικών στοιχείων με βάση τις αναδιανεμητικές τους ιδιότητες σε μεγάλα δείγματα δεδομένων γλώσσας. Η βασική ιδέα της διανεμητικής σημασιολογίας μπορεί να συνοψιστεί στη λεγόμενη Distributional Hypothesis: 'Γλωσσική αντικείμενα με παρόμοιες κατανομές έχουν παρόμοια σημασία' ή με άλλα λόγια, 'λέξεις που τείνουν να εμφανίζονται μαζί σε κομμάτια του κειμένου, τείνουν να έχουν παρόμοιο νόημα'.

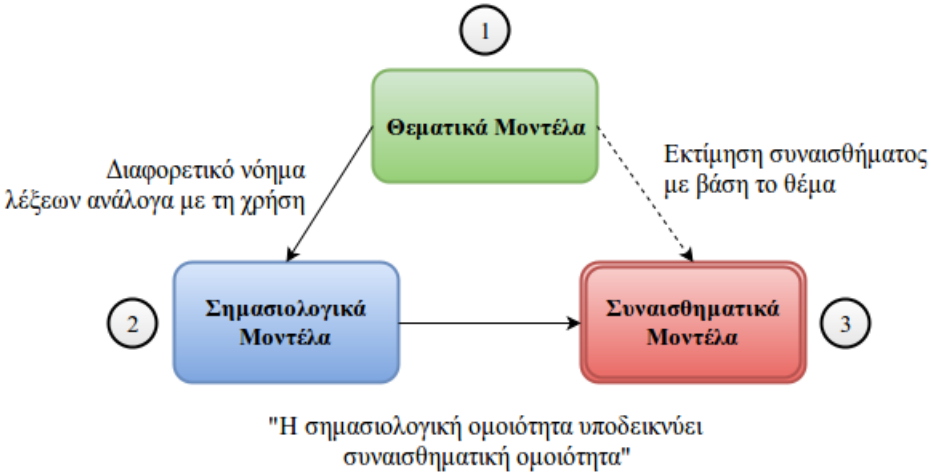
Τα Διανεμητικά Σημασιολογικά μοντέλα (Distributional Semantic Models (DSMs)), δηλαδή τα μοντέλα που προσπαθούν να κωδικοποιήσουν το νόημα των λέξεων και μεγαλύτερων τμημάτων κειμένου σε μια μαθηματική αναπαράσταση ενός διανύσματος, συνδέονται στενά με την συναισθηματική ανάλυση κειμένου. Τα DSMs μπορούν να συνδυαστούν με μεθόδους συναισθηματικής ανάλυσης κειμένου, προκειμένου να συλλάβουν το συναίσθημα των λέξεων. Η αναπαράσταση των λέξεων ως σημασιολογικά διανύσματα χαρακτηριστικών μπορεί να χρησιμοποιηθεί, όχι μόνο ως πρόσθετα χαρακτηριστικά για την ταξινόμηση συναισθήματος, αλλά και ως μέσο μέτρησης του συναισθήματος μέσα από την εκτίμηση της ομοιότητας, συγγένειας, συνωνυμίας και πολλαπλές άλλες σχέσεις μεταξύ των λέξεων, με τη βοήθεια της γραμμικής άλγεβρας, καθιστώντας το ένα ισχυρό εργαλείο για αναγνώριση πολικότητας προτάσεων.

1.3 Σκοπός & Συνεισφορά της Εργασίας

Ο κύριος στόχος αυτής της εργασίας είναι να διερευνήσει έναν προσαρμοστικό αλγόριθμο Σημασιολογικών Μοντέλων για τον προσδιορισμό της πολικότητας των προτάσεων ανάμεσα σε δύο κατηγορίες, θετική ή αρνητική.

Υπάρχουν τρία μοντέλα που συνδυάζονται στην προτεινόμενη προσέγγιση, Θεματικά Μοντέλα, Σημασιολογικά Μοντέλα και Συναισθηματικά Μοντέλα, χρησιμοποιώντας κάποιες βασικές παραδοχές, όπως φαίνεται στο σχήμα 1.2. Το ζευγάρι ένα-δύο είναι άμεσα συνδεδεμένο με το γεγονός ότι οι λέξεις μπορούν να έχουν πολλαπλές σημασίες σε διάφορους τομείς, το ζεύγος δύο-τρία συνδέεται άμεσα με την παραδοχή ότι 'σημασιολογική ομοιότητα συνεπάγεται συναισθηματική ομοιότητα' και το ζεύγος ένα-τρία είναι έμμεσα συνδεδεμένο μέσα από την παραδοχή ότι πριν από την εκτίμηση της πολικότητας μια φράσης που μπορούμε πρώτα να προσδιορίσουμε τα θέματα που περιέχει.

Το παρούσα τεχνική χρησιμοποιεί η λανθάνουσα μέθοδο Dirichlet Κατανομής (Latent Dirichlet Allocation (LDA)). Η Θεματική Μοντελοποίηση είναι σε θέση να εξάγει σημαντικές πληροφορίες από αδόμητα κείμενα, με το διαχωρισμό εγγράφων στα θέματα που μπορεί να



Σχήμα 1.2: Βασικές περιοχές ενασχόλησης της εργασίας.

περιέχουν. Επιπλέον, η σημασιολογική αναπαράσταση των λέξεων διερευνάτε σε κάθε θέμα, με σκοπό την εξαγωγή κατανοητής πληροφορίας για τις σχέσεις μεταξύ των λέξεων, και ιδίως τον προσδιορισμό της ομοιότητας τους. Η προτεινόμενη προσέγγιση βασίζεται στην προσαρμογή του σημασιολογικού χώρου της κάθε πρότασης με βάση το θέμα, με στόχο να περιγράψει καλύτερα το νόημά της. Τέλος, προκειμένου να προσδιορίσει το συναίσθημα στις προτάσεις, εξετάζεται μια αντιστοίχιση από το σημασιολογικό στο συναισθηματικό χώρο, αξιοποιώντας την υπόθεση ότι 'σημασιολογική ομοιότητα συνεπάγεται συναισθηματική ομοιότητα'.

Η συμβολή της παρούσας διατριβής έγκειται στη βελτίωση της δυαδικής ταξινόμησης σε επίπεδο προτάσεων. Ειδικότερα, στόχος μας είναι όχι μόνο να βελτιωθεί η ακρίβεια ταξινόμησης αλλά και τη συσχέτιση από την άποψη της βαθμολογίας θετικού/αρνητικού συναισθήματος σε επίπεδο προτάσεων. Η κύρια διαφορά μεταξύ των θεματικών μοντέλων και των μοντέλων που βασίζονται στα λεξικά, είναι ότι μπορούν να επιτύχουν λεξιλογική αποσαφήνιση, προσδιορίζοντας το σημασιολογικό πλαίσιο της κάθε πρότασης. Αντίθετα, οι μέθοδοι ταξινόμησης με λεξικά, αποτυγχάνουν να συλλάβουν το πλαίσιο των λέξεων που περιλαμβάνονται σε μια πρόταση και μπορεί να δώσουν παραπλανητικά αποτελέσματα. Η αξιολόγηση της ομοιότητας σε λέξεις-ζευγάρια και ο εντοπισμός της πολικότητας σε προτάσεις, δείχνει ότι η ακρίβεια ταξινόμησης και η συσχέτιση μπορεί να βελτιωθεί με την ενσωμάτωση συγκεκριμένων σημασιολογικών μοντέλων.

Αυτό το σύστημα προσαρμογής σημασιολογικών μοντέλων ήταν μέρος του συστήματος που υποβλήθηκε στο διαγωνισμό SemEval-2016,Task 4: Sentiment Analysis in Twitter, in subtasks A: Message Polarity Classification and B: Tweet classification according to a two-point scale, όπου έλαβε την πρώτη θέση στο subtask B (Palogiannidi et al. (2016)).

1.4 Διάρθρωση της Εργασίας

Αυτή η διπλωματική εργασία είναι οργανωμένη σε επτά κεφάλαια και ένα παράρτημα.

Στο κεφάλαιο 2 μια μεγάλη περιγραφή της βιβλιογραφίας παρέχεται, η οποία χωρίζεται σε τέσσερα τμήματα, σύμφωνα με τα πεδία τα οποία ασχολείται η διατριβή. Το πρώτο τμήμα, περιγράφει τις διαφορετικές προσεγγίσεις ανάλυσης συναισθήματος που ερευνήθηκαν, με σκοπό την ανίχνευση της πολικότητας (συνήθως τρεις κατηγορίες, θετική-αρνητική-ουδέτερη) των προτάσεων και κριτικών (μικρές παράγραφοι). Το δεύτερο τμήμα, εισάγει τα Διανεμητικά Σημασιολογικά Μοντέλα που χρησιμοποιούνται ευρέως για να ανακαλύψουν τη σημασιολογική ομοιότητα των λέξεων. Στην τρίτη ενότητα, περιγράφεται η εξέλιξη των Τεχνικών Θεματικής Μοντελοποίησης μοντελοποίησης στην πάροδο του χρόνου και οι state-of-the-art αλγόριθμοι που χρησιμοποιούνται στις μέρες μας. Τέλος, η τέταρτη ενότητα είναι μια βιβλιογραφική ανασκόπηση σχετικά με τις τεχνικές Sentiment Analysis που ενσωματώνουν τις τεχνικές Θεματικής Μοντελοποίησης ως μια σύγκριση με αυτή που προτείνεται σε αυτή τη διατριβή.

Στο κεφάλαιο ;; η Topic Modeling τεχνική που χρησιμοποιείται ως μέρος του αλγορίθμου περιγράφεται. Η λειτουργικότητα της τεχνικής LDA εξηγείται σε μεγάλο βαθμό σε θεωρητικό και πρακτικό επίπεδο, μαζί με την στατιστική δομή της.

Στο κεφάλαιο ;; δύο διαφορετικά μοντέλα Επεξεργασίαω Φυσικής Γλώσσας περιγράφονται. Στο πρώτο μέρος, Σημασιολογικά μοντέλα, που έχουν ως στόχο να αναπαραστήσουν την έννοια των λέξεων, ως μαθηματικές δομές διανυσματικών χώρων, αναλύονται μαζί με το μοντέλο Word2Vec που χρησιμοποιήθηκε για τα πειράματα της εργασίας. Στο δεύτερο μέρος, ένα υπάρχον Συναισθηματικό μοντέλο, που προσπαθεί να συλλάβει τη συναισθηματική πληροφορία των λέξεων αναλύεται. Αυτό το μοντέλο εκμεταλλεύεται ένα υπάρχον συναισθηματικό λεξικό για να προβλέψει το συναίσθημα των άγνωστων λέξεων με τη δημιουργία μιας αντιστοίχισης από το σημασιολογικό στο συναισθηματικό χώρο.

Το Κεφάλαιο ;; είναι το πιο σημαντικό μέρος αυτού του έργου, αφού περιγράφει το νέο αλγόριθμο που έχει ως στόχο να βελτιώσει την κατηγοριοποίηση των προτάσεων σε θετικές ή αρνητικές, χρησιμοποιώντας τα προαναφερθέντα εργαλεία. Η ιδέα του Σημασιολογικού μοντέλου προσαρμογής (Semantic Model Adaptation (SMA)) παρουσιάζεται και ιδίως πώς ο σημασιολογικός χώρος της πρότασης μπορεί να αναπαρασταθεί ως ένα μείγμα πολλών μοντέλων θεματικών-σημασιολογικών.

Το Κεφάλαιο ;; περιλαμβάνει όλα τα αποτελέσματα και την πειραματική διαδικασία για τα διάφορα πειράματα που διεξήχθησαν, σε διαφορετικά σύνολα δεδομένων. Πρώτον, τα θεματικά μοντέλα που δημιουργήθηκαν αξιολογήθηκαν με διαφορετικές μεθόδους ανάλυσης. Δεύτερον, η απόδοση της SMA τεχνικής εξετάζεται σε ζεύγη λέξεων για εκτίμηση σημασιολογικής ομοιότητας και σε επίπεδο προτάσεων για τον εντοπισμό της πολικότητας.

Το Κεφάλαιο ;; περιλαμβάνει μια περίληψη του έργου αυτού, καθώς και τα συμπεράσματα που προέκυψαν από τα διαφορετικά πειράματα. Επιπλέον, μια βελτιωμένη έκδοση του αλγορίθμου SMA παρουσιάζεται ως εν εξελίξει εργασία. Τέλος, κατευθύνσεις για πιθανή μελλοντική

εργασία αναλύονται.

Στο παράρτημα ;; το εργαλείο Gensim περιγράφεται και πιο συγκεκριμένα δύο από τις βασικές εφαρμογές του, που χρησιμοποιήθηκαν σε αυτή την εργασία, ο αλγόριθμος LDA για Θεματική Μοντελοποίηση και το σημασιολογικό μοντέλο Word2Vec. Οι απαραίτητες εντολές για προ-επεξεργασία δεδομένων και η κατασκευή των μοντέλων παρουσιάζονται μαζί με κάποιες εξόδους των εντολών κατά τη διάρκεια της εκτέλεσης.

Κεφάλαιο 2

Προϋπάρχουσα Εργασία

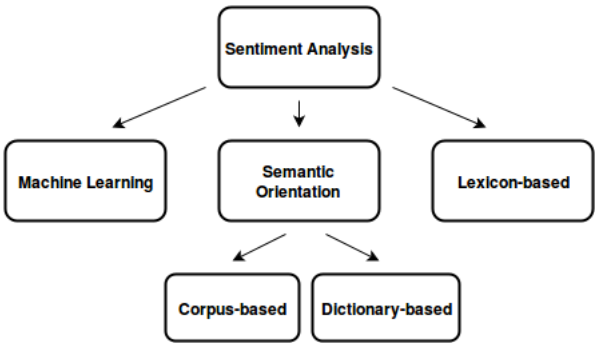
2.1 Συναισθηματική Ανάλυση Προτάσεων

Οι περισσότερες από τις δημοσιευμένες εργασίες, επικεντρώνονται στην ανίχνευση συναίσθημα στην αναθεώρηση-παραγράφου επίπεδο, ως μια προσπάθεια να προσδιορίσει αν η συνολική εξέφρασε γνώμη είναι θετική, αρνητική ή ουδέτερη. Ωστόσο, η άμεση αύξηση των Social Media έχει οδηγήσει τους ερευνητές να καταδικάσει επίπεδο εξόρυξης συναίσθημα. Οι τεχνικές και για τα δύο επίπεδα διερευνηθούν εδώ, καθώς υπάρχουν στενά συνδεδεμένες.

Οι προσεγγίσεις που ακολουθείται στην ανάλυση συναίσθημα, μπορούν να ταξινομηθούν σε 3 μεγάλες κατηγορίες: α) Μηχανική μάθηση προσεγγίσεις, υπό την επίβλεψη (χρησιμοποιώντας επισημασμένα δεδομένα) και χωρίς επίβλεψη (με μικρή (ημι-εποπτευόμενης) ή χωρίς την ένδειξη δεδομένων) που καθιστούν σαφή χρήση των ταξινομητών, β) Σημασιολογικού προσεγγίσεις προσανατολισμό, οι οποίες βασίζονται σε θετικές και αρνητικές λέξεις συναίσθημα που περιέχονται στο κείμενο της αξιολόγησης και γ) προσεγγίσεις Λεξιςον που βασίζεται χρησιμοποιούν τις υφιστάμενες σχολιασμένη λεξικά συναίσθημα των λέξεων και μικρές φράσεις. Η δεύτερη κατηγορία μπορεί να χωριστεί σε ι) τεχνικές corpus-based που προσπαθούν να προσδιορίσουν λέξη σχέσεις που βασίζονται στη συνεργασία τους, εμφάνιση στο κείμενο, και ιι) τεχνικές λεξικό-βάση που χρησιμοποιούν τα συνώνυμα, αντώνυμα και τις αισθήσεις λέξη αποθηκεύονται σε λεξιλογικό βάσεις δεδομένων που βοηθούν καθορίζουν συναισθήματα λέξη. Επιπλέον, οι συνδυασμοί και υβριδικές προσεγγίσεις από τις ανωτέρω τεχνικές που χρησιμοποιούνται συνήθως.

2.1.1 Τεχνικές Μηχανικής Μάθησης

Ξεκινώντας με τις προσεγγίσεις Μηχανική Μάθηση, στόχος τους είναι να εκτιμηθεί η πολικότητα των ποινών ή σχόλια συνήθως ένα πρόβλημα ταξινόμησης δύο ή τρεις-κατηγορίας. Τα γενικά βήματα της διαδικασίας είναι τα εξής: Πρώτον αποκτήσετε ένα corpus των δεδομένων εκπαίδευσης, δεύτερον αντιπροσωπεύουν κάθε έγγραφο / ποινή ως ένα διάνυσμα χαρακτηριστικών γνωρισμάτων, να μάθουν ένα μοντέλο από αυτά τα δεδομένα από την εκπαίδευση ενός ταξινομητή και στη συνέχεια να ταξινομήσει τις αθέατες δεδομένων / δοκιμών σε μια κατηγορία χρησιμοποιώντας το εκπαιδευμένο ταξινομητή.



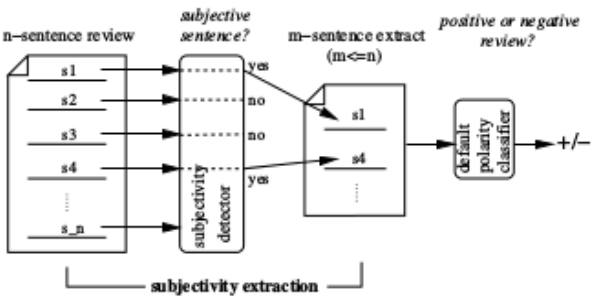
Σχήμα 2.1: Sentiment Analysis categories.

Το πρωτότυπο έργο για την ανίχνευση συναίσθημα σε σχόλια, παρουσίασε Pang et al. (2002). Στην προσέγγισή τους, πειραματίστηκαν με τρεις διαφορετικούς ταξινομητές, στην Κριτικές μια ταινία σύνολο δεδομένων, για ένα δυαδικό πρόβλημα ταξινόμησης (θετικά-αρνητικά): Naive Bayes (NB), μέγιστης εντροπίας (ME) και Support Vector Machines (SVM).

Classifier	Formula
Naive Bayes	$P_{NB}(c\|d) := \frac{P(c)\cdot(\prod_{i=1}^{m} P(f_i\|c)_i^n(d))}{P(d)}$
Maximum Entropy	$P_{ME}(c\|d) := \frac{1}{Z(d)} exp(\sum_{i=1} \lambda_{i,c} F_{i,c}(d,c))$ $F_{i,c}(d,c') = \begin{cases} 1 & \text{if } n_i(d) > 0 \text{ and } c = c' \\ 0 & \text{other} \end{cases}$
SVM	$\overrightarrow{w} = \sum_j a_j c_j \overrightarrow{d_j}, a_j \geq 0$

Table 2.1: Basic classifiers tested in Pang et al. (2002).

Τα χαρακτηριστικά που χρησιμοποιήθηκαν ποικίλες από unigrams (μεμονωμένες λέξεις) σε επίθετα και πολλαπλούς συνδυασμούς, καθώς και η αναπαράσταση του διανύσματος χαρακτηριστικών διέφερε από δυαδική μορφή (ύπαρξη - ανυπαρξία μιας λέξης-χαρακτηριστικό γνώρισμα) με τη συχνότητα (της λέξης-χαρακτηριστικό γνώρισμα στο η περίληψη). SVM αποδείχθηκε να δώσει την καλύτερη απόδοση σε σύγκριση με NB, αλλά συνολικά, κατέληξαν στο συμπέρασμα ότι αυτοί οι ταξινομητές έχουν καλύτερες επιδόσεις στο θέμα με βάση την κατηγοριοποίηση και όχι την ταξινόμηση συναίσθημα. Στη συνέχεια, Pang and Lee (2004) χρησιμοποιείται ανίχνευσης υποκειμενικότητας ως προ βήμα για τη βελτίωση των επιδόσεων (εικόνα 2.2). Ειδικότερα, εκπαίδευσαν έναν ανιχνευτή υποκειμενικότητα να ταξινομήσει ποινές σε υποκειμενικά και αντικειμενικά, απορρίψτε το υποκειμενικό και να περάσει το στόχο σε ένα ταξινομητή. Έδειξαν ότι τα αποτελέσματα της υποκειμενικότητας, εργασία ως ένα καλύτερο και καθαρότερο εισόδου στην ταξινομητή από το πρωτότυπο έγγραφο, καθώς περιέχουν περισσότερο συναισθηματική (υποκειμενική) πληροφορίες.



Σχήμα 2.2: Polarity classification via subjectivity detection Pang and Lee (2004).

Προκειμένου να επεκτείνει την ανάλυση συναισθήματος, Pang and Lee (2005) διερευνηθεί το πρόβλημα των πέντε κλίμακας πολικότητας επανεξέτασης (όπως ϛιτε ϛϛηλερ2005ιμπορτανϛε), χρησιμοποιώντας ένα SVM multi-class (ένας-εναντίον-όλα (Ο΄Α», Παλινδρόμηση (Σ΄Ρ) και τη μονάδα μέτρησης επισήμανση (σχέση μεταξύ των ϛτοιχείων και ετικέτες) αλγόριθμοι. Ως προέκταση Goldberg and Zhu (2006) εκπροσωπούνται έγγραφα σε ένα γράφημα και χωρίς ετικέτες δεδομένων, μέσω της σύνδεσής τους με τη χρήση ενός μέτρου απόσταση από τα χαρακτηριϛτικά του εγγράφου. Ο στόχος αυτής της προσέγγισης ήταν να χρησιμοποιήσει μη επισημασμένα δεδομένα για την εκπαίδευση, όπως επισημαίνονται αυτά είναι λίγα, η οποία αποδείχθηκε να δώσει την καλύτερη απόδοση από ό, αγνοώντας μη επισημασμένη δεδομένων. Προκειμένου να προσδιορίσει το κλίμα για τις διάφορες πτυχές που περιέχονται σε μία κριτική, Shimada and Endo (2008) πειραματίϛτηκε με Σ΄Μϛ, Σ΄Ρ, μέγιϛτης εντροπίας και το μέτρο ομοιότητας συνημιτόνου αλγόριθμους, μαζί με διάφορες μεθόδους εξαγωγής χαρακτηριϛτικών.

2.1.2 Χώρος Χαρακτηριστικών

Επιπλέον, ένας άλλος σημαντικός παράγοντας σε αυτά τα μοντέλα είναι το είδος των χαρακτηριϛτικών που χρησιμοποιούνται. Όπως Μαϛηινε κλίνει προσεγγίσεις χρησιμοποιούν βασικά ταξινομητές, προκειμένου να εκτιμηθεί το συναίσθημα σε γραπτές πηγές, τα δεδομένα που χρειάζονται πρέπει να έχουν καταγραφεί στη μορφή ενός φορέα. Σε γενικές γραμμές, αυτοί οι φορείς μπορεί να περιέχει λεξιλογικά ή μη λεξιλογικές πληροφορίες καθώς και αριθμητικές τιμές.

Χαρακτηριϛτικά που ανήκουν στην πρώτη κατηγορία, μπορεί να προέρχονται από οποιαδήποτε πηγή του κειμένου. Υποθέτοντας ότι ένα λεξιλόγιο εξάγεται από το σύνολο δεδομένων εκπαίδευσης, οι λέξεις που περιέχει μπορεί να χρησιμοποιηθεί ως χαρακτηριϛτικά. Πιο συγκεκριμένα, οι λέξεις που εκφράζουν το συναίσθημα, τις προθέσεις, πολύ μεγάλες λέξεις, ο αριθμός ενός λέξεις δυνατόν αισθήσεις (ϛιτε μαλανδρακιϛ2013διϛτριβυτιοναλ), ακόμα και να σταματήσει-λέξεις (λέξεις όπως Ι ΄,΄ η) μπορεί να κωδικοποιηθεί σε μια τσάντα -Από-λέξεις (ΒΟΩ) μορφή και χρησιμεύουν ως χαρακτηριϛτικά γνωρίσματα που αντιπροσωπεύουν ένα έγγραφο. Όσον αφορά τη δεύτερη κατηγορία, μη-λεξιλογικό χαρακτηριϛτικά μπορεί να είναι μέρος του λόγου (ΠΟΣ) ετικέτες (επίθετα, ρήματα, επιρρήματα), σημεία στίξης, εμοτιϛονς, διευθύνσεις ΥΡΛ ή αναφορές, που προέρχονται κυρίως από τις κοινωνικές πηγές των μέσων

ενημέρωσης, όπως το Τωιττερ. Τέλος, αριθμητική χαρακτηριστικά περιλαμβάνουν: δυαδική αναπαράσταση (παρουσία ή όχι του όρου αυτού στο έγγραφο), η συχνότητα εμφάνισης (συχνότητα του όρου αυτού στα έγγραφα), ΤΦ-ΙΔΦ βαθμολογίες (συχνότητα όρου - αντίστροφη συχνότητα έγγραφο), των στατιστικών αξιών των συναισθηματικών βαθμολογίες (μιν, μαξ, μέσου όρου, στδ), ετικέτες και τις αξίες που παράγεται από Αναδιανεμητικές Σημασιολογικού Μοντέλα διασποράς.

Απόλυτα που σχετίζονται με την κατηγορία λεξικό που βασίζεται, πολλαπλών τεχνικών μηχανικής μάθησης χρησιμοποιούν λεξικά συναίσθημα για να βελτιώσουν τις επιδόσεις τους. Osherenko and André (2007) πρότεινε να μειωθεί η διάσταση της λέξης διανύσματα χαρακτηριστικών γνωρισμάτων και να χρησιμοποιούν μόνο μικρά σύνολα των συναισθηματικών λέξεων όπως χαρακτηριστικά λαμβάνεται από ένα λεξικό. Η τεχνική αυτή, αξιολογείται με Σ΄Μ, δεν έδειξε καμία βελτίωση στην απόδοση της αναγνώρισης, αλλά μπορεί να συναχθεί το συμπέρασμα ότι η μείωση το χαρακτηριστικό μπορεί να επιτευχθεί χωρίς να χάσει σημαντικές πληροφορίες. Τέλος, Lin et al. (2012) διερεύνησε τη χρήση των διαφορετικών χαρακτηριστικών σε καθήκοντα κατάταξης.

2.2 Τεχνικές Σημασιολογικής Κατεύθυνσης

Σημασιολογικές προσεγγίσεις προσανατολισμό έχουν αποδειχθεί για να βελτιώσει τις επιδόσεις, καθώς βασίζονται στις σχέσεις μεταξύ των λέξεων στο κείμενο. Σε γενικές γραμμές, τα state-of-the-art μεθόδους που βασίζονται στην παρατήρηση ότι παρόμοιες λέξεις γνώμη τείνουν να εμφανίζονται μαζί σε ένα σώμα. Ως εκ τούτου, μπορεί κανείς να υποθέσει ότι, αν εμφανίζονται συχνά στο ίδιο πλαίσιο που θα εκφράσει το ίδιο συναίσθημα. Αυτή είναι μια στατιστική προσέγγιση που δείχνει ότι η συχνή εμφάνιση των λέξεων σε ένα έγγραφο σημαίνει εγγύτητα τους στο νόημα. Η πρώτη ιδέα προτάθηκε από τον Turney (2002). Στο έργο τους, το αμοιβαίας πληροφόρησης (ΠΜΙ) κριτήριο Ποιντ-σοφός χρησιμοποιήθηκε για να εκφράσει τη στατιστική εξάρτηση μεταξύ των δύο λέξεων. Συγκεκριμένα, μέτρησαν το συναίσθημα πολικότητα μιας λέξης β ως η διαφορά μεταξύ της αξίας ΠΜΙ της λέξης v με λέξη ‘ άριστα ’ και της αξίας ΠΜΙ της λέξης v με τη λέξη ‘κακή’ (εξίσωση ;;). Επιπλέον, σε Turney and Littman (2003) ένα σύνολο θετικών και αρνητικών λέξεων χρησιμοποιείται αντί μόνο ἐξαιρετική’ και ‘φτωχών΄. Ως εκ τούτου, η διαφορά του ποσού ΠΜΙ υπολογίζεται . Ως επέκταση σε αυτό το έργο, Chaovalit and Zhou (2005) χρησιμοποιείται για τη μηχανή αναζήτησης Google και μετριέται ο αριθμός των επισκέψεων επέστρεψε για ένα ζευγάρι λέξεις ερωτήματος, προκειμένου να εκτιμηθεί καλύτερα η σχέση τους.Επιπλέον, Read and Carroll (2009) επεκταθεί η προσέγγιση αυτή με τη χρήση του Σημασιολογικού χώρους και Αναδιανεμητικές Ομοιότητα ως ασθενώς-εποπτεύονται μεθόδους. Απαιτούν μόνο μια μεγάλη συλλογή των μη επισημασμένη δεδομένων που προκύπτουν μια μικρότερη απόδοση δομαιν-εξαρτώμενη. Στο ίδιο πλαίσιο, Baroni and Lenci (2010) παρουσιάζονται στην εργασία τους το πώς οι πληροφορίες σχετικά με τις σχέσεις των λέξεων μπορεί να εξαχθεί από ένα corpus ως πλειάδες λέξη-λινκ-λέξη που διοργανώνονται σε τανυστής τρίτης τάξης. Αυτό το είδος της αναπαράστασης επιτρέπει τις πληροφορίες που πρέπει να χρησιμοποιείται σε πολλές ερ-

γασίες, όπως (όπως αναφέρεται στο έγγραφο) μοντελοποίηση αποφάσεων ομοιότητα λέξης, ανακαλύπτοντας συνώνυμα, η έννοια κατηγοριοποίηση, προβλέποντας Selectional προτιμήσεις των ρημάτων, την επίλυση των προβλημάτων άναλογία, την ταξινόμηση των σχέσεων μεταξύ των ζευγών λέξεων, την ποιότητα της συγκομιδής δομές με μοτίβα ή παράδειγμα ζεύγη, προβλέποντας τις χαρακτηριστικές ιδιότητες των εννοιών, και την ταξινόμηση των ρημάτων σε κατηγορίες εναλλαγή. Περισσότερες πληροφορίες σχετικά με το Σημασιολογικό μοντέλα μπορούν να βρεθούν στο κεφάλαιο 2, Τμήμα 2.2 και το Κεφάλαιο 4.

$$PMI(word_1, word_2) = log_2\Big(\frac{P(word_1 \bigcap word_2}{P(word_1)P(word_2)}\Big) \tag{2.1}$$

$$SO(phrase) = PMI(phrase, \text{“}excellent\text{”}) - PMI(phrase, \text{“}poor\text{”}) \tag{2.2}$$

$$SO - PMI(word) = \sum_{\substack{pword \\ \in Pwords}} PMI(word, pword) - \sum_{\substack{nword \\ \in Nwords}} PMI(word, nword) \tag{2.3}$$

Σε ό,τι αφορά τη δεύτερη υποκατηγορία των σημασιολογικών καθηκόντων προσανατολισμού, το πιο γνωστό εργαλείο είναι ΩορδΝετ (Fellbaum (1998)). Είναι μια μεγάλη βάση δεδομένων που παρέχει διαφορετικά είδη πληροφοριών σχετικά με τη σχέση μεταξύ των λέξεων, καθώς και πολλαπλές αισθήσεις τους, η οποία μπορεί να βοηθήσει στην πρόβλεψη συναίσθημα. Η δυνατότητα να αποσαφηνίζουν τις αισθήσεις των λέξεων με ΩορδΝετ οδηγεί στον εντοπισμό των γνωμοδοτήσεων στο κείμενο. Kamps et al. (2004) πρότεινε να χρησιμοποιηθεί η σχετική μικρότερη απόσταση διαδρομής του συνωνυμο, ενώ Kim and Hovy (2004), συνιστάται να αποκτήσετε μια λίστα με λέξεις συναισθήματος με την επέκταση ενός αρχικού συνόλου με συνώνυμα και αντώνυμα αυτών των λέξεων από το ΩορδΝετ. στο έργο τους, η πολικότητα της λέξης καθορίζεται από τη σχετική καταμέτρηση των θετικών και αρνητικών συνώνυμα αυτής της λέξης. στην **?** ένα συναίσθημα-Αναλψζερ προσπαθεί να εξαγάγει το συναίσθημα για ένα θέμα από το ονλινε έγγραφα. δεν εντοπίζουν η συνολική συναίσθημα σε μία κριτική, αλλά καθορίζουν το συναίσθημα για μια ακριβή διάσταση αναφέρεται στην επανεξέταση, από την εξαγωγή χαρακτηριστικών για θέματα, διαθέτει περίπου το κλίμα και τις σχέσεις μεταξύ ατόμων και τα συναισθήματα με τη βοήθεια ενός λεξικού και μιας βάσης δεδομένων μοτίβο συναίσθημα. Επίσης, Andreevskaia and Bergler (2006) παρουσίασε μια μέθοδο για την εξαγωγή επίθετα συναίσθημα που φέρουν από το WordNet χρήση του Προγράμματος εξόρυξη συναισθήματος Tag (STEP).

2.2.1 Lexicon-based approaches

Τέλος, λεξικό προσεγγίσεις χρησιμοποιούν τα υπάρχοντα λεξικά με τη σωστή πολικότητα γνώμη για τις λέξεις και μικρές φράσεις. Μερικά από τα πιο γνωστά λεξικά παρουσιάζονται στον παρακάτω πίνακα. Αυτές οι συναισθηματικές λεξικά μπορεί να χρησιμοποιηθεί για να προβλέψει το συναίσθημα που εκφράζεται σε ένα έγγραφο ή μια πρόταση, με τον προσδιορισμό των οιηματίας λέξεις που περιέχουν και να υπολογίζουν το τελικό συναίσθημα με τη

Affective Lexica	Source
General Inquirer	http://www.wjh.harvard.edu/~inquirer/
WordNet-Affect	http://wndomains.fbk.eu/wnaffect.html
Senti-WordNet	Esuli and Sebastiani (2006)
Affective Norms for English Words	Bradley and Lang (1999)
Dictionary of Affect of Language	http://www.hdcus.com/
Bing Liu Opinion Lexicon	http://www.cs.uic.edu/~liub/FBS/
AFINN	Nielsen (2011)

Table 2.2: List of most known existing affective dictionaries

βοήθεια απλών αλγορίθμων βασίζεται σε κανόνες (Zhu et al. (2009)). Στο έργο Taboada et al. (2011) χρησιμοποιήσετε το Σημασιολογικό Προσανατολισμός Υπολογιστής (ΣΟ-ΎΑΛ), η οποία εκμεταλλεύεται λεξικών των λέξεων σχολιασμένη με σημασιολογικό προσανατολισμό τους (πολικότητα και δύναμη), και ενσωματώνει την εντατικοποίηση και την άρνηση, προκειμένου να προβλέψουμε το συναίσθημα στο κείμενο. Σε γενικές γραμμές, όλες οι προσεγγίσεις τείνουν να χρησιμοποιούν λεξικά συναίσθημα που διαθέτει για να εξαγάγετε νέες βαθμολογίες πολικότητα για αόρατο λέξεις. Η προσέγγιση του Malandrakis et al. (2011a) οποίο χρησιμοποιεί μια υπάρχουσα λεξιλόγιο όχι μόνο να προβλέψει λόγια βαθμολογίες, αλλά και για να καθοριστεί η πολικότητα των ποινών αναλύεται περαιτέρω στο κεφάλαιο 4. Τέλος, Jijkoun et al. (2010) παρουσιάσει μια μέθοδο για την αυτόματη δημιουργία του θέματος ειδικά λεξικά υποκειμενικότητας από ένα γενικού σκοπού πολικότητα λεξικό που επιτρέπουν στους χρήστες να γίνουν αντιληπτά σημείων υποκειμενική επί του θέματος πληροφοριών σε ένα σύνολο των σχετικών εγγράφων.

2.3 Διανεμητικά Σημασιολογικά Μοντέλα

Harris ήταν ο πρώτος στην ερώτηση αν είναι δυνατόν να διαμορφώσει τη γλώσσα ως διανομή κερδών (Harris (1954)). Στο έγγραφό του, ο ίδιος υπογραμμίζει ότι η γλώσσα περιέχει διάφορα αντικείμενα (π.χ. λέξεις, φράσεις) με συγκεκριμένες ιδιότητες (χαρακτηριστικά) που μπορούν να κωδικοποιηθούν σε διανεμητική δομή, μετρώντας την εμφάνιση κάποιων αντικειμένων με τους άλλους. Ως εκ τούτου, η κατανομή ενός στοιχείου μπορεί να παρασταθεί ως το άθροισμα των άλλων στοιχείων που περιβάλλει, δηλ ως φορέας του οποίου κάθε θέση αντιστοιχεί σε πόσες φορές αυτά τα στοιχεία, όπου μάρτυρες μαζί (στο ίδιο περιβάλλον). Αυτά είναι γνωστά ως μοντέλα διάνυσμα χώρου . Κάποιος μπορεί να υποθέσει, ότι αν το μοντέλο λέξεις ως φορείς αυτοί οι αναπαραστάσεις μπορούν να εκφράσουν νοήματα τους. Ωστόσο, αυτή η παράσταση είναι μια ιδιότητα της γλώσσας, ενώ το νόημα μιας λέξης είναι η ερμηνεία ενός ατόμου, και μπορεί ενδεχομένως να αλλάξει. Αναδιανεμητικές σημασιολογία ευνοούν τη

χρήση της γραμμικής άλγεβρας ως υπολογιστικό εργαλείο και αναπαράστασης πλαισίου. Η βασική προσέγγιση είναι η συλλογή πληροφοριών κατανομής σε φορείς υψηλών διαστάσεων, και να καθορίσει κατανομής / σημασιολογικής ομοιότητας από άποψη ομοιότητας φορέα. Όπως αναφέρθηκε προηγουμένως, η κύρια παραδοχή της ΔΣΜς είναι ότι ‘ Ομοιότητα πλαίσιο συνεπάγεται ομοιότητας του νοήματος ». Με την αξιοποίηση των ιδιοτήτων της γραμμικής άλγεβρας, πολλαπλές λειτουργίες μπορεί να εφαρμοστεί στις ή μεταξύ φορέων, προκειμένου να εξαχθούν διαφορετικά είδη πληροφοριών. Υπάρχει μια πλούσια ποικιλία των υπολογιστικών μοντέλων εφαρμογής αναδιανεμητική σημασιολογία. Στην έρευνα του Verma et al. (2011) Vector Space Μοντέλα εξετάζονται για γενική κειμένου σημασιολογική επεξεργασία, με την άμεση εξόρυξη γλωσσικά χαρακτηριστικά από μια συλλογή κειμένων. Το μέτρο της ομοιότητας μεταξύ δύο λέξεων μπορεί να υπολογιστεί μέσω της ομοιότητας των δύο διανυσμάτων (σειρές) στη μήτρα. Baroni and Lenci (2010) περιγράφουν λεπτομερώς την κατασκευή των διανεμητικών σημασιολογικών μοντέλων ως διανεμητικές πλαίσιο μνήμη που είναι σε θέση να εξάγει όλες τις σημασιολογικές πληροφορίες από το σώμα και να το αποθηκεύσει σε τρίτες τανυστές σειρά. Δείχνουν ότι αυτό το είδος των πληροφοριών που μπορούν να χρησιμοποιηθούν σε διάφορες εργασίες και τα αποτελέσματα της αξιολόγησης συγκρίνουν ΔΣΜς με state-of-the-art μεθόδων για συγκεκριμένα καθήκοντα. Στην ίδια κατεύθυνση, Iosif and Potamianos (2010) απεικονίζουν μια μέθοδο για να υπολογίσουμε την σημασιολογική ομοιότητα μεταξύ των λέξεων σε μια ανεξέλεγκτη τρόπο με τη βοήθεια του ωεβ-συγκομισθεί δεδομένων. Επιπλέον, οι διαφορετικοί τύποι των χαρακτηριστικών της έρευνας, που δείχνουν ότι μετρώντας τον αριθμό των συν-εμφανίσεις των λέξεων σε ένα παράθυρο (λίγα λόγια πριν και μετά τη λέξη ενδιαφέροντος - πλαίσιο μέθοδος που βασίζεται) παρέχει καλύτερη απόδοση κατά τη μέτρηση της συσχέτισης με την ανθρώπινη σχολιασμούς, σε σύγκριση να υπολογίζουμε σε μεγαλύτερες ενότητες (μέθοδος συνύπαρξη βάση). Τα τελευταία χρόνια, όλο και περισσότερες διαφορετικές μεθόδους εισήχθη για διανεμητικών μοντέλα κατασκευής. Εκτός από τις πιο κοινές μεθόδους καταμέτρηση με βάση το πλαίσιο, τα πιο δημοφιλή χρήση νευρωνικών δικτύων, όπως το μοντέλο Word2Vec (Mikolov et al. (2013)) και άλλες προσεγγίσεις που βασίζονται σε νευρωνικά δίκτυα (Zeng et al. (2014), Mnih and Kavukcuoglu (2013)) για την αξιολόγηση της ομοιότητας και άλλη σχέση μεταξύ των λέξεων. Το μοντέλο γάντι (Pennington et al. (2014)), είναι ένα μοντέλο παλινδρόμησης λογ-διγραμμική που συνδυάζει τα πλεονεκτήματα των δύο μεγάλων οικογενειών μοντέλο στη βιβλιογραφία: παγκόσμια παραγοντοποίηση της μήτρας και των τοπικών μεθόδων παράθυρο πλαίσιο.

2.4 Τεχνικές Θεματικής Μοντελοποίησης

Στους τομείς της μηχανικής μάθησης και επεξεργασία φυσικής γλώσσας, ένα μοντέλο το θέμα (TM) θεωρείται ως ένα είδος στατιστικής-παραγωγική μοντέλο που βοηθά στην ανακάλυψη αφηρημένη ‘ θέματα »που συμβαίνουν σε μια μεγάλη συλλογή εγγράφων. Στην βιβλιογραφία, μια παραγωγικό μοντέλο είναι ένα Bayesian δίκτυο το οποίο παρέχει μια περιγραφή ενός παρατηρείται το φαινόμενο. περιγράφει τις εξαρτήσεις μεταξύ των τυχαίων μεταβλητών σε ένα δίκτυο, μαζί με τη σχέση αιτίου και αποτελέσματος. μοντέλων θέμα είναι πιθανοτικά μον-

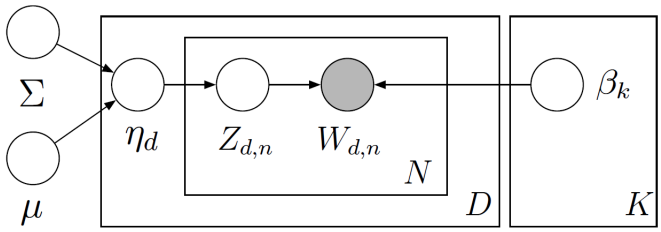
τέλα παραγωγική. Προσπαθούν να διαμορφώσει μια πηγή κειμένου και συνήθως η διαδικασία παραγωγής αυτής της πηγής μέσω ενός Bayesian δικτύου. Διαισθητικά, δεδομένου ότι ένα έγγραφο είναι για ένα συγκεκριμένο θέμα, θα περίμενε κανείς τα λόγια συγκεκριμένους τομείς να εμφανίζεται στο έγγραφο περισσότερο ή λιγότερο συχνά. Από την άλλη πλευρά, είναι προφανές ότι κάθε έγγραφο αποτελείται από πολλαπλές θέματα σε διαφορετικές αναλογίες. Έτσι, το θέμα μοντελοποίηση, προσπαθεί να συλλάβει τις πληροφορίες που περιέχονται σε μεγάλες συλλογές κειμένων με τη βοήθεια των στατιστικών στοιχείων. Η πιο γνωστή από τις τεχνικές που είναι, βασίζονται στην παραδοχή της ανταλλακτικής ικανότητας (Aldous (1985)): Η σειρά των λέξεων και των εγγράφων μπορεί να αγνοηθεί. Αυτές οι τεχνικές που περιλαμβάνονται επίσης στο έγγραφο του Blei (2012) και παρουσιάζονται παρακάτω. Το πρώτο μοντέλο που θα μελετηθεί ήταν η Λανθάνουσα Σημασιολογική Ευρετηρίαση (ΛΣΙ), επίσης γνωστή ως Λανθάνουσα Ανάλυση Σημασιολογικού (ΛΣΑ). Αυτό αναλύθηκε με Papadimitriou et al. (1998) και ήταν αρχικά χρησιμοποιηθεί σε εργασίες εξαγωγής πληροφορίας. Στοχεύει στη δημιουργία φορέων εγγράφων για να εκπροσωπήσει τη σημασιολογική τους πλαίσιο και πιο συγκεκριμένα, για τον εντοπισμό κρυφών σχέσεων μεταξύ των εγγράφων και των λέξεων. Τα βήματα αυτού του αλγορίθμου είναι τα εξής: Κατ άρχάς, μια 2-διαστάσεων μήτρα δημιουργείται, όπου οι σειρές αντιστοιχούν στις λέξεις σε μια συλλογή εγγράφων (α λεξιλόγιο) και οι στήλες αντιστοιχούν στα έγγραφα αυτής της συλλογής. Κάθε στοιχείο στη μήτρα έχει μια τιμή συχνότητας που αντιπροσωπεύει τον αριθμό των φορών που υπάρχει αυτή η λέξη σε κάθε έγγραφο. Αυτή η αναπαράσταση των εγγράφων (ως φορείς των συχνοτήτων λέξης) ονομάζεται τσάντα-οφ-λέξεις μοντέλο (ΒΟΩ). Η υπόθεση όπου ΛΣΑ βασίζεται στο ότι είναι λέξεις που είναι σημασιολογικά σχετικές, τείνουν να συνυπάρχουν σε παρόμοια κομμάτια του κειμένου. Στη συνέχεια, μια διάσταση μειώνοντας γραμμική προβολή, την τεχνική αποσύνθεση ιδιαζουσών ιδιοτιμών (Σ˝Δ), εφαρμόζεται για τη μείωση της διάστασης των γραμμών ή των στηλών, διατηρώντας την άλλη διάσταση. Υποτίθεται ότι η ομοιότητα μεταξύ εγγράφων ή λέξεις είναι καλύτερα εκτιμάται στο λανθάνουσα χώρο από το αρχικό, όπως όταν λέξη ή έγγραφο φορείς ομαδοποιούνται μια κρυμμένη έννοια υπονοείται. Λογικά, ‘ τα έγγραφα που μοιράζονται συχνά συνυπάρχουσες όροι θα έχουν παρόμοια αντιπροσώπευση στη λανθάνουσα χώρο, ακόμη και αν δεν έχουν άποψη στην κοινή », όπως αναφέρεται ρητά στην δημοσίευση. Με άλλα λόγια, ΛΣΑ κατορθώνει να μειώσει το θόρυβο σε ένα κείμενο και να εντοπίσει λέξεις που αναφέρονται στο ίδιο θέμα. Αν και αυτή η τεχνική ήταν επιτυχής σε τομείς όπως η ευρετηρίαση (Deerwester

$$
\begin{array}{cccccccc}
 & \boldsymbol{X} & & \boldsymbol{U} & & \boldsymbol{\Sigma} & & \boldsymbol{V}^T \\
 & (\mathbf{d}_j) & & & & & & (\hat{\mathbf{d}}_j) \\
 & \downarrow & & & & & & \downarrow \\
(\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} & = & (\hat{\mathbf{t}}_i^T) \rightarrow & \left[\begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} \dots \begin{bmatrix} \\ \mathbf{u}_l \\ \\ \end{bmatrix} \right] & \cdot & \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{bmatrix} & \cdot & \begin{bmatrix} [& \mathbf{v}_1 &] \\ & \vdots & \\ [& \mathbf{v}_l &] \end{bmatrix}
\end{array}
$$

Σχήμα 2.3: Singular Value Decomposition technique used in LSI for dimensionality reduction.

et al. (1990)) αντιμετώπισε προβλήματα λόγω της ασταθούς στατιστικό υπόβαθρο του. Ως εκ τούτου, ήταν αργότερα βελτιώθηκε στο Πιθανοθεωρητικές Λανθάνουσα Σημασιολογική Ευρετηρίαση (PLSI) ή Πιθανοθεωρητικές Λανθάνουσα Σημασιολογική Ανάλυση (PLSA), από Hofmann (1999). Είναι βασική ιδέα είναι ότι ένα έγγραφο μπορεί να θεωρηθεί ως ένα μείγμα μοντέλο, όπου τα συστατικά του μίγματος είναι multinomial τυχαίες μεταβλητές και μπορεί να ερμηνευθεί ως θέματα. Ως εκ τούτου, κάθε λέξη μπορεί να προέρχονται από ένα θέμα και διαφορετικές λέξεις σε ένα έγγραφο μπορεί να παραχθεί από διάφορα θέματα. Είναι γενικά μια μέθοδος που μπορεί να παρατηρηθεί από δύο διαφορετικές οπτικές γωνίες. Πρώτον, ως λανθάνουσα Μεταβλητή Model, του οποίου η δομή βασίζεται σε ένα στατιστικό μοντέλο, το οποίο ονομάζεται μοντέλο πτυχή. Οι κρυφές μεταβλητές αυτού του μοντέλου αντιστοιχούν σε θέματα ή έννοιες και σχετίζονται με τις παρατηρούμενες μεταβλητές που αντιπροσωπεύουν τα έγγραφα και τα λόγια σε μια συλλογή κειμένων. Κάθε έγγραφο αναπαρίσταται ως ένα πιθανοτικό μίγμα μυλτινομιαλ μεταβλητές και συνεπώς εκφράζεται ως διανομή πάνω θέματα. Δεύτερον, ως τεχνική Matrix παραγοντοποίηση. Ομοίως για ΛΣΑ, που αποσκοπεί στη μείωση της διάστασης της μήτρας συνύπαρξη μεταξύ των λέξεων (σειρές) και έγγραφα (στήλες), που ονομάζεται το έγγραφο διάρκειας μήτρα. Ωστόσο, PLSA βασίζεται σε μια πιο σταθερή στατιστική και πιθανοτική ερμηνεία από ΛΣΑ, που χρησιμοποιεί μόνο έναν μαθηματικό τύπο. Στο σχήμα ;; η δομή της πτυχής-μοντέλο PLSA εμφανίζεται. Εν ολίγοις, PLSA είναι μια παραγωγική διαδικασία για έγγραφα που επιλέγει πρώτον ένα έγγραφο (από τη συλλογή), με μια ορισμένη πιθανότητα, και για κάθε λέξη σε αυτό το έγγραφο επιλέγει 1) ένα θέμα από μια πολυωνυμική κατανομή δεδομένο το έγγραφο και 2) λέξη από μια πολυωνυμική κατανομή δεδομένο το θέμα. παραμέτρων του μοντέλου μπορεί να εκτιμηθεί με την εύρεση των τιμών που μεγιστοποιούν την προγνωστική πιθανότητα για τις παρατηρούμενες εμφανίσεις λέξη μέσω του αλγορίθμου Προσδοκία Μεγιστοποίηση (EM). Από την άλλη πλευρά, εάν η δεύτερη προοπτική προτιμάται, προκειμένου να μειωθεί το μέγεθος του συν-περιστατικό έγγραφο διαρκείας μήτρας (N χ M), μια τεχνική αποσύνθεση εφαρμόζεται σε τρεις διαφορετικές μήτρες: ϒ μήτρα με φορείς λέξη, Δ διαγώνιος πίνακας με ιδιόμορφες τιμές και φορείς έγγραφο ˝. Ωστόσο, PLSA δεν παρέχει καμία πιθανολογική μοντέλο σε επίπεδο εγγράφου. Αυτό σημαίνει ότι καμία παραγωγική διαδικασία λαμβάνει χώρα προκειμένου να εξαχθούν οι αναλογίες μίγματος που αποτελούν το έγγραφο (π.χ. ένα έγγραφο είναι 50 % για τα ζώα, το 30 % για τα τρόφιμα, το 20 % για τα αυτοκίνητα). Ως αποτέλεσμα, πάσχει από 2 σημαντικά προβλήματα: α) ο αριθμός των παραμέτρων αυξάνεται γραμμικά με το μέγεθος σώμα, οδηγώντας σε overfitting προβλήματα και β) δεν υπάρχει καμία ένδειξη για το πώς να αποδίδεται πιθανότητα σε νέα αθέατες έγγραφα. Για το λόγο αυτό, η πιο γνωστή τεχνική για θέμα Μοντελοποίηση εισήχθη από Blei et al. (2003), η λανθάνουσα Διριχλετ Κατανομή (ΛΔΑ). ΛΔΑ προσπαθεί να ξεπεράσει τα προβλήματα της PLSA με τη μετατροπή των μεταβλητών μοντέλα σε κρυφό τυχαίες μεταβλητές. Με αυτόν τον τρόπο, αυτές οι μεταβλητές που δεν συνδέονται με το εκπαιδευμένο σετ και, κατά συνέπεια, δεν περιορίζονται. Επιπλέον, υπάρχει μια αλλαγή στις κατανομές που είναι τώρα Διριχλετ. Ως εκ τούτου, ΛΔΑ επιτυγχάνει καλύτερη γενίκευση και αποφεύγει υπερπροσαρμογής. Για μια πιο λεπτομερή ανασκόπηση της ΛΔΑ βλέπε Κεφάλαιο 3. Η επέκταση της ΛΔΑ ήρθε τέσσερα χρόνια αργότερα και πάλι με Blei and Lafferty (2007),

η Συναφές θέμα Μοντέλο (ΈΤΜ) και το μηχάνημα Παςηιγκο Κατανομή (Li and McCallum (2006)). Αυτό το μοντέλο έχει στόχο να βελτιώσει ΛΔΑ που δεν μοντελοποιεί τη συσχέτιση του περιστατικού μεταξύ θέματα. Αν και LDA υποθέτει ότι τα θέματα που καθορίζονται πριν από την παραγωγή των εγγράφων, και δεν υπάρχει καμία συσχέτιση μεταξύ τους, η διαίσθηση πίσω από CTM είναι ότι μία λανθάνουσα θέμα μπορεί να συσχετιστεί με την παρουσία του άλλου. Ως παράδειγμα, ένα θέμα για το ' διακοπές ' είναι πιο πιθανό να είναι, επίσης, για ' εστιατόρια ' από περίπου ' γενετική '. Μια σημαντική διαφορά μεταξύ CTM και LDA είναι ότι οι αναλογίες θέμα προέρχονται από μια λογιστική κανονικό πριν και όχι ένα Dirichlet. Καθώς η Διριςηλετ είναι μια διανομή που υποθέτει ότι τα συστατικά είναι σχεδόν ανεξάρτητη, η υλικοτεχνική κανονική μπορεί να μοντελοποιήσει την εξάρτηση μεταξύ των συσκευών. Κάτω από αυτό το σκορ, το CTM είναι σε θέση να συλλάβει όλα τα πιθανά θέματα που δεν είναι ανεξάρτητες και ως εκ τούτου υποστηρίζει περισσότερα θέματα.



Σχήμα 2.4: Graphical model representation of the CTM (from Blei and Lafferty (2007)).

Οι άλλες επεκτάσεις του ΛΔΑ παρέχεται μέσω των ετών. Σε εργασίες όπου η σειρά των λέξεων σημασία σε ένα έγγραφο (όπως η παραγωγή γλώσσα) η υπόθεση των αμελητέα σειρά των λέξεων δεν είναι κατάλληλη. Μια πιθανή λύση εισήχθη από Griffiths et al. (2004) ως ένα μοντέλο που εναλλάσσει ΛΔΑ και ένα πρότυπο Ηιδδεν Μαρκο Μοδελ (HMM). Παρά το γεγονός ότι ο χώρος παράμετρος είναι μεγαλύτερο, η απόδοση μοντελοποίηση γλώσσα βελτιώνεται. Επιπλέον, όταν πρέπει να αναλυθεί, όπου τα θέματα αλλάζουν με την πάροδο του χρόνου μακροχρόνια συλλογές, ένα δυναμικό μοντέλο θέμα εισήχθη από Blei and Lafferty (2006). Το θέμα είναι τώρα μια σειρά από διανομές πάνω από λέξεις, και είναι εύκολο να παρακολουθείτε πώς αλλάζει την πάροδο του χρόνου. Για την περίπτωση όπου πρέπει να ανακαλύψει ο αριθμός των θεμάτων (που δεν θεωρείται γνωστό και σταθερό από την αρχή ως το απλό ΛΔΑ), ένα Βαψεσιαν μοντέλο μη παραμετρική θέμα διαμορφώθηκε από Teh et al. (2006), η διαδικασία Ιεραρχική Διριςηλετ (HDP) . Σε αυτό το μοντέλο, ο αριθμός των θεμάτων καθορίζεται από τα δεδομένα κατά τη διάρκεια της οπίσθιας εξαγωγής συμπερασμάτων και νέα έγγραφα μπορεί να παρουσιάζουν προηγουμένως αόρατο θέματα. Για να χαλαρώσετε την υπόθεση ότι κάθε λέξη είναι πιθανό να συμβεί σε οποιοδήποτε θέμα, ένα σφαιρικό μοντέλο θέμα επιτρέπει λέξεις για να είναι απίθανο σε ένα θέμα (Reisinger and Mooney (2010)). Σε πολλές ρυθμίσεις κειμένου που πρέπει να συμπεριληφθούν πρόσθετα μετα-δεδομένα, άλλες τεχνικές έχουν εισαχθεί. Για κάθε λέξη σε ένα έγγραφο που να επηρεάζει την πιθανότητα ότι η λέξη εμφανίζεται στο έγγραφο, ' εκρηκτική ' μοντέλα θέμα έχουν αναπτυχθεί από Doyle and Elkan (2009). Αραιή μοντέλα θέμα που επιβάλλουν τη δομή στις διανομές θέμα (Wang

and Blei (2009)), μοντέλα Ρελατιοναλ Θέμα (Chang and Blei (2010)) υποθέτουν ότι κάθε έγγραφο διαμορφώνεται ως το ΛΔΑ και ότι οι δεσμοί μεταξύ εγγράφων εξαρτάται από την απόσταση μεταξύ τους αναλογίες και, τέλος, Συγγραφέας-θέματος μοντέλα (Rosen-Zvi et al. (2004)) επιτρέπουν συμπέρασμα σχετικά με τους συγγραφείς, καθώς και τα έγγραφα.

2.5 Συναισθηματική Ανάλυση χρησιμοποιώντας ΘΜ

Οι προαναφερθείσες προσεγγίσεις για την ανάλυση συναισθήματος σε επίπεδο πρότασης, ιδίως εκείνων που χρησιμοποιούν συναισθηματική λεξικά, δεν περιλαμβάνουν πληροφορίες σχετικά με το πλαίσιο όπου εμφανίζονται οι λέξεις μιας πρότασης. Για να ξεπεραστεί αυτό το πρόβλημα η έρευνα επικεντρώθηκε στο πώς μπορούν να χρησιμοποιηθούν τα θέματα και το θέμα μοντέλα για να βοηθήσουν την εκτίμηση του εξέφρασε το συναίσθημα της ποινής από το συνδυασμό των δύο. Στο έργο του **?** συναίσθημα και το θέμα ανιχνεύονται μέσω ερωτήματα των χρηστών. Συνδυάζουν τα μοντέλα ενδιαφέρον συναίσθημα και μοντέλα θέμα συνάφεια με παραμέτρους του μοντέλου υπολογίζεται από δεδομένα εκπαίδευσης, λαμβάνοντας υπόψη το θέμα της εξάρτησης του συναισθήματος. Η πιο γνωστή προσέγγιση η οποία ήταν σε θέση να ανιχνεύσει όχι μόνο θέματα, αλλά και συναισθήματα σε κριτικές και εισήχθη από **?**. Στην προσέγγισή τους, προσπαθούν να εντοπίζουν τα θέματα σε ένα άρθρο, συνδέουν κάθε θέμα με τη σωστή πολικότητα συναίσθημα, και το μοντέλο κάθε θέμα με τις αντίστοιχες συναισθήματα του, χρησιμοποιώντας ένα μοντέλο θέμα-συναίσθημα-μείγμα. Βασίζονται στην υπόθεση ότι ένα έγγραφο μπορεί να περιέχει διάφορα θέματα και κάθε θέμα αποτελείται από διαφορετικά συναισθήματα χρησιμοποιώντας πολυωνυμική κατανομή. Αναλυτικότερα, το χωρίζει τις λέξεις στο έγγραφο σε δύο κύριες κατηγορίες: κοινές λέξεις της αγγλικής γλώσσας και το θέμα λέξεις, ενώ ο πρώην μπορεί να είναι θετική, ουδέτερη ή αρνητική. Με βάση αυτή την παραγωγική διαδικασία ο αλγόριθμος είναι σε θέση να 1) να μάθουν γενικά μοντέλα συναίσθημα (ένα για θετική και μία αρνητική γνώμη), 2) Απόσπασμα μοντέλα θέμα και καλύψεις συναίσθημα, 3) το μοντέλο του κύκλου και το συναίσθημα της δυναμικής θέμα τη ζωή. Στο έργο του **?** κοινή μοντέλο συναίσθημα και θέματα (ΘΣΤ) προτείνεται. Σε αυτό το μοντέλο, υπάρχουν διακριτές θέμα και το συναίσθημα ετικέτες. Με τη βοήθεια του Διριχλετ διανομές, οι δύο επιπλέον λανθάνουσες μεταβλητές ενσωματωθεί και η παραγωγική διαδικασία είναι ως εξής: Για κάθε έγγραφο, επιλέξτε μια κατανομή με παράμετρο ‘γ’, για κάθε ετικέτα συναίσθημα επιλέξετε μια κατανομή με παράμετρο ‘α’, για κάθε λέξη στο έγγραφο, επιλέξτε μια ετικέτα συναίσθημα, ένα θέμα και μια λέξη από τη διανομή πάνω από λέξεις που ορίζεται από το θέμα και την ετικέτα συναίσθημα. Αυτό το μοντέλο είναι εντελώς χωρίς επίβλεψη και δεν απαιτεί επισημασμένο δεδομένων.

Ένα άλλο μοντέλο που προτείνει **?** είναι ένα μοντέλο ενοποίησης πτυχή-συναίσθημα. Αναλύει το πρόβλημα του πώς εκφράζονται συναισθήματα για διαφορετικές πτυχές. Πρώτον ένα μοντέλο πιθανολογική παραγωγική Ποινή-LDA εισάγεται η οποία υποθέτει ότι όλες οι λέξεις σε μια μόνο φράση που παράγεται από μία άποψη. Στη συνέχεια, μια επέκταση αυτού του μοντέλου γίνεται, το μοντέλο ενοποίησης Aspect-Κλίματος (Asum), η οποία ενσωματώνει την πτυχή και το συναίσθημα μαζί με το μοντέλο αισθήματα προς διάφορες πτυχές. Είναι παρόμοιο

με το μοντέλο JST, αλλά περιορίζει λόγια να προέρχονται από το ίδιο μοντέλο γλώσσα. **?**, προτείνει δύο μοντέλα συναίσθημα θέμα να συνδέσει λανθάνουσα θέματα με προκλητών τα συναισθήματα των αναγνωστών. Το πρώτο μοντέλο που αποτελεί επέκταση της υφιστάμενης Εποπτευόμενοι Θέμα μοντέλο, δημιουργεί μια σειρά από θέματα από τα λόγια, ακολουθούμενο από δειγματοληψία συναισθήματα από κάθε θέμα. Το δεύτερο μοντέλο δημιουργεί θέματα από την κοινωνική συναισθήματα άμεσα. Τέλος, **?** είναι το πιο παρόμοιο έγγραφο με αυτό το έργο, καθώς επίσης να χρησιμοποιήσετε ένα πρότυπο μείγμα. Οι κύριες διαφορές είναι: 1) ότι το έργο μας, χρησιμοποιείται ένα μοντέλο μείγμα ομοιότητα (όχι μια συναισθηματική ένα) και στη συνέχεια το συναίσθημα προβλέπεται, χρησιμοποιώντας μια υπάρχουσα συναισθηματική λεξικό, 2) Η ομαδοποίηση βασίζεται στην LDA και όχι στο K-means και 3) έχουμε εκπαιδευτεί διαφορετικές σημασιολογικές μοντέλα και όχι SVMs πράγμα που σημαίνει ότι η προσέγγισή μας είναι χωρίς επίβλεψη, μέχρι τη χρήση του λεξικού.

Κεφάλαιο 3

Θεματική Μοντελοποίηση

3.1 Θεωρητική Εξήγηση

Όπως ο κύριος στόχος αυτής της εργασίας είναι να μην διερευνήσει σε βάθος πώς λειτουργεί το LDA, αλλά βασικά να κατανοήσουν τον τρόπο που λειτουργεί για το έργο μας, θα περιγράψουμε είναι η διαίσθηση και τη γνωστική κίνητρο πρώτον [1]. LDA είναι ένα αλγόριθμος που προτάθηκε από τον Blei et al. (2003) που εξάγει θεματικές ενότητες από συλλογές κειμένων. Ας σκεφτούμε το παρακάτω παράδειγμα.

Παράδειγμα 3.1. *Assume the we have the following sentences:*
1. "I like to eat fish and vegetables"
2. "Cats eat fish"
3. "Fish are pets"

LDA είναι σε θέση να ταξινομήσει τις λέξεις αυτές καταδίκη σε θέματα. Κάποιος μπορεί να παρατηρήσετε ότι οι πράσινες λόγια μιλούν για 'τρόφιμα' και οι μπλε λέξεις είναι περίπου 'κατοικίδια ζώα'.

So, overall these sentences contain the following topics:
Sentence 1: 100% food
Sentence 2: 66.7% pets, 33.3% food
Sentence 3: 100% pets

And each topic can contain the following words:
Topic 'food': 40% eat, 40% fish, 20% vegetables
Topic 'pets': 80% fish, 20% cats

Τώρα, σε ένα θεωρητικό τρόπο, μπορούμε να περιγράψουμε πώς λειτουργεί αυτή η διαδικασία. Ας φανταστούμε ότι έχουμε μια μεγάλη συλλογή εγγράφων και θέλουμε να προσδιορίσει τα διάφορα θέματα, θέματα ή έννοιες που περιέχει. Ας υποθέσουμε ότι κάθε έγγραφο που

[1] ηττπς://τεδυνδερωοοδ.ςομ/2012/04/07/τοπις-μοδελινγ-μαδε-θυστ-σιμπλε-ενουγη/

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Figure 3.1: Example of Topics produced by LDA from Steyvers and Griffiths (2007).

περιγράφει πολλαπλά θέματα (όπως το ‘πρόταση 2’) στο προηγούμενο παράδειγμα) και ότι κάθε θέμα περιέχει λέξεις με ορισμένες πιθανότητες, σύμφωνα με την εμφάνισή τους στο θέμα (μεγαλύτερη πιθανότητα να δείχνει πιο συχνές / κοινή λέξη για το θέμα). Στο προηγούμενο παράδειγμα, το θέμα 1 μιλά για τα κατοικίδια ζώα και έτσι η λέξη ‘ψάρια’ θα είναι πολύ κοινά σε όλα τα έγγραφα που περιέχουν αυτό το θέμα και, κατά συνέπεια, το θέμα αυτό θα παράγει τη λέξη ‘ψάρια’ με μεγάλη πιθανότητα. Επιπλέον, οι ίδιες λέξεις μπορεί να ανήκει σε περισσότερα από ένα θέμα (‘ψάρια’ και πάλι σύμφωνα με το παράδειγμα, ανήκει σε δύο θέματα). Φυσικά είναι αδύνατο να γνωρίζουμε από την αρχή που είναι αυτά τα θέματα, σε περίπτωση χιλιάδες έγγραφα. Ο στόχος του LDA είναι να κάνουν πίσω τη διαδικασία της γραφής ένα έγγραφο και να βρείτε τα θέματα που δημιουργείται αυτό το έγγραφο. Για παράδειγμα, σκεφτείτε 3.1. Αν κάποιος δίνει ίσες πιθανότητες στις δύο πρώτα θέματα, που θα δημιουργήσει το έγγραφο των συνομιλιών για ένα πρόσωπο του οποίου η αντίληψη των χρωμάτων επηρεάστηκε από τη μεγάλη κατανάλωση φαρμάκων. Η υπόθεση είναι ότι ο συγγραφέας έχει ήδη προ-αποφασίσει τα θέματα που πρόκειται να γράψουν. Ωστόσο, οι αστάθμητοι παράγοντες είναι πάρα πολλά για να εξαγάγετε άμεσα αυτά τα θέματα. Για να ξεπεραστεί αυτό το πρόβλημα, μπορούμε να υποθέσουμε ότι γνωρίζουμε ποια θέματος παράγει όλες τις λέξεις, εκτός από μια τυχαία λέξη β στο έγγραφο d. Το μόνο πράγμα που έχει απομείνει για να απαντήσει είναι, πώς μπορούμε να αποφασίσουμε για την οποία το θέμα ανήκει αυτή η λέξη. Για να απαντήσουμε σε αυτό το ερώτημα, μπορούμε να θεωρήσουμε α) η συχνότητα αυτής της λέξης σε ένα θέμα T (λέξεις μπορεί να συμβαίνουν συχνά σε περισσότερα από ένα θέμα) και β) πόσο κοινό είναι αυτό το θέμα T στο έγγραφο. Για παράδειγμα, αν η λέξη ‘τράπεζα’ είναι πολύ συχνή σε έγγραφα που μιλούν για ‘χρήμα’ (θέμα T), υπάρχει μια πιθανότητα ότι ‘τράπεζας ανήκει στο’ θέμα χρημάτων ¨. Ωστόσο, ‘τράπεζας μπορεί επίσης να συμβαίνουν συχνά σε ένα θέμα για το’ νερό ¨. Γι αυτό είναι πολύ σημαντικό να επιλέξετε το ‘θέμα χρημάτων’ εάν το έγγραφο πήραμε τη λέξη ‘τράπεζα’ από τις συνομιλίες περισσότερα για το ‘χρήματα’ από ό, τι ‘νερό¨. Συνοψίζοντας όλες αυτές τις σκέψεις μαζί, μια λογική προσέγγιση είναι η εξής: Για κάθε πιθανό θέμα T, πολλαπλασιάστε τη συχνότητα αυτής της λέξης b στο T από τον αριθμό των λέξεων στο έγγραφο d που ανήκουν ήδη σε T (χωρίς να λαμβάνεται υπόψη λέξη b). Τέλος, διαιρέστε με το συνολικό αριθμό των λέξεων στο έγγραφο d για να ληφθεί μια πιθανότητα. Το αποτέλεσμα

θα αντιπροσωπεύει την πιθανότητα ότι αυτή η λέξη προήλθε από T. Ο μαθηματικός τύπος περιγράφεται παρακάτω:

$$P(T, w|d) = \frac{\text{freq of word w in T} + \beta_w}{\text{total tokens in T} + \beta} \cdot \{\text{\# of words in d that belong to T} + \alpha\} \quad (3.1)$$

Για να γίνει αυτή η φόρμουλα περισσότερο κατανοητό, στην πραγματικότητα, η δεύτερη θητεία στο προϊόν, περιγράφει την αναλογία του θέματος T στο έγγραφο δ (π.χ. έγγραφο d είναι σχετικά με το θέμα T 20 %). Με αυτό τον τρόπο, τόσο η δύναμη του θέματος T στο έγγραφο d και της λέξης β στο θέμα T θεωρούνται. Οι παράμετροι β_w, β και α ονομάζεται ηψπερπαραμετερς και να εξασφαλίσει ότι υπάρχει κάποια πιθανότητα ότι η λέξη W ανήκει στο θέμα T, ακόμη και αν δεν υπάρχει σύνδεση μεταξύ τους. Αυτό σημαίνει ότι δεν θα υπάρξουν μηδέν πιθανότητες, μόνο ένα πολύ μικρό σε μια τέτοια περίπτωση. Στο τέλος, θα πρέπει να κάνουμε αυτή τη διαδικασία παραγωγική. Χρησιμοποιώντας τον τύπο 3.1 χρειαζόμαστε κάποιες αρχικές μεταβλητές. Αυτό το βήμα γίνεται τυχαία από απλά μαντέψουν όπου κάθε λέξη ανήκει. Στη συνέχεια, περνούν όλα τα έγγραφα της συλλογής, λέξη προς λέξη και τοποθέτησή του σε κάθε λέξη ένα νέο πιθανότητα για κάθε θέμα. Κατά τη διάρκεια αυτής της διαδικασίας, θα παρατηρήσετε ότι οι λέξεις που ήταν κοινή σε θέματα θα ομαδοποιηθούν, καθώς και θέματα που είναι κοινά σε έγγραφα. Ως αποτέλεσμα, τα θέματα των συγκεκριμένων λέξεων (πιθανότερος αυτοί) θα σχηματίζονται και γίνονται όλο και πιο συνεπής.

3.2 Στατιστική Εξήγηση

Εκτός από το την απλή-θεωρητική προσέγγιση που περιγράφεται στο προηγούμενο τμήμα, είναι αναγκαία μια περισσότερο τυπική. LDA είναι μέθοδος που μπορεί να δημιουργήσετε έγγραφα, δεδομένου συγκεκριμένα θέματα και αδόμητων κείμενο. Επιλέγοντας το παράδειγμα που προτείνεται στο Blei (2012) (Σχήμα 3.2), ένα άρθρο με τίτλο ‘ ναζητώντας Βαρε (Γενετις) ανάγκες της ζωής ¨είναι σχετικά με τη χρήση ανάλυσης δεδομένων για τον καθορισμό του αριθμού των γονιδίων ενός οργανισμού χρειάζεται για να επιβιώσει. εξετάζοντας προσεκτικά αυτό το άρθρο, μπορεί κανείς να εντοπίσει λόγια για ἀνάλυση των δεδομένων¨, ἐξελικτική βιολογία¨, ἀγενετική¨. Όπως παρατηρείται το έγγραφο αυτό περιέχει πολλά θέματα σε διαφορετικές αναλογίες και το γεγονός ότι οι συνομιλίες είναι για αυτά τα θέματα θα μπορούσε να μας βοηθήσει να ταξινομήσει ότι σε μεγαλύτερες συλλογές. LDA προσπαθεί να συλλάβει αυτή την διαίσθηση και την παραγωγή εγγράφων. στην στατιστική γλώσσα, κάθε έγγραφο αναπαρίσταται ως διανομή κερδών πάνω θέματα και κάθε θέμα ως διανομή πάνω από λέξεις. Φυσικά, το θέμα για το ἀγενετική’ περιέχει λέξεις που σχετίζονται με τη γενετική με υψηλή πιθανότητα, καθώς και το θέμα για το ἐξελικτική βιολογία’ περιέχει σχετικές λέξεις με υψηλή πιθανότητα μία βασική παραδοχή που γίνεται:. *ο αριθμός των θεμάτων έχει αποφασιστεί πριν έχει δημιουργηθεί τα δεδομένα*. Ή με άλλα λόγια, ότι ο συγγραφέας της συλλογής έχει αποφασίσει ποια τα θέματα θα είναι πριν άρχισε να γράφει. Επιπλέον, όλα τα έγγραφα στο μετοχικό συλλογή όλα τα θέματα, αλλά τα παρουσιάζουν σε διαφορετικές αναλογίες.

Περνώντας όλα τα κείμενα στη συλλογή η παραγωγική διαδικασία φαίνεται παρακάτω.

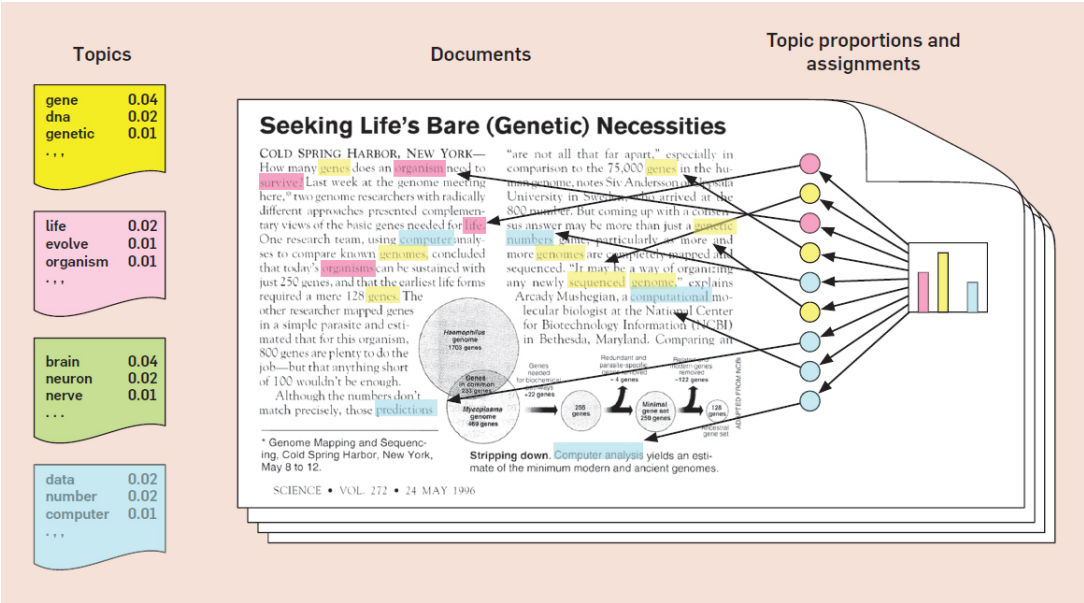


Figure 3.2: Explanation of the intuition behind LDA, Blei (2012).

Ο στόχος του LDA είναι να ανακαλύψουν αυτά τα θέματα σε μια συλλογή με αυτοματοποιημένο τρόπο. Οι παρατηρήσεις στο μοντέλο είναι τα έγγραφα, ενώ τα θέματα και πιο συγκεκριμένα οι κατανομές τους (έγγραφο-θέμα και θέμα-λέξη) είναι κρυμμένες μεταβλητές. Κατά συνέπεια, το πρόβλημα που πρέπει να λυθεί είναι πώς να προβλέψει την κρυμμένη διανομής με βάση τις παρατηρήσεις. Ή σε ένα πιο αφηρημένο τρόπο ' Τι είναι το κρυφό δομή που πιθανόν δημιουργείται το παρατηρηθεί συλλογή· ¨. Αυτό είναι ένα κλασικό πρόβλημα της εξαγωγής συμπερασμάτων. Σε πιθανοτικά μοντέλα θέμα τα δεδομένα αντιμετωπίζονται ως προήλθαν από μια παραγωγική διαδικασία που περιλαμβάνει κρυφές μεταβλητές. Αυτή η διαδικασία ορίζει μια κοινή κατανομή πιθανότητας πάνω τόσο την παρατηρούμενη και τα κρυμμένα τυχαίες μεταβλητές. Αυτή η κατανομή πιθανότητας χρησιμοποιείται για να συμπεράνει τις λανθάνουσες μεταβλητές και να δημιουργήσουν μια υπό όρους (ή οπίσθια) διανομής. Έτσι, το πρόβλημα συμπέρασμα είναι ένα υπολογιστικό πρόβλημα της εύρεσης αυτό εξαρτάται από τη διανομή των θεμάτων.

3.2.1 Συμβολισμός και Ορολογία

Εδώ ορίζουμε βασική ορολογία, προκειμένου να εξηγήσει τα βήματα του αλγορίθμου.

- One word is defined as a unigram from a certain vocabulary of size V
- A word is represented as a vector of zero elements, with 1 only to the position of this given word (e.g. $w_3 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & ... & 0 \end{bmatrix}$)
- We define a document as a group of N words $\mathbf{d} = \{w_1, w_2, ..., w_N\}$
- We define a corpus as a collection of M documents $\mathbf{C} = \{d_1, d_2, ..., d_M\}$

- A topic β_k is a distribution over the vocabulary
- θ_d is defined as the distribution of topics for document d and $\theta_{k,d}$ is the topic proportion of topic β_k in document d
- z_d is the topic-word distribution for document d and $z_{d,n}$ is the topic assignment for word w_n in document d

3.2.2 Αλγόριθμος

Για να περιγράψει επίσημα τον αλγόριθμο, η παραγωγική διαδικασία που LDA βασίζεται στο είναι η ακόλουθη διαδικασία:
For each document **d** in a corpus **C**,

1. Choose N $\sim$ Poisson(ξ)
2. Choose $\theta \sim$ Dir(α)
3. For each of the N words in the document:
   (a) Choose a topic $z_i \sim$ Multinomial(θ)
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a Multinomial probability conditioned on the topic z_n

The joint distribution of the hidden and observed variables is described below:

$$p(t_{1:T}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(t_i) \prod_{d=1}^{D} p(\theta_d) \Big(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|t_{1:K}, z_{d,n}) \Big) \tag{3.2}$$

Σύμφωνα με την ανωτέρω εξίσωση, παρατηρούνται πολλαπλές εξαρτήσεις. Η $Z_{d,n}$ εξαρτάται από το έγγραφο-θέμα της διανομής θ_d, η λέξη $w_{d,n}$ εξαρτάται από την ανάθεση θέματος $Z_{d,n}$ και όλα τα θέματα β. Αυτές οι εξαρτήσεις απεικονίζονται καλύτερα στο σχήμα ;;.

Distribution	Probability Mass Function (PMF)
Dirichlet	$\frac{1}{B(\alpha)} \prod_{i=1}^{M} x_i^{(\alpha_i - 1)}, B(\alpha) = \frac{\prod_{i=1}^{M} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{M} \alpha_i)}$
Poisson	$\exp^{-\lambda} \sum_{i=0}^{M} \frac{\lambda^i}{i!}$
Multinomial	$\frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^{M} p_i^{x_i}$

Table 3.1: Distributions used in LDA table.

3.2.3 Τεχνικές Εξαγωγής Συμπεράσματος

Για να υπολογίσουμε το οπίσθιο διανομή του θέματος δομής το Θεώρημα Bayes χρησιμοποιείται:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \tag{3.3}$$

Ο αριθμητής είναι η από κοινού διανομή όλων των τυχαίων μεταβλητών και ο παρονομαστής είναι η οριακή πιθανότητα των παρατηρήσεων, η οποία είναι η πιθανότητα να δει την παρατηρούμενη συλλογή των εγγράφων στο πλαίσιο οποιασδήποτε δομής θέμα. Ωστόσο, ο δυνατός αριθμός θέματος δομών είναι εκθετικά μεγάλος καθιστώντας έτσι δύσκολη την συμπεράνουμε τις λέξεις που σχετίζονται με ένα συγκεκριμένο θέμα και τα θέματα που συζητούνται σε ένα δεδομένο έγγραφο.

αλγορίθμων μοντελοποίησης θέμα προσπαθούν να βρουν τρόπους για να εκτιμηθεί η επιθυμητή οπίσθια πιθανότητα, ότι το λεγόμενο *μεθόδους προσέγγιση συμπέρασμα* (Σχήμα 3.3). Μπορούν γενικά εμπίπτουν σε δύο κατηγορίες: 1) αλγόριθμους δειγματοληψίας με βάση και 2) Μεταβολική αλγορίθμων.

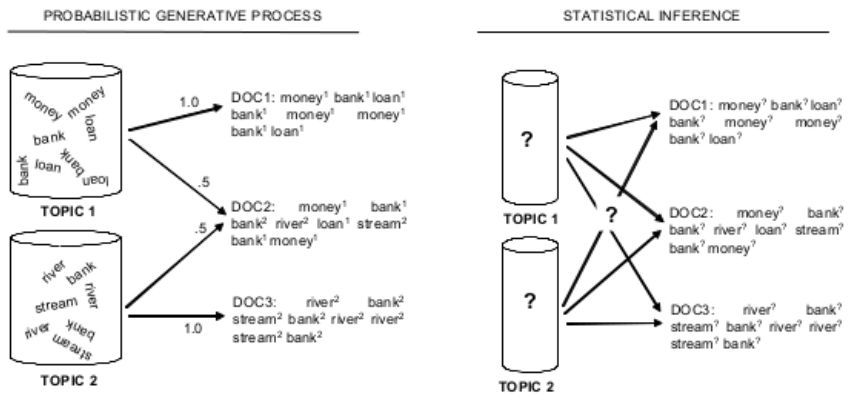


Figure 3.3: Generative process along with the problem of inference (source: Steyvers and Griffiths (2007))

Μεταβολική μεθόδους από την άλλη πλευρά, αντί να προσεγγίζει την οπίσθια με τα δείγματα, ας υποθέσουμε ότι υπάρχει μια οικογένεια κατανομών επί των κρυφών δομή (κατώτερο-όρια σχετικά με την πιθανότητα λογ) και βρείτε ποια μέλος είναι πλησιέστερα προς το επιθυμητό οπίσθια. Έτσι, το πρόβλημα συμπέρασμα είναι πλέον ένα πρόβλημα βελτιστοποίησης. Για να αποκτήσει αυτές τις οικογένειες, μπορεί κανείς να απλοποιήσει το μοντέλο πλάκα-συμβολισμός του σχήματος ;; αφαιρώντας τις άκρες μεταξύ θ, z και β. Αν το W κόμβοι απομακρύνονται, καθώς, το προκύπτον μοντέλο θα έχει ελεύθερη Μεταβολική παραμέτρους. Οι βέλτιστες τιμές που βρέθηκαν με την ελαχιστοποίηση της απόκλισης Kullback-Leibler (KL).

Κεφάλαιο 4

Σημασιολογικά & Συναισθηματικά Μοντέλα

4.1 Διανεμητικά Σημασιολογικά Μοντέλα

Οι ιδέες της τυπικής σημασιολογίας, συζητήθηκε επίσης στο κεφάλαιο 2, περιλαμβάνουν την απεικόνιση φράσεις, τις προτάσεις και τα λόγια όσον αφορά τα μοντέλα σετ-θεωρητική. Η κύρια διαίσθηση πίσω από σημασιολογία είναι ότι ο κόσμος αποτελείται από αντικείμενα με ιδιότητες και υπάρχουν σχέσεις μεταξύ τους. έχουν τα μοντέλα που αποτυπώνουν αυτήν την διαίσθηση έχουν αναπτυχθεί και αποδειχθεί ότι είναι χρήσιμο για την υπολογιστική σημασιολογία. Υπάρχει μια τάση να χρησιμοποιούν διανυσματικών χώρων, προκειμένου να εκπροσωπήσει την έννοια των ποινών, καθώς παρέχουν ένα φυσικό μηχανισμό για τον υπολογισμό της ομοιότητας και της απόστασης. Με τον τρόπο αυτό το βήμα της σύγκρισης λέξεων και μεγαλύτερες λεκτικές μονάδες γίνεται πολύ αποτελεσματική.

4.1.1 Ορισμός

Μοντέλα όπως αυτές που αναφέρθηκαν παραπάνω, θεωρούνται ένα είδος μοντέλων γνώσης. Αυτά τα μοντέλα, που έχουν αναπτυχθεί εδώ και δεκαετίες, είναι επίσης γνωστά ως χώροι φορέα, σημασιολογική χώρων ή χώρων λέξη. Όπως αναφέρεται στην Baroni and Lenci (2010) Αναδιανεμητικές Σημασιολογικού Μοντέλα (ΔΣΜς) στηρίζονται σε κάποια έκδοση της κατανομής υπόθεση (Harris (1954), **?**, **?**): ‘ λέξεις που εμφανίζονται σε παρόμοια περιβάλλοντα τείνουν να έχουν παρόμοια σημασία ’. Με άλλα λόγια, ‘ ο βαθμός της σημασιολογικής ομοιότητας μεταξύ δύο λέξεις μπορεί να μοντελοποιηθεί ως συνάρτηση του βαθμού επικάλυψης μεταξύ γλωσσικά πλαίσια τους¨. Μοντέλα χώρου Vector (VSMs) αντιπροσωπεύουν (εμβεδ) λέξεις σε ένα συνεχές διανυσματικό χώρο όπου σημασιολογικά παρόμοια οι λέξεις αντιστοιχίζονται σε κοντινά σημεία (ἐίναι ενσωματωμένα σε κοντινή απόσταση μεταξύ τους¨). Συνήθως, η (BOW) μορφή τσάντα-οφ-λέξεων που χρησιμοποιούνται για την αναπαράσταση φορέα. Αυτό σημαίνει ότι η σειρά των λέξεων σε ένα έγγραφο δεν έχει σημασία, μόνο τη συχνότητα εμφάνισής τους. Προφανώς, αυτό είναι μια αόριστη παραδοχή, δεδομένου ότι η έννοια μιας

λέξης είναι άρρηκτα συνδεδεμένη με τη σειρά των λέξεων γύρω από αυτό. Ωστόσο, αυτή η απλή προσέγγιση φαίνεται να έχει καλά αποτελέσματα, δεδομένου ότι συλλαμβάνει γενικά το πλαίσιο της λέξης. Στο τέλος της κατασκευής φορέων θα είμαστε σε θέση να συγκρίνουμε αυτά τα διανύσματα, χρησιμοποιώντας μια μετρική απόστασης, συνημίτονο ή Ευκλείδεια για παράδειγμα, και την εκτίμηση της ομοιότητας τους. Αναδιανεμητικές σημασιολογικά μοντέλα διαφέρουν κυρίως σε σχέση με τις ακόλουθες παραμέτρους:

- Τύπος στοιχείου Πλαίσιο (περιφέρειες κείμενο εναντίον της γλωσσικής στοιχεία)
- Παράθυρο στοιχείο Πλαίσιο (μέγεθος, επέκταση, κλπ)
- στάθμιση συχνότητας (π.χ. εντροπία, το σημείο-σοφός αμοιβαίας ενημέρωσης, κλπ)
- Μείωση Διάσταση στοιχείο (π.χ. τυχαία ευρετηρίαση, μοναδική αξία αποσύνθεση, κ.λπ.)
- μέτρο ομοιότητας (π.χ. συνημίτονο ομοιότητα, Jaccard, κλπ)

4.1.2 Λειτουργικότητα

Σε γενικές γραμμές, για την κατασκευή αυτών των φορέων, α) ένα σώμα και β) ένα λεξιλόγιο απαιτούνται. Αν το λεξιλόγιο δεν παρέχεται κάποιος μπορεί να εξαχθεί από το σύνολο των υπαρχόντων. Ως πρώτο βήμα, θα μπορούσε κανείς να σκεφτεί να εκπροσωπεί κάθε λέξη ως φορέας με τις συχνότητες της συνύπαρξη με όλες τις λέξεις στο λεξιλόγιο. Αυτό σημαίνει ότι εάν το λεξιλόγιο περιέχει 1000 λόγια, κάθε λέξη θα εκπροσωπείται ως φορέας 1000Δ. Κάθε στοιχείο αυτού του φορέα θα περιέχει τη συχνότητα της συνύπαρξης αυτών των δύο λέξεων (ωορδ ενδιαφέροντος (γραμμή) και το στοιχείο-λέξη (στήλη» σε ολόκληρο το σώμα. Ωστόσο, αυτή είναι μια πολύ ευρεία έννοια της ομοιότητας λέξεων (οι φορείς φαίνεται να παραχθεί τυχαία), και για το λόγο αυτό θα πρέπει να περιορίσετε σε αυτή την αναζήτηση για να βρει μια *τοπική* ομοιότητα. Για να επιτευχθεί αυτό, θεωρούμε μόνο τη συνύπαρξη με λίγα λόγια σε κάθε πλευρά της λέξης-στόχου, επίσης γνωστή ως `παράθυρο». Εικόνα 4.1 δείχνουν ένα παράδειγμα αυτής της κατασκευής από ένα ςορπυς παιχνίδι χρησιμοποιώντας ένα παράθυρο μεγέθους 1 (εξετάσει μόνο 1 λέξη από την αριστερή και 1 από τη δεξιά της λέξης-στόχου). Οι σειρές αντιπροσωπεύουν τις λέξεις-στόχο και οι στήλες είναι οι λέξεις λεξιλόγιο που χρησιμοποιούνται ως πλαίσιο-χαρακτηριστικά. Δεν είναι υποχρεωτικό για τις λέξεις-στόχους που πρέπει να περιλαμβάνονται στο λεξιλόγιο, μόνο για να υπάρχουν στο σώμα. Ως επόμενο βήμα, η ομοιότητα μεταξύ αυτών των φορέων μπορεί να υπολογιστεί χρησιμοποιώντας την γωνία συνημίτονο μεταξύ τους.

$$sim(\overrightarrow{v_1}, \overrightarrow{v_2}) = \frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{\|\overrightarrow{v_1}\| \cdot \|\overrightarrow{v_2}\|} \tag{4.1}$$

Ως αποτέλεσμα, για το παραπάνω παράδειγμα, έχουμε:
sim(αυτοκίνητο, αυτοκίνητο) $= 4/\sqrt{21} = 0,87$
sim(αυτοκίνητο, ποδόσφαιρο) $= 0/\sqrt{12} = 0$
sim(αυτοκίνητο, ποδόσφαιρο) $= 1/\sqrt{24} = 0,2$

An automobile is a wheeled motor vehicle used for transporting passengers .
A car is a form of transport, usually with four wheels and the capacity to carry around five passengers .
Transport for the London games is limited , with spectators strongly advised to avoid the use of cars .
The London 2012 soccer tournament began yesterday , with plenty of goals in the opening matches .
Giggs scored the first goal of the football tournament at Wembley , North London .
Bellamy was largely a passenger in the football match , playing no part in either goal .

Term vocab: ⟨*wheel, transport, passenger, tournament, London, goal, match*⟩

$$\begin{array}{l} \\ \textit{automobile} \\ \textit{car} \\ \textit{soccer} \\ \textit{football} \end{array} \begin{array}{c} \begin{array}{ccccccc} \textit{wheel} & \textit{transport} & \textit{passenger} & \textit{tournament} & \textit{London} & \textit{goal} & \textit{match} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 2 & 1 \end{pmatrix} \end{array}$$

Σχήμα 4.1: Example of term-term BOW matrix, using term co-occurrence frequency and window = 1. Image from Clark (2012).

sim(αυτοκίνητο, ποδόσφαιρο) $= 1/\sqrt{28} = 0,19$
sim(αυτοκίνητο, ποδόσφαιρο) $= 2/\sqrt{56} = 0,27$
sim(ποδόσφαιρο, ποδόσφαιρο) $= 5/\sqrt{32} = 0,88$

Η μήτρα συχνότητα όρος διάρκειας συνύπαρξη μπορεί να κατασκευαστεί με διαφορετικές αξίες, όπως αμοιβαίας πληροφόρησης (ΠΜΙ) βαθμολογίες Ποιντ-σοφός ($PMI(x, y) = log\ fracp(x, y)p(x)p(y)$), TF-IDF βαθμολογίες, δυαδική (1-0) για την παρουσία ή όχι του όρου χαρακτηριστικό στο πλαίσιο κλπ τεχνικές μείωσης μπορούν να εφαρμοστούν και στην περίπτωση που το μέγεθος του λεξιλογίου είναι πολύ μεγάλο. Ένα μειονέκτημα της τεχνικής Σ΄Δ (η πιο κοινή) είναι ότι οι προκαλούμενες κρυμμένες διαστάσεις είναι δύσκολο να ερμηνευθούν, ενώ διανύσματα βάσης, μπορεί να σχετίζεται με την εννοιολογική ιδιότητες και δεδομένου ψυχο διερμηνεία ως αναφορά σε Βαρονι ανδ Λενςι (2010). Επίσης, εκτός από την ομοιότητα συνημίτονου, μπορούν να χρησιμοποιηθούν πολλαπλοί τύποι άλλες μετρήσεις ομοιότητας, όπως ο συντελεστής Πεαρσον, συντελεστής Θαςςαρδ, συντελεστής Τανιμοτο και την Ευκλείδεια. Οι συντελεστές αυτοί συνοψίζονται στον ακόλουθο πίνακα. μοντέλα διανυσματικό χώρο, μπορούν να χωριστούν σε δύο κατηγορίες, α) μετρούν με βάση και β) μέθοδοι πρόβλεψης. Όπως αναφέρεται στην Baroni and Lenci (2010), μέθοδοι μετράνε με βάση τον υπολογισμό των στατιστικών στοιχείων του πόσο συχνά κάποια λέξη συν-συμβαίνει και με τα λόγια του γείτονά της σε ένα μεγάλο σώμα κειμένου, και στη συνέχεια να χαρτογραφήσει αυτά μετράνε Στατιστικά κάτω σε ένα μικρό, πυκνό φορέα για κάθε λέξη. Προβλεπτικά μοντέλα

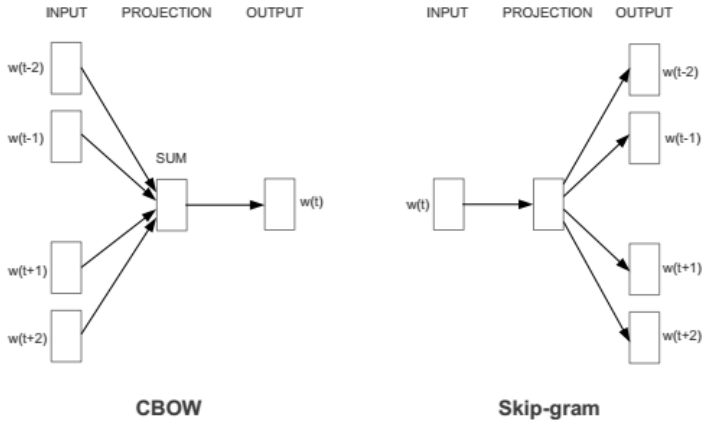
Similarity Type	Formula
Cosine	$cos(\theta) = \frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{\lVert\overrightarrow{v_1}\rVert \cdot \lVert\overrightarrow{v_2}\rVert}$
Euclidean	$d(v_1, v_2) = \sqrt{\sum_{i=1}^{N} (v_{1i} - v_{2i})^2}$
Pearson	$\rho_{v_1,v_2} = \frac{cov(v_1,v_2)}{\sigma_{v_1}\sigma_{v_2}}$
Jaccard	$J = \frac{M_{11}}{M_{01}+M_{10}+M_{11}}$
Tanimoto	$T(\overrightarrow{v_1}, \overrightarrow{v_2}) = \frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{\lVert\overrightarrow{v_1}\rVert^2 + \lVert\overrightarrow{v_2}\rVert^2 - \overrightarrow{v_1} \cdot \overrightarrow{v_2}}$

Table 4.1: List of similarity metrics for word-vectors comparison.

προσπαθήσουμε άμεσα να προβλέψει μια λέξη από τους γείτονές της όσον αφορά την εκμάθηση μικρά, πυκνά φορείς ενσωμάτωσης (θεωρείται παραμέτρων του μοντέλου). Προηγουμένως, περιγράψαμε την πρώτη κατηγορία και συνεχίζουμε με το πιο γνωστό μοντέλο από τη δεύτερη.

4.1.3 Word2Vec

Word2vec είναι ένα ιδιαίτερα υπολογιστικά αποδοτικό μοντέλο πρόβλεψης για την εκμάθηση της υψηλής ποιότητας embeddings λέξη από τις πρώτες κείμενο και είναι αυτός που χρησιμοποιείται σε αυτή τη διατριβή. Εισήχθη από Mikolov et al. (2013) και στην εργασία τους δείχνουν πώς νευρωνικό δίκτυο μοντέλα γλώσσα με βάση ξεπερνούν σημαντικά μοντέλα N-γραμ. Ο πρώην, μπορεί επίσης να εκπαιδευτεί σε μεγαλύτερες ποσότητες δεδομένων και εκατομμύρια λέξεις στο λεξιλόγιο. Δεδομένου ότι επιπλέον αναφέρουν, διαπιστώθηκε ότι η ομοιότητα των παραστάσεων λέξη πηγαίνει πέρα από την απλή συντακτική κανονικότητες. Χρησιμοποιώντας μια λέξη οφφσετ τεχνική όπου οι απλές αλγεβρικές πράξεις που εκτελούνται στα φορείς λέξη, φάνηκε για παράδειγμα, ότι η $vector(\text{“}king\text{”}) - vector(\text{“}Man\text{”}) + vector(\text{“}Woman\text{”})$ οδηγεί σε μια φορέα που είναι ο πλησιέστερος προς τον φορέα εκπροσώπησης της λέξης “$Queen$”. Η γενική δομή βασίζεται στην πιθανολογική μεταδράσεως νευρωνικό δίκτυο γλώσσα μοντέλου, το οποίο έχει προταθεί στο **?**. Αποτελείται από είσοδο, προβολή, απόκρυψη και στρώματα εξόδου. Στο στρώμα εισόδου, οι N προηγούμενες λέξεις κωδικοποιημένα σε ένα ζεστό κωδικοποίησης, όπου ˝ είναι το μέγεθος του λεξιλογίου. Αυτό σημαίνει ότι κάθε λέξη αντιπροσωπεύεται ως διάνυσμα μηδενικά με ένα μόνο στη θέση της λέξης αυτής. Το στρώμα εισόδου στη συνέχεια προβάλλεται σε ένα στρώμα προβολής P που έχει dimentionality NxD, χρησιμοποιώντας μια κοινή μήτρα προβολής. Σε Word2Vec ιεραρχική Softmax μπορούν επίσης να χρησιμοποιηθούν, όπου το λεξιλόγιο αναπαρίσταται ως Huffman δυαδικό δένδρο. Υπάρχουν 2 πιθανές αρχιτεκτονικές, ο Skipgram μοντέλο και η συνεχής τσάντα-οφ-λέξεις (CBOW). Οι παράμετροι της Word2Vec εξηγείται λεπτομερώς στο **?**.



Σχήμα 4.2: CBOW and Skipgram architectures of Word2Vec.

Skip-Gram

Το μοντέλο skipgram προσπαθεί να προβλέψει το πλαίσιο μιας λέξης-στόχου, δίνεται αυτή η λέξη. Στην πραγματικότητα, μαθαίνει δύο ξεχωριστές εμβεδδινγς για κάθε λέξη w, η λέξη ενσωμάτωσης και πλαίσιο ενσωμάτωσης. Αυτά είναι κωδικοποιημένα σε δύο μήτρες, μήτρα λέξη W και της μήτρας πλαίσιο C. Οι σειρές του πρώτου είναι η λέξη φορέων με διάσταση d $(1 \times \delta)$ (μια λέξη στο λεξιλόγιο) και τις στήλες του δεύτερου είναι τα διανύσματα πλαίσιο $(\delta \times 1)$. Το μοντέλο σκιπγραμ, προβλέπει κάθε γειτονική λέξη σε ένα παράθυρο πλαίσιο L, να κερδίσει $= [w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}]$. Στο τέλος, θέλουμε να υπολογίσουμε $P(w_k|w_j)$ (w_j είναι ο στόχος, w_k είναι το πλαίσιο). Η διαδικασία είναι η εξής:

1. Το διάνυσμα εισόδου x είναι ένα ονε-ηοτ για τη λέξη-στόχο w_j
2. Προβλέψτε την πιθανότητα για κάθε ένα για τα 2Λ λέξεις νίκη
   (α′) Επιλέξτε το φορέα από W και πολλαπλασιάστε το με x για την παραγωγή του πίνακα προβολής. Το στρώμα προβολή θα είναι η ενσωμάτωση των w_j (v_j)
   (β′) Υπολογίστε το γινόμενο $v_j \cdot c_k$ που δίνει τις βαθμολογίες εξόδου για κάθε λέξη του λεξιλογίου
   (γ′) Κανονικοποίηση των γινομένων σε πιθανότητες, χρησιμοποιώντας μια λειτουργία soft-max

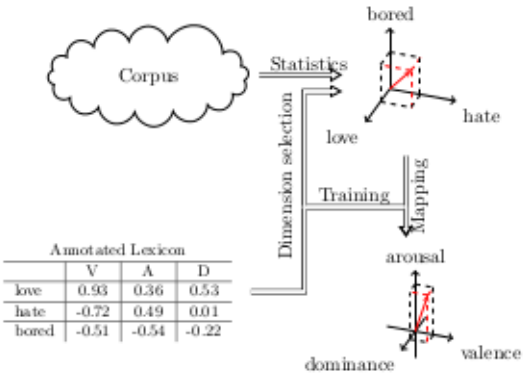
CBOW

CBOW είναι ένα είδωλο της skipgram, καθώς προβλέπει τη λέξη-στόχο με δεδομένες τις λέξεις πλαίσιο. Η κύρια διαφορά είναι στον υπολογισμό του στρώματος προβολής. Και πάλι υπάρχουν 2 δ-διαστάσεων embeddings, οι λέξεις και οι περιβάλλοντα που αποτελούν τις μήτρες W και C (μεταφέρθηκε σε σύγκριση με σκιπγραμ αυτά). Τώρα θέλουμε να προβλέψουμε $P(w_k, w_j)$ (w_k είναι ο στόχος, w_j είναι το πλαίσιο). Το στρώμα εισόδου είναι συνδεδεμένο με το στρώμα προεξοχή από $C = |V| xdmtraplasio$. Αυτή η μήτρα επαναλαμβάνεται μεταξύ

κάθε διάνυσμα ένα καυτό είσοδο και το στρώμα προβολής. Ο πολλαπλασιασμός του φορέα x και μήτρα C θα δώσει 2 φορείς πλαίσιο, του οποίου παίρνουμε το μέσο όρο και το σχηματισμό του στρώματος προβολής. Στη συνέχεια, το στρώμα προβολή πολλαπλασιάζεται με τη μήτρα λέξη W με αποτέλεσμα τις βαθμολογίες ˝ για κάθε λέξη στο λεξιλόγιο. Τέλος, αυτές οι βαθμολογίες μετατραπεί σε μια πιθανότητα χρησιμοποιώντας ένα στρώμα μαλακής-μαξ.

4.2 Συναισθηματική Μοντέλο

Το μοντέλο που προτείνει Malandrakis et al. (2011a) χρησιμοποιεί μια υπάρχουσα συναισθηματική λεξικού που ορίζει την σημασιολογική προσανατολισμό των λέξεων (unigrams) σε συνεχή συναισθηματική βαθμολογίες. Η κύρια υπόθεση του μοντέλου αυτού είναι ότι ‘ σημασιολογική ομοιότητα μπορεί να μεταφραστεί σε συναισθηματική ομοιότητα »και, επομένως, μια χαρτογράφηση από το σημασιολογικό στη συναισθηματική χώρο είναι χτισμένο. Ως αποτέλεσμα, δεδομένης της σημασιολογικής ομοιότητας μεταξύ δύο λόγια, η αξιολόγηση συναισθηματική του μία δεδομένη την άλλη μπορεί να καθοριστεί. η ιδέα βασίστηκε στο έργο του **?** που χρησιμοποιούνται για ένα σύνολο λέξεων με γνωστές συναισθηματικές αξιολογήσεις, γνωστό ως *λέξεις σπόρων*. Τότε η σημασιολογική ομοιότητα μιας λέξης σπόρων και μια νέα λέξη υπολογίζεται με τη βοήθεια μιας κατανομής σημασιολογικό μοντέλο και χρησιμοποιούνται για την εκτίμηση της συναισθηματικής βαθμολογίας από την νέα λέξη. Οι αρχικές λέξεις σπόρος μπορεί να είναι οι ετικέτες των συναισθηματικών κατηγορίες (‘ θυμό ¨,` ευτυχία¨), δηλαδή με συνεχή βαθμολογίες κλπ το τελικό βήμα είναι να συνδυάσει τα αποτελέσματα των λέξεων σπόρων, προκειμένου να εξαχθεί η βαθμολογία του αόρατου λέξη.

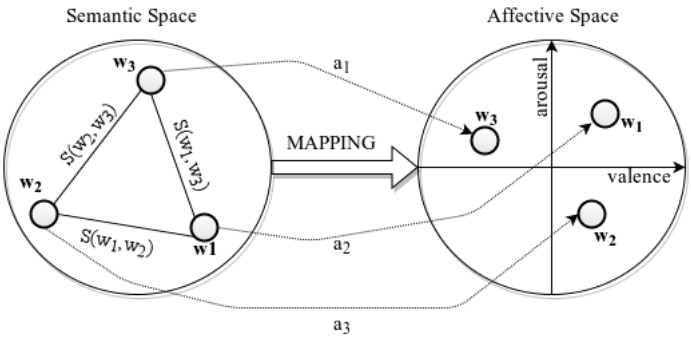


Σχήμα 4.3: Semantic-Affective model as proposed in Malandrakis et al. (2014).

4.2.1 Από το Σημασιολογικό στο Συναισθηματικό Χώρο

Σε γενικές γραμμές, σημασιολογικές μοντέλα μπορεί να θεωρηθεί ως μια γνωστική διαδικασία ανάλυσης της ανθρώπινης συμπεριφοράς, όταν πρόκειται για την κατανόηση της γλώσσας. Ως εκ τούτου, φαίνεται η ανάγκη για ένα μοντέλο που μπορεί να συλλάβει το συναίσθημα. Ο Σημασιολογικός-συναισθηματική Μοντέλο (ΣΑΜ) έχει ως στόχο να γεφυρώσει το χάσμα

μεταξύ της έννοιας και της συναισθηματικής χώρο, καθώς και να μεταφέρει τη γνώση από τη μια περιοχή στην άλλη. Η υπόθεση στην οποία βασίζεται μπορεί να συνοψιστεί στη φράση 'σημασιολογική ομοιότητα συνεπάγεται συναισθηματική ομοιότητα'. Η έννοια παρουσιάζεται στο σχήμα 4.3. Το μοντέλο υποθέτει την ύπαρξη δύο χώρων α) σημασιολογική χώρος, όπου όλες οι λέξεις είναι κάπως συνδέονται μεταξύ τους με σημασιολογική σημασία τους και β) μια συναισθηματική χώρος, όπου οι λέξεις που αντιπροσωπεύεται από συναισθηματική σημασία τους ως σημεία σε 3 διαστάσεων χώρο του σθένους , διέγερση και την κυριαρχία (που περιγράφεται στο ρεφ Κεφάλαιο 1). Ένα μεγάλο γενικό σώμα χρησιμοποιείται για να κατασκευαστεί το σημασιολογικό μοντέλο από το οποίο ομοιότητες μεταξύ των ζευγαριών των λέξεων θα εξαχθούν και να χρησιμοποιηθούν στο ΣΑΜ. Ο στόχος είναι να βρούμε την κατάλληλη χαρτογράφηση, προκειμένου να εκτιμηθεί συναισθηματική βαθμολογίες για αθέατες λέξεις που χρησιμοποιούν και τις δύο θέσεις. Η διαδικασία αυτή περιγράφεται στο σχήμα 4.4.



Σχήμα 4.4: Semantic-Affective Mapping concept.

4.2.2 Επισημείωση Λέξεων

Προκειμένου να εκτιμηθεί η βαθμολογία μιας λέξης, το μοντέλο χρησιμοποιεί μια υπάρχουσα σχολιασμένο λεξικού απ΄όπου μόνο ένα σύνολο δεδομένων είναι στην πραγματικότητα χρησιμοποιείται (λέξεις σπόρων). Οι συναισθηματικές βαθμολογίες που συνήθως βρίσκονται στη διάσταση σθένους σε μια σειρά από $[-1, 1]$ (από πολύ θετικές έως πολύ αρνητική), *από την άποψη αναγνώστη*. Επιπλέον, το σημασιολογικό μοντέλο κατασκευάστηκε χρησιμοποιώντας ένα σώμα χρησιμοποιείται για την εξαγωγή λέξη-ζεύγη σκορ ομοιότητας. Τέλος, το σημασιολογικό-συναισθηματική χαρτογράφηση υπολογίζεται χρησιμοποιώντας τόσο το λεξικό και τις βαθμολογίες ομοιότητας σε μια διαδικασία κατάρτισης. Οι αξιολογήσεις της νέας λέξης, υπολογίζεται ως ο γραμμικός συνδυασμός των αξιολογήσεων των λέξεων σπόρων. Η ακόλουθη εξίσωση συνοψίζει τα παραπάνω:

$$v(w_j) = \alpha_0 + \sum_{n=1}^{N} \alpha_i \cdot v(w_i) \cdot d(w_i, w_j) \tag{4.2}$$

όπου w_j είναι η λέξη που σημαίνει να χαρακτηρίσει, $w_1, ..., w_N$ είναι τα λόγια των σπόρων, $n(w_i)$ είναι η βαθμολόγηση σθένος για τη λέξη σπόρων w_i, $alpha_i$ είναι το βάρος που αντι-

στοιχεί στη λέξη w_i και $d(w_i, w_j)$ είναι ένα μέτρο της σημασιολογικής ομοιότητας μεταξύ των λέξεων w_i και w_j. Τα βάρη $alpha_i$ είναι βασικά το σημασιολογικό-συναισθηματική χαρτογράφηση περιγράψαμε πριν και μπορεί να μάθει από την επίλυση ενός γραμμικού συστήματος. Αν έχουμε ένα σώμα με K σχολιασμένη λέξεις, και να επιλέξει N τους ως σπόροι ($N < K$), μπορούμε να υπολογίσουμε τα βάρη α_i του μοντέλου. Μπορούμε να χρησιμοποιήσουμε την εξίσωση ;; για να δημιουργήσετε K γραμμικών εξισώσεων με $N + 1$ άγνωστες μεταβλητές. Αυτές οι παράμετροι α μπορούν να λάβουν τη βέλτιστη αξία τους χρησιμοποιώντας Τουλάχιστον Μέση Πλατείες (ΛΜΣ) ή τακτοποιηθεί έκδοση γνωστή ως Ριδγε Παλινδρόμηση (PP). Η μέθοδος ΛΣΕ περιγράφεται στην εξίσωση 4.3.

$$\mathbf{X} \cdot \alpha = \mathbf{y}, \text{ ωηερε } \hat{\alpha}_{lse} = argmin_{\alpha_{lse}} ||\mathbf{y} - \mathbf{X}\alpha_{lse}||^2 \tag{4.3}$$

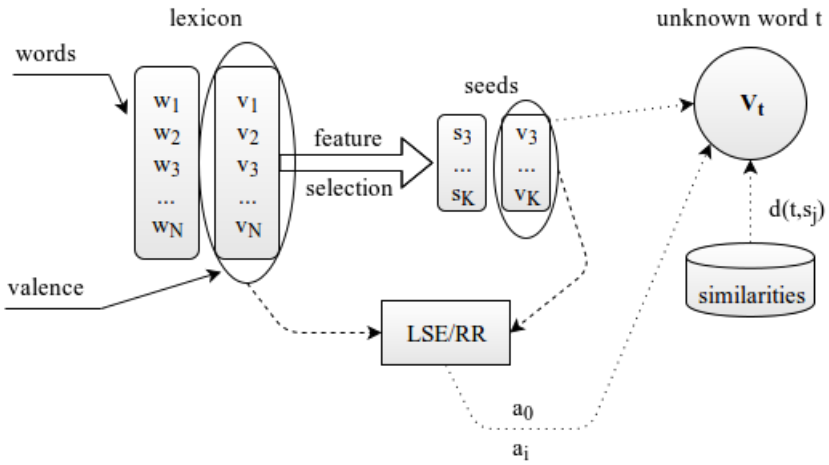
Από την άλλη πλευρά, για τη μέθοδο PP, ένα επιπλέον παράμετρος λ εισάγεται που λειτουργεί ως παράγοντας νομιμοποίησης που χρησιμοποιούνται προσπαθεί να ταιριάζει καλύτερα μια καμπύλη στα δεδομένα (Εγκυκλοπαίδεια συναισθηματικών βαθμολογίες) όπως φαίνεται στην εξίσωση 4.4 .

$$\hat{\alpha}_{rr} = argmin_{\alpha_{rr}} \Big(||\mathbf{y} - \mathbf{X}\alpha_{rr}||^2 + \lambda ||\alpha_{rr}|| \Big) \tag{4.4}$$

$\mathbf{\Xi}$ είναι ένας πίνακας που περιέχει όλα τα ζεύγη ομοιότητα μεταξύ των λέξεων σπόρων και το σύνολο της συναισθηματικής λεξικό του μεγέθους $K \times (N + 1)$, $alpha$ είναι οι συντελεστές (χαρτογράφηση), φορέας του μεγέθους $((N + 1) \times 1)$ και y είναι το διάνυσμα με τις συναισθηματικές αποτελέσματα για ολόκληρο το λεξικό του μεγέθους $(K + 1) \times 1$.

$$\begin{bmatrix} 1 & d(w_1, w_1)v(w_1) & \cdots & d(w_1, w_N)v(w_N) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & d(w_K, w_1)v(w_1) & \cdots & d(w_K, w_N)v(w_N) \end{bmatrix} \cdot \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} = \begin{bmatrix} 1 \\ v(w_1) \\ \vdots \\ v(w_K) \end{bmatrix} \tag{4.5}$$

όπου $\alpha_1, ..., \alpha_N$ είναι τα N βάρη λέξη σπόρων και α_0 είναι μια πρόσθετη παράμετρος που λειτουργεί ως συναισθηματική προκατάληψη στη λέξη σπόρων που. Το βασικό κίνητρο πίσω από την ιδέα της κατάρτισης αυτών των παραμέτρων, είναι το γεγονός ότι η σημασιολογική ομοιότητα δεν αντικατοπτρίζουν πλήρως τη σημασία μιας λέξης σπόρων για σθένους υπολογισμούς, όπως αναφέρεται στην Malandrakis et al. (2011a). Που απεικονίζει ένα παράδειγμα, ας υποθέσουμε ότι η λέξη ἀγάπη» είναι η λέξη που θέλουμε να εκτιμήσουμε την επίδραση των και υπάρχουν δύο λέξεις σπόρων, για παράδειγμα `μήλο ΄,` η ζωή »που έχουν την ίδια σημασιολογική ομοιότητα με` αγάπη ». Χρησιμοποιώντας την παραδοχή ότι ‘ σημασιολογικής ομοιότητας συνεπάγεται συναισθηματική ομοιότητα », τόσο οι βαθμολογίες των λέξεων αυτών θα περιλαμβάνεται στο άθροισμα με το ίδιο βάρος. Ωστόσο, δεν είναι οι δύο λέξεις πρέπει να σταθμίζονται με τον ίδιο, καθώς μπορεί να εκφράζουν εντελώς διαφορετικά περιβάλλοντα ή είναι σαφείς . Αυτό θα έχει ως αποτέλεσμα μια παραπλανητική γενικό συναίσθημα για τη λέξη ενδιαφέροντος. Η μέθοδος μηχανικής μάθησης ΛΜΣ μας επιτρέπει να συλλάβει ακριβώς αυτό διαίσθηση. Για να επιλέξετε τους σπόρους λέξεις (N) που θα χρησιμοποιηθεί για να εκτιμηθεί μια άγνωστη λέξη, μια μέθοδος επιλογής χαρακτηριστικών γίνεται. Ειδικότερα, στοχεύουμε



Σχήμα 4.5: word affective score estimation procedure using SAM.

ότι το υποσύνολο που συλλέγονται είναι ισορροπημένη, πράγμα που σημαίνει ότι το άθροισμα των βαθμολογιών κάθε επιμέρους λέξη θα πρέπει να είναι κοντά στο μηδέν. Το σημασιολογικό μοντέλο μπορεί να είναι οποιαδήποτε αναδιανεμητική σημασιολογικό μοντέλο (όπως αυτά που περιγράφονται στην προηγούμενη ενότητα και κεφάλαιο 2 και τη μετρική ομοιότητας που χρησιμοποιείται στο μοντέλο μπορεί να είναι οποιοδήποτε μέτρο ομοιότητας μεταξύ των φορέων των δύο λέξεων. Επιπλέον, πυρήνες μπορεί να εφαρμοστεί σε αυτό το μετρικό σύστημα για την καλύτερη απόδοση, όπως προτείνεται στο Malandrakis et al. (2011b). Σε γενικές γραμμές, η συναισθηματική λεξικό που χρησιμοποιείται για τη μέθοδο αυτή είναι η συναισθηματική Πρότυπα για αγγλικές λέξεις (MIA NEA) προτείνει Bradley and Lang (1999) το οποίο περιέχει 1034 αγγλικές λέξεις με τα αποτελέσματα σθένους, διέγερση και την κυριαρχία μεταξύ [-1,1]. Επιπλέον, τα πέντε βασικά συναισθήματα περιλαμβάνονται επίσης σε αυτό το λεξικό. Ένα παράδειγμα για το πώς η εκπαίδευση γίνεται παρουσιάζεται παρακάτω: Εν ολίγοις, η

Order	w_i	$v(w_i)$	α_i	$v(w_i)x\alpha_i$
1	triumphant	0.96	0.34	0.33
2	love	0.93	0.78	0.73
3	paradise	0.93	2.07	1.93
4	loved	0.91	1.72	1.57
5	joy	0.9	2.57	2.31
6	rape	-0.94	2.56	-0.09
7	suicide	-0.94	2.56	-2.5
8	funeral	-0.9	0.09	-0.08
9	cancer	-0.88	0.77	-0.68
10	rejected	-0.88	0.83	-0.73
-	(offset)	1	-0.11	-0.11

Table 4.2: Example of training with 10 seed words from ANEW, using LSE.

προαναφερθείσα μοντέλο υπολογίζει μια νέα αθέατες λέξη χρησιμοποιώντας τα αποτελέσματα των υφιστάμενων λέξεις σε διαφορετικές αναλογίες, με βάση το μέτρο ή συγγένειας τους και εκπαιδεύσιμα συντελεστή βαρύτητας. Γενικά, μπορεί να υπολογίσει γενική συναισθηματική βαθμολογίες πολύ καλά, ή να προσαρμόζουν τους σε ένα συγκεκριμένο πλαίσιο με την εκπαίδευση των βαρών συντελεστή με ένα άλλο σώμα (Malandrakis et al. (2011b)).

4.2.3 Επισημείωση προτάσεων

Fusion

Προχωρώντας προς το επίπεδο ποινής, η αρχή της ςομποσιτιοναλιτψ (**?**) αναφέρει ότι η έννοια της ποινής είναι το άθροισμα της την έννοια των μερών της. Έτσι, προκειμένου να εκτιμηθεί η συναισθηματική βαθμολογία μιας πρότασης με τη χρήση των συναισθηματικών βαθμολογίες των λέξεων του μπορεί να επιτευχθεί χρησιμοποιώντας τρία μοντέλα σύντηξης, όπως προτείνεται στο Malandrakis et al. (2011a).

1. Linear Fusion

$$v_1(s) = \frac{1}{N}\sum_{i=1}^{N} v(w_i) \tag{4.6}$$

   Σε αυτό το σχέδιο συγχώνευσης, απλά ο μέσος όρος ενός φράσεις λέξεων αποτελέσματα θα προκύψει η συνολική βαθμολογία για αυτή την πρόταση. Αυτό εξίσου βάρη όλων των λέξεων και τείνει να δώσουν χαμηλότερες βαθμολογίες σε προτάσεις που περιέχουν πολλές ουδέτερες λέξεις.

2. Weighted Fusion

$$v_2(s) = \frac{1}{\sum_{i=1}^{N} |v(w_i)|}\sum_{i=1}^{N} v(w_i)^2 \cdot sign(v(w_i)) \tag{4.7}$$

   Αυτή είναι μια κανονικοποιημένη μέση σταθμισμένη, στην οποία οι λέξεις με υψηλότερες βαθμολογίες συναισθηματική σταθμισμένο περισσότερο.

3. Non-linear Max Fusion

$$v_3(s) = max_i(|v(w_i)|) \cdot sign(v(w_z)), \text{ ωηερε } z = argmax_i(|v(w_i)|) \tag{4.8}$$

   Τέλος, ένα μη-γραμμικό σύντηξης που χαρακτηρίζει την πρόταση με το ίδιο σκορ ως η υψηλότερη λέξη στην πρόταση.

Ταξινόμηση

Επιπλέον, προκειμένου να προβλέψει την συναισθηματική βαθμολογία για μια πρόταση η μέθοδος state-of-the art περιλαμβάνει τη χρήση ενός ταξινομητή. Πιο συγκεκριμένα, εάν λαμβάνονται οι συναισθηματικές βαθμολογίες για κάθε λέξη σε μια πρόταση, είναι μερικά στατιστικά στοιχεία εξάγεται και εργασία ως χαρακτηριστικά. Με αυτόν τον τρόπο, κάθε πρόταση αναπαρίσταται ως ένα διάνυσμα χαρακτηριστικών γνωρισμάτων, με τις τιμές των

στατιστικών στοιχείων από τα λόγια βαθμολογίες του. Η κύρια διαφορά μεταξύ αυτής της προσέγγισης και της σύντηξης είναι ότι η πρώτη απαιτεί δεδομένα εκπαίδευσης, προκειμένου να εκπαιδεύσουν τον ταξινομητή, ενώ η δεύτερη είναι εντελώς χωρίς επίβλεψη σε ό, τι αφορά την αξιολόγηση. Τα στατιστικά στοιχεία που μπορούν να χρησιμοποιηθούν, όπως προτείνεται στο Malandrakis et al. (2011b) είναι:

- length (cardinality)
- max
- min
- mean
- variance
- standard deviation
- range (max - min)
- amplitude
- sum

Κανονικοποιημένη εκδοχή από αυτά τα χαρακτηριστικά μπορούν να δημιουργηθούν, σύμφωνα με το μέρος-οφ-Σπεεςη Ταγγινγ (ΠΟΣ Ταγγινγ). Για παράδειγμα, διαθέτουν μαξ πάνω από όλα ουσιαστικά διαιρείται με το μέγιστο πάνω από όλες τις μάρκες. Αν λάβουμε υπόψη τα ουσιαστικά, επιρρήματα, επίθετα και ρήματα όπως ετικέτες ΠΟΣ ενδιαφέροντος, καταλήγουμε με 27 λειτουργίες, το οποίο σημαίνει ένα διάνυσμα 27Δ χαρακτηριστικό για κάθε πρόταση. Σε μια πιο απλή περίπτωση, μπορούμε να εξάγουμε απλές λειτουργίες και τους ομαλοποιήθηκαν με τη διάρκεια της ποινής, που οδηγεί σε ένα φορέα 18Δ. Ένας ταξινομητής μπορεί στη συνέχεια να εκπαιδευτεί χρησιμοποιώντας κάποια επισημαίνονται φράσεις και χρησιμοποιείται για να προβλέψει τις συναισθηματικές βαθμολογίες των νέων προτάσεων. Naive Bayes Δέντρο, Random Forest, SVM είναι οι πιο κοινές περιπτώσεις. Η ταξινόμηση είναι συνήθως δυαδική, μεταξύ θετικών και αρνητικών φράσεις, αλλά το πρόβλημα 3-class μπορεί να πραγματοποιηθεί επίσης, χρησιμοποιώντας την ουδέτερη τάξη επιπλέον.

Κεφάλαιο 5

Προσαρμοστικό Μοντέλο

5.1 Κίνητρο

Όσο απλό λεξιλόγιο με βάση τα μοντέλα αδυνατούν να συλλάβουν το πλαίσιο όπου μια φράση ανήκει, μέθοδοι θέμα μοντέλων προσπαθούν να ενσωματώσουν το θέμα της γνώσης, με στόχο να οδηγήσει σε μια πιο εξειδικευμένη ανίχνευση του εξέφρασε το συναίσθημα σε μια πρόταση. Αυτό μπορεί να θεωρηθεί ως διαδικασία ταξινόμησης δύο βημάτων, όπου αφενός η ποινή έχει ταξινομηθεί σε ένα θέμα και, δεύτερον, σε μια θετική ή αρνητική τάξη συναίσθημα. Με βάση το γεγονός ότι ' σημασιολογικής ομοιότητας συνεπάγεται συναισθηματική ομοιότητα »η ιδέα είναι να χωρίσει μια μεγάλη συλλογή από έγγραφα στα θέματα που αποτελείται από και να εκπαιδεύσει πολλαπλά σημασιολογικά μοντέλα σε αυτούς τους συγκεκριμένους τομείς. Το σχέδιο είναι να συνδυάσει ένα διαφορετικό υποσύνολο αυτών των μοντέλων σε ένα μοντέλο μίγμα, για κάθε πρόταση δοκιμής. στην πραγματικότητα, οι σημασιολογικές ομοιότητες των λέξεων σε έναν τομέα που υπολογίζονται εκ νέου (για κάθε θέμα) και, στη συνέχεια συγχωνεύθηκαν για να δημιουργήσουν νέες σημασιολογικές ομοιότητες, μοναδικό για κάθε πρόταση. Μετά την προσαρμογή του σημασιολογικό χώρο για κάθε θέμα , είμαστε σε θέση να χρησιμοποιήσουν το Semantic-συναισθηματικές υπόδειγμα που περιγράφεται στο κεφάλαιο 4 για να συνδέσετε τη σημασιολογική και τη συναισθηματική χώρο, εκτιμούν τις συναισθηματικές βαθμολογίες των λέξεων μιας φράσης και, στη συνέχεια, την εκτίμηση της συναισθηματικής βαθμολογίας από ολόκληρη την πρόταση.

Η ιδέα αυτή δικαιολογείται από το γεγονός ότι μέσω αυτής της διαδικασίας μπορεί να επιτευχθεί μια μορφή της λέξης αποσαφήνισης. Για παράδειγμα, αν μια πρόταση μιλά για τους υπολογιστές, είναι πιθανό να περιέχουν τη λέξη `μήλο», αλλά αυτό μπορεί επίσης να είναι η περίπτωση, αν πρόκειται για τα τρόφιμα. Μετά την κατάρτιση ενός μοντέλου θέμα, θα είναι σε θέση να διακρίνουν ότι μια πρόταση με τη λέξη `μήλο» είναι για τα φρούτα και τα κατατάσσουν αυτό με το θέμα που σχετίζεται με τα τρόφιμα (όπως είδαμε στο κεφάλαιο 3). Για μια νέα πρόταση, τα θέματα που είναι πιθανό να είναι περίπου θα επιλεγούν και οι ομοιότητες που απαιτούνται στην εξίσωση ;; θα λαμβάνονται από τα σημασιολογικά μοντέλα από αυτά τα θέματα.

Συνοψίζοντας, μπορούμε να χρησιμοποιήσουμε πολλαπλά σημασιολογικά μοντέλα από δι-

άφορους τομείς, που η ποινή μπορεί ενδεχομένως να περιέχει, προκειμένου να κατασκευάσουν με μεγαλύτερη ακρίβεια σημασιολογικό χώρο της.

5.2 Topic Modeling

5.2.1 Εκπαίδευση LDA

Τα δεδομένα που απαιτούνται για το έργο αυτό είναι ένα μεγάλο σώμα. Μπορεί να είναι μια γενική ένας (κατεβάσει από το διαδίκτυο (π.χ. Wikipedia» ή μια συλλογή που σχηματίζεται από κριτικές (π.χ. σχόλια Ταινία).

Θα χρειαστείτε δύο εκδόσεις αυτού του corpus. Ένας που περιέχει έγγραφα (από τις σκέψεις σε μεγαλύτερες λεξικές μονάδες) και στο άλλο το περιέχει προτάσεις. Το πρώτο σώμα θα χρησιμοποιηθεί ως πρώτη ύλη για την Λανθάνουσα Dirichlet αλγόριθμος κατανομής σε αθέτηση της μορφή προϋποθέτει ότι κάθε έγγραφο που περιέχει πολλά θέματα. Στην περίπτωση των ποινών η υπόθεση αυτή δεν μπορεί να σταθεί (πόσα θέματα μπορεί μια μόνο φράση περιέχει·), έτσι ώστε τα έγγραφα θα πρέπει να είναι η είσοδος. Το σώμα ποινές θα χρησιμοποιηθεί προκειμένου να κατασκευάσει τα θέματα. Η κύρια διαίσθηση πίσω από αυτή την ιδέα είναι ότι τα μικρότερα κομμάτια του εγγράφου (προτάσεις), ενδεχομένως θα μιλήσουμε για ένα θέμα, οπότε η κατάταξη μπορεί να είναι αυστηρή (1 φράση → 1 θέμα).

Έτσι ώστε να δημιουργήσετε θέματα, δίνουμε ως είσοδο στο ΛΔΑ το σώμα του εγγράφου. Η εργαλειοθήκη που μας έδωσε τη δυνατότητα να συνεργαστούμε με μεγάλες ποσότητες δεδομένων είναι η Γενσιμ Τοολβοξ, που δημιουργήθηκε από Ρεηὐρεκ ανδ Σοθκα (2010). Περισσότερες πληροφορίες σχετικά με αυτήν την εργαλειοθήκη περιλαμβάνονται στο προσάρτημα.

Οι είσοδοι του ΛΔΑ είναι:
- Ένα λεξιλόγιο
- Ένα σώμα στην τσάντα-οφ-λέξεις (BOW) μορφή

Εξάγουμε ένα λεξιλόγιο από το σώμα. Αν είναι δυνατόν να στολισμένα πολύ μεγάλο σπάνια όρους. Η μορφή ΒΟΩ είναι κατασκευασμένο με βάση αυτό το λεξιλόγιο, για κάθε έγγραφο στο σώμα. Αυτό σημαίνει, ότι κάθε έγγραφο αναπαρίσταται ως ένα ˝-διάστατο διάνυσμα με τη συχνότητα του κάθε στοιχείου (λέξη στο λεξιλόγιο) ως αξία.

Οι έξοδοι του ΛΔΑ είναι 2 διανομές,
- Διανομή των εγγράφων θέματα
- Κατανομή του θέματος λέξεις

Το πρώτο, περιλαμβάνει τα θέματα που κάθε έγγραφο είναι πιθανό να περιέχει και την πιθανότητα τους. Για παράδειγμα, το 40 % Θέμα 1 (π.χ. τρόφιμα), το 30 % Θέμα 5 (π.χ. ζώα) και 30 % Θέμα 9 (π.χ. μουσική). Η δεύτερη, περιλαμβάνει τις λέξεις που κάθε θέμα είναι πιθανό να περιέχουν, εκχωρώντας τους μια πιθανότητα. Ένα παράδειγμα φαίνεται στον πίνακα 5.1 παρόμοια με το σχήμα 3.1.

Χρησιμοποιώντας αυτές τις δύο κατανομές, τότε μπορούμε να κατατάξει κάθε *φράση* από τη μεγάλη συλλογή (φράσεις corpus) στα θέματα που επιλέξαμε για να δημιουργήσει. Η ιδέα

Topic 1		Topic 2		Topic 3		Topic 4	
space	0.06	recipe	0.07	music	0.06	medical	0.09
astronomy	0.05	recipes	0.07	guitar	0.06	treatment	0.08
moon	0.05	food	0.06	piano	0.05	health	0.07
astrology	0.04	cooking	0.06	classical	0.06	disease	0.07
star	0.04	chicken	0.06	jazz	0.045	cancer	0.06
solar	0.03	chocolate	0.06	sheet	0.04	symptoms	0.06
NASA	0.03	kitchen	0.06	dance	0.03	pain	0.06
earth	0.02	cheese	0.05	free	0.02	causes	0.04
mars	0.02	cake	0.02	opera	0.01	surgery	0.04
news	0.02	home	0.01	band	0.01	answers	0.03

Table 5.1: Example of Topic Modeling output, distribution words in a topic, using the top 10 most probable words.

είναι να εφαρμόσει το εκπαιδευμένο μοντέλο ΛΔΑ σε κάθε σώμα-ποινή, και να λαμβάνουν το επιστρέψετε τα θέματα που μπορεί να ανήκει, καθώς και τις πιθανότητες τους.

Πρώτον, η ποινή θα πρέπει να εκπροσωπούνται ως ένα διάνυσμα τσάντα-οφ-λέξεις (ΒΟΩ), με τις λέξεις του λεξιλογίου.

Ένα παράδειγμα φαίνεται παρακάτω:

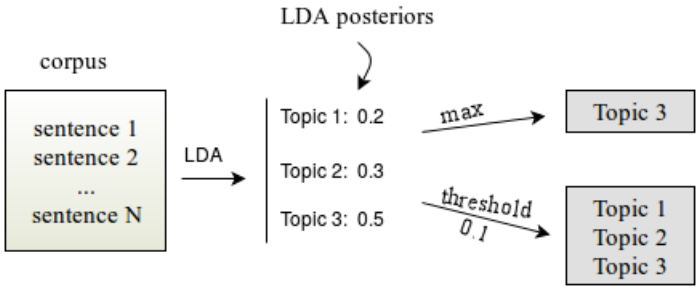
Παράδειγμα 5.2. *Assume that our sentence of interest is:*
s = "Test to predict breast cancer relapse is approved."
After applying a trained LDA model on 60 topics, the result is:
Topic 1: 0.687, Topic 5: 0.174
This means that this sentence, has 68.7% chance to belong to topic 1 and 17.4% chance to belong to topic 5. The contents of these topics, in terms of the 10 most probable words they contain, are:
Topic #1: health, medical, cancer, treatment, disease, symptoms, pain, information, medicine, causes
Topic #5: drug, health, online, drugs, marijuana, alcohol, effects, smoking, addiction, sleep

Ωε ςαν νοτιςε τηατ βοτη τοπιςς ταλκ αβουτ ηεαλτη, τηε φιρστ ας α γενεραλ δομαιν ανδ τηε σεςονδ ας μορε σπεςιφις ον δρυγς. Γιεν τηε σεντενςε, ιτ σεεμς τηατ ΛΔΑ ςορρεςτλψ ςλασσιφιες ιτ ιντο τηεσε τοπιςς, γιινγ μορε ωειγητ το τηε φιρστ ονε.

5.2.2 Κατασκευή Topics

Το επόμενο βήμα είναι να συλλέξει τις κατάλληλες προτάσεις από τη συλλογή corpus και να οικοδομήσουν το θέμα-clusters. Μπορούμε να εκτελέσει μια αυστηρή ομαδοποίηση, το οποίο σημαίνει να ταξινομήσει κάθε πρόταση μόνο σε ένα θέμα ή να εκτελέσει soft clustering, πράγμα που σημαίνει την ταξινόμηση κάθε πρόταση με πολλαπλά θέματα. Στην πρώτη περίπτωση, η

ποινή έχει ταξινομηθεί με το θέμα που δίνει τη μέγιστη οπίσθια πιθανότητα. Στη δεύτερη, η ποινή έχει ταξινομηθεί σε όλα τα θέματα που έχουν οπίσθια πάνω από ένα όριο.



Σχήμα 5.1: Topic construction diagram.

Αν μια φράση ανήκει σε όλα τα θέματα με την ίδια πιθανότητα δεν κατατάσσεται πουθενά καθώς θεωρείται πολύ ασαφής. Επιπλέον, εάν ΛΔΑ αποτύχει να επιστρέψει μια πιθανά θέματα για μια πρόταση η φράση αυτή είναι, επίσης, δεν κατατάσσεται πουθενά. Εικόνα 5.1 περιγράφει τη διαδικασία θέματα κατασκευών.

Μετά την κατασκευή θέμα-clusters, ένα ξεχωριστό σημασιολογικό μοντέλο έχει εκπαιδευτεί για κάθε θέμα. Για το σκοπό αυτό, διαφορετικά ΔΣΜς μπορεί να εκπαιδευτεί, αλλά χρησιμοποιήσαμε το εργαλείο Word2Vec ως πρότυπο state-of-the art. Εμείς συντονιστείτε τις παραμέτρους του (κεφάλαιο 6) και εκτίμηση των ομοιοτήτων μεταξύ των λέξεων για κάθε θέμα που χρησιμοποιεί την υπο-σώματα που δημιουργήθηκε στο προηγούμενο βήμα.

5.3 Προσαρμογή Σημασιολογικών Μοντέλων

Ο αλγόριθμος μπορεί να περιγραφεί παρακάτω:

1. Transform a test sentence in BOW format
2. Apply a trained LDA model to this sentence
3. Receive topics along with their posterior probabilities for this sentence
4. For each word in the sentence and a seed word (with known affective score), extract the similarity between the two from all the Topics that the sentence can be classified
5. Multiply each similarity with the topic-posterior and calculate the average over all topics

The combination of these models is summarized in the equation below:

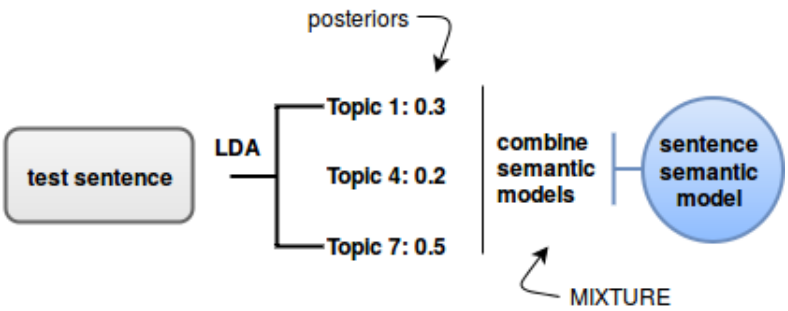
$$sim(w_j, seed_i) = \frac{\sum_{t=1}^{T} p(top_t|s) \cdot sim_t(w_j, seed_i)}{\sum_{t=1}^{T} p(top_t|s)} \tag{5.1}$$

where $s = \{w_0, w_1, ..., w_j, ..., w_n\}]$ as a test sentence, w_j is a word in the sentence, $seed_i$ is a seed word, with known affective score, $p(top_t|s)$ is the topic posterior probability

(probability for sentence s to contain topic t) and $sim_t(w_j, seed_i)$ is the similarity obtained from topic t between the words w_j and $seed_i$.

Συνολικά, το μοντέλο βάρη κάθε ομοιότητα με το θέμα οπίσθια και στο τέλος παράγει ένα νέο ομοιότητα για ένα δεδομένο ζεύγος.

Παράδειγμα 5.3. *Assume that the sentence of interest is the one in example 5.1 and that we use as seed words, the ones shown in table 4.2 from Chapter* **??**. *In order to estimate the new-adapted similarity between the word 'breast' (word in the sentence) and 'cancer' (seed word) we extract the similarity between these words from topic 1 and topic 5.*
$sim_1(breast, cancer) = 0.32, sim_5(breast, cancer) = 0.25$
$p(t = 1|s) = 0.687, p(t = 1|s) = 0.174$
$sim_{final}(breast, cancer) = (0.32 * 0.687 + 0.25 * 0.174)/(0.687 + 0.174) = 0.31$



Σχήμα 5.2: Mixture of Semantic models diagram.

Με παρόμοιο τρόπο, υπολογίζουμε τις ομοιότητες μεταξύ όλων των λέξεων μιας φράσης και τα λόγια των σπόρων που επιλέξαμε.

5.3.1 Εκτίμηση Συναισθήματος

Τέλος, χρησιμοποιείται το σημασιολογικό-συναισθηματική μοντέλο 4.2 για να προβλέψει την συναισθηματική βαθμολογία για κάθε λέξη στην πρόταση. Η βαθμολογία φράση-επίπεδο υπολογίζεται χρησιμοποιώντας είτε τη συγχώνευση ή το σύστημα ταξινόμησης, όπως περιγράφεται στο κεφάλαιο 4.

Πιο αναλυτικά, η διαδικασία επιλογής των σπόρων χρησιμοποιείται για να εξάγει μια σειρά από λέξεις σπόρων που θα χρησιμοποιηθούν για την εκτίμηση της συναισθηματικής βαθμολογίας των λέξεων μιας φράσης, χρησιμοποιώντας το προσαρμοσμένο σημασιολογικής ομοιότητας που ήταν το αποτέλεσμα του μείγματος.

$$v_{adapted}(w_j) = \alpha_0 + \sum_{n=1}^{N} \alpha_i \cdot v(w_i) \cdot d_{adapted}(w_i, w_j) \tag{5.2}$$

where $d_{adapted}(w_i, w_j)$ is equal to the $sim(w_j, seed_i)$ from 5.1.

Κεφάλαιο 6

Πειραματική Διαδικασία & Αποτελέσματα

6.1 Περιγραφή Δεδομένων

Corpus

Το κύριο σώμα που χρησιμοποιήθηκε για τα πειράματα ήταν μια γενική, ιστοσυγκομιδής ςορπυς κατεβάσει από το διαδίκτυο. Για να δημιουργήσετε το σώμα, πρώτον, ένα λεξιλόγιο χτίστηκε. Το λεξιλόγιο της αγγλικής συσκευάζονται στην ορθογράφος Ασπελλ για τα αγγλικά, που περιέχει 135.433 λέξεις χρησιμοποιήθηκε για να θέσουν ένα ερώτημα για κάθε λέξη στη μηχανή αναζήτησης Yahoo! και να συλλέγουν αποσπάσματα (σύντομη εκπρόσωπος αποσπάσματα του εγγράφου εμφανίζεται κάτω από το αποτέλεσμα) από τα κορυφαία 500 αποτελέσματα . Κάθε απόσπασμα συνήθως αποτελείται από 2 φράσεις: τίτλος και το περιεχόμενο. Το τελικό ςορπυς περιέχει περίπου 116 εκατομμύρια ποινές. Το κείμενα προ-επεξεργάστηκε:

1. Tokenized: Split corpus to tokens

2. Numbers, stop-words (words like 'i', 'to', 'the' etc), punctuation removed

3. Duplicate lines removed

Το τελικό ςορπυς (μετά την επεξεργασία) περιέχει μοναδικές προτάσεις. Για να δημιουργήσετε ένα ςορπυς των εγγράφων, φράσεις του σώματος ήταν συνεχόμενα διαδοχικά σε ομάδες των 100. Αυτό σημαίνει ότι οι πρώτοι 100 που συλλέγονται ποινές, θα μπορούσε να θεωρηθεί ως το πρώτο έγγραφο, η δεύτερη 100 φράσεις θα ήταν το δεύτερο έγγραφο και ούτω καθεξής . Επειδή το σώμα κατασκευάστηκε με πολλαπλά ερωτήματα των λέξεων με τη συλλογή των πρώτων 500 αποτελέσματα για κάθε ερώτημα, αυτές οι 100 προτάσεις θα περιλαμβάνει κατά πάσα πιθανότητα πολλαπλά παραδείγματα διαφορετικών αισθήσεων μιας λέξης και έτσι μπορούν να σχηματίσουν ένα έγγραφο σχετικά με αυτό το ερώτημα. Στο τέλος έχουμε 900 χιλιάδες έγγραφα.

Λέξεις

Όσον αφορά την λεξικά λέξεις που χρησιμοποιούνται, όπως η συναισθηματική ανθρώπου-σχολιασμένο λεξικό χρησιμοποιούμε την συναισθηματική Πρότυπα για αγγλικές λέξεις (MIA NEA) σύνολο δεδομένων (Bradley and Lang (1999)), το οποίο περιέχει 1034 βαθμολογίες λέξη σε συνεχή διαστάσεις σθένους, διέγερση και την κυριαρχία στο [-1,1]. Αυτό το σύνολο δεδομένων δημιουργήθηκε ως μέρος ενός ψυχολογικού πειράματος ήταν τα θέματα (φοιτητές) κλήθηκαν να βαθμολογήσουν το πώς αισθάνθηκαν κατά την ανάγνωση κάθε λέξη.

Lexicon	# words	Dimension	Scores Range
ANEW	1034	Valence Arousal Dominance	[-1,1]

Table 6.1: ANEW affective lexicon.

Επιπλέον, προκειμένου να εκτιμηθεί η ομοιότητα μεταξύ των λέξεων, υπάρχουν πολλά ζεύγη σημασιολογική σύνολα δεδομένων ομοιότητα. Χρησιμοποιούμε δύο από τα σύνολα δεδομένων πιο καλά γνωρίζουν περιγράφονται στον πίνακα 6.2:

Dataset	Reference	Number of pairs	Scores range
WS-353	Finkelstein et al. (2001)	353	[0,10]
MEN	Bruni et al. (2014)	3000	[0,50]

Table 6.2: Word-pairs similarity datasets.

Προτάσεις

Προκειμένου να εκτιμηθούν οι συναισθηματικές βαθμολογίες για ποινές, χρησιμοποιούμε δύο κύριες σύνολο δεδομένων, ενός γενικού και Twitter μία. Η πρώτη είναι η SemEval 2007 πρωτοσέλιδα σύνολο δεδομένων. Περιέχει 250 φράσεις όπως δεδομένα εκπαίδευσης και 1000 φράσεις όπως στοιχεία για τις δοκιμές, με συνεχή βαθμολογίες σθένους σε [-100 100] (επανκλιμακωθεί στο [-1,1] για πειράματα έξω). Επιπλέον, το ίδιο σύνολο δεδομένων σχολιασμένο για τα έξι βασικά συναισθήματα (θυμός, αηδία, φόβο, χαρά, λύπη, έκπληξη) σε [0,100]. Για τα πειράματά μας χρησιμοποιούμε μόνο τις βαθμολογίες σθένος του συνόλου δεδομένων. Το δεύτερο σύνολο δεδομένων από την SemEval 2016 Ανάλυση Συναισθήματος στο Twitter εργασία 4, και πιο συγκεκριμένα από την Ομάδα B: ‘ Τωεετ ταξινόμηση σύμφωνα με μια δύο-σημείο κλίμακας ¨Αποτελούνταν από 3938 τωεετς ως σύνολο εκπαίδευσης και 10551 tweets ως σύνολο δοκιμών. , με δυαδικό ετικετών (1 για τη θετική και -1 για αρνητικό).

Dataset	# Sentences	Dimension	Ratings
News Headlines SemEval-2007	250 train 1000 test	Valence Emotion	[-1,1] [0,100]
Twitter SemEval-2016 Task B	3938 train 10551 test	Binary	-1, 1

Table 6.3: Sentences datasets for polarity classification.

6.2 Ανάλυση Topics

Για να αξιολογηθούν τα θέματα που κατασκευάσαμε με την βοήθεια του ΛΔΑ, μια ανάλυση πρέπει να πραγματοποιηθεί. Οι παρακάτω ενότητες περιγράφουν τα αποτελέσματα κάθε βήμα του αλγορίθμου, πριν το μοντέλο μείγμα.

LDA Training

Χρησιμοποιώντας τα έγγραφα ςορπυς ως είσοδο εκπαιδεύουμε το μοντέλο ΛΔΑ του Γενσιμ Εργαλειοθήκη με τις ακόλουθες παραμέτρους:

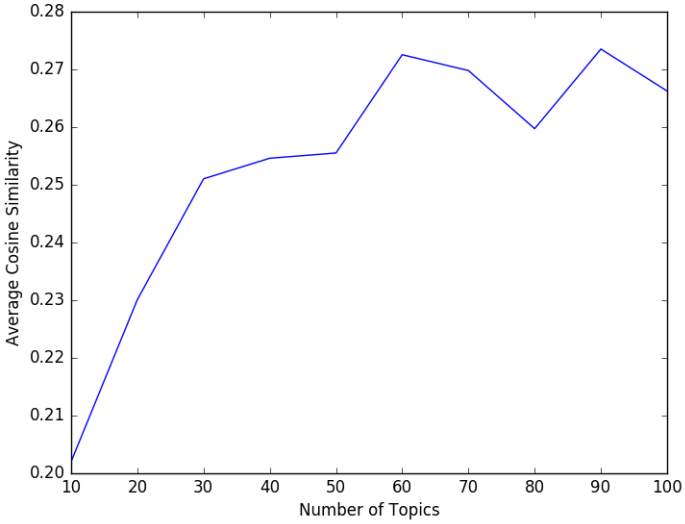
- Number of topics from 10 to 100
- Number of iterations = 200

Πρώτον, θα πρέπει να αξιολογήσει την ποιότητα των θεμάτων που παράγει ΛΔΑ. Μία από τις εξόδους είναι η κατανομή θέμα-λέξη για κάθε θέμα. Χρησιμοποιώντας την κορυφή $n = 10$ πιο συνηθισμένες λέξεις σε κάθε θέμα, όπως επέστρεψε από την ΛΔΑ $t = \{w_1, w_2, ..., w_{10}\}$, μπορούμε να μετρήσουμε την ενδο-θέμα συνοχής ως η μέση ομοιότητα μεταξύ όλων των ζευγαριών (45 ζεύγη αν έχουμε 10 λέξεις από κάθε θέμα) όπως προτείνεται από Newman et al. (2010) και το μέσο όρο όλων των θεμάτων $t \in t$ (εξίσωση 6.1). Λόγος φορείς που θα βοηθήσουν στον υπολογισμό της σημασιολογικής ομοιότητας μπορεί να επιτευχθεί με την κατάρτιση word2vec στο σύνολό ιστοσυγκομιδής ςορπυς που περιγράφεται παραπάνω (φράσεις 90M).

$$Coh_{groupT} = \frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{\substack{1 \leq i \leq n \\ i+1 \leq j \leq n}} sim(w_i, w_j)}{\binom{n}{2}} \tag{6.1}$$

Το μέτρο αυτό περιγράφει τον τρόπο συνεκτικό μπορεί να θεωρηθεί ένα θέμα, με βάση τις σχέσεις ανάμεσα στις λέξεις που την χαρακτηρίζουν. Μια υψηλότερη μέσος όρος βαθμολογίας ομοιότητας θα δείχνουν ότι τα περισσότερα θέματα σε αυτήν την ομάδα είναι καλά ορίζονται από LDA, ενώ χαμηλή βαθμολογία ομοιότητας θα σημαίνει ότι πολύ γενικά θέματα παρήχθησαν. Ο αριθμός των θεμάτων που δίνουν την υψηλότερη μέση ομοιότητα μπορεί να θεωρηθεί ως το καλύτερο δυνατό αριθμό θεμάτων για να εξαγάγετε από το σώμα, όπως οι λέξεις που αντιπροσωπεύουν είναι η πιο στενή σχέση σε σύγκριση με άλλα θέματα.

Παρατηρώντας το σχήμα 6.1, μπορούμε να παρατηρήσουμε ότι σε γενικές γραμμές 60 και 90 θέματα φαίνεται να είναι η πιο συνεκτική, καθώς σύμφωνα με το προηγούμενο θεώρημα,



Σχήμα 6.1: Intra-topic coherence measured after LDA training.

οι λέξεις που τις χαρακτηρίζουν έχουν μεγάλη ομοιότητα μεταξύ τους. Παρά το γεγονός ότι μπορούμε να αναμένουμε ο αριθμός των θεμάτων για να επιτευχθεί η καλύτερη δυνατή απόδοση στα παρακάτω πειράματα, αυτό είναι κάτι που εξαρτάται από τον αριθμό των δεδομένων που συλλέγονται για κάθε θέμα στην ομάδα. Μια άλλη μετρική που μπορεί να χρησιμοποιηθεί είναι η μέτρηση της ενδο-θέμα της συνοχής, που σημαίνει ότι η πραγματική απόσταση μεταξύ αυτών των θεμάτων. Υψηλότερη απόσταση σημαίνει ότι τα θέματα είναι ανεξάρτητα και μοναδικά ενώ η χαμηλή απόσταση εμφανίζει περιορισμένη δύναμη των θεμάτων διάκριση από ΛΔΑ. Ωστόσο, αν ένα θέμα είναι ένα δευτερεύον θέμα του άλλου θα είναι πολύ στενή και, κατά συνέπεια, η απόστασή τους θα είναι πολύ χαμηλή. Ο καλύτερος τρόπος για να μετρηθεί πραγματικά το μεταξύ το θέμα συνοχής είναι να απεικονίσει αυτά τα μοντέλα. Για να το κάνουμε αυτό, χρησιμοποιούμε την εργαλειοθήκη [1] η οποία είναι σε θέση να ανιχνεύει όχι μόνο την οριακή θέμα διανομές αλλά και την σημαντικότητα του λεξιλογίου σε κάθε θέμα. Για λόγους πληρότητας μπορούμε να εισαγάγει την έννοια της σαλιεςψ και του διακριτικού χαρακτήρα (**?**). Για μια δεδομένη λέξη β, θα πρέπει να υπολογίσουμε της δεσμευμένης πιθανότητας $P(T|w)$, η πιθανότητα που παρατηρείται λέξη β δημιουργήθηκε από λανθάνουσα θέμα T. Μπορούμε επίσης να υπολογίσουμε την *posterior* Π (T) ως η συνολική πιθανότητα θέματος T στη συλλογή. Ο διακριτικός χαρακτήρας ενός λέξη ορίζεται ως εξής:

$$Distinctiveness(w) = \sum_{T} P(T|w) log \frac{P(T|w)}{P(T)} = KL(P(T|w)||P(T)) \tag{6.2}$$

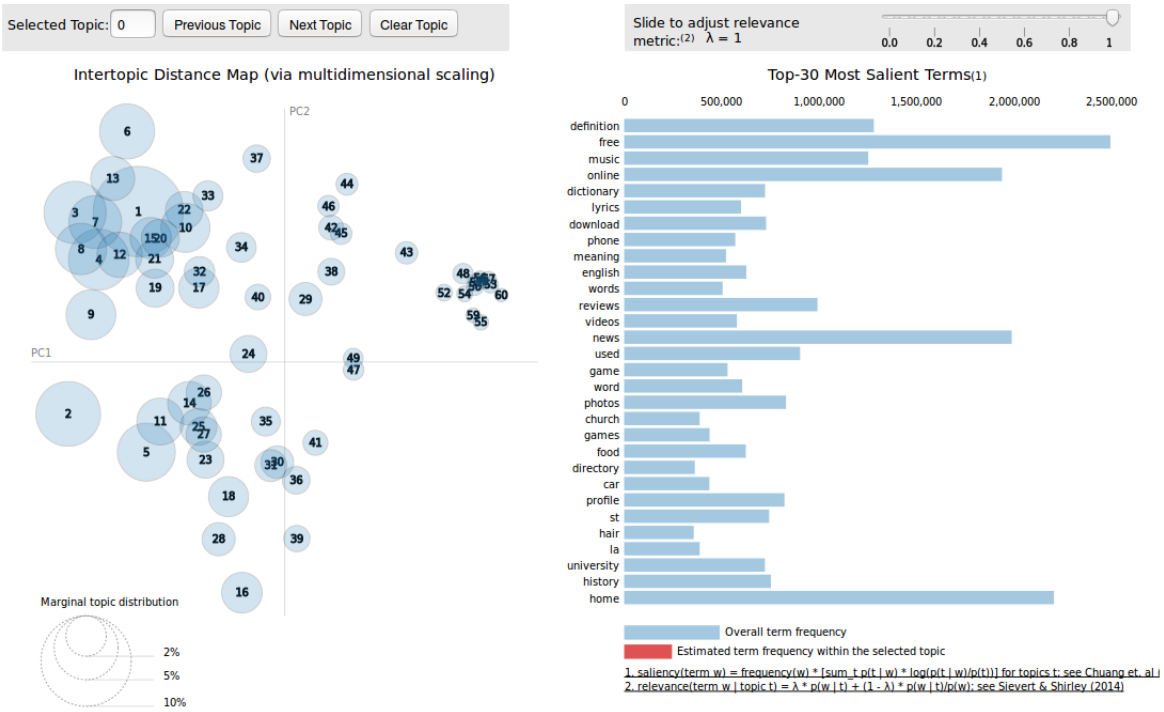
η οποία είναι η Kullback-Leibler (KL) απόκλιση (Kullback and Leibler (1951)) μεταξύ, η κατανομή θέμα $P(T|w)$ δοθεί η λέξη W, και την οριακή κατανομή θέμα $P(T)$, η πιθανότητα ότι οποιαδήποτε τυχαία λέξη έχει συνταχθεί από το θέμα T. Τα μέτρα διακριτικός λέξη πόσο

[1] ηττπς://ςραν.ρ-προθεςτ.οργ/ωεβ/παςκαγες/ΛΔΑις/ινδεξ.ητμλ

κατατοπιστική είναι μια λέξη για το θέμα σε σχέση με μια τυχαία επιλεγμένη λέξη.

$$Saliency(w) = P(w) \cdot Distinctiveness(w) \tag{6.3}$$

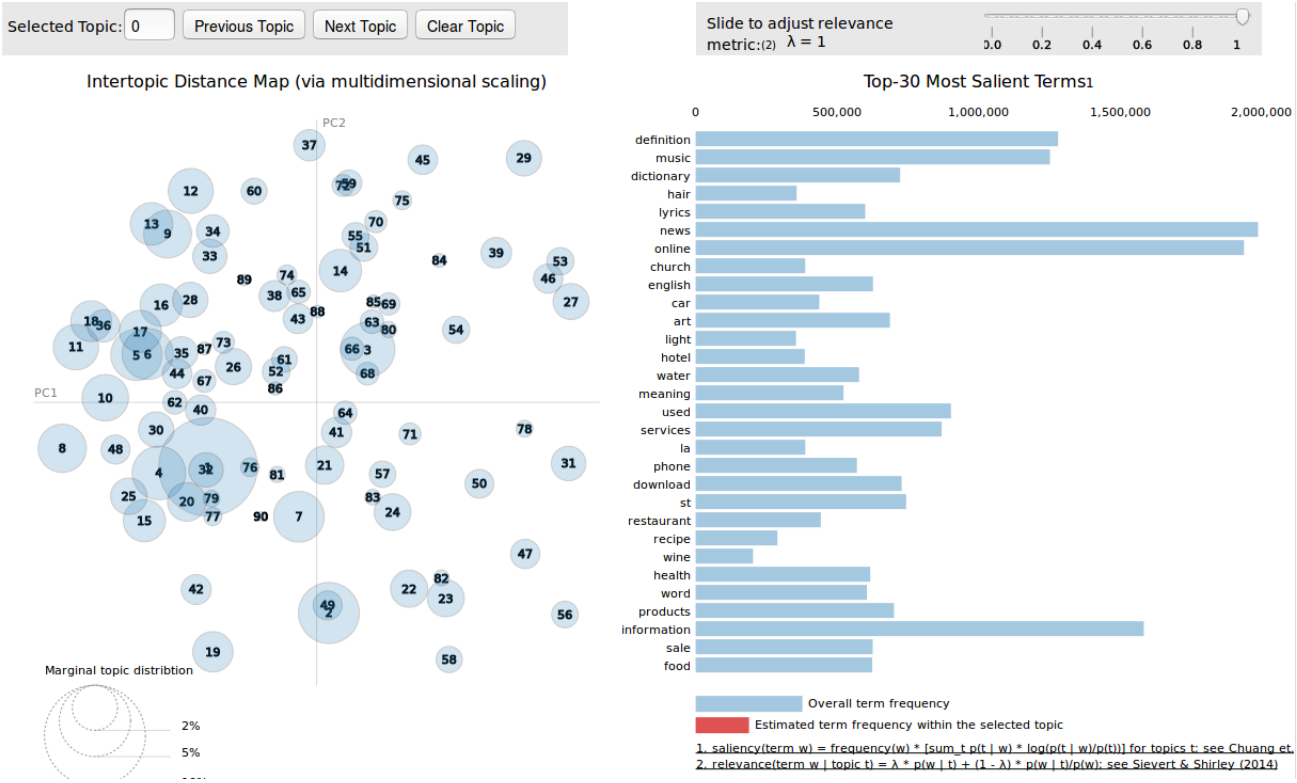
Μιας λέξης W σταθμίζοντας τη συχνότητα του από τον διακριτικό χαρακτήρα του. Σε σύγκριση με την κατάταξη ανά συχνότητα $P(w)$, η κατάταξη με την σημαντικότητα θα τιμωρήσει τις λέξεις από κοινού σε διάφορα θέματα, όπως αυτά θα έχουν χαμηλό διακριτικό χαρακτήρα, και να ενισχύσει τις λέξεις που είναι καλοί προάγγελοι ενός θέματος, καθώς θα έχουν ένα υψηλό διακριτικό χαρακτήρα. Ένα παράδειγμα για μια πιθανή ταξινόμηση φαίνεται στα σχήματα ρεφ 60Τ, ρεφ 90 τόνοι, μαζί με μια υπερ-σύνδεση για μια διαδραστική οπτική.



Σχήμα 6.2: Example of visualization for 60 topics (*60 Topics Example Link*).

Κατασκευή Topics

Το επόμενο βήμα, είναι να κατασκευάσει κάθε θέμα κατατάσσοντας προτάσεις από το σώμα σε διάφορα συμπλέγματα θέμα. Τα θέματα μπορούν να ταξινομηθούν χρησιμοποιώντας ένα όριο οπίσθια ή να ταξινομηθεί σε ένα μόνο θέμα, το θέμα με τη μέγιστη οπίσθια πιθανότητα. Αυτό σημαίνει ότι μια πρόταση έχει ταξινομηθεί σε περισσότερα από ένα θέματα, εάν η πιθανότητα να ανήκουν σε αυτό το θέμα είναι $> thres$ ή έχει ταξινομηθεί μόνο για την πιο πιθανή θέμα. Ο αριθμός των προτάσεων που συλλέγονται με κάθε μέθοδο ανά θέμα παρουσιάζονται στον πίνακα 6.4 χρησιμοποιώντας ένα οπίσθιο όριο του 0,1. Όπως μπορούμε να παρατηρήσουμε τον αριθμό των προτάσεων που συλλέγονται με τη μέγιστη πιθανότητα είναι πολύ χαμηλό, ειδικά για μεγάλο αριθμό θεμάτων. Το μέγεθος των θεμάτων είναι ένας σημαντικός παράγοντας για την εκπαίδευση των θέμα-ΔΣΜς. Για το λόγο αυτό, για τα ακόλουθα πειράματα χρησιμοποιούμε το οπίσθιο όριο του 0,1, το οποίο μας δίνει τη δυνατότητα να εκτελέσει



Σχήμα 6.3: Example of visualization for 90 topics (*90 Topics Example Link*).

	0.1 Posterior			Max Posterior		
Topics	Min	Max	**Mean**	Min	Max	**Mean**
10	10.8 M	20.7 M	14.6 M	5.2 M	12.9 M	8.9 M
20	4.9 M	15 M	7.7 M	2.3 M	10.1 M	4.4 M
30	2.4 M	11.9 M	5.4 M	830 K	7.4 M	2.9 M
40	2.3 M	8.4 M	4.3 M	692 K	5 M	2.2 M
50	1.2 M	7.7 M	3.6 M	338 K	4.8 M	1.7 M
60	1.2 M	9.4 M	3.1 M	292 K	5.2 M	1.4 M
70	604 K	5.3 M	2.7 M	99 K	3.3 M	1.2 M
80	297 K	6.2 M	2.4 M	44 K	3.4 M	1.1 M
90	412 K	6.4 M	2.2 M	56 K	3.6 M	987 K
100	277 K	7.4 M	2 M	42 K	4.6 M	888 K

Table 6.4: Statistics for number of sentences collected per topic using a posterior probability threshold = 0.1 and the max posterior probability.

μια ομαλή ομαδοποίηση όπως είμαστε σε θέση να εισπράξει περίπου διπλάσιο αριθμό προτάσεων όταν χρησιμοποιείτε τη μέγιστη οπίσθια. Με βάση την υπόθεση ότι εάν μια ποινή έχει ταξινομηθεί σε όλα τα θέματα με την ίδια πιθανότητα είναι πολύ ασαφής για να ταξινομηθούν σε ένα θέμα, θα υπάρξουν προτάσεις που δεν θα ταξινομηθεί οπουδήποτε. Όπως μπορούμε να

Topics	# sentences
10	828360
20	828360
30	828360
40	828362
50	828361
60	828370
70	828382
80	828382
90	828416
100	828430

Table 6.5: Number of sentences not classified.

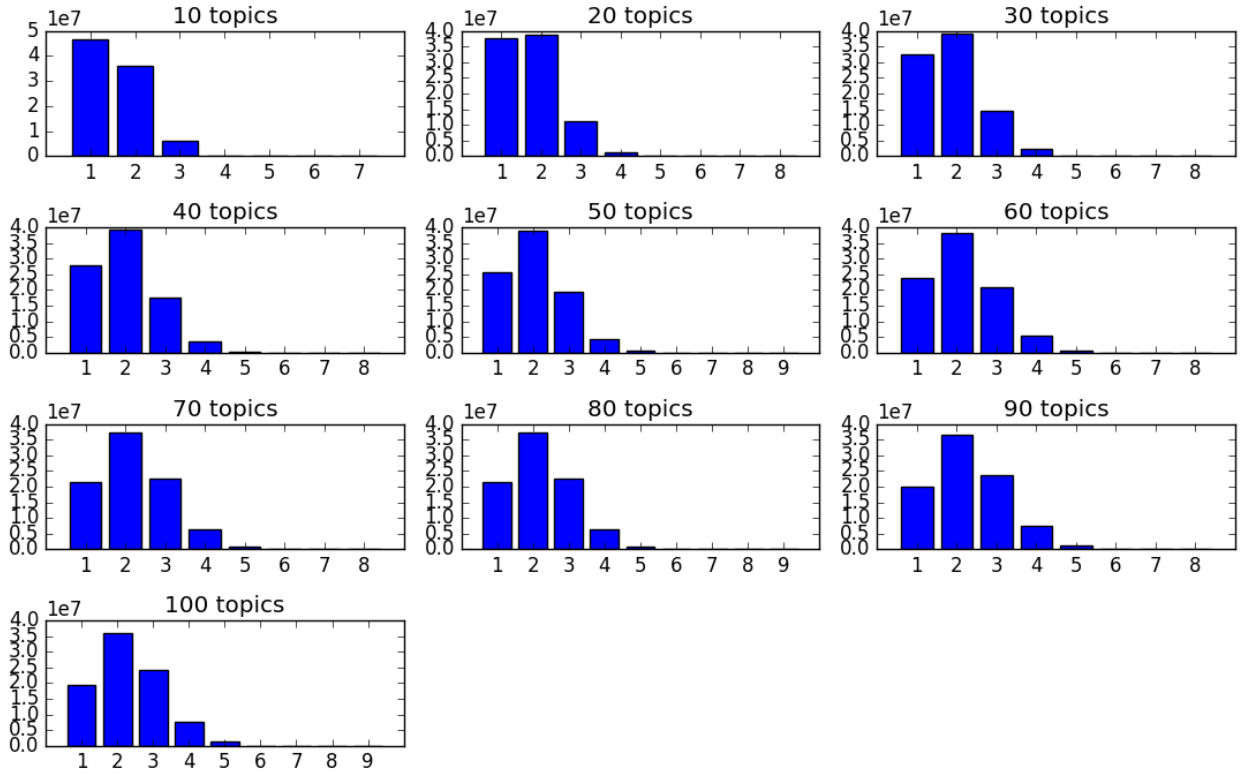
παρατηρήσουμε από τον πίνακα 6.5, ο αριθμός των 'θυνκ' φράσεις παραμένει σχεδόν ανέπαφη μέσα από όλους τους αριθμούς των θεμάτων, που δείχνουν ότι είναι πιθανόν κάποιες τυπικές φράσεις θορύβου, τα οποία συλλέγονται κατά τη διάρκεια της διαδικασίας κατασκευής σωμάτων που πρέπει να αγνοηθεί . Είναι χρήσιμο τώρα να παρατηρήσουμε πώς πολλές προτάσεις κατατάσσονται σε ένα, δύο, τρία ή Τ θέματα. Εικόνα **;;** δείχνει διαφορετικά οικόπεδα γραμμή με τον αριθμό των θεμάτων που μια φράση μπορεί να ταξινομηθεί μαζί με αυτές τις προτάσεις μετράνε.

Σημασιολογικά Μοντέλα

Στο στάδιο αυτό, ένα διαφορετικό σημασιολογικό μοντέλο έχει εκπαιδευτεί για κάθε θέμα. Προκειμένου να αποφασίσει ποια είναι τα καλύτερα παράμετροι για την εκπαίδευση των σημασιολογικά μοντέλα, έχουμε πραγματοποιήσει ένα πείραμα με τα μεγέθη διαφορετικά σώματα, τις διαστάσεις φορέα και μεγέθη παραθύρων, χρησιμοποιώντας την αρχιτεκτονική ΈΒΟΩ της Ωορδ2ές και 5 επαναλήψεις. Ειδικότερα, εξάγουμε τυχαία φράσεις από το γενικό, ιστοσυγκομιδής ςορπυς και αποτελούν υπο-σώματα. Δεδομένου ότι αυτή η διαδικασία είναι τυχαία, κάνουμε το ίδιο πείραμα 5 φορές και να πάρετε το μέσο όρο βαθμολογίας συσχέτισης Σπεαρμαν κατά την αξιολόγηση στο σύνολο δεδομένων MEN.

6.3 Σημασιολογική Ομοιότητα

Τώρα μπορούμε να προχωρήσουμε στην αξιολόγηση του μοντέλου μίγματος, που εκτελεί τη σημασιολογική χώρο προσαρμογή της κάθε πρότασης. Καθώς εργαζόμαστε στο χώρο της έννοια, μπορούμε επίσης να ελέγξετε την απόδοση αυτού του μοντέλου σε μια σημασιολογική εργασία ομοιότητας σε επίπεδο λέξης, δηλαδή τη μέτρηση της ομοιότητας μεταξύ των ζευγαριών των λέξεων. Προκειμένου να εφαρμόσει το εκπαιδευμένο μοντέλο ΛΔΑ σε αυτά τα σύνολα δεδομένων, θεωρούμε ως ποινή ένα απλό ζευγάρι, $frsh = w_1w_2 2$. Άλλες πιθανές λύσεις μπορούν να εξεταστούν οι οποίες παρουσιάζονται στο κεφάλαιο 7. Και πάλι, μετράμε



Σχήμα 6.4: Number of sentences classified per number of topics.

τον αριθμό των θεμάτων κάθε ζεύγος έχει ταξινομηθεί (εικόνα ;;) και φαίνεται ότι τα περισσότερα ζεύγη κατατάσσονται σε ένα θέμα, ενώ πολύ λιγότερο ταξινομούνται σε δύο θέματα και τα τρία θέματα. Αυτό είναι το επιθυμητό αποτέλεσμα, όπως είναι λογικό ότι είναι πολύ δύσκολο ένα ζεύγος λέξεων για να εκφράσουν περισσότερες από μία έννοια. Ως εκ τούτου, μπορούμε να υποθέσουμε ότι οι λέξεις που δεν είναι τόσο όμοια μπορούν να ταξινομηθούν σε περισσότερα από ένα θέμα όπως ΛΔΑ δεν μπορεί να εκτιμήσει ένα συγκεκριμένο θέμα για αυτούς.

Dataset	# High Sims (≥ 0.75)	# Median Sims ($[0.25, 0.75)$)	# Low Sims (< 0.25)
WS-353	94	227	32
MEN	595	1764	641

Table 6.6: Number of pairs for different similarity ranges for each dataset.

Και για τα δύο σύνολα δεδομένων, παρατηρείται το ίδιο μοτίβο. Για το σύνολο δεδομένων MEN, μπορούμε να παρατηρήσετε μια βελτίωση της τάξεως του 10 % για την υψηλή ζεύγη ομοιότητας, ενώ για χαμηλό και μεσαίο αυτά η βελτίωση είναι της τάξης του 3 % για το μεσαίο ομοιότητες και λίγο τα θέματα και 5-7 % για χαμηλή ομοιότητες. Αυτή είναι μια αναμενόμενη συμπεριφορά. Συγκεκριμένα, γνωρίζουμε ότι δύο λέξεις αποτελούν ήδη ένα πλαίσιο. ΛΔΑ θα εκτιμήσει τις πιο πιθανό θέματα (πλαίσιο), ότι ένα ζευγάρι των λέξεων μπορεί να ανήκει. Αν

Topics	High Sims (> 0.75)	Median Sims	Low Sims (< 0.25)	Overall
10	0.2185	**0.6341**	0.2993	**0.7977**
20	0.2394	0.6227	0.2788	0.7972
30	0.2279	0.6109	**0.3183**	0.7881
40	**0.2448**	0.5578	0.3098	0.7502
50	0.2105	0.5598	0.2979	0.7512
60	0.2355	0.5522	0.2836	0.7492
70	0.2233	0.5312	0.2856	0.7398
80	0.2144	0.5341	0.2871	0.7337
90	0.2075	0.5388	0.2685	0.7368
100	0.2034	0.5173	0.2917	0.7151
No Topics	**0.1568**	0.5998	**0.2459**	**0.7728**

Table 6.7: Spearman correlation for MEN dataset, using different similarity ranges.

Topics	High Sims (> 0.75)	Median Sims	Low Sims (< 0.25)	Overall
10	0.5030	**0.5256**	0.0632	**0.7212**
20	0.5188	0.5021	0.2570	**0.7248**
30	**0.5536**	0.4920	0.1858	**0.7279**
40	0.4940	0.4807	0.2413	0.6673
50	0.4697	0.4460	0.0013	0.6590
60	0.4624	0.3950	0.0545	0.6332
70	0.4938	0.4612	0.3419	0.6538
80	0.4837	0.4750	0.2313	0.6913
90	0.5056	0.4603	0.2189	0.6729
100	0.4916	0.3877	0.2910	0.6428
No Topics	0.5014	0.4990	-0.0951	**0.7030**

Table 6.8: Spearman correlation for WordSim353 dataset, using different similarity ranges.

αυτές οι λέξεις έχουν στενή σχέση (υψηλή ομοιότητα) θα είμαστε σε θέση να προσδιορίσει το ακριβές θέμα στο οποίο συν-συμβαίνουν με υψηλή πιθανότητα. Ως ένα συγκεκριμένο σημασιολογικό μοντέλο είχε εκπαιδευτεί για αυτό το θέμα, η ομοιότητα μεταξύ αυτών των λέξεων θα είναι πολύ υψηλό. Ως αποτέλεσμα, η τελική ομοιότητα ζεύγος θα είναι καλύτερη σε αυτόν τον τομέα (αν και δεν μπορεί να είναι υψηλότερη), σε σύγκριση με το βασικό μοντέλο που χρησιμοποιεί μια σημασιολογική μοντέλο εκπαιδεύονται σε μία γενική ϱορπυς και επηρεάζεται από πολλούς άλλους τομείς στο σώμα. Επιπλέον, εάν οι δύο λέξεις είναι εντελώς ανόμοια, η ΛΔΑ θα έχουν πρόβλημα ταξινόμησης τους σε ένα θέμα που περιέχει και τα δύο. Για το λόγο αυτό, το πιο πιθανό σενάριο είναι να επιστρέψει ότι ανήκουν σε όλα τα θέματα με την ίδια πιθανότητα. Σε κάθε θέμα, αυτές οι λέξεις θα έχουν ένα πολύ χαμηλό σκορ ομοιότητας, όπως είναι ανόμοια, και το μείγμα θα τους εκχωρήσετε μια τιμή μεταξύ της χειρότερης και

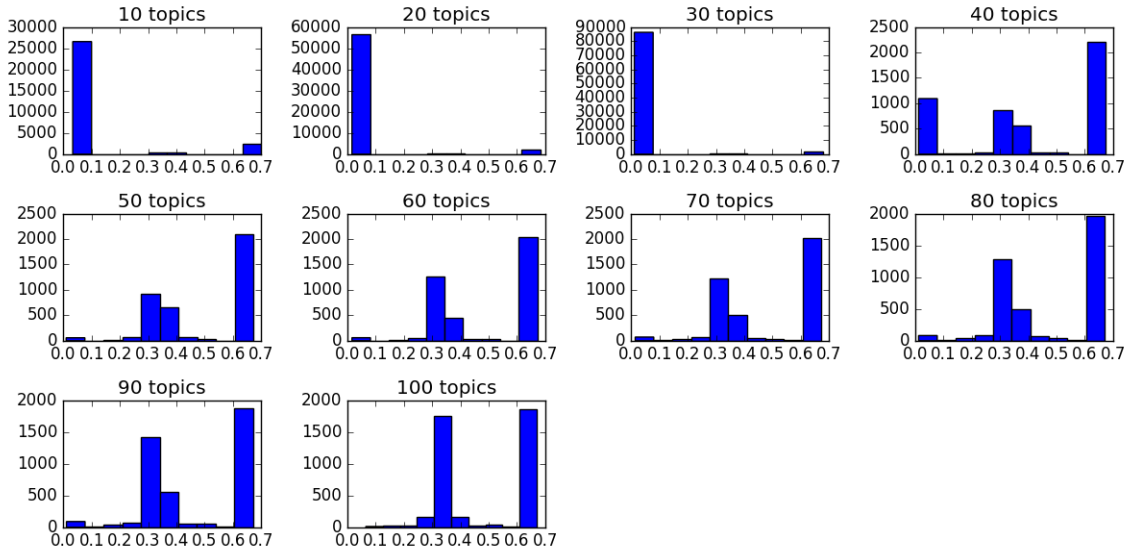
στην καλύτερη περίπτωση ομοιότητας τους, οδηγώντας σε μια πιο ακριβής υπολογισμός της πραγματικής τους ομοιότητα. Για το σύνολο δεδομένων ΩΣ-353, τα χαμηλά ομοιότητες είναι πολύ λίγες και ως εκ τούτου δεν μπορούμε να εξηγήσουμε αυτά τα αποτελέσματα. Προκει-

	High Sims		Median Sims		Low Sims	
Topics	# pairs	Posteriors Std	# pairs	Posteriors Std	# pairs	Posteriors Std
10	595	0.1976	1764	0.193	641	0.1751
20	595	0.1437	1764	0.1386	641	0.1212
30	595	0.1171	1764	0.1124	641	0.0984
40	42	0.0379	354	0.0287	404	0.0156
50	48	0.0487	420	0.0319	409	0.0166
60	36	0.0432	458	0.0252	435	0.0132
70	55	0.0425	454	0.0262	422	0.0116
80	70	0.0491	507	0.0291	429	0.0122
90	68	0.049	595	0.0297	436	0.0093
100	74	0.0402	574	0.0261	449	0.0111

Table 6.9: Number of pairs for MEN dataset, that are classified to more than one topic along with the standard deviation of their posterior probabilities for the topics they belong.

μένου να ενισχυθεί αυτή η εξήγηση, πίνακα ρεφ ςλασσ-ένα δείχνει τον αριθμό των ζευγαριών που είχαν ταξινομηθεί σε περισσότερες από ένα θέμα για τις διάφορες κλίμακες της ομοιότητας. Έχει παρατηρηθεί ότι για υψηλές ομοιότητες (595 ζεύγη) μόνο λίγα έχουν ταξινομηθεί σε περισσότερα από ένα θέμα. Για τις χαμηλές ομοιότητες, περίπου όλοι κατατάσσονται σε περισσότερα από ένα θέμα (όπως αναμενόταν) και για τη διάμεσο το αποτέλεσμα εξαρτάται από τις συνδέσεις μεταξύ των δύο λέξεων και εάν μπορούν να ανήκουν σε ένα υπάρχον θέμα. Είναι επίσης σημαντικό να σημειωθεί ότι για 10, 20 και 30 θέματα, όλα τα ζεύγη κατατάσσονται σε όλα τα θέματα. Αυτό είναι λογικό, διότι τα θέματα αυτά είναι αρκετά γενικές και ακόμη και όροι που είναι ιδιαίτερα σχετικό θα αποκτήσει μια πιθανότητα 0,1 για το υπόλοιπο των θεμάτων εκτός από εκείνο που μπορεί περισσότερο πιθανώς ανήκουν. Αυτό φαίνεται επίσης στο σχήμα ρεφ μετα-άνδρες όπως τα ιστογράμματα για τους άνδρες σύνολο δεδομένων πιθανότητες κατά τη διάρκεια της ταξινόμησης ΛΔΑ. Η υψηλότερη οπίσθια πιθανότητα ένα ζευγάρι μπορεί να λάβει είναι 0.67 και μπορεί κανείς να υποθέσει ότι το μοντέλο ΛΔΑ είναι προκατειλημμένη. Ωστόσο, αυτό συμβαίνει μόνο για το σύνολο δεδομένων ζευγών όπως θα δειχθεί στη συνέχεια.

Τέλος, λαμβάνοντας υπόψη τον αριθμό των θεμάτων στη συνολική απόδοση, η βελτίωση παρατηρείται για 30 και 40 θέματα, κάτι που έρχεται σε αντίθεση με τη συνοχή αυτών των μοντέλων, όπως μετρήθηκε στο προηγούμενο τμήμα. Αυτό είναι κάτι που σχετίζεται άμεσα με τον αριθμό των δεδομένων που συλλέγονται για τα θέματα αυτά. Όπως φαίνεται στους πίνακες ρεφ ΙΤΕ-άνδρες, ρεφ ΙΤΕ-ΩΣ με την αύξηση του αριθμού των επαναλήψεων για την κατασκευή της κατανομής σημασιολογικό μοντέλο φαίνεται να βελτιώνει τις υψηλές ομοιότητες μόνο για 60 θέματα (αλλά όχι η συνολική βαθμολογία) στην ΩΣ-353 σύνολο δεδομένων,



Σχήμα 6.5: Posteriors histogram for MEN dataset.

ενώ στους άνδρες το σύνολο δεδομένων βελτίωση παρατηρείται για τα μεσαία ομοιότητες και κατά συνέπεια τη συνολική βαθμολογία. Για ένα μικρό αριθμό θεμάτων (συνήθως 10-20), η συνολική απόδοση μπορεί να ξεπεράσει τη γραμμή βάσης (καθώς αυξάνει τη μέση εκτίμηση ομοιότητες). Αυτό μπορεί να εξηγηθεί αν λάβουμε υπόψη τα εξής: Κάθε ένα από τα 10 θέματα, είναι αρκετά μεγάλο (φράσεις 15M) έτσι ώστε όλες οι σημασιολογικές μοντέλα μπορούν να εκπαιδευτούν επαρκώς. Επιπλέον, η ομαδοποίηση του σώματος σε 10 ασαφή θέματα, αφαιρώντας `φράσεις θυνκ», μειώνει το θόρυβο στο σώμα. Στην πραγματικότητα, κάποια ενίσχυση γίνεται. Τα ασαφή θέματα μπορεί να εκτιμήσει και την ομοιότητα μεταξύ όλων των ειδών των λέξεων (πολύ παρόμοια, δεν είναι τόσο παρόμοια, ανόμοια) και ο συνδυασμός των διαφορετικών ομοιότητες ως αποτέλεσμα μια καλύτερη εκτίμηση του συνόλου δεδομένων.

6.4 Εκτίμηση Συναισθήματος

6.4.1 Baseline Training

Για την ΣΜΑ, οι συντελεστές α εκπαιδεύονται στο σύνολό λεξικού εκ νέου, με την παράμετρο λ και θα παραμείνει το ίδιο για το σύστημα βασικής γραμμής (που δεν ενσωματώνει θέματα) και το ΣΜΑ, όπως σε αυτό το μοντέλο θα προσαρμοστούν *μόνο* το σημασιολογικό χώρο και όχι η χαρτογράφηση μεταξύ της έννοιας και της συναισθηματικής χώρο.

6.4.2 Semantic Model Adaptation

News Headlines

Οι μεταγενέστεροι ιστόγραμμα για τα 1000 φράσεις που φαίνεται στην εικόνα ρεφ τεστ-ποστ. Μπορούμε να παρατηρήσουμε ότι για τις 10 και 20 θέματα, η πλειοψηφία των ποστεριορς

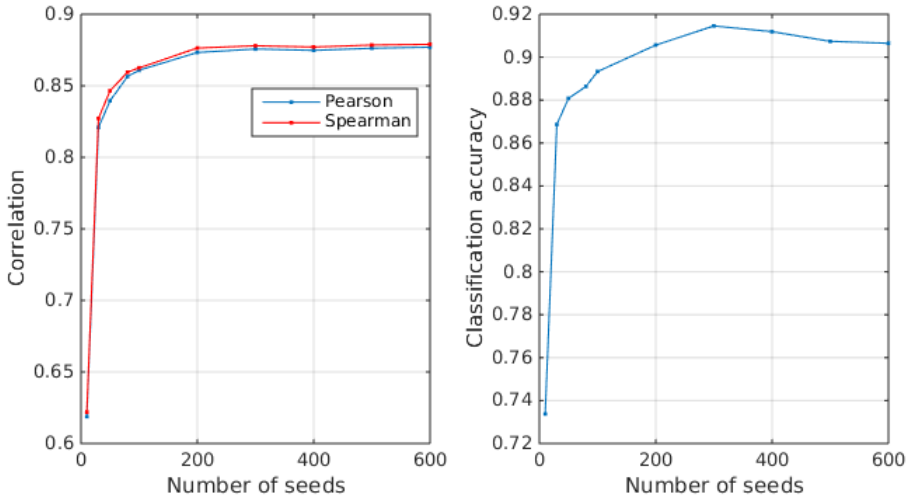
	Iterations	High Sims	Median Sims	Low Sims	Overall
30 topics	1	0.5529	0.4430	0.1555	0.7020
	5	0.5536	0.4920	0.1858	0.7279
	10	0.5430	0.4622	0.1092	0.7087
	15	0.5439	0.4538	0.1678	0.7003
60 topics	1	0.4464	0.3437	0.0707	0.5957
	5	0.4624	0.3950	0.0545	0.6332
	10	0.4996	0.3747	0.0907	0.6158
	15	0.5019	0.3790	0.0683	0.6129
90 topics	1	0.4111	0.4297	0.1329	0.6078
	5	0.5056	0.4603	0.2189	0.6729
	10	0.4746	0.4562	0.1963	0.6468
	15	0.4722	0.4747	0.1423	0.6659

Table 6.10: Different number of iterations when trained the semantic model for 30, 60 and 90 topics for WS-353 dataset.

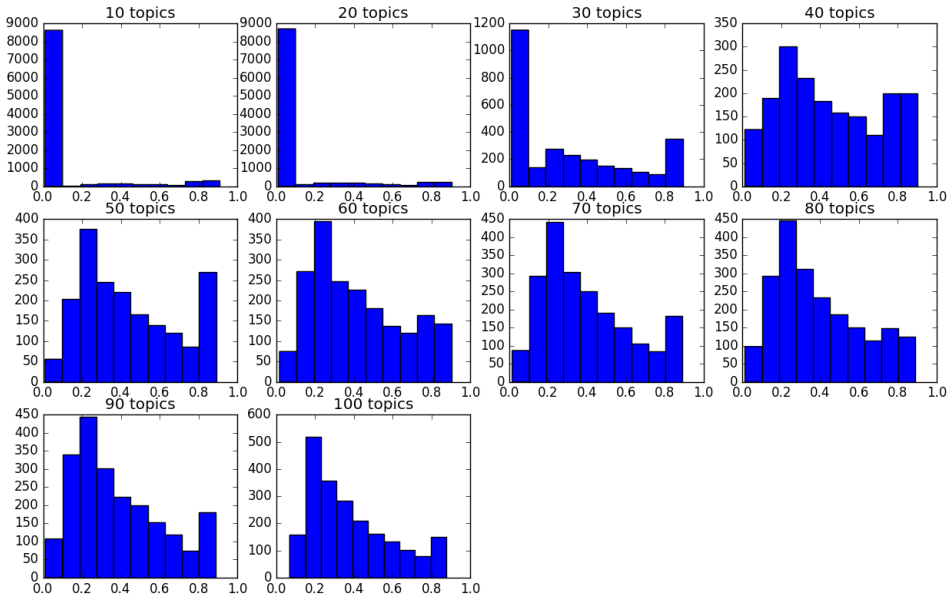
	Iterations	High Sims	Median Sims	Low Sims	Overall
30 topics	1	0.1684	0.5837	0.2750	0.7596
	5	0.2279	0.6109	0.3183	0.7881
	10	0.2109	0.6036	0.2951	0.7792
	15	0.1977	0.5986	0.2991	0.7738
60 topics	1	0.1688	0.4873	0.2204	0.6678
	5	0.2355	0.5522	0.2836	0.7492
	10	0.2186	0.5602	0.2566	0.7521
	15	0.2135	0.5602	0.2335	0.7512
90 topics	1	0.1531	0.4308	0.1665	0.6305
	5	0.2075	0.5388	0.2685	0.7368
	10	0.1954	0.5457	0.2627	0.7451
	15	0.2017	0.5546	0.2660	0.7440

Table 6.11: Different number of iterations when trained the semantic model for 30, 60 and 90 topics for MEN dataset.

αποκτήσει πολύ μικρές τιμές, ενώ όταν τα θέματα αυξήσει όλες οι πιθανές περιοχές που καλύπτονται. Η συμπεριφορά είναι αναμενόμενη καθώς όταν τα θέματα είναι πολύ γενικό και αφηρημένο, οι περισσότεροι ποινές θα μπορούσαν να ταξινομηθούν σε όλα αυτά και, κατά συνέπεια, όλα τα θέματα αποκτήσει μια κοινή μικρή πιθανότητα (0.1), εκτός από την πιο πιθανή θέμα. Από την άλλη πλευρά, όταν τα θέματα γίνονται πιο συγκεκριμένοι, ΛΔΑ μπορεί να τους διακρίνει καλύτερα και ταξινομεί τις φράσεις με μεγαλύτερη ακρίβεια οδηγούν σε μεγαλύτερες οπίσθια πιθανότητες.

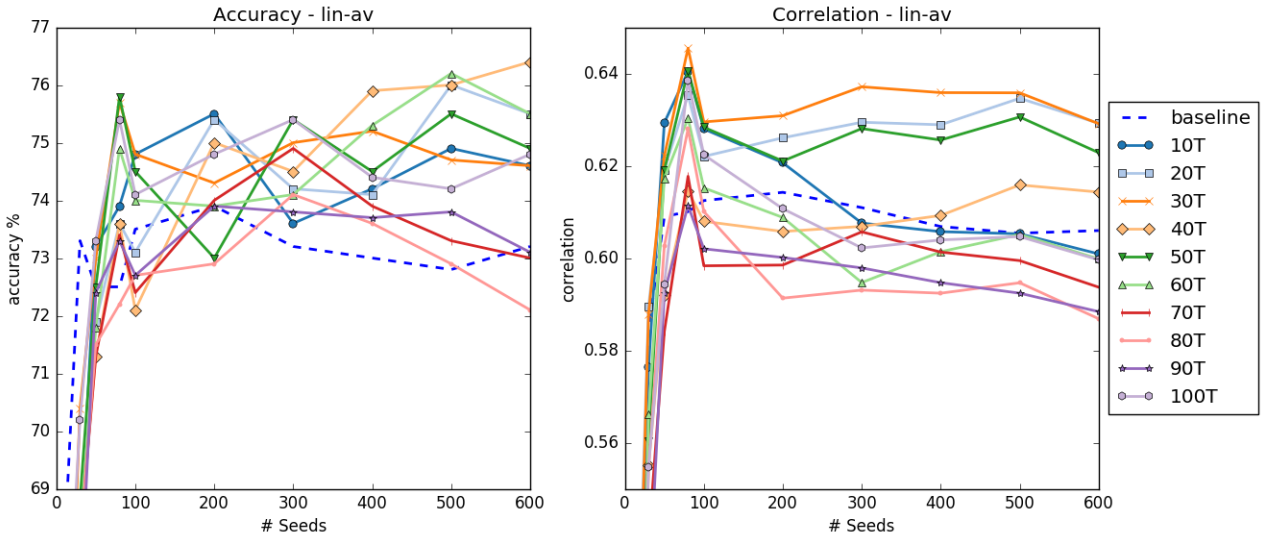


Σχήμα 6.6: 10-fold Cross Validation on ANEW dataset using Ridge Regression and $\lambda = 0.35$.

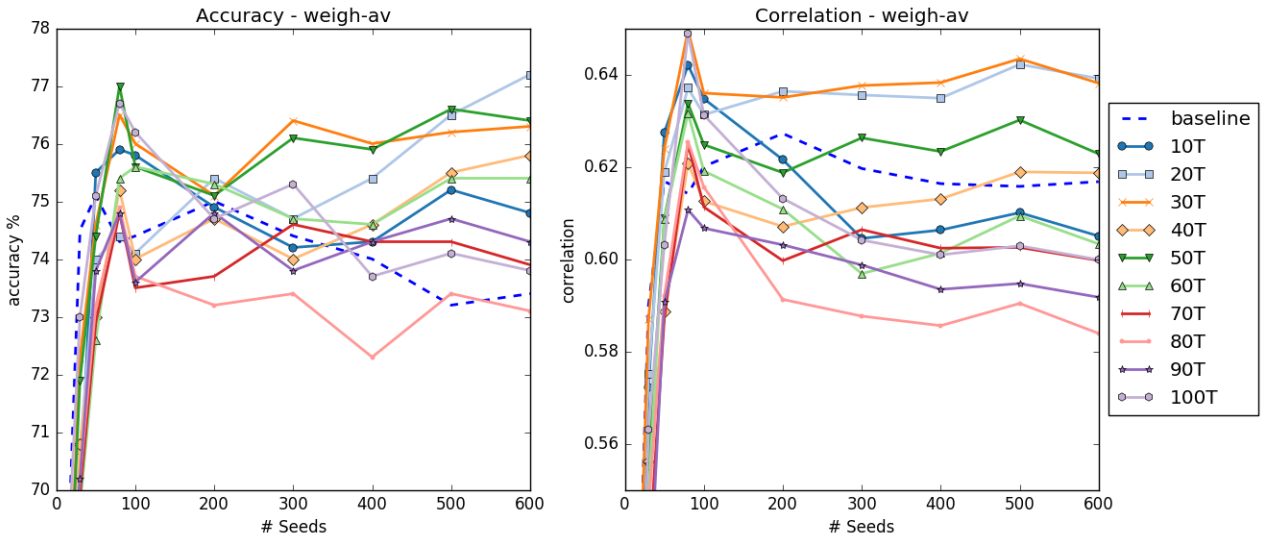


Σχήμα 6.7: Posteriors histogram for 1000 news headlines sentences.

Τα αποτελέσματα που παρουσιάζονται, δείχνουν μια σημαντική βελτίωση σε σύγκριση με το σύστημα βασικής γραμμής (χωρίς θέματα) που χρησιμοποιεί μόνο ένα γενικό σημασιολογικό μοντέλο. Είναι σημαντικό να σημειωθεί ότι τόσο η ακρίβεια και η συσχέτιση λαμβάνει μέγιστες τιμές τους κατά τη χρήση 80 λέξεις σπόρων. Αυτό σημαίνει ότι ίσως μια διαδικασία επιλογής των σπόρων πρέπει να πραγματοποιηθεί, προκειμένου να εντοπίσει λέξεις σπόρος που είναι το θέμα-συγκεκριμένες. Στην πραγματικότητα, όταν εφαρμόζουμε τη μέθοδο επιλογής σπόρων (όπως περιγράφεται στο κεφάλαιο ρεφ Κεφάλαιο 4) επιλέγουμε τις λέξεις με υψηλή απόλυτη

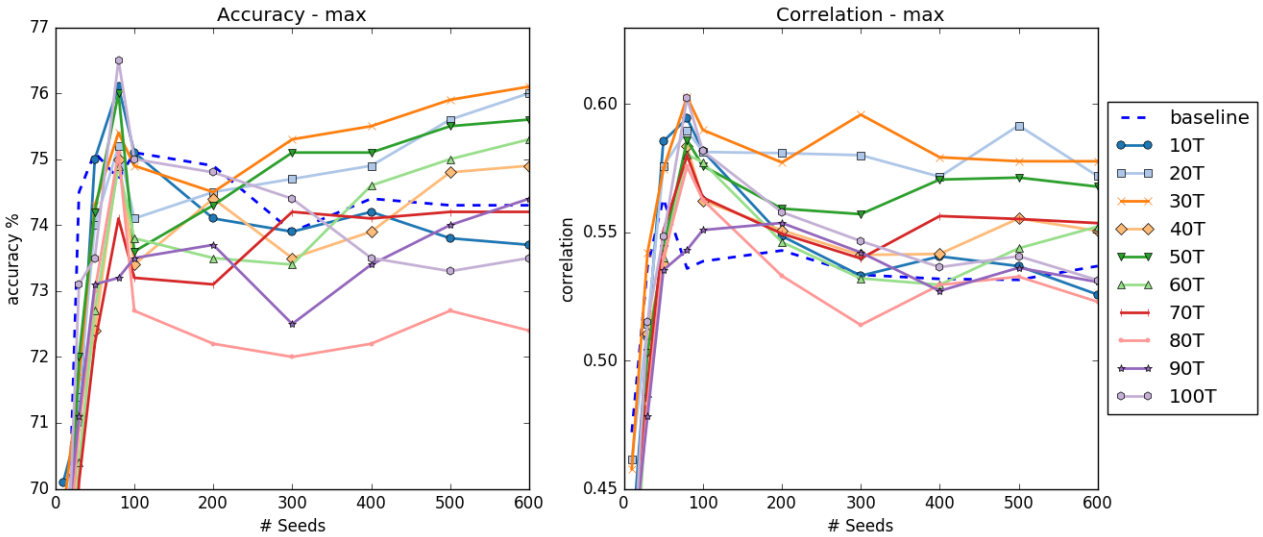


Σχήμα 6.8: Accuracy and Spearman Correlation using linear fusion.

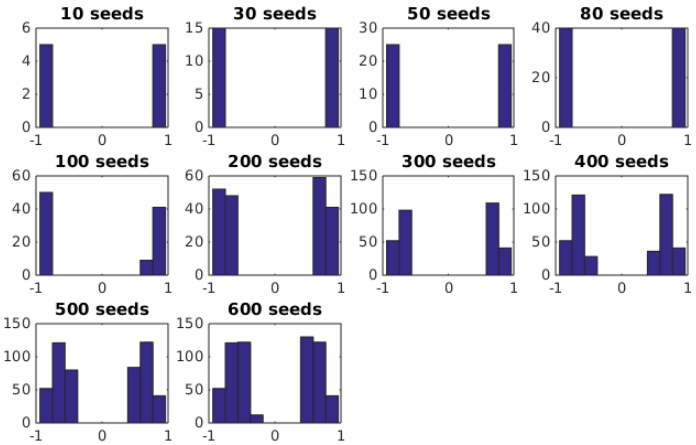


Σχήμα 6.9: Accuracy and Spearman Correlation using weighted fusion.

βαθμολογία σθένους. Σε γενικές γραμμές, μπορούμε εύλογα να υποθέσουμε, ότι τα λόγια με πολύ υψηλές βαθμολογίες σθένους είναι σαφείς, πράγμα που σημαίνει ότι δεν μπορούν να έχουν περισσότερες από μία λογική και είναι πιθανό να υπάρχουν σε όλα τα θέματα (φυσικά υπάρχουν και εξαιρέσεις, όπως `καρκίνου»: τα αποτελέσματα -0.8 σθένους η οποία μπορεί να είναι η ασθένεια ή το σημάδι ωροσκόπιο). Με την παρατήρηση αυτών των αποτελεσμάτων είναι πιθανό ότι 80 λέξεις σπόρος είναι ο καλύτερος αριθμός σαφείς συλλέγονται λέξεις που μπορούν να χρησιμοποιηθούν για την εκτίμηση ενός άγνωστη λέξη. Αυτό μοιάζει με το έργο του ¨ciτε τυρνεψ2002τηυμβς η οποία χρησιμοποίησε τις λέξεις ‘ εξαιρετική ¨και` `κακή’ ως σημείο αναφοράς προκειμένου να εκτιμηθεί το συναίσθημα της επανεξέτασης. Εν ολίγοις, μια καλύτερη προσέγγιση θα περιλαμβάνει ως λέξεις σπόρων θέμα-συγκεκριμένες λέξεις που είναι



Σχήμα 6.10: Accuracy and Spearman Correlation using non-linear max fusion.



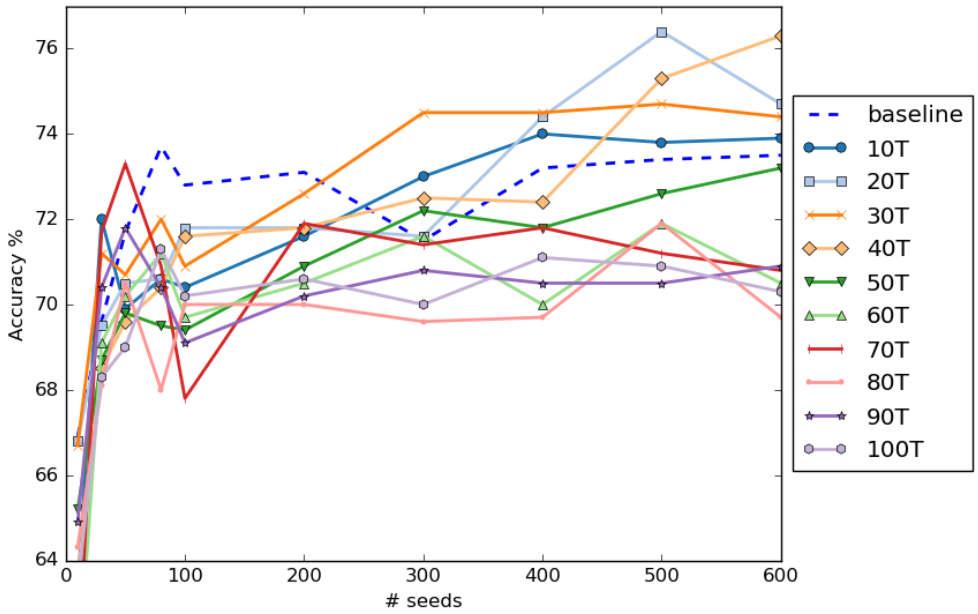
Σχήμα 6.11: Histograms for seed words valence scores obtain from ANEW.

σε θέση να χαρακτηρίσουν μια νέα λέξη σε ένα συγκεκριμένο τομέα. Τα ιστογράμματα λέξεις σπόρων φαίνεται στο σχήμα ;;. Επιπλέον, και πάλι τα 30 θέματα που φαίνεται να έχουν καλές επιδόσεις για όλα τα συστήματα σύντηξης, ενώ το καλύτερο αποτέλεσμα επιτυγχάνεται για 20 θέματα και 600 λέξεις σπόρων. Η σταθμισμένη πρόγραμμα σύντηξης δίνει γενικά τα καλύτερα αποτελέσματα, δεδομένου ότι τα βάρη περισσότερο η ' σημαντικό ¨λέξεις σε ένα ςορπυς, εκείνοι που έχουν μεγαλύτερη απόλυτη βαθμολογία σθένους. Ως εκ τούτου, αγνοεί τις λέξεις με χαμηλές βαθμολογίες (όπως στοπ-λέξεις ή ονόματα) και επικεντρώνεται σε εκείνους που δίνουν περισσότερες συναισθηματικές πληροφορίες. Όσον αφορά τη συσχέτιση, η μέγιστη σύστημα σύντηξης δίνει τα χειρότερα αποτελέσματα ως Νέα τίτλοι είναι προτάσεις γενικού σκοπού που περιέχουν εντυπωσιακές λέξεις με υψηλή βαθμολογία σθένους, αλλά δεν αντιπροσωπεύουν πάντα την πραγματική έννοια της φράσης. Και πάλι, ο λόγος ότι οι μικρές θέματα φαίνεται να αποδίδουν καλά μπορεί να είναι α) ότι αυτή είναι η βέλτιστη διακριτή σειρά θεμάτων ςορπυς

μας μπορούν να περιέχουν και β) τη μείωση του θορύβου που εκτελείται. Πίνακας ρεφΒΕΣΤΙ συνοψίζει τα καλύτερα αποτελέσματα που παρουσιάζονται στα σχήματα.

		Linear Fusion		Weighted Fusion		Max Fusion	
Topics	Seeds	CA (%)	CC	CA (%)	CC	CA (%)	CC
10	80	74.8	0.6370	76.1	0.5945	76.1	0.5631
20	600	75.2	0.6257	**77.2**	0.6391	76	0.5719
30	80	75.7	**0.6455**	76.5	**0.6496**	75.4	0.6028
40	600	76.4	0.6143	75.9	0.6173	74.9	0.5506
50	80	**75.8**	0.6406	**77**	0.6337	76	0.5859
60	500	76.2	0.6051	75.5	0.6078	75	0.5438
100	80	75.6	0.6366	76.7	0.6489	76.5	0.6024
No Topics	200	73.9	0.6142	75.1	0.6272	74.9	0.5429

Table 6.12: Results summary (CA: Classification Accuracy, CC: Classification Correlation (Spearman)) for sentence polarity detection using SMA.

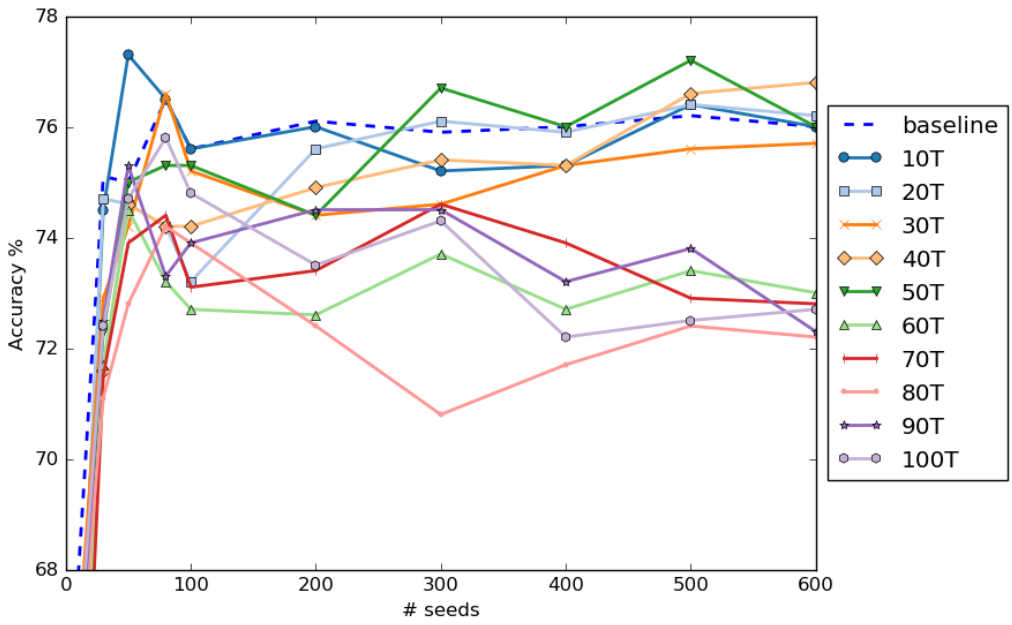


Σχήμα 6.12: Classification accuracy using Linear Logistic Regression classifier.

Twitter

$$recall = \frac{TP}{TP + FN}, precision = \frac{TP}{TP + FP}, F1 = 2 \cdot \frac{\text{πρεςισιον} \cdot \text{ρεςαλλ}}{\text{πρεςισιον} + \text{ρεςαλλ}} \tag{6.4}$$

Όπως μπορούμε να παρατηρήσουμε η ακρίβεια ταξινόμησης είναι γενικά υψηλό και για 20 Θέματα ξεπερνά τη γραμμή βάσης (Δεν Θέματα). Στην πραγματικότητα, το σύνολο δεδομένων



Σχήμα 6.13: Classification accuracy using Naive Bayes classifier.

Topics	Accuracy (%)	Av. Recall	Av. Precision	Av. F1 score
10	82.98	0.672	0.7795	0.70005
20	**84.72**	0.734	0.791	**0.755**
30	83.41	0.673	**0.7955**	0.703
40	82.85	0.689	0.7645	0.713
50	83.42	0.678	0.79	0.708
60	82.75	0.6625	0.7795	0.6905
70	83.94	0.72	0.7785	0.742
80	82.51	0.71	0.7495	0.726
90	83.07	0.6635	0.792	0.693
100	76.44	**0.754**	0.6955	0.708
No Topics	83.27	0.6925	0.7745	0.7185
Tweester	0.862	0.797	-	0.799

Table 6.13: Classification Results for SemEval-2016 Twitter Task B dataset.

είναι προκατειλημμένη προς θετική τωεετς, όπως φαίνεται στον πίνακα 6.14. Λαμβάνοντας

	# Total Tweets	# Positive	# Negative
Train	3938	3260	678
Test	10551	8212	2339

Table 6.14: Number of positive and negative tweets in the SemEval-2016 Task 4, Subtask B dataset.

υπόψη αυτό, το μοντέλο αποδίδει καλά στην ανίχνευση των αρνητικών ποινών και, επομένως, με αποτέλεσμα χαμηλό μέσο μέσες τιμές βαθμολογίας F1 ανάκληση και. Σε γενικές γραμμές, τα σύνολα δεδομένων Twitter θεωρείται το πιο δύσκολο να ταξινομήσει, λόγω της άτυπης γλώσσα τους, υνσυαλ εκφράσεις και ποικίλο λεξιλόγιο. Κατά συνέπεια, κατά την κατάρτιση ενός μοντέλου LDA σε ένα γενικό σώμα που περιέχει ‘ κανονική ¨φράσεις και όχι Τωεετς, ένα μεγάλο μέρος των πληροφοριών χάνεται, όπως κατά τη διαδικασία αξιολόγησης δεν θεωρούμε τα σημεία στίξης, emoticons, επίμηκες λέξεις κλπ, χαρακτηριστικά που το παιχνίδι ένα σημαντικό ρόλο στα σύνολα δεδομένων των κοινωνικών μέσων μαζικής ενημέρωσης. Επιπλέον, το σώμα και το σύνολο δεδομένων twitter δημιουργήθηκαν σε εντελώς διαφορετικές χρονικές περιόδους. Για παράδειγμα, η ταινία ‘ Ant-Man ¨δεν περιέχεται ως όνομα της ταινίας στο σώμα μας. Αυτό καθιστά επίσης η κατάταξη των tweets συνόλου δεδομένων πιο δύσκολη. Ωστόσο, η απόδοση του μοντέλου είναι πολύ ενθαρρυντικά παρά όλα αυτά τα εμπόδια που περιγράφηκαν παραπάνω, δείχνοντας ότι μπορεί να έχει δυναμικό κατά τη χρήση των κατάλληλων στοιχείων για την εκπαίδευση του μοντέλου ΛΔΑ.

Κεφάλαιο 7

Επίλογος

7.1 Ανακεφαλαίωση & Συμπεράσματα

Αυτή η διατριβή συζήτησε την ανάλυση συναισθήματος των προτάσεων σε ένα δυαδικό πρόβλημα ταξινόμησης χρησιμοποιώντας συνεχείς συναισθηματικούς χώρους, Διανεμητικά Σημασιολογικά Μοντέλα και τεχνικές Θεματικής Μοντελοποίησης. Προτείναμε ένα μοντέλο που έχει ως στόχο να βελτιώσει την ακρίβεια ταξινόμησης στην πρόβλεψη της πολικότητας σε επίπεδο προτάσεων και τη συσχέτισης της ομοιότητας λέξεων-ζευγών με ανθρώπινες τιμές. Οι κυριότεροι μηχανισμοί που έχουν ενσωματωθεί σε αυτό το σύστημα, είναι μια προσέγγιση Θεματικής Μοντελοποίησης, που μπορεί να βοηθήσει στη διάκριση των συγκεκριμένων τομέων ή θεμάτων που περιλαμβάνονται σε μια μεγάλη συλλογή από έγγραφα, τα Διανεμητικά σημασιολογικά μοντέλα, που προσπαθούν να αντιπροσωπεύουν το νόημα των λέξεων με τη μορφή μαθηματικών διανυσμάτων, εκμεταλλευόμενα τις ιδιότητες της γλώσσας, και ένα Συναισθηματικό μοντέλο που βασίζεται στην υπόθεση ότι η 'σημασιολογική ομοιότητα συνεπάγεται συναισθηματική ομοιότητα¨, η οποία λειτουργεί ως γέφυρα μεταξύ του χώρου του νοήματος και του χώρου του συναισθήματος.

Αναλυτικότερα, σε αυτή την εργασία, μια νέα προσέγγιση προσαρμογής εισήχθη η οποία ενσωματώνει τη γνώσης των θεματικών ενοτήτων με σκοπό να προσαρμόσει το σημασιολογικό χώρο της κάθε πρότασης. Το μοντέλο βασίζεται στην ιδέα ότι, προκειμένου να εκτιμηθεί η πολικότητα ή η έννοια της πρότασης, πρέπει πρώτα να προσδιορίσει το πλαίσιο (θέμα) όπου ανήκει. Με αυτόν τον τρόπο είμαστε σε θέση να ξεπεράσουμε το πρόβλημα της αποσαφήνισης κατά την εκτίμηση της συναισθηματικής βαθμίδας των λέξεων, αφού μια λέξη μπορεί να εκφράσει μια διαφορετική έννοια και, κατά συνέπεια, ένα διαφορετικό συναίσθημα, ανάλογα με τη χρήση της σε μια μεγαλύτερη λεξική μονάδα.

Τα βασικά βήματα του αλγορίθμου είναι, α) εξόρυξη θεμάτων από ένα web-harvested corpus χρησιμοποιώντας την τεχνική LDA, β) δημιουργία θεμάτων-clusters κατατάσσοντας τις φράσεις του corpus στις εξαγόμενες θεματικές ενότητες χρησιμοποιώντας ένα εκπαιδευμένο μοντέλο LDA και γ) ένα διαφορετικό σημασιολογικό μοντέλο για κάθε θέμα-cluster. Στο τελικό στάδιο, ο σημασιολογικός χώρος μιας πρότασης προσαρμόζεται, δηλαδή οι ομοιότητες ανάμεσα στις λέξεις της φράσης και ένα υπάρχον συναισθηματικό. Η τελική μήτρα ομοιότητας

για μια πρόταση παράγονται χρησιμοποιώντας ένα σταθμισμένο μείγμα των σημασιολογικών μοντέλων από τα θέματα στα οποία μπορεί να περιέχει, και το οποίο αποφασίζεται από την εφαρμογή ενός εκπαιδευμένου μοντέλου LDA σε αυτή την πρόταση.

Τα αποτελέσματα που προκύπτουν από τη εφαρμογή δοκιμών σε ζεύγη για την εκτιμίηση της σημασιολογικής ομοιότητας μεταξύ των λέξεων, αποκαλύπτουν την ευρωστία του αλγορίθμου, ιδίως κατά τη μέτρηση της ομοιότητας μεταξύ πολύ συσχετισμένων ζευγαριών. Δείξαμε ότι ένα καλά εκπαιδευμένο μοντέλο LDA μπορεί να εντοπίσει με χαμηλή αβεβαιότητα ακόμη και τα σχετικά θέματα για ζεύγη λέξεων. Επιπλέον, συνάγεται το συμπέρασμα ότι η συνολική απόδοση μπορεί να βελτιωθεί όταν οι θεματικές ενότητες περιέχουν αρκετά δεδομένα για την καλή εκπαίδευση του σημασιολογικού μοντέλου, αν και σε γενικές γραμμές η απόδοση του κάθε μοντέλου εξαρτάται από τις παραμέτρους του. Όταν χρησιμοποιούνται σημασιολογικά μοντέλα πρόβλεψης, όπως το Word2Vec, ο αριθμός των επαναλήψεων είναι ο κύριος παράγοντας που οδηγεί σε μεγαλύτερο ή μικρότερο βαθμό προσαρμογής, αλλά εισάγει επίσης μια μορφή υπερπροσαρμογής (overfitting) που μπορεί να οδηγήσει σε κακή απόδοση.

Η απόδοση του μείγματος του σημασιολογικού μοντέλου σε ταξινόμηση προτάσεων με βάση το συναίσθημα αποδεικνύει ότι αυτή η διαδικασία των δύο σταδίων ταξινόμησης, δηλαδή του προσδιορισμού των θεμάτων τη πρότασης και στη συνέχεια του συναισθήματός της, μπορεί να βελτιώσει όχι μόνο την αναγνώριση πολικότητας, αλλά επίσης υπολογίζει με μεγαλύτερη ακρίβεια το πραγματικό valence score της πρότασης. Παρά το γεγονός ότι μια μεγάλη βελτίωση επιτυγχάνεται κατά την αξιολόγηση για γενικά σύνολα δεδομένων (π.χ. πρωτοσέλιδα), τα μηνύματα των social media (π.χ. Twitter) είναι ένα πολύ πιο δύσκολο πρόβλημα και απαιτείται ειδική εκπαίδευση σε συγκεκριμένα δεδομένα προκειμένου να επιτευχθούν καλύτερα αποτελέσματα. Ωστόσο, αυτό το γενικό μοντέλο δείχνει μια αξιοπρεπή συμπεριφορά στο έργο αυτό, παρά την απώλεια σημαντικών πληροφοριών και η άγνοια τεχνοτροπικών χαρακτηριστικών που είναι συχνά σε τέτοιου είδους δεδομένα.

Εν ολίγοις, τα θεματικά σε συνδυασμό με τις μεθόδους συναισθηματικής ανάλυσης κειμένου φαίνονται πολύ ελπιδοφόρα για την ανάλυση συναισθήματος των προτάσεων, δεδομένου ότι προσφέρουν ένα χωρίς επίβλεψη τρόπο προσαρμογής ένα σημασιολογικό μοντέλο σε πολλαπλά καθήκοντα.

7.2 Εν εξελίξει Εργασία

7.2.1 Semantic-affective model adaptation (SAMA)

Η βελτίωση του αλγορίθμου (SMA) προτείνεται εδώ, που θα μπορούσε να οδηγήσει σε καλύτερα συναισθηματικά αποτελέσματα. Αν και δεν υπάρχουν τα αποτελέσματα της αξιολόγησης ακόμη για αυτόν τον αλγόριθμο, το κίνητρο και τα βήματά του περιγράφονται ως μέρος της εν εξελίξει εργασίας.

Κίνητρο

Σε αυτή την προσέγγιση, προσπαθούμε να προσαρμόσουμε τόσο το συναισθηματικό χώρο και το σημασιολογικό χώρο. Αυτό σημαίνει ότι οι συναισθηματικές τιμές τώρα υπολογίζονται εκ νέου επιπρόσθετα με τις σημασιολογικές ομοιότητες όπως στην SMA. Το κίνητρο πίσω από αυτή την ιδέα είναι ότι ο προηγούμενος αλγόριθμος δεν ενσωματώνει καμία γνώση συναισθήματος όταν χρησιμοποιεί τις θεματικές ενότητες και ως εκ τούτου θα μπορούσε να οδηγήσει σε καλύτερη απόδοση σημασιολογικής ομοιότητας από συναισθηματική εκτίμηση. Όπως φαίνεται στη βιβλιογραφία, τα μοντέλα που αποτυπώνουν το συναίσθημα και το θέμα την ίδια στιγμή, οδηγούν σε καλύτερα αποτελέσματα. Έτσι, η κύρια ιδέα, δεν είναι μόνο η εξαγωγή διαφορετικών λέξεις πυρήνων για ένα θέμα, αλλά και το να προσαρμοστούν οι συναισθηματικές βαθμολογίες αυτών των λέξεων ανάλογα με τη θεματική ενότητα. Στη συνέχεια, μπορούμε να τα χρησιμοποιήσουμε στο Semantic-Affective μοντέλο (εξίσωση ;;), για την εκτίμηση πολλαπλών θεματικών-συναισθηματικών σκορ για τις λέξεις μιας πρότασης. Το τελικό σκορ μιας λέξης παράγεται χρησιμοποιώντας ένα μοντέλο μείγμα των θεματικών-συναισθηματικών βαθμολογιών. Ξανά, η εκτίμηση τους συναισθήματος της πρόταση βασίζεται σε ένα σύστημα μείξης ή ταξινόμησης.

Αλγοριτημ

Σε γενικές γραμμές, η διαδικασία Θεματικής Μοντελοποίησης μέχρι το βήμα που ένα διαφορετικό σημασιολογικό μοντέλο χτίζεται σε κάθε θέμα-cluster είναι η ίδια όπως και στο Σημασιολογικό μοντέλο προσαρμογής (SMA). Η βασική αλλαγή είναι η συναισθηματική προσαρμογή και το μοντέλο μείγμα.

Τα βήματα του αλγορίθμου περιγράφονται παρακάτω:

1. Για κάθε topic:
   (α′) Εντόπισε topic-seed λέξεις με βάση δύο κριτήρια
   - Μεγάλη συχνότητα στη θεματική ενότητα
   - Ζυγισμένο set (άθροισμα valence score κοντά στο μηδέν)

   (β′) Δημιουργία μιας μήτρας ομοιοτήτων $N \times N$ που περιέχει τις σημασιολογικές ομοιότητες μεταξύ των topic-seeds
2. Όσο $< k$:
   (α′) Εκπαίδευσε τους α συντελεστές από το semantic-affective model (εξίσωση 4.2) για τα επιλεγμένα topic-seeds

   (β′) Επαναυπολόγισε τα valence scores των topic-seeds χρησιμοποιώντας το semantic-affective model:

$$v_{topic-adapted}(seed_j) = \alpha_{0,topic} + \sum_{i=1}^{N} \alpha_{i,topic} v_{anew}(w_i) sim_{topic}(seed_j, w_i)$$

3. Εφαρμογή LDA σε κάθε πρόταση και έλεγχος των εξαγόμενων θεματικών ενοτήτων

4. Για κάθε λέξη της πρότασης ($w_{sentence}$):

   - Χρησιμοποίησε την σημασιολογική ομοιότητα μεταξύ αυτής της λέξης και της topic-seed $d(w_{sentence}, w_{topic-seed})$ από ένα γενικό DSM
   - Εκτίμησε το valence score της λέξης μέσω του SAM για κάθε topic:

$$v_{topic}(w_{sentence}) = a_{0,topic} + \sum_{i=1}^{N} a_{i,topic} v_{topic-adapted}(w_{topic-seed}) d(w_{sentence}, w_{topic-seed})$$

5. Συγχώνευσε τα διαφορετικά valence scores για αυτή τη λέξη, χρησιμοποιώντας topic-posteriors για το τελικό score της λέξης

$$v_{final}(w_{sentence}) = \frac{\sum_{t=1}^{T} p(t_i|sentence) \cdot v_t(w_{sentence})}{\sum_{t=1}^{T} p(t_i|sentence)}$$

7.3 Μελλοντικές Επεκτάσεις

Το μοντέλο που προτείνεται στην παρούσα διατριβή, περιέχει μια πληθώρα διαφορετικών παραμέτρων που πρέπει να συντονιστούν προκειμένου να βελτιωθεί η γενική απόδοση του μοντέλου. Σε όλα τα πειράματα που περιγράφονται στο κεφάλαιο ;;, συγκεκριμένες υλοποιήσεις για Θεματική Μοντελοποίηση και σημασιολογικά μοντέλα χρησιμοποιήθηκαν λόγω του τεράστιου όγκου των δεδομένων που παράγονται από αυτά τα μοντέλα, για όλα τα πιθανά θέματα.

Κάποιες πιθανές ιδέες για κάθε τομέα περιγράφονται παρακάτω.

Topic Modeling

- Η μέθοδος θεματικής μοντελοποίησης (LDA) μπορεί εύκολα να αντικατασταθεί, έτσι ώστε να να εφαρμοστεί άμεσα σε φράσεις και όχι έγγραφα. Διαφορετικές υλοποιήσεις μπορούν να βρεθούν στο διαδίκτυο ([1], [2]).

- Η απλή μέθοδος κατασκευής clusters που χρησιμοποιήσαμε μπορεί να βελτιωθεί. Ένα παράδειγμα είναι να ταξινομηθεί κάθε πρόταση σε όλα τα θέματα που μπορεί να ανήκει και να αποθηκευτεί η πληροφορία πιθανότητας για χρήση στο σημασιολογικό μοντέλο. Αυτό είναι ένα δυνατό 'weigted' soft clustering. Εναλλακτικά, μπορεί να χρησιμοποιηθεί καμία τεχνική θεματικής μοντελοποίησης μπορεί να μην χρησιμοποιηθεί, αλλά άλλες μέθοδοι ταξινόμησης (όπως η K-Means) μπορεί να πάρει τη θέση του.

[1] ηττπς://γιτηυβ.ςομ/δατχυοςνγυψεν/ΛΦΤΜ/
[2] ηττπς://γιτηυβ.ςομ/μινγηυι/Τωιττερ-ΛΔΑ

Semantic Models

- Άλλα μοντέλα μπορούν να αντικαταστήσουν τη Word2Vec, όπως κλασσικά context-based τεχνικές χρησιμοποιώντας PPMI features ή το Glove package [3].

- Για όλα τα θέματα, οι παράμετροι του σημασιολογικού μοντέλου έχουν καθοριστεί. Μια πιθανή επέκταση θα ήταν να χρησιμοποιήσουμε διαφορετικές παραμέτρους για κάθε θέμα (αν και μπορεί να είναι δύσκολο για μεγάλο αριθμό θεμάτων). Η απόδοση μπορεί να αυξηθεί συντονίζοντας τις ακόλουθες σημαντικές παραμέτρους της Word2Vec είναι:

  – vector dimension

  – window size

  – number of iterations

  – algorithm (CBOW, SkipGram)

- Επιπλέον, όσον αφορά την εξομάλυνση των διανυσμάτων χαρακτηριστικών που εξάγεται από το σημασιολογικό μοντέλο, εκτός από την ομαλοποίηση ως προς l2 νόρμα (μοναδιαίο διάνυσμα) που χρησιμοποιήσαμε, z-score ή min-max ομαλοποίηση μπορεί να βελτιώσει την απόδοση.

Semantic Model Adaptation

Σχετικά με το πρώτο μοντέλο που προσπαθεί να προσαρμόσει το σημασιολογικό χώρο κάθε πρότασης ως ένα μείγμα συγκεκριμένων σημασιολογικών-θεματικών μοντέλων, τα παρακάτω μπορούν να εξεταστούν:

- Μια τεχνική seed selection αφού παρατηρείται ότι το μοντέλο λειτουργεί καλύτερα όταν ο αριθμός των seeds είναι μικρός.

- Ένα διαφορετικό μοντέλο μείγματος, που μπορεί να ζυγίζει περισσότερο μεγαλύτερες σημασιολογικές ομοιότητες μεταξύ ζευγαριών λέξεων.

- Μια άλλη ιδέα είναι να γίνει προσαρμογή μόνο συγκεκριμένων λέξεων κάθε κάθε θεματικής ενότητας. Με άλλα λόγια, όλες οι άλλες λέξεις-διανύσματα που είναι γενικού περιεχομένου θα εκπαιδεύονται πάνω σε ένα γενικό σημασιολογικό μοντέλο και μόνο οι θεματικά ειδικευμένες θα προέρχονται από συγκεκριμένα θεματικά σημασιολογικά μοντέλα.

- Η ιδέα που αφορά το έργο της σημασιολογικής ομοιότητας μεταξύ ζευγαριών λέξεων, είναι να δίνουμε σαν είσοδο στο LDA μόνο δύο λέξεις σαν πρόταση. Αυτό μπορεί να γίνει εναλλακτικά με τη χρήση του WordNet. Για κάθε λέξη ενδιαφέροντος, μπορούμε να πάρουμε τον ορισμό αυτής της λέξης για όλες τις πιθανές της έννοιες. Όλοι οι

[3] http://νλπ.στανφορδ.εδυ/προθεςτς/γλοε/

πιθανοί συνδυασμοί ορισμών μπορούν να δημιουργηθούν μεταξύ δύο λέξεων. Η τελική ομοιότητα των λέξεων θα προκύπτει από την υπόθεση μέγιστης έννοιας: 'Η ομοιότητα δύο λέξεων είναι η μέγιστη ομοιότητα μεταξύ των εννοιών τους'.

Παράρτημα Α΄

Gensim Toolbox

Το Gensim Toolbox [1] δημιουργήθηκε από τον Řehůřek and Sojka (2010) Και είναι μια συλλογή από προγράμματα σε Python. Το όνομά του προέρχεται από το 'Generate Similar¨, όπως ήταν αρχικά χρησιμοποιήθηκε για τη συλλογή των πιο παρόμοια είδη για ένα συγκεκριμένο άρθρο. Σε αυτήν την εργασία, χρησιμοποιήσαμε τόσο τα εργαλεία θέμα Μοντελοποίηση, καθώς και την εφαρμογή των Word2Vec που προσφέρει σε python. Αυτά τα δύο εργαλεία που περιγράφονται παρακάτω, με παραδείγματα από τα tutorials της σελίδας.

Α΄.1 Topic Modeling

Για να χρησιμοποιήσετε τη θεματική μοντελοποίηση Gensim, θα πρέπει πρώτα να μετατρέψουν τη συλλογή εισόδου εγγράφων. Το πρώτο βήμα είναι η κατασκευή ενός λεξικού:

```
import gensim, logging
from gensim import corpora, models, similarities

logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.INFO)

# load a list with known stopwords
stoplist = [i.strip() for i in open('my_en_stopwords.txt', 'r')]

# collect statistics about all tokens (reads file line by line)
dictionary = corpora.Dictionary((line.lower().split() for line in open('mycorpus.txt','r')), prune_at=None)

# remove stop words and words that appear only once
stop_ids = [dictionary.token2id[stopword] for stopword in stoplist if stopword in dictionary.token2id]

# this method, removes stopwords from dictionary
dictionary.filter_tokens(bad_ids=stop_ids)

# Remove too infrequent and frequent words
dictionary.filter_extremes(no_below=3, no_above=0.8)

# remove gaps in id sequence after words that were removed
dictionary.compactify()

# store the dictionary, for future reference
dictionary.save('corpus.dict')
dictionary.save_as_text('corpus.dict_txt',sort_by_word=True)
```

[1] ηττπς://ραδιμρεηυρεκ.ςομ/Γενσιμ/

Αν τυπώσουμε το λεξικό:

```
print(dictionary)
print(dictionary.token2id)
```

```
Dictionary(72 unique tokens: [u'heavy', u'amish', u'says', u'deal', u'porn']...)
{u'heavy': 0, u'amish': 1, u'says': 2, u'deal': 3, u'porn': 4, u'insists': 60, u'lebanon': 6, u'dead': 7, u'two': 26, u'b
rain': 9, u'say': 10, u'set': 11, u'kill': 12, u'ex': 13, u'satellites': 71, u'madonna': 14, u'ship': 15, u'linked': 16,
u'marathon': 17, u'yankee': 18, u'brazil': 19, u'plot': 20, u'police': 21, u'cancer': 22, u'record': 64, u'tv': 23, u'pho
tographer': 24, u'vegetables': 25, u'three': 8, u'tropical': 27, u'iraqi': 28, u'attack': 29, u'adoption': 30, u'storm':
31, u'scientists': 32, u'terror': 33, u'women': 65, u'happy': 36, u'viking': 37, u'life': 38, u'good': 39, u'crash': 40,
u'sinks': 41, u'missing': 42, u'confusion': 43, u'troops': 55, u'gaza': 35, u'internet': 45, u'admits': 46, u'dies': 47,
u'china': 48, u'child': 49, u'russian': 50, u'drinking': 51, u'iraq': 52, u'killed': 53, u'vaccine': 54, u'man': 34, u'ba
ghdad': 61, u'mars': 56, u'prize': 57, u'wins': 58, u'study': 59, u'flu': 5, u"men's": 44, u"alzheimer's": 62, u'victory'
: 66, u'uk': 67, u'lose': 68, u'report': 69, u'found': 70, u'finds': 63}
```

Για μια νέα πρόταση, ο μετασχηματισμός να υποκύψει χρησιμοποιώντας το λεξιλόγιο παρουσιάζεται παρακάτω:

```
new_doc = "life on mars"
new_vec = dictionary.doc2bow(new_doc.lower().split()) # 'on' is a stoword and so removed
print(new_vec)
```

```
[(38, 1), (56, 1)]
```

Gensim, είναι σε θέση να χειριστεί μεγάλες συλλογές, φορτώνοντας μόνο μία γραμμή κάθε φορά, αποφεύγοντας τη μεγάλη χρήση μνήμης. Αυτό μπορεί να επιτευχθεί με τη χρήση ενός iterator, που μετατρέπει κάθε γραμμή σε μορφή BOW, χρησιμοποιώντας το λεξικό που δημιουργήθηκε παραπάνω. Μπορούμε να ονομάσουμε αυτό iterator, και να αποθηκεύσετε σώμα μας σε μορφή Matrix Market (χρησιμοποιούνται, καθώς και άλλες μορφές).

```
class MyCorpus(object):
    def __iter__(self):
        for line in open('mycorpus.txt'):
            # assume there's one document per line, tokens separated by whitespace
            yield dictionary.doc2bow(line.lower().split())

corpus = MyCorpus()
print(corpus)

corpora.MmCorpus.serialize('corpus.mm', corpus)
```

```
<__main__.MyCorpus object at 0x7f71a7f8f7d0>
```

```
print list(corpus)[25] # print line 26 of the corpus
```

```
[(8, 1), (15, 1), (42, 1), (50, 1), (69, 1), (70, 1)]
```

Το επόμενο βήμα είναι να φορτώσει το σώμα και το λεξικό και να εφαρμόσει μια μεταμόρφωση. Στην περίπτωσή μας χρησιμοποιήσαμε τον μετασχηματισμό LDA και πιο συγκεκριμένα την εφαρμογή LDAMulticore που είναι σε θέση να τρέξει παράλληλα:

```
# load dictionary and corpus
dictionary = corpora.Dictionary.load('corpus.dict')
corpus = corpora.MmCorpus('corpus.mm')

# apply LDA transformation to the BOW corpus
model = gensim.models.ldamulticore.LdaMulticore(corpus, num_topics=2, id2word=dictionary, workers=2, iterations=100)

# once the model is ready, save it
model.save('corpus_2T.lda')
```

Τέλος, αφού το μοντέλο έχει εκπαιδευτεί, για κάθε νέα πρόταση μπορούμε να ελέγξουμε, όπου το εκπαιδευμένο μοντέλο τους κατατάσσει ως εξής:

Σε γενικές γραμμές, οι παράμετροι του LDA είναι οι ακόλουθοι.

```
# load model
model = models.LdaModel.load('corpus_2T.lda',mmap='r')

# transform new sentence
sentence = "life on mars"
transf_sent = dictionary.doc2bow(sentence.lower().split())
print transf_sent
```

```
[(38, 1), (56, 1)]
```

```
lda_sent = model[transf_sent]
print lda_sent
```

```
[(0, 0.19657055121821196), (1, 0.80342944878178812)]
```

Parameters

num_topics: the number of topics to be created
id2word: the dictionary that will be used for the topic construction
aplha, eta: hyperparameters of document-topic and topic-word distributions. The default values are 1/number of topics.
minimum_probability: Topics that give posterior probability below this number are ignored (default is 0.001).
iterations: number of passes for each chunk of documents

Α΄.2 Word2Vec

Με τον ίδιο τρόπο, word2vec μπορεί να φορτώσει μία γραμμή κάθε φορά και στη συνέχεια, εφαρμόζοντας το word2vec αλγόριθμο. Υπάρχουν πολλές παράμετροι που μπορούν να ρυθμιστούν, όπως αυτές περιγράφονται παρακάτω.

```
from gensim import corpora, models, similarities, logging
from gensim.models.word2vec import Word2Vec

logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.INFO)

class MySentences(object):
    def __iter__(self):
        for line in open('mycorpus.txt'):
            yield line.lower().split()

sentences = MySentences() # a memory-friendly iterator

model = Word2Vec(sentences, size=300, window=5, workers=7, sg=0, hs=0, negative=5,
                 cbow_mean=1, iter=5, min_count=5, max_vocab_size=1000000)
model.init_sims(replace=False) # cannot be trained after this temp!

model.save('corpus_w2v.model')
```

Parameters

size: Dimension of the feature vector
window: Context window around the term of interest (w words left, w words right)
workers: number of threads to be used
sg: if 1 skip-gram algorithm is used, if 0 cbow algorithm is used
hs: if 1 hierarchical softmax is used as inference method, if 0 and negative is non-zero, negative sampling is used
negative: if > 0, negative sampling is used and the number is equal to the "noise words"

that should be drawn (usually between 5-20)
cbow_mean: if 0, the sum of the context word vectors will be used, if 1 the mean will be used
iter: the number of epochs over the corpus
min_count: minimum frequency of words. Words with frequency below this value will be ignored.
max_vocab_size: maximum vocabulary size created during training. The infrequent words are cropped.

Bibliography

David J Aldous. *Exchangeability and related topics*. Springer, 1985.

Alina Andreevskaia and Sabine Bergler. Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *EACL*, volume 6, pages 209–216, 2006.

Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

David M Blei and John D Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology, University of Florida, 1999.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(1-47), 2014.

Jonathan Chang and David M Blei. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, pages 124–150, 2010.

Pimwadee Chaovalit and Lina Zhou. Movie review mining: A comparison between supervised and unsupervised classification approaches. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 112c–112c. IEEE, 2005.

Stephen Clark. Vector space models of lexical meaning. *Handbook of Contemporary Semantics, Wiley-Blackwell, à paraître*, 2012.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

Gabriel Doyle and Charles Elkan. Accounting for burstiness in topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 281–288. ACM, 2009.

Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM, 2001.

Andrew B Goldberg and Xiaojin Zhu. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52. Association for Computational Linguistics, 2006.

Thomas L Griffiths, Mark Steyvers, David M Blei, and Joshua B Tenenbaum. Integrating topics and syntax. In *NIPS*, volume 4, pages 537–544, 2004.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

Elias Iosif and Alexandros Potamianos. Unsupervised semantic similarity computation between terms using web documents. *IEEE Transactions on Knowledge and Data Engineering*, 22(11):1637–1647, 2010.

Valentin Jijkoun, Maarten de Rijke, and Wouter Weerkamp. Generating focused topic-specific sentiment lexicons. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 585–594. Association for Computational Linguistics, 2010.

Jaap Kamps, MJ Marx, Robert J Mokken, M de Rijke, et al. Using wordnet to measure semantic orientations of adjectives. 2004.

Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Wei Li and Andrew McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584. ACM, 2006.

Yuming Lin, Jingwei Zhang, Xiaoling Wang, and Aoying Zhou. Sentiment classification via integrating multiple feature presentations. In *Proceedings of the 21st International Conference on World Wide Web*, pages 569–570. ACM, 2012.

Nikolaos Malandrakis, Alexandros Potamianos, Kean J Hsu, Kalina N Babeva, Michelle C Feng, Gerald C Davison, and Shrikanth Narayanan. Affective language model adaptation via corpus selection. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4838–4842. IEEE, 2014.

Nikos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. Emotiword: Affective lexicon creation with application to interaction and multimedia data. In *Computational Intelligence for Multimedia Understanding*, pages 30–41. Springer, 2011a.

Nikos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth S Narayanan. Kernel models for affective lexicon creation. In *In Proceedings of INTERSPEECH*, pages 2977–2980, 2011b.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems*, pages 2265–2273, 2013.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

F. Å. Nielsen. Afinn, 2011.

Alexander Osherenko and Elisabeth André. Lexical affect sensing: Are affect dictionaries necessary to analyze affect? In *Affective Computing and Intelligent Interaction*, pages 230–241. Springer, 2007.

Elisavet Palogiannidi, Athanasia Kolovou, Fenia Christopoulou, Filippos Kokkinos, Elias Iosif, Nikolaos Malandrakis, Harris Papageorgiou, Shrikanth Narayanan, and Alexandros Potamianos. Tweester at semeval-2016 task 4: Sentiment analysis in twitter using

semantic-affective model adaptation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, US*, 2016.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

Christos H Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 159–168. ACM, 1998.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43, 2014.

Robert Plutchik. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.

Jonathon Read and John Carroll. Weakly supervised techniques for domain-independent sentiment classification. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 45–52. ACM, 2009.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA, 2010.

Joseph Reisinger and Raymond Mooney. A mixture model with sharing for lexical semantics. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182. Association for Computational Linguistics, 2010.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

Harold Schlosberg. Three dimensions of emotion. *Psychological review*, 61(2):81, 1954.

Kazutaka Shimada and Tsutomu Endo. Seeing several stars: A rating inference task for a document containing several evaluation criteria. In *Advances in Knowledge Discovery and Data Mining*, pages 1006–1014. Springer, 2008.

Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Peter D Turney. 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*, pages 417–424. Association for Computational Linguistics, 2002.

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

Sudha Verma, Sarah Vieweg, William J Corvey, Leysia Palen, James H Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. Natural Language Processing to the Rescue? Extracting" Situational Awareness" Tweets During Mass Emergency. In *ICWSM*. Citeseer, 2011.

Chong Wang and David M Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *Advances in neural information processing systems*, pages 1982–1989, 2009.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344, 2014.

Jingbo Zhu, Muhua Zhu, Huizhen Wang, and Benjamin K Tsou. Aspect-based sentence segmentation for sentiment summarization. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 65–72. ACM, 2009.