ΑΛΕΞΑΝΔΡΕΙΟ ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB INTELLIGENCE

Ανάλυση Συναισθήματος Κριτικών Προϊόντων με Μεθόδους Μηχανικής Μάθησης σύμφωνα με τα Κυριότερα Χαρακτηριστικά τους

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΑΣΗΜΙΝΑΣ ΖΑΪΜΗ

Επιβλέπων : Κωνσταντίνος Γουλιάνας
Αναπληρωτής Καθηγητής Α.Τ.Ε.Ι.Θ.

Θεσσαλονίκη, Ιούνιος 2018

ΑΛΕΞΑΝΔΡΕΙΟ ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB INTELLIGENCE

Ανάλυση Συναισθήματος Κριτικών Προϊόντων με Μεθόδους Μηχανικής Μάθησης σύμφωνα με τα Κυριότερα Χαρακτηριστικά τους

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΑΣΗΜΙΝΑΣ ΖΑΪΜΗ

Επιβλέπων : Κωνσταντίνος Γουλιάνας
Αναπληρωτής Καθηγητής Α.Τ.Ε.Ι.Θ.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 30 Ιουνίου 2018.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Κωνσταντίνος Γουλιάνας Αναπληρωτής Καθηγητής Α.Τ.Ε.Ι.Θ.	Κωνσταντίνος Διαμαντάρας Καθηγητής Α.Τ.Ε.Ι.Θ.	Παναγιώτης Αδαμίδης Καθηγητής Α.Τ.Ε.Ι.Θ.

Θεσσαλονίκη, Ιούνιος 2018

(Υπογραφή)

.....

Ασημίνα Ζαΐμη

Μηχανικός Πληροφορικής Α.Τ.Ε.Ι.Θ.

Περίληψη

Με την ολοένα και αυξανόμενη ανάπτυξη χρήσης του Παγκοσμίου Ιστού (World Wide Web) αυξάνεται επαγωγικά και το περιεχόμενο που παράγεται από τους χρήστες του (User Generated Content - UGC), το οποίο έχει αποδειχθεί πολλάκις ότι επηρεάζει σε σημαντικό βαθμό την καθημερινότητα των ανθρώπων, γεγονός που καθιστά την αποτελεσματική ανάλυση αυτού του περιεχομένου ζωτικής σημασίας, με έντονο ενδιαφέρον της επιστημονικής, βιομηχανικής ακόμη και πολιτικής κοινότητας. Στις περιπτώσεις που το περιεχόμενο προς ανάλυση ως προς την γνώμη που εκφέρει εντοπίζεται με την μορφή κειμένου, εισχωρείται η έννοια της Ανάλυσης Συναισθήματος (Sentiment Analysis) του πεδίου της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing - NLP), η οποία αποσκοπεί στην αυτόματη αναγνώριση υποκειμενικής πληροφορίας από γραπτές πηγές. Στην παρούσα διπλωματική εργασία αποσαφηνίζεται η Ανάλυση Συναισθήματος ως ένα προκλητικό πρόβλημα και ερευνάται η επίλυσή του με μεθόδους Μηχανικής Μάθησης (Machine Learning), εφαρμοσμένες πάνω σε κριτικές προϊόντων. Επομένως, το πρόβλημα αντιμετωπίζεται ως πρόβλημα ταξινόμησης κειμένων κριτικών προϊόντων σε κάποια συναισθηματική κλάση, εστιάζοντας στα χαρακτηριστικά γνωρίσματα του προϊόντος για τα οποία εκφράζεται ο σχολιαστής. Η προσέγγιση σε τέτοιου είδους προβλήματα είναι το πεδίο που εξετάζουμε και συνήθως υλοποιείται σε επίπεδο πρότασης ή λέξης και η ανάλυση αυτού του επιπέδου αναφέρεται σαν Ανάλυση Συναισθήματος βασισμένη σε λέξεις - κλειδιά (Aspect Based Sentiment Analysis - ABSA) που έχει ως στόχο τόσο τον εντοπισμό των λέξεων – κλειδιών της πρότασης (χαρακτηριστικά γνωρίσματα προϊόντος), όσο και την αποτίμηση του συναισθήματος που φέρουν. Επισημαίνονται οι δυσκολίες επίλυσης του προβλήματος και πιθανοί αλγόριθμοι Μηχανικής Μάθησης ως λύσεις σε αυτές, καθώς περιγράφεται βηματικά η ολοκληρωμένη διαδικασία επίλυσης του προβλήματος. Στο τέλος παρουσιάζουμε μεγάλο αριθμό σχετικών εργασιών που επιχείρησαν να επιλύσουν το εν λόγω πρόβλημα, παραθέτοντας μια υποκειμενική σύγκριση με σημαντικές παρατηρήσεις. Μία από τις παρατηρήσεις που καθορίζεται και το τελικό συμπέρασμα της διπλωματικής αυτής είναι ότι κανένα υλοποιημένο σύστημα της ABSA δεν έχει καταφέρει να επιλύσει εξολοκλήρου το πρόβλημα, γεγονός που διαλευκαίνει την αντιξοότητα διαχείρισης της φυσικής γλώσσας.

Λέξεις Κλειδιά: Ανάλυση Συναισθήματος, Μηχανική Μάθηση, Ανάλυση Συναισθήματος βασισμένη στις λέξεις – κλειδιά, Ανάλυση Συναισθήματος σε κριτικές προϊόντων, Ανάλυση Συναισθήματος σε επίπεδο πρότασης, Εξόρυξη Γνώμης

Abstract

With the growing use of the World Wide Web, the user-generated content (UGC) is inductively increased, which has been proven to have a significant impact on people's everyday lives, that is why making its effective analysis vital, with a keen interest in the scientific, industrial and even political community. In cases where the content to be analyzed in terms of its opinion is in the form of text, the concept of Sentiment Analysis of the Natural Language Processing (NLP) field is introduced, which aims at the automatic recognition of subjective information from written sources. In this diploma thesis, Sentiment Analysis is described as a challenging problem and its solution is investigated through Machine Learning methods, applied to product reviews. Therefore, the problem is treated as a problem of classifying product reviews texts in an emotional class, focusing on the product attributes for which the reviewer is expressed. Approach to such problems is the field that we examine and it is usually implemented at the sentence or aspect level, and the analysis of this level is referred to as an Aspect Based Sentiment Analysis (ABSA) aimed at both identifying the aspects (product attributes) in the sentence, as well as the sentiment they carry. We identify the difficulties of solving such problems and possible Machine Learning algorithms as solutions, as the comprehensive problem solving process is described step by step. Finally, we present a large number of related work that attempted to resolve this problem by quoting a subjective comparison with important observations. One of the remarks and the final conclusion of this diploma is that none implemented ABSA system has succeeded in solving the problem altogether, a fact that discovers the adversity of managing the natural language.

Keywords: Sentiment Analysis, Machine Learning, Aspect Based Sentiment Analysis, Sentiment Analysis on product reviews, Sentence Level Sentiment Analysis, Opinion Extraction

Ένα μεγάλο ευχαριστώ σε όσους με υποστήριξαν και έδειξαν κατανόηση.
Στην μητέρα μου, για την οποία καμία αφιέρωση δεν είναι αρκετή.

Πίνακας περιεχομένων

Κατάλογος Εικόνων

Κατάλογος Πινάκων

1

Εισαγωγή

1.1 Κίνητρο

Οι επιλογές μας στην καθημερινότητα, που καθορίζουν και την προσωπικότητά μας, επηρεάζονται κατά ένα αρκετά μεγάλο ποσοστό από απόψεις άλλων, έτσι εύκολα εντοπίζουμε την σημαντικότητα αυτών των απόψεων και τον λόγο που το συναίσθημα που εκφράζουν σε συνδυασμό με τις αξιολογήσεις τους, έχουν χτίσει μια ολόκληρη επιστήμη, ονόματι Ανάλυση Συναισθήματος (Sentiment Analysis) ή Εξόρυξη Γνώμης (Opinion Mining) ή Εκτίμηση Διάθεσης. Σε γενικές γραμμές η Ανάλυση Συναισθήματος έχει ως στόχο να καθορίσει τη στάση που έχει ένας ομιλητής ή συγγραφέας ή κριτής για κάποιο συγκεκριμένο θέμα και την συνολική του άποψη περιγραφόμενη συχνά σε κάποιο έγγραφο. Η δυνατότητά μας να σημασιολογούμε τα δεδομένα, δηλαδή να εκφέρουμε προσωπική αξιολογική άποψη, δικαιολογεί και το χαρακτηρισμό του ανθρώπου ως έλλογου όντος, που μας κάνει να ξεχωρίζουμε από τις υπόλοιπες υπάρξεις της γης, έτσι εξηγείται και η δύναμη της άποψης που εύλογα εγκαθίδρυσε την επιστήμη της Ανάλυσης Συναισθήματος για την διαχείριση και αξιολόγησή της.

Το εφαλτήριο του κλάδου της Ανάλυσης Συναισθήματος εκτιμάται κατά την ραγδαία εξέλιξη των Κοινωνικών Δικτύων (Social Media) στον Παγκόσμιο Ιστό (World Wide Web), γύρω στο 2001, από όπου τεράστιος αριθμός απόψεων, μέσω σχολίων, αναρτήσεων και κριτικών που εξετάζονται ώστε να γίνει η κατάλληλη αξιοποίηση της γνώσης που παράγουν και της δημιουργικής επεξεργασίας ιδεών και απόψεων, κυρίως για εμπορικούς σκοπούς. Εδώ αξίζει

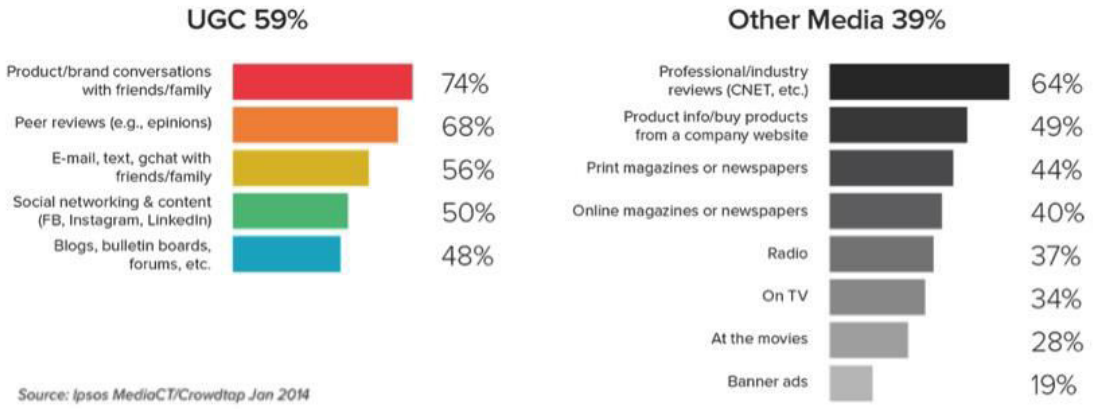
να σημειωθεί ο αριθμός των ενεργών χρηστών, δύο μόνο των πιο δημοφιλή κοινωνικών δικτύων, αν και αποτελούν μόνο ένα μικρό μέρος των Web 2.0 εφαρμογών, σύμφωνα με επίσημα στατιστικά Απριλίου 2018 [1]. Το Facebook κατέχει πάνω από 2.200 εκατομμύρια ενεργούς χρήστες, ενώ το Twitter πάνω από 330 εκατομμύρια με περισσότερες από 500 εκατομμύρια δημοσιεύσεις ημερησίως, γεγονός που υποδηλώνει τον τεράστιο όγκο δεδομένων που διακινείται στο διαδίκτυο.

Οι κριτικές (reviews) αποτελούν μια ιδιαίτερα διαδεδομένη μορφή ιδιοπαραγόμενου περιεχομένου τις οποίες έχουν ενσωματώσει πολλά ηλεκτρονικά καταστήματα στα προϊόντα τους, αλλά παράλληλα υπάρχουν και ειδικά διαμορφωμένες διαδικτυακές πλατφόρμες εξολοκλήρου καθιερωμένες για συγγραφή ή άντληση κριτικών. Αυτές μπορούν να εντοπιστούν σε διάφορα πεδία ενδιαφέροντος, όπως ταινίες (IMDB), βιβλία (Amazon), καταναλωτικά προϊόντα (Epinions), τουριστικές υπηρεσίες (TripAdvisor) κ.α. Έχει αποδειχθεί ότι το μεγαλύτερο ποσοστό των καταναλωτών συμβουλεύονται τις απόψεις άλλων, κάθε μορφής, πριν προβούν στην αγορά κάποιου προϊόντος ή υπηρεσίας. Το ενδιαφέρον των σχολίων, μολαταύτα, δεν τραβάει μόνο απλούς χρήστες του διαδικτύου και πιθανούς πελάτες, αλλά και οργανισμών / επιχειρήσεων, στις οποίες στηρίζονται για να πραγματοποιήσουν έρευνα αγοράς που τα αποτελέσματά της θα βοηθήσουν στην διευθέτηση την εταιρική τους ταυτότητας και την σχεδίαση των προωθητικών στρατηγικών τους.

Οι κύριοι παράγοντες για την βιωσιμότητα και εξέλιξη του κλάδου της Ανάλυσης Συναισθήματος είναι η διαθεσιμότητα του μεγάλου αριθμού δεδομένων, αναφερόμενου και ως περιεχόμενο που παράγεται από τους χρήστες (User generated content - UGC), μαζί με την πρόοδο των μεθόδων Μηχανικής Μάθησης (Machine Learning) πάνω στην Επεξεργασία Φυσικής Γλώσσας (NLP) για την διαχείρισή τους. Η ζωτικότητά του από την άλλη, φαίνεται ξεκάθαρα από το χτίσιμο νεοσύστατων επιχειρήσεων στηριζόμενες στην Ανάλυση Συναισθήματος, τις νέες βιομηχανικές δραστηριότητες από επιχειρήσεις κολοσσούς κ.α. Στην Εικόνα 1 [2], μιλώντας με νούμερα, αποδεικνύεται η τεράστια επιρροή του περιεχομένου που παράγεται από τους χρήστες (UGC) από τα ποσοστά εμπιστοσύνης των χρηστών στο εκάστοτε μέσο ενημέρωσης.

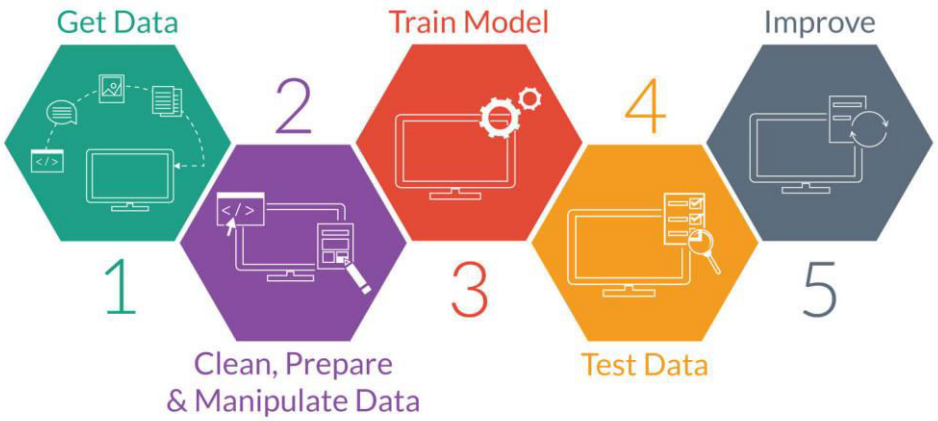
[1] Διαθέσιμα από: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users

[2] Τα δεδομένα για το διάγραμμα της Εικόνας 1 βασίζονται σε μια μελέτη του Yotpo που έγινε το 2014. Περισσότερες πληροφορίες από: https://www.yotpo.com/blog/reviews-increase-social-media-conversions/



Εικόνα 1: Ποσοστά εμπιστοσύνης των μέσων ενημέρωσης [3]

Ωστόσο, ο τεράστιος αυτός όγκος δεδομένων είναι δύσκολα διαχειρίσιμος και αποτελεί εμπόδιο για την αποτελεσματική αξιοποίησή του. Συνεπώς, δημιουργείται η ανάγκη για εργαλεία λογισμικού, ικανά να ανιχνεύουν, να κατηγοριοποιούν και να αποδελτιώνουν με χρήσιμο τρόπο τις γνώμες αυτές ώστε να αξιοποιηθούν κατάλληλα, καθώς αυτός είναι και ο τελικός στόχος αυτής της διπλωματικής εργασίας. Οι τεχνικές Μηχανικής Μάθησης είναι ένας τρόπος που μπορεί να βοηθήσει στην επίτευξη αυτού του στόχου, υλοποιώντας ένα μοντέλο πρόβλεψης. Η τυπική διαδικασία που ακολουθεί ένα μοντέλο πρόβλεψης, φαίνεται στην παρακάτω Εικόνα 2, του Eunxu [4].



Εικόνα 2: Βήματα μοντέλου πρόβλεψης

1.2 Αντικείμενο διπλωματικής εργασίας

Στο γενικό του πλαίσιο, το πρόβλημα αυτής της διπλωματικής εργασίας είναι η Ανάλυση Συναισθήματος με μεθόδους Μηχανικής Μάθησης. Εάν παρατηρήσουμε όμως

[3] Τα δεδομένα για το διάγραμμα της Εικόνα 1 βασίζονται σε μια μελέτη του Yotpo που έγινε το 2014. Περισσότερες πληροφορίες από: https://www.yotpo.com/blog/reviews-increase-social-media-conversions/

[4] Διαθέσιμη από: https://steemit.com/@eunxu

προσεκτικότερα τον τίτλο της διπλωματικής «*Ανάλυση Συναισθήματος Κριτικών Προϊόντων με Μεθόδους Μηχανικής Μάθησης σύμφωνα με τα Κυριότερα Χαρακτηριστικά τους*», μπορούμε να σκιαγραφήσουμε το ακριβές πρόβλημα που κληθήκαμε να επιλύσουμε και σε ποιόν συγκεκριμένο τομέα της Ανάλυσης Συναισθήματος επικεντρωνόμαστε. Στόχος μας, λοιπόν, είναι η εξαγωγή του συνολικού συναισθήματος κριτικών προϊόντων μέσω του συναισθήματος για κάθε ξεχωριστό χαρακτηριστικό του γνώρισμα [5].

Πιο συγκεκριμένα, η διπλωματική αυτή εκτελεί μια λεπτομερής έρευνα, που συμπεριλαμβάνει εντοπισμό, ανάλυση και σύγκριση των καταλληλότερων και αποδοτικότερων μεθόδων Μηχανικής Μάθησης που έχουν δοκιμαστεί και χρησιμοποιούνται για να εντοπίσουν τα χαρακτηριστικά γνωρίσματα του προϊόντος στα οποία αναφέρεται ο σχολιαστής στο κείμενό του, αλλά και το συναίσθημα που εκφράζει για αυτά, με σκοπό να ανακαλυφθεί η συνολική άποψη της κριτικής και κατά επέκταση του προϊόντος. Καλύπτεται πλήρως το απαραίτητο θεωρητικό υπόβαθρο της Ανάλυσης Συναισθήματος και αντιμετωπίζεται σαν πρόβλημα ταξινόμησης συναισθήματος, ώστε να προσεγγιστεί και να επιλυθεί με μεθόδους Μηχανικής Μάθησης με απώτερο σκοπό την κατασκευή ενός ολοκληρωμένου, αυτοματοποιημένου συστήματος σύγκρισης προϊόντων μέσω των κριτικών που έχουν γραφτεί για αυτά, ικανού να αποδίδει σε μεγάλο αριθμό δεδομένων ώστε να αξιοποιηθεί η γνώση που παράγει. Συμπεραίνουμε ότι η επίλυση του εν λόγω προβλήματος χωρίζεται σε δύο κύρια στάδια (υπο - προβλήματα):

1. τον εντοπισμό των χαρακτηριστικών γνωρισμάτων του προϊόντος και
2. την εξόρυξη του συναισθήματος για καθένα από αυτά.

Εστιάζοντας στο κάθε υπο – πρόβλημα ξεχωριστά, αποσαφηνίζονται οι δυσκολίες τους και προτείνονται τεχνικές Μηχανικής Μάθησης ικανές να τα επιλύσουν. Στη συνέχεια, περιγράφεται αναλυτικά μια τυπική διαδικασία αντιμετώπισης του προβλήματος που ακολουθείται (καθώς δεν υπάρχει προκαθορισμένη) και κλείνουμε, παρουσιάζοντας μεγάλο αριθμό σχετικών εργασιών, μέσω δημοσιεύσεων, που επιχείρησαν να επιλύσουν το εν λόγω πρόβλημα συνολικά με Μηχανική Μάθηση και βάσει λεξικών, οι οποίες συνήθως περιγράφουν την μεθοδολογία που ακολούθησαν και αξιολογώντας την απόδοσή της προτείνουν το μοντέλο τους. Τέλος, συγκρίνοντας τις μεθόδους που χρησιμοποιεί η κάθε μία και παρατηρώντας την αποδοτικότητά της, καταλήγουμε σε χρήσιμα συμπεράσματα τα οποία παραθέτουμε και αναφερόμαστε σε πιθανές μελλοντικές επεκτάσεις της διπλωματικής εργασίας μας, καθώς ο συγκεκριμένος ερευνητικός τομέας ήταν αδύνατον να εξεταστεί και να καλυφθεί στα πλαίσια και την έκταση μιας διπλωματικής εργασίας.

[5] Ως χαρακτηριστικό γνώρισμα θεωρείται οποιοδήποτε στοιχείο του προϊόντος το χαρακτηρίζει σε σχέση με όμοιά του προϊόντα, που έχει κάποια τιμή που το καθιστά συγκρίσιμο (π.χ. ταχύτητα επεξεργαστή).

Το κάτωθεν παράδειγμα πιθανής εξεταζόμενης κριτικής ίσως βοηθήσει στον ευκολότερο εντοπισμό της έννοιας του προβλήματος στο οποίο γίνεται συνεχώς αναφορά και ξεδιαλύνει το αντικείμενο της διπλωματικής εργασίας. Η κριτική «*This Asus laptop is the best for gaming as its graphics is incredible and its CPU faster than anyone*» θεωρούμε ότι εκφράζει συνολικό θετικό προσανατολισμό για το προϊόν στο οποίο αναφέρεται (*Asus laptop*), έχοντας εντοπίσει τα χαρακτηριστικά γνωρίσματα στα οποία αναφέρεται (*graphics, CPU*) και τον θετικό προσανατολισμό του καθενός ξεχωριστά.

1.3 Συνεισφορά

Ο εντοπισμός, η εκτενής μελέτη, η αξιολόγηση και η τελική σύγκριση των διαθέσιμων μεθόδων Μηχανικής Μάθησης που εξετάζονται σε αυτή την διπλωματική εργασία, μας καθιστά σε θέση να προτείνουμε τις καλύτερες / καταλληλότερες που μπορούν να εφαρμοστούν για την επιτυχημένη επίλυση του προβλήματος της Ανάλυσης Συναισθήματος των κριτικών προϊόντος εστιάζοντας στα χαρακτηριστικά γνωρίσματά του. Τα τελικά συμπεράσματα εξέτασης των σχετικών εργασιών που εντοπίζονται στο χώρο, βοηθούν στην αποσαφήνιση του προβλήματος επομένως και στην διευκόλυνση επίλυσής του. Επιπλέον, η λεπτομερής και αξιόλογη αυτή έρευνα αξιοποιώντας ερευνητικά πρωτότυπο υλικό είναι σε θέση να αποτελέσει την αφορμή και τα θεμέλια ώστε να δημιουργηθεί ένα αυτοματοποιημένο, ολοκληρωμένο και κερδοφόρο σύστημα σύγκρισης προϊόντων που θα αντιμετωπίζει όλα τα υπο - προβλήματα που αναφέραμε παραπάνω ταυτόχρονα, προσφέροντας όλες τις απαραίτητες γνώσεις. Η βοήθεια ανάπτυξης ενός τέτοιου συστήματος αποτελεί πεδίο ενδιαφέροντος της ερευνητικής περιοχής της Ανάλυσης Συναισθήματος που θα βοηθήσει τόσο τους υποψήφιους αγοραστές αυτών των προϊόντων στις τελικές τους αποφάσεις, όσο και τις εταιρείες των προϊόντων να καταλάβουν τους καταναλωτές τους και να χαράξουν τις στρατηγικές τους προωθήσεις. Ως εκ τούτου, η συνεισφορά της διπλωματικής γίνεται καταφανής και ουσιώδες.

1.4 Οργάνωση κειμένου

Η διάρθρωση / οργάνωση της παρούσας διπλωματικής εργασίας στο υπόλοιπο της έχει ως εξής:

- Στο Κεφάλαιο 2 καλύπτουμε το απαραίτητο θεωρητικό υπόβαθρο που χρειάζεται κάθε αναγνώστης αυτής της διπλωματική εργασίας ώστε να του είναι εύκολο να την παρακολουθήσει και να την κατανοήσει. Αρχικά αποσαφηνίζεται ο τομέας της Μηχανικής Μάθησης και στη συνέχεια η Ανάλυση Συναισθήματος με μια συνοπτική βιβλιογραφική επισκόπηση της ερευνητικής περιοχής της Ανάλυσης Συναισθήματος.

- Στο Κεφάλαιο 3 παρουσιάζουμε τις σχετικές εργασίες που επιλύουν το πρόβλημα που κληθήκαμε να αντιμετωπίσουμε και τις συγκρίνουμε.
- Στο Κεφάλαιο 4 αναφερόμαστε στα συμπεράσματα που εξήχθησαν έπειτα της ολοκλήρωσης αυτής της διπλωματικής εργασίας και στις δυνατότητες επέκτασής της.
- Στο Κεφάλαιο 5 παρουσιάζουμε τις πηγές γνώσης / βιβλιογραφικές αναφορές που εξετάσαμε ώστε να υλοποιήσουμε την εν λόγω διπλωματική εργασία.

2

Θεωρητικό υπόβαθρο

2.1 Μηχανική Μάθηση

Η **Μηχανική Μάθηση** (Machine Learning), όπως περιγράφει και το όνομα της, είναι η διαδικασία δημιουργίας μηχανών ικανών να μαθαίνουν και να βελτιώνουν την απόδοσή τους μέσω της αξιοποίησης προγενέστερης γνώσης και εμπειρίας, αλλά και με την εκμετάλλευση αυτών έτσι ώστε να μπορούν να παίρνουν όσο το δυνατό σωστότερες αποφάσεις για την επίτευξη καλύτερων αποτελεσμάτων. Θεωρείται υπο - πεδίο της επιστήμης των υπολογιστών. Σχηματίστηκε έχοντας σαν βάση τη μελέτη της Αναγνώρισης Προτύπων [6] και της υπολογιστικής θεωρίας μάθησης στην Τεχνητή Νοημοσύνη [7], καθώς αποτελεί βασικό συστατικό αυτής. Εκτός αυτών, είναι άρρητα συνδεδεμένη τόσο με την υπολογιστική στατιστική όσο και με την Γνωσιακή Επιστήμη [8] και μαθηματική βελτιστοποίηση.

Ένα πρόβλημα μάθησης θεωρεί ένα σύνολο *n* δειγμάτων από δεδομένα

[6] Η Αναγνώριση Προτύπων (Pattern Recognition) είναι ένα επιστημονικό πεδίο με στόχο την ανάπτυξη αλγορίθμων για την αυτοματοποιημένη απόδοση κάποιας τιμής ή διακριτικού στοιχείου σε εισαγόμενα δεδομένα, συνήθως κωδικοποιημένα ως αλληλουχίες αριθμών. [1]

[7] Ο όρος Τεχνητή Νοημοσύνη (Artificial intelligence - AI) αναφέρεται στον κλάδο της πληροφορικής ο οποίος ασχολείται με τη σχεδίαση και την υλοποίηση υπολογιστικών συστημάτων που μιμούνται στοιχεία της ανθρώπινης συμπεριφοράς τα οποία υπονοούν έστω και στοιχειώδη ευφυΐα όπως μάθηση, προσαρμοστικότητα, εξαγωγή συμπερασμάτων, κατανόηση από συμφραζόμενα κ.λπ. [2]

[8] Η Γνωσιακή Επιστήμη (Cognitive Science) είναι το επιστημονικό πεδίο που ασχολείται με τη μελέτη του νου και αντλεί γνώσεις και ερευνητική μεθοδολογία από τις γνωστικές νευροεπιστήμες, τη γνωστική ψυχολογία, την τεχνητή νοημοσύνη, τη γλωσσολογία και τη φιλοσοφία του νου. [3]

$$D = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\}$$

και προσπαθεί να μάθει ιδιότητες άγνωστων δεδομένων. Το σύνολο αυτό ονομάζεται σύνολο δεδομένων εκπαίδευσης (training dataset) και όπως καταλαβαίνουμε από το όνομά του χρησιμοποιείται για την εκπαίδευση του συστήματος Μηχανικής Μάθησης, με τη βοήθεια ενός αλγορίθμου, μαθαίνοντας εμπειρικά. Το κάθε δείγμα $\boldsymbol{x_i}$ ονομάζεται χαρακτηριστικό ή γνώρισμα (feature) και μπορεί να είναι μεμονωμένο χαρακτηριστικό (single feature) ή διάνυσμα (feature vector). Στη συνέχεια το μοντέλο πρόβλεψης χρησιμοποιεί ένα σύνολο δεδομένων τα οποία δεν έχει συναντήσει κατά την εκπαίδευσή του, το λεγόμενο σύνολο δεδομένων δοκιμής (testing dataset) και με βάση τις προβλέψεις που κάνει πάνω σε αυτά, αξιολογείται η επίδοση του αλγορίθμου.

Τα δεδομένα που τροφοδοτούν το σύστημα, κάθε $\boldsymbol{x_i}$ δηλαδή, κατά την εκπαίδευση μπορεί να συνοδεύονται από κάποια επιθυμητή, από τον αλγόριθμο, πρόβλεψη, οπότε η μάθηση σε αυτή την περίπτωση ονομάζεται **εποπτευόμενη** ή επιβλεπόμενη ή με επίβλεψη (supervised), η οποία μπορεί να επιλύσει προβλήματα:

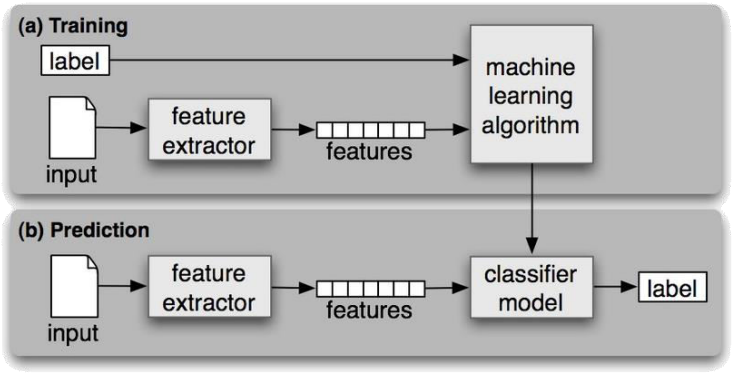
- ταξινόμησης (Classification), όπου τα δεδομένα εισόδου ανήκουν σε δύο ή περισσότερες κλάσεις και το σύστημα Μηχανικής Μάθησης προβλέπει την κλάση άλλων νέων άγνωστων προς αυτό δεδομένων. Το πρόβλημα αυτό εντάσσεται στην κατηγορία της Αναγνώρισης Προτύπων, στην οποία στηρίχθηκε η Μηχανική Μάθηση και άπτει μεγάλο αριθμό εφαρμογών. Μερικές από αυτές είναι η οπτική αναγνώριση προσώπων και χαρακτήρων μέσω της μηχανικής όρασης (computer vision), η αναγνώριση ομιλίας και μουσικής και πάνω στον τομέα της Επεξεργασίας Φυσικής Γλώσσας (NLP), η αυτόματη συνόψιση κειμένου, η μηχανική μετάφραση, η εξαγωγή ονοματικών οντοτήτων (Named Entity Recognition), η συντακτική κατασκευή συντακτικού δέντρου πρότασης, το φιλτράρισμα ανεπιθύμητης αλληλογραφίας, η Ανάλυση Συναισθήματος κ.α.
- παλινδρόμησης (Regression), όπου το σύστημα καλείται να εκτιμήσει την έξοδο που αντιστοιχεί σε ένα πρότυπο εισόδου και αναζητείται μέσα από ένα συνεχές σύνολο τιμών. Τέτοιου είδους προβλήματα μπορεί να είναι για παράδειγμα η εκτίμηση της θερμοκρασίας με βάση τις τιμές της υγρασίας υψομέτρου και πίεσης του αέρα, ή η πρόβλεψη του μήκους ενός ζώου σαν συνάρτηση της ηλικίας και του βάρους του.

  Με την ανάλυση παλινδρόμησης (Regression Analysis) εξετάζεται η σχέση μεταξύ δύο ή περισσοτέρων μεταβλητών με σκοπό την πρόβλεψη των τιμών της μιας, μέσω των τιμών της άλλης (ή των άλλων). Σε κάθε πρόβλημα παλινδρόμησης διακρίνονται:
  - Οι άγνωστες παράμετροι συσχέτισης που δηλώνονται ως β (διάνυσμα).
  - Οι ανεξάρτητες ή ελεγχόμενες μεταβλητές (Independent / Predictor variables) X (διάνυσμα).
  - Η εξαρτώμενη ή απόκρισης μεταβλητή (Dependent / Response variable) Y.

Ένα μοντέλο παλινδρόμησης συσχετίζει το Y σε μία συνάρτηση παλινδρόμησης των X και β. $Y \simeq F(X, \beta)$

$$E\,(Y|X) = f\,(X, \beta)$$

Σε πειραματικές έρευνες, ανεξάρτητη μεταβλητή X είναι εκείνη την οποία μπορούμε να ελέγξουμε, δηλαδή να καθορίσουμε τις τιμές της, ενώ εξαρτημένη μεταβλητή Y είναι εκείνη στην οποία αντανακλάται το αποτέλεσμα των μεταβολών στις ανεξάρτητες μεταβλητές. Ωστόσο, σε μη πειραματικές έρευνες (δειγματοληψίες) η διάκριση μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών δεν είναι πάντοτε σαφής, διότι καμία μεταβλητή δεν είναι ελεγχόμενη αλλά όλες είναι τυχαίες.



Εικόνα 3: Βασικό μοντέλο εποπτευόμενης Μηχανικής Μάθησης

Στην αντίθετη περίπτωση, που το σύστημα δεν έχει ενημέρωση για την πρόβλεψη που πρέπει να κάνει, η μάθηση ονομάζεται **μη εποπτευόμενη** ή μη επιβλεπόμενη ή χωρίς επίβλεψη (unsupervised). Ο σκοπός σε τέτοιου είδους προβλήματα μπορεί να είναι:

- η συσταδοποίηση (clustering), δηλαδή η ανακάλυψη ομοιοτήτων μεταξύ των προτύπων εισόδου (σε συστάδες), η οποία εκτελείται χωρίς προγενέστερη εκπαίδευση του ταξινομητή [9],
- η εκτίμηση πυκνότητας (Density Estimation), δηλαδή η κατανομή των δεδομένων στον χώρο εισόδου,
- η συμπίεση δεδομένων (Data Compression) (ή μείωση διαστάσεων), στην οποία ο όγκος δεδομένων μεγάλων διαστάσεων αντικαθίσταται από δεδομένα μικρότερης διάστασης για διευκόλυνση διαχείρισής τους.

Στην υβριδική περίπτωση που μερικά μόνο από τα δεδομένα συνοδεύονται από επιθυμητή πρόβλεψη, μιλάμε για την **ημι - εποπτευόμενη** ή ημι - επιβλεπόμενη μάθηση (Semi - Supervised). Η παραπάνω διαδικασία σε μερικές περιπτώσεις μπορεί να γίνει με αυτοματοποιημένο τρόπο, όπως για παράδειγμα στην Ανάλυση Συναισθήματος σε κριτικές

[9] Η έννοια του ταξινομητή (classifier) συναντάται στην επίλυση προβλημάτων ταξινόμησης και χαρακτηρίζεται ως ένα μαθηματικό εργαλείο το οποίο είναι υπεύθυνο στο να παίρνει τις αποφάσεις ταξινόμησης και να αναθέτει ετικέτες στα αντικείμενα εισόδου.

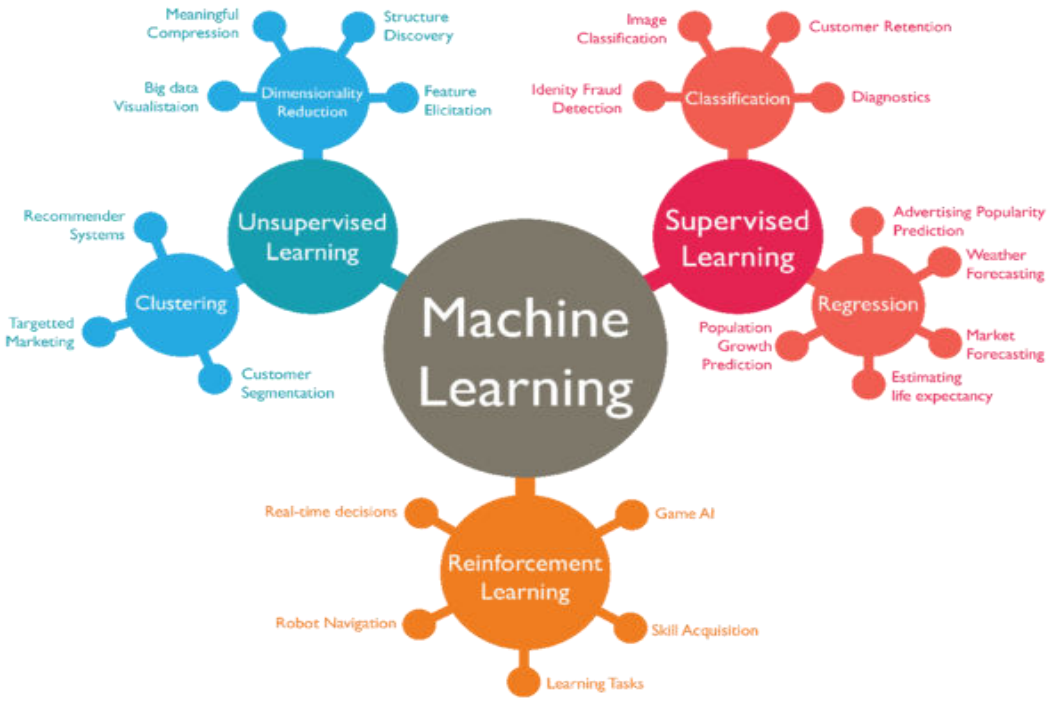
στο διαδίκτυο όπου κάθε χρήστης μαζί με την κριτική του αφήνει και κάποια βαθμολογία (Rating), π.χ. αριθμό αστεριών.

Τέλος, υπάρχει μία ακόμη κατηγορία Μηχανικής Μάθησης, η **ενισχυτική** (Reinforcement), κατά την οποία το σύστημα μαθαίνει την επιθυμητή πρόβλεψη μέσω συνεχούς αλληλεπίδρασης με το περιβάλλον και την ύπαρξη ενός κριτή που «τιμωρεί» ή «επιβραβεύει», όπου τα δεδομένα δεν συνοδεύονται από την επιθυμητή πρόβλεψη. Έτσι, επιλέγει σε κάθε κατάσταση την συμπεριφορά αυτή που θα οδηγήσει στη μεγαλύτερη δυνατή «ανταμοιβή» με βάση αυτά που έχει να μάθει. Η ενισχυτική μάθηση βρίσκει εφαρμογή στην ανάπτυξη στρατηγικής σε παιχνίδια, την αλληλεπίδραση με ανθρώπους και άλλους παρόμοιους τομείς αλλά όχι στην ταξινόμηση κειμένου, συνεπώς, στη συνέχεια αυτής της διπλωματικής εργασίας δεν θα μας απασχολήσει αυτού του είδους η μάθηση, θα επικεντρωθούμε μόνο με τις τρεις πρώτες προαναφερθείσες κατηγορίες Μηχανικής Μάθησης.

Γενικότερα, η μη εποπτευόμενη μάθηση πλεονεκτεί έναντι της εποπτευόμενης λόγω της ευκολότερης και ταχύτερης εύρεσης μεγαλύτερου αριθμού δεδομένων για την εκπαίδευση [10] του αλγορίθμου, καθώς δεν απαιτούν την συνοδεία επιθυμητής πρόβλεψης, δηλαδή έναν άνθρωπο που να τα χαρακτηρίσει χειροκίνητα ως προς την επιθυμητή πρόβλεψη. Η εποπτευόμενη μάθηση ωστόσο με επαρκή δεδομένα, συνήθως αποδίδει καλύτερα. Συνήθως τα επισημασμένα αυτά δεδομένα είναι δυσεύρετα, λόγω της απαίτησης για ανθρώπινη παρέμβαση, σε αντίθεση με αυτά χωρίς την πρόβλεψη που υπάρχουν σε αφθονία.

Μερικοί επιπλέον αξιοσημείωτοι τομείς εφαρμογής της Μηχανικής Μάθησης είναι στο εμπόριο π.χ. για πρόβλεψη της ζήτησης, στην προώθηση αγαθών (marketing) π.χ. για βελτιστοποίηση διαφημίσεων και συστήματα συστάσεων, στην υγεία π.χ. για πρόβλεψη ασθένειας, στις τηλεπικοινωνίες π.χ. ανάλυση αρχείων συστήματος και πρόβλεψη σφαλμάτων καθώς και στην οικονομία π.χ. για πρόβλεψη πιστωτικής βαθμολογίας. Όπως εύκολα διαπιστώνουμε είναι μια επιστήμη που βρίσκει εφαρμογή σε μεγάλο αριθμό κλάδων που επηρεάζουν την καθημερινότητά μας, άρα διαπιστώνεται και η σπουδαιότητά της. Στην κάτωθεν Εικόνα 4 διακρίνονται μερικοί ακόμη τομείς που βρίσκει εφαρμογή η εκάστοτε κατηγορία Μηχανικής Μάθησης.

[10] Με τον όρο εκπαίδευση στο πεδίο της Μηχανικής Μάθησης, εννοούμε την παρουσία προτύπων (στοιχείων του συνόλου εκπαίδευσης) στο σύστημα πρόβλεψης (με ή χωρίς στόχους, ανάλογα τον τύπο μάθησης) με σκοπό την ρύθμιση των παραμέτρων του, ώστε να βελτιώνει την λειτουργία αναγνώρισης.



Εικόνα 4: Κατηγορίες και Εφαρμογές Μηχανικής Μάθησης [4]

Προσεγγίζοντας με τη Μηχανική Μάθηση τα δύο υπο – προβλήματα που τελικά διαχωρίστηκε το βασικό μας πρόβλημα, το πρώτο υπο – πρόβλημα του εντοπισμού των κυριότερων χαρακτηριστικών γνωρισμάτων του προϊόντος θεωρείται κυρίως μη εποπτευόμενης μάθησης, καθώς δεν μπορούμε να γνωρίζουμε εκ των προτέρων τα χαρακτηριστικά γνωρίσματα του προϊόντος στα οποία θα αναφερθεί η κριτική. Αντιθέτως, το δεύτερο υπο – πρόβλημα της εξόρυξης του συναισθήματος για κάθε ξεχωριστό χαρακτηριστικό γνώρισμα, συνηθέστερα, αντιμετωπίζεται ως εποπτευόμενης μάθησης, καθώς τα δεδομένα απαιτούν να συνοδεύονται από κάποια επιθυμητή πρόβλεψη για τον αλγόριθμο. Καθότι εδώ αυτό που απασχολεί είναι ο προσανατολισμός του συναισθήματος (Sentiment Orientation - SO), η ταξινόμηση γίνεται συνήθως σε δύο γνωστές κλάσεις που αναπαριστούν το θετικό ή αρνητικό συναίσθημα, αλλά μπορεί να επεκταθεί συμπεριλαμβάνοντας και το ουδέτερο συναίσθημα, ακόμη και περισσότερων όπως θα εξετάσουμε στη συνέχεια, δυσκολεύοντας το έργο των ταξινομητών.

Όσον αφορά την επίλυση στο πρώτο υπο – πρόβλημα της εξαγωγής λέξεων - κλειδιών γίνεται αναλυτική αναφορά στο Κεφάλαιο 2.4.3. Στο δεύτερο υπο - πρόβλημα βρίσκει εφαρμογή η Ανάλυση Συναισθήματος η οποία αποσαφηνίζεται πλήρως στο επόμενο Κεφάλαιο 2.2 και γίνεται η σύνδεσή της με την Μηχανική Μάθηση στο Κεφάλαιο 2.3, οι μέθοδοι της οποίας μας ενδιαφέρει να φέρουν εις πέρας το υπο – πρόβλημα αυτό.

2.2 Ανάλυση Συναισθήματος

2.2.1 Ορισμός Ανάλυσης Συναισθήματος

Το συναίσθημα (Sentiment) καθίσταται το μέσο με το οποίο μπορεί να αποδοθεί τιμή στον υποκειμενικό χαρακτήρα της φυσικής γλώσσας και περικλείει, τόσο την πολικότητα όσο και την ένταση. Η Ανάλυση Συναισθήματος (Sentiment Analysis) ή Εξόρυξη Γνώμης (Opinion Mining), επομένως, έχει ως στόχο την αναγνώριση στοιχείων υποκειμενικότητας όπως απόψεις, γνώμες και συναισθήματα εκφρασμένα μέσα από μη δομημένα κείμενα σε φυσική γλώσσα.

Τυπικά, ένα συναίσθημα *S* μπορεί να οριστεί ως ένα ζευγάρι παραμέτρων $S = (p,i)$, όπου:

- η παράμετρος *p* αναφέρεται στον προσανατολισμό (orientation) ή πολικότητα (polarity) του συναισθήματος και στη συνηθέστερη περίπτωση λαμβάνει τις δύο τιμές θετικό ή αρνητικό.
- η παράμετρος *i* αναφέρεται στην ένταση (intensity) του συναισθήματος, δηλαδή στο πόσο ισχυρό ή ασθενές είναι και το πλήθος των τιμών που μπορεί να λάβει η ένταση, έστω *I*, επιλέγεται κατά περίπτωση.

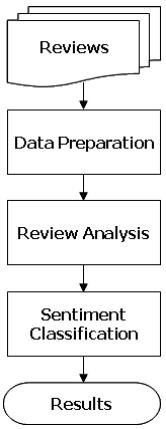
Οι διαφορετικές τιμές που επιτρέπεται να λάβει το συναίσθημα αναφέρονται ως κλάσεις συναισθήματος, ως εκ τούτου, το σύνολο των κλάσεων συναισθήματος απαρτίζεται από *2I* κλάσεις, οι μισές εκ των οποίων έχουν θετικό και οι άλλες μισές αρνητικό προσανατολισμό. Οι κλάσεις συναισθήματος εδώ αποτελούν εκφράσεις υποκειμενικότητας μοντελοποιώντας τα αποτελέσματα που παράγει ένας ταξινομητής κειμένου, προσεγγίζοντας την Ανάλυση Συναισθήματος από τη σκοπιά της Μηχανικής Μάθησης, όπου αντιπροσωπεύει ένα μεγάλο **πρόβλημα ταξινόμησης κειμένου** (ή εξόρυξη γνώμης). Η βασική τους διαφορά όπως επισημαίνει ο Sebastiani (2002) είναι ότι η ταξινόμηση κειμένου αναφέρεται στην αντιστοίχιση κειμένων φυσικής γλώσσας σε θεματικές κατηγορίες ή κλάσεις οι οποίες ανήκουν σε ένα προκαθορισμένο σύνολο με βάση τους στόχους του εκάστοτε προβλήματος, ενώ η Ανάλυση Συναισθήματος αναφέρεται σε ένα μικρότερο σύνολο κατηγοριών διότι επικεντρώνεται κυρίως στην κατάταξη ενός κειμένου ως προς την πολικότητά του, άρα οι κατηγορίες είναι ανεξάρτητες της θεματολογίας του προβλήματος (π.χ. θετική – αρνητική κατηγορία).

Η Ανάλυση Συναισθήματος καθίσταται μια από τις πιο δραστήριες περιοχές έρευνας στην Επεξεργασία Φυσικής Γλώσσας (NLP), καθώς οι απόψεις των ανθρώπων εκφραζόμενες σε μεγάλο βαθμό πλέον και μέσω του διαδικτύου έχουν αναμφισβήτητα έντονο αντίκτυπο στην κοινωνία. Οι απόψεις που εξετάζονται εντοπίζονται σε μορφή κριτικών, συζητήσεων σε τόπος δημόσιας συζήτησης (Forum) και προσωπικές ιστοσελίδες όπου καταγράφονται

απόψεις σε τακτική βάση, σχολίων και δημοσιεύσεων στα μέσα κοινωνικής δικτύωσης (Social Media) και συλλέγονται συνήθως από τις ενδιαφερόμενες επιχειρήσεις με web crawlers [11]. Εν τούτοις, πολλές επιχειρήσεις δεν αρκούνται στις ανωτέρω εξωτερικές πηγές πληροφόρησης και χτίζουν δικές τους εσωτερικές πηγές με σκοπό να συλλέξουν μεγαλύτερο αριθμό δεδομένων. Για παράδειγμα συλλέγουν γνώμες καταναλωτών μέσω ηλεκτρονικών μηνυμάτων (e - mail), SMS (Short Message Service), άμεσα μηνύματα [12] και τηλεφωνικές επικοινωνίες αλλά και ενσωματώνοντας δημοσκοπήσεις και φόρμες συμπλήρωσης σχολίων / κριτικών μέσα στις ιστοσελίδες ή τα ηλεκτρονικά καταστήματά τους. Οι εσωτερικές αυτές πηγές συγκεντρώνουν κυρίως σχόλια για τα προϊόντα τους αλλά και την γενική εικόνα της εταιρικής τους ταυτότητας / εμπορικού σήματος (brand name) και τα αξιοποιούν για να βελτιώσουν τις πελατειακές τους σχέσεις και να λάβουν γνώση για την γνώμη που επικρατεί όσον αφορά τα προϊόντα και υπηρεσίες τους και τις συμπεριφορές του κοινού πάνω στην εταιρική τους φήμη. Γενικότερα, η κατάλληλη επεξεργασία, η εξαγωγή ορθών αποτελεσμάτων και η αξιολόγηση τέτοιου είδους πληροφοριών είναι ζωτικής σημασίας, καθώς μπορεί να οδηγήσει σε συγκλονιστικά αποτελέσματα σε κάθε τομέα, ξεκινώντας από τα καταναλωτικά προϊόντα και υπηρεσίες και φτάνοντας έως την υγεία, την ψυχολογία, την ενημέρωση, τις μετοχές και την οικονομία, τα κοινωνικά γεγονότα αλλά ακόμη και τον τομέα την πολιτικής επηρεάζοντας τα αποτελέσματα πολιτικών εκλογών. Συνεπώς, κατανοούμε την τεράστια επιρροή των απόψεων των άλλων στις ζωές μας και στην κοινωνία ευρύτερα και την ανάγκη για αποτελεσματική αξιοποίησή τους. Τα βήματα ενός τυπικού μοντέλου Ανάλυσης Συναισθήματος είναι (1) η συλλογή των κειμένων, (2) η διαμόρφωσή τους ώστε να έχουν όλα την ίδια μορφή που απαιτεί ο αλγόριθμος που θα τα χρησιμοποιήσει, (3) η ανάλυσή τους, (4) η αποτίμηση του συναισθήματος και (5) τα τελικά αποτελέσματα αποτίμησης. Η διαδικασία αυτή απεικονίζεται στην Εικόνα 5.

[11] Ο web crawler (ή spider) «διαβάζει» / «σκανάρει» τον κώδικα ιστοσελίδων, γραμμένων κατά κύριο λόγο στην γλώσσα σήμανσης HTML, συνήθως για τον σκοπό της ευρετηριοποίησης ιστού (web spidering). [5]

[12] Η τεχνολογία άμεσων μηνυμάτων (Instant Messaging - IM) είναι ένας τύπος συνομιλίας συνήθως μεταξύ δύο χρηστών, που προσφέρει μετάδοση κειμένου σε πραγματικό χρόνο μέσω του Διαδικτύου. [6]



Εικόνα 5: Τυπικό μοντέλο Ανάλυσης Συναισθήματος κριτικών, Leung (2008)

2.2.2 Τομείς Εφαρμογής Ανάλυσης Συναισθήματος

Οι χρήσεις της Ανάλυσης Συναισθήματος αυξάνονται με ραγδαίους ρυθμούς καθώς γίνεται αντιληπτή η χρησιμότητά της, με συνέπεια νέοι τομείς να εμφανίζονται ολοένα και περισσότερο λόγω της ζωτικότητας του κλάδου. Η πληθώρα των τομέων εφαρμογής που αναφέρονται κάτωθεν μπορεί να επιβεβαιώσει την παραπάνω εκδοχή. Μερικοί λοιπόν σημαντικοί τομείς εφαρμογής της Ανάλυσης Συναισθήματος είναι οι εξής:

- τα Συστήματα Συστάσεων (Recommender Systems), τα οποία θεωρούνται μια υποκατηγορία συστημάτων φιλτραρίσματος σημαντικότητας πληροφοριών που επιδιώκει να προβλέψει την προτίμηση ενός χρήστη για ένα αντικείμενο σύμφωνα με τα ενδιαφέροντά του. Τα δεδομένα που έχουν να διαχειριστούν τα συστήματα συστάσεων αφορούν διάφορες συσχετίσεις μεταξύ των τριών ειδών οντοτήτων χρήστες, αντικείμενα και συναλλαγές, Ricci κ.α (2015). Στον χώρο του ηλεκτρονικού εμπορίου εντοπίζονται σαν πελάτης και προϊόν. Οι συναλλαγές από την άλλη, μπορεί να είναι αξιολογήσεις που συλλέγονται είτε άμεσα (π.χ. δίνοντας μια τιμή σε μια βαθμολογική κλίμακα πέντε αστεριών για ένα προϊόν) είτε έμμεσα (π.χ. από το ιστορικό αγορών του χρήστη). Τα συστήματα συστάσεων γίνονται ολοένα και πιο δημοφιλή τα τελευταία χρόνια και χρησιμοποιούνται σε ποικίλους τομείς, όπως ψυχαγωγία, ειδήσεις, βιβλία, ερευνητικά άρθρα, ερωτήματα αναζήτησης και προϊόντα / υπηρεσίες γενικότερα [7].
- η ανίχνευση ανεπιθύμητης αλληλογραφίας (Opinion Spam Detection) στο διαδίκτυο. Το spamming της γνώμης αναφέρεται σε «παράνομες» δραστηριότητες που προσπαθούν να παραπλανήσουν τους αναγνώστες ή τα συστήματα αυτοματοποιημένης ανάλυσης συναισθημάτων. Τα μηνύματα ανεπιθύμητης αλληλογραφίας έχουν πολλές μορφές, για παράδειγμα ψεύτικες κριτικές (ή ψευδείς κριτικές), ψεύτικα σχόλια, ψεύτικα ιστολόγια (blogs), ψευδείς δημοσιεύσεις σε κοινωνικά δίκτυα, παραπλανητικά μηνύματα ηλεκτρονικού ταχυδρομείου κ.α.

Σύμφωνα με τον Liu (2012), οι απόψεις του διαδικτύου χρησιμοποιούνται όλο και περισσότερο στην πράξη από τους καταναλωτές, τους οργανισμούς και τις επιχειρήσεις για τη λήψη αποφάσεων, έτσι το spamming της γνώμης θα γίνει ανεξέλεγκτο και εξελιγμένο ενώ η ανίχνευση σχολιασμών spam ή απόψεων θα γίνεται ολοένα και πιο κρίσιμη.

- η στοχευμένη τοποθέτηση διαφημίσεων (Advertising Opinion Mining) σε ιστοσελίδες σύμφωνα με τις προτιμήσεις του επισκέπτη κατά την πλοήγησή του στο διαδίκτυο.
- η διαχείριση των σχολίων (Feedback management) από επιχειρήσεις, όπου μέσω της χρήσης της Ανάλυσης Συναισθήματος γίνονται περισσότερο κατανοητές οι προτιμήσεις των καταναλωτών με σκοπό την βελτιωμένη εξυπηρέτησή τους αλλά και την γενικότερη βελτίωση της φήμης της επιχείρησης εντοπίζοντας τα συναισθήματα τόσο των πελατών όσο και των εργαζομένων και των επενδυτών τους. Θεωρείται πλέον το σύγχρονο μέσο επικοινωνίας μιας επιχείρησης με τους πελάτες της καθώς έχει αντικαταστήσει παλιά μέσα διαχείρισης πελατειακών σχέσεων (Customer relationship management - CRM) και δείχνει πολλά υποσχόμενος τομέας. Αντιθέτως, η έλλειψη της διαδικτυακής παρουσίας μιας επιχείρησης θεωρείται ανησυχητική καθώς αυτό προδιαθέτει έλλειψη ενδιαφέροντος προς τους καταναλωτές της, άρα και έλλειψη εμπιστοσύνης. Χαρακτηριστικά παραδείγματα παγκοσμίου φήμης επιχειρήσεων που χρησιμοποιούν την ανάλυση συναισθήματος για να καθορίσουν τις στρατηγικές τους προωθήσεις είναι η Cisco Systems [13], η Kia [14] κ.α
- στον τομέα της Επιχειρηματικής Ευφυΐας (Business intelligence - BI) με σκοπό την σχεδίαση νέων προϊόντων και υπηρεσιών. Η αποτελεσματική «ανάγνωση» της αγοράς με τα κατάλληλα εργαλεία προβλέπει μελλοντικές ανάγκες και απαιτήσεις συνεπώς είναι ένα πολύ σημαντικό συμβουλευτικό εργαλείο για όλες τις σύγχρονες επιχειρήσεις. Για παράδειγμα, εάν με βάση την ανάλυση συναισθήματος τεράστιου αριθμού δημόσιων αναρτήσεων σε μέσα κοινωνικής δικτύωσης προβλέπεται ότι το κόκκινο χρώμα θα είναι η νέα τάση μόδας στο ρουχισμό για την επόμενη εποχή, αναμένεται οι μεγαλύτερες επιχειρήσεις μόδας και σχεδιαστές να το ενσωματώσουν στα σχέδιά τους της επόμενης σαιζόν (season).
- στον τομέα της ψυχολογίας, έχοντας αντικαταστήσει τις παραδοσιακές μορφές έρευνας των ερωτηματολογίων και των ακαδημαϊκών συνεντεύξεων με τεχνικές ανάλυσης συναισθήματος, οι οποίες ευεργετούν την επιστήμη καθώς οι αναλύσεις που αφορούν την ψυχολογία απαιτούν μεγάλο όγκο δεδομένων, ο πλούτος των

[13] Η Cisco Systems, Inc. (CSCO) είναι η μεγαλύτερη πολυεθνική εταιρεία δικτύωσης στον κόσμο.

[14] Η Kia Motors είναι θυγατρική εταιρεία της Hyundai Motors, μία από τις μεγαλύτερες αυτοκινητοβιομηχανίες παγκοσμίως.

οποίων γίνεται πλέον άμεσα διαθέσιμος. Για παράδειγμα ο εντοπισμός της κατάθλιψης και πιο συγκεκριμένα της προδιάθεσης για αυτοκτονία σε ορισμένους ανθρώπους είναι ζωτικής σημασίας.

- στον τομέα της δημοσιογραφίας, καθώς μεγάλος αριθμός κολοσσών ενημέρωσης, όπως το CNN [15] και το Twitter [16], εστιάζουν στην ανάλυση συναισθήματος ώστε να κατανοήσουν τον τρόπο σκέψης των ανθρώπων με σκοπό την ενημέρωσή τους, όπως αναφέρει ο Sam Petulla στο ηλεκτρονικό του άρθρο [8].
- στον οικονομικό τομέα γενικότερα, όπου η αγορά μετοχών επηρεάζεται άμεσα από άρθρα και απόψεις κυρίως στα κοινωνικά δίκτυα. Η επίδραση των ειδήσεων σχετικά με το εμπόριο των τιμών και συναλλαγών είναι ασύμμετρη ως προς το χρόνο, καθώς ειδήσεις που προξενούν θετικό συναίσθημα έχει αποδειχθεί ότι σχετίζονται με μεγάλες αυξήσεις τιμών σε σχετικά σύντομο χρονικό διάστημα, ενώ ειδήσεις που προξενούν αρνητικό συναίσθημα συνδέονται με μειώσεις των τιμών για ένα πιο παρατεταμένο χρονικό διάστημα.
- στην πολιτική. Μεγάλος αριθμός εκλογικών εκστρατειών στηρίζεται στις απόψεις των ψηφοφόρων κυρίως από τα μέσα κοινωνικής δικτύωσης, επομένως η ανάλυση συναισθήματος παίζει και εδώ καθοριστικό ρόλο, καθώς μπορεί να τροποποιήσει ακόμη και τα αποτελέσματα πολιτικών εκλογών.
- στις δημοσκοπήσεις. Η αντικατάσταση των παραδοσιακών μορφών δημοσκοπήσεων (π.χ. τηλεφωνικές) με την ανάλυση συναισθήματος στον τεράστιο αριθμό απόψεων που εντοπίζονται πλέον στο διαδίκτυο, κυρίως έπειτα της εντυπωσιακής αύξησης των μέσων κοινωνικής δικτύωσης, οδηγεί σε ταχύτερη και λιγότερο δαπανηρή εξόρυξη της κοινής γνώμης σε διάφορα θέματα ενδιαφέροντος. Η εικασία αυτή επιβεβαιώνεται έχοντας παραδεχτεί μεγάλες εταιρείες δημοσκοπήσεων ότι και αυτές με τη σειρά τους χρησιμοποιούν εργαλεία της ανάλυσης συναισθήματος για βελτιωμένα αποτελέσματα.

2.2.3 Πρόβλημα Ανάλυσης Συναισθήματος

Στο γενικό πλαίσιο, η διάκριση ενός συναισθήματος από κάποιον άλλο είναι ένα αμφισβητούμενο ζήτημα στην έρευνα των συναισθημάτων και στη συναισθηματική επιστήμη (Affective Science). Ο τομέας που ασχολείται με την ανάπτυξη ευφυών συστημάτων που αντιλαμβάνονται και κωδικοποιούν το συναίσθημα είναι γνωστός ως Συναισθηματική Υπολογιστική (Affective Computing) και παραδείγματα τέτοιων συστημάτων παρουσιάζονται στο επόμενο Κεφάλαιο 2.2.4. Για την οικοδόμηση τέτοιων συστημάτων είναι

[15] Το CNN (Cable News Network) είναι ένα από τα μεγαλύτερα ειδησεογραφικά τηλεοπτικά κανάλια.

[16] Το Twitter είναι ένας ιστοχώρος κοινωνικής δικτύωσης και υπηρεσία ειδήσεων που επιτρέπει στους χρήστες του να στέλνουν και να διαβάζουν σύντομα μηνύματα που ονομάζονται tweets.

απαραίτητη η αναπαράσταση των συναισθημάτων σε μια κλίμακα κατανοητή από τους υπολογιστές. Τα συναισθήματα μπορούν να γίνουν αντιληπτά σε δισδιάστατο ή τρισδιάστατο χώρο. Τόσο ο Wilhelm Wundt [17] που θεωρείται ένας από τους ιδρυτές της σύγχρονης ψυχολογίας, όσο και ο Schlosberg (1954) προτείνουν τον διαχωρισμό του συναισθήματος σε τρεις διαστάσεις. Οι βασικές συναισθηματικές διαστάσεις που προκύπτουν είναι οι εξής:

- Σθένος (Valence), η διάσταση που καθορίζει το αίσθημα ευχαρίστησης ή δυσαρέσκειας και κυμαίνεται μεταξύ πολύ θετικών και πολύ αρνητικών τιμών. Για παράδειγμα, τόσο ο *θυμός* όσο και ο *φόβος* είναι δυσάρεστα συναισθήματα και βρίσκονται υψηλά στην κλίμακα δυσαρέσκειας.
- Διέγερση (Arousal), η διάσταση που εκφράζει τον βαθμό που ένα άτομο είναι σε θέση να πράξει, μετρά την ένταση του συναισθήματος. Για παράδειγμα, ενώ και ο *θυμός* και η *οργή* είναι δυσάρεστα συναισθήματα, η *οργή* έχει υψηλότερη διεγερτική κατάσταση (ή ένταση).
- Κυριαρχία (Dominance), η διάσταση που αντιπροσωπεύει τον κυρίαρχο χαρακτήρα του συναισθήματος. Για παράδειγμα, ενώ και ο *θυμός* και ο *φόβος* είναι δυσάρεστα συναισθήματα, ο *θυμός* είναι ένα κυρίαρχο συναίσθημα ενώ ο *φόβος* ένα υποτακτικό συναίσθημα.

Κατά συνέπεια, εστιάζοντας στην έκφραση της γνώμης με γραπτό λόγο, θεωρούμε το πρόβλημα της Ανάλυσης Συναισθήματος ως πρόβλημα κατηγοριοποίησης της γνώμης που ως στόχο έχει να κατατάξει ένα έγγραφο ή ένα απόσπασμα εγγράφου σε κάποια διάσταση (ταξινόμηση κειμένου). Η απλούστερη εκδοχή του προβλήματος είναι δυαδικής απόφασης, όπου η εκτίμηση γίνεται χρησιμοποιώντας μόνο την διάσταση σθένους για δύο κατηγορίες, την θετική που προβλέπει υψηλό σθένος και την αρνητική με χαμηλό σθένος. Εντούτοις, μια τόσο απλουστευμένη μοντελοποίηση συχνά δεν καλύπτει τις ανάγκες απαιτητικών συστημάτων Ανάλυσης Συναισθήματος, έτσι πολλές φορές χρειάζεται η επέκταση της συναισθηματικής κλίμακας σε τρεις κατηγορίες, συμπεριλαμβανομένης της ουδέτερης με μηδενικό σθένος ή και περισσότερες, ανάλογα τις ανάγκες της εφαρμογής. Συνεπώς στις περιπτώσεις αυτές οι γνώμες διακρίνονται όχι μόνο ως προς τον προσανατολισμό τους (σθένος), αλλά και ως προς την έντασή τους (διέγερση), μιλώντας με όρους Μηχανική Μάθησης. Τα περισσότερα μοντέλα αναπαράστασης του συναισθήματος χρησιμοποιούν το σθένος και την κυριαρχία καθώς η διέγερση συσχετίζεται έντονα με την διέγερση και ταυτίζονται.

Συμπερασματικά από όσα προαναφέρθηκαν, το συναίσθημα καθίσταται ένα πολύπλοκο πεδίο ανάλυσης, καθώς πρέπει να αποσαφηνιστεί μέσα από τις ποικίλες ιδιαιτερότητές του. Όταν αυτό δε εκφράζεται με γραπτό λόγο, το πεδίο του εξετάζουμε δηλαδή, μπορεί να

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

προσεγγιστεί αξιοποιώντας εργαλεία και τεχνικές από τις περιοχές της Ανάκτησης Πληροφορίας (Information Retrieval - IR), της Μηχανικής Μάθησης (Machine Learning - ML), της Υπολογιστικής Γλωσσολογίας (Computational Linguistics) και γενικότερα της Επεξεργασίας Φυσικής Γλώσσας (NLP). Στην παρούσα διπλωματική εργασία επικεντρωνόμαστε στις μεθόδους Μηχανικής Μάθησης, τις εντοπίζουμε, τις αναλύουμε και συγκρίνοντάς τες ξεχωρίζουμε τις καταλληλότερες που χρησιμοποιούνται στον τομέα της Ανάλυσης Συναισθήματος πάνω σε κριτικές προϊόντων.

Ένα από τα σημαντικότερα εμπόδια που συναντώνται κατά την χρήση της Μηχανικής Μάθησης στην Ανάλυση Συναισθήματος είναι η εξάρτηση του εκπαιδευμένου υπολογιστικού συστήματος από τον συγκεκριμένο θεματικό τομέα, αλλά και τον τύπο των δεδομένων σύμφωνα με τα οποία έγινε η εκπαίδευσή του (genre & domain dependence). Αυτό έχει ως αποτέλεσμα την χαμηλή απόδοση του συστήματος σε περιπτώσεις εισαγωγής κειμένων προς ανάλυση, διαφορετικής δομής και τομέα ενδιαφέροντος, όπως επισημαίνουν εύστοχα οι Andreevskaia και Bergler (2008) Πέραν αυτού, εντοπίζεται πληθώρα τομέων για τους οποίους είναι δύσκολο να βρεθεί ικανοποιητικό υλικό εκπαίδευσης (training data) για το σύστημα, αλλά ακόμη πιο δύσκολο και χρονοβόρο η δημιουργία ενός νέου συνόλου δεδομένων με σκοπό να χρησιμοποιηθεί για την εκπαίδευση του συστήματος. Αυτή η επίπονη διαδικασία, αλλά τις περισσότερες φορές μη αναπόφευκτη, απαιτεί την συλλογή ενός ικανοποιητικού μεγέθους δεδομένων και πολλές φορές τον σχολιασμό αυτών, που ενέχει τον κίνδυνο της υποκειμενικότητας του σχολιαστή, πρόκληση που αναλύεται στα επόμενα κεφάλαια. Εντούτοις, υπάρχει η επιλογή χρήσης έτοιμων συνόλων δεδομένων που μπορούν να χρησιμοποιηθούν αντί της δημιουργίας νέων, τα οποία παρουσιάζονται στο Κεφάλαιο 2.2.6.

2.2.4 Λογισμικά Ανάλυσης Συναισθήματος

Η χρήση των ευφυών συστημάτων στα οποία αναφερθήκαμε στα προηγούμενα κεφάλαια που επιχειρούν να επιλύσουν το πρόβλημα της Ανάλυσης Συναισθήματος, κερδίζουν ολοένα και περισσότερο έδαφος με την εξέλιξη των τεχνικών Επεξεργασίας Φυσικής Γλώσσας (NLP). Τα τελευταία χρόνια δε, μπορούν να εντοπιστούν και ελεύθερα λογισμικά (free software) στο Διαδίκτυο, πέραν των επί πληρωμής επιχειρηματικών συστημάτων. Αν και το καθένα χρησιμοποιεί διαφορετικές μεθόδους και αλγορίθμους, όλα έχουν κοινό στόχο, την αποτελεσματικότερη επίλυση στο πρόβλημα της Ανάλυσης Συναισθήματος και συνήθως η διαδικασία που ακολουθούν για να το επιτύχουν είναι κάτι που δεν φανερώνεται από τον κατασκευαστή με λεπτομέρειες καθώς αποτελεί την "μυστική συνταγή" της επιτυχίας του.

Κατά κανόνα διατίθενται σε μορφή APIs [18] ώστε να μπορούν να ενσωματωθούν σε οποιαδήποτε εφαρμογή. Ενδεικτικά παρακάτω παρουσιάζουμε τα πιο ευρέως χρησιμοποιούμενα κατά τα σημερινά χρόνια, καθώς εντοπίστηκε μια πληθώρα αυτών που θα ήταν περιττό και κουραστικό να αναφερθούμε σε όλα, καθώς δεν είναι το αντικείμενο μελέτης της παρούσας διπλωματικής εργασίας, καθώς έπεται ο με τα παραθέτει συγκεντρωτικά σύμφωνα με το εύρος γλωσσών που καλύπτουν και την έξοδο που παρέχει το καθένα.

- **Semantria** [19]. Ηλεκτρονική εφαρμογή που διατίθεται και σε API και χρησιμοποιεί εργαλεία κειμένου για να εκτελέσει Ανάλυση Συναισθήματος σε tweets [20], το Facebook, σε κριτικές, σε έρευνες, σε σχόλια ή σε περιεχόμενα επιχειρήσεων. Επιτρέπει την ανάλυση οποιουδήποτε κειμένου μέχρι 16.384 χαρακτήρες με σύγχρονες τεχνολογίες και είναι πλήρως εξοπλισμένο, για να υποστηρίζει παραπάνω από δέκα γλώσσες Αγγλικά, Γαλλικά, Πορτογαλικά, Ισπανικά, Γερμανικά, Κινέζικα, Ιταλικά, Κορεατικά, Ιαπωνικά, Ολλανδικά. Από την ανάλυση του κειμένου, εκτός του αρνητικού, θετικού ή ουδέτερου συνολικού συναισθήματος που προκύπτει, μαζί με την αντίστοιχη βαθμολογία, εξάγονται οι οντότητες που εμφανίζονται σε αυτό, οι κατηγορίες στις οποίες ανήκουν οι οντότητες, τα θέματα στα οποία αναφέρεται το κείμενο και η περίληψή του. Για την ανάλυση του συναισθήματος χρησιμοποιείται συντακτική ανάλυση του κειμένου και έπειτα εντοπίζονται οι φράσεις με συναισθηματικό περιεχόμενο. Στο τέλος η βαθμολογία προκύπτει από το συνδυασμό των βαθμολογιών των επιμέρους φράσεων. Το γεγονός που το κάνει να ξεχωρίζει είναι η δυνατότητα παραμετροποίησης, καθώς η κατηγοριοποίηση και η εξαγωγή οντοτήτων μπορεί εύκολα να εκπαιδευτεί ώστε να ταιριάζει με τα ειδικά λεξιλόγια κάθε περίπτωσης.
- **Skyttle** [21]. Είναι ένα API που επιστρέφει τις κατηγορίες θετικών, αρνητικών και ουδέτερων συναισθημάτων σε επίπεδο φράσεων. Υποστηρίζει τέσσερις γλώσσες Αγγλικά, Γαλλικά, Γερμανικά και Ρώσικα. Η ανάλυσή του περιλαμβάνει, συν τοις άλλοις, τον υπολογισμό ποσοστών για τις κατηγορίες συναισθήματος που εμφανίζονται και τον εντοπισμό των λέξεων - κλειδιών (keywords) του κειμένου, τον σχολιασμό των κειμένων για το συναίσθημα και τις λέξεις-κλειδιά καθώς και την εύρεση του συναισθήματος που σχετίζεται με οντότητες και λέξεις κλειδιά.

[18] Ένα API (Application Programming Interface) είναι η διεπαφή των προγραμματιστικών διαδικασιών που παρέχει ένα λειτουργικό σύστημα, βιβλιοθήκη ή εφαρμογή προκειμένου να επιτρέπει να του γίνονται αιτήσεις από άλλα λογισμικά ή/και ανταλλαγή δεδομένων [9].

[19] Διαθέσιμο από: https://www.lexalytics.com/

[20] Τα tweets είναι μικρής έκτασης κείμενα που αντιπροσωπεύουν μια άποψη δημοσιευμένα στο κοινωνικό δίκτυο Twitter.

[21] Διαθέσιμο από: http://www.getsentiment.io/

- **Bittext** [22]. Είναι ένα API που υλοποιεί ανάλυση κειμένου σε επτά γλώσσες με τη μέθοδο Deep Linguistic Analysis η οποία του επιτρέπει να εξάγει συναίσθημα, έννοιες και κατηγορίες από το κείμενο.
- **OpenDover** [23]. Είναι μια εφαρμογή που διατίθεται και σε API και επιτρέπει την εξαγωγή χαρακτηριστικών συναισθήματος από ιστολόγια, συστήματα διαχείρισης περιεχομένου, ιστοσελίδες κι άλλες ποικίλες πηγές. Χρησιμοποιεί σημασιολογική τεχνολογία για το συναισθηματικό καθορισμό των κειμένων. Δίνει την δυνατότητα επιλογής ανάμεσα στις δύο κατηγορίες ανάλυσης: (1) βασισμένη σε οντολογίες και (2) βασισμένη στο συναίσθημα αλλά και την χειροκίνητη επιλογή μέχρι 6 επιθυμητών αντικείμενων με μεγαλύτερη έμφαση στην ανάλυση.
- **Sentigem** [24]. Μια ηλεκτρονική πλατφόρμα που υλοποιεί ανάλυση συναισθήματος σε Αγγλικά κείμενα. Είναι ένα API εύκολο στη χρήση του που υπολογίζει το συνολικό συναίσθημα του κειμένου καθώς και το συναίσθημα των επιμέρους φράσεων στις κατηγορίες θετικό / αρνητικό / ουδέτερο συναίσθημα.
- **Sentiment viz** [25]. Είναι μια εφαρμογή που πραγματοποιεί Ανάλυση Συναισθήματος στο κοινωνικό δίκτυο Twitter και οπτικοποιεί τα αποτελέσματά της. Για την εκτίμηση του συναισθήματος χρησιμοποιείται ένα λεξικό γνώμης στο οποίο αναζητείται κάθε tweet και συνδυάζοντας τις λέξεις ευχαρίστηση και διέγερση επέρχεται το τελικό συναίσθημα του tweet. Το λεξικό γνώμης παρέχει τα μέτρα σθένος και διέγερσης για περίπου 10.680 αγγλικές λέξεις και κάθε λέξη βαθμολογείται σε κλίμακα που κυμαίνεται από 1 έως 9. Οι υπολογιστικές του μέθοδοι για την εκτίμηση συναισθήματος περιλαμβάνουν αλγόριθμους Μηχανικής Μάθησης όπως προσεγγίσεις Naive Bayes, SVM και Μέγιστης Εντροπίας ή συνδυασμούς κοινής λογικής σκέψης και συναισθηματικής οντολογίας.
- **SentiStrength** [26]. Ένα εργαλείο που εκτιμά την ένταση των θετικών και αρνητικών συναισθημάτων σε κάποιο μικρό κείμενο, ακόμα και για ανεπίσημη καθομιλουμένη γλώσσα. Εκτός από την πολικότητα κάθε κειμένου υπολογίζεται και η αντίστοιχη ισχύ του συναισθήματος με εύρος τιμών 1 έως 5. Έχει πολύ καλή ακρίβεια για μικρά κείμενα που εξάγονται από τα κοινωνικά δίκτυα στην αγγλική γλώσσα, εξαιρώντας τα κείμενα που περιέχουν πολιτικό περιεχόμενο. Άλλες γλώσσες που μπορεί να διαχειριστεί είναι τα: Φινλανδικά, Γερμανικά, Ολλανδικά, Ισπανικά. Ρωσικά, Πορτογαλικά, Γαλλικά, Αραβικά, Πολωνικά, Περσικά, Σουηδικά, Ελληνικά, Ουαλίας, Ιταλικά και Τουρκικά. Χρησιμοποιεί ένα σύνολο από 2600 σχόλια και μία λίστα με 2310 θετικούς και

[22] Διαθέσιμο από: https://www.bitext.com/text-analysis-api-2/
[23] Διαθέσιμο από: http://demo.opendover.nl/
[24] Διαθέσιμο από: https://sentigem.com/
[25] Διαθέσιμο από: https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/
[26] Διαθέσιμο από: http://sentistrength.wlv.ac.uk/

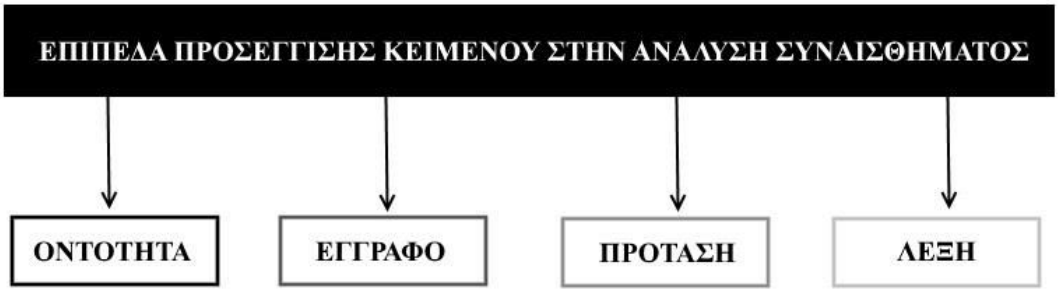
αρνητικούς όρους ταξινομημένους ως προς την πολικότητά τους μαζί με την αντίστοιχη ισχύ τους. Στην λίστα συμπεριλαμβάνονται emoticons [27], όροι άρνησης, λέξεις που αυξάνουν ή μειώνουν την ισχύ του συναισθήματος των συμφραζόμενων όρων, καθώς στην τελευταία έκδοσή του ενισχύθηκε η λίστα με ιδιωματισμούς και την έννοια της ενίσχυσης της πολικότητας.

Πίνακας 1: Σύγκριση έτοιμων Λογισμικών Ανάλυσης Συναισθήματος

	Αριθμός Υποστηριζόμενων Γλωσσών	Εξαγόμενο Αποτέλεσμα
Bittext	7	συναίσθημα & έννοιες & κατηγορίες λέξεων – κλειδιών
OpenDover	1	επιλογή ανάλυσης σε οντολογίες ή συναίσθημα
Semantria	10	οντότητες, κατηγορίες οντοτήτων, θέματα & περίληψη κειμένων, φράσεις συναισθήματος
Sentigem	1	συνολικό συναίσθημα κειμένου & επιμέρους φράσεων
Sentiment Viz	1	συναισθηματική οντολογία
SentiStrength	16	πολικότητα & ισχύ συναισθήματος σε εύρος τιμών 1-5
Skyttle	4	λέξεις – κλειδιά, οντότητες, συναίσθημα & κατηγορίες λέξεων – κλειδιών

2.2.5 Επίπεδα προσέγγισης προβλήματος Ανάλυσης Συναισθήματος

Το πρόβλημα της Ανάλυσης Συναισθήματος ως πρόβλημα ταξινόμησης κειμένων απόψεων στη Μηχανική Μάθηση, διερευνάται κατά κύριο λόγο σε τέσσερα επίπεδα: σε επίπεδο οντότητας (Entity level), σε επίπεδο εγγράφου (Document level), σε επίπεδο πρότασης (Sentence level) και λέξης [28] (Aspect / Feature level), τα οποία εξετάζουμε αναλυτικότερα παρακάτω.



Εικόνα 6: Επίπεδα προσέγγισης κειμένου στην Ανάλυση Συναισθήματος

[27] Τα emoticons είναι ακολουθίες χαρακτήρων που έχουν καθοριστεί να εκφράζουν μια έννοια (π.χ. *:)* σημαίνει *χαρούμενος*).

[28] Ως features θεωρούνται οι διάφορες όψεις ενός θέματος για τις οποίες μπορεί να εκφράζει γνώμη ένα έγγραφο. Συναντώνται στην βιβλιογραφία με τους όρους γνώρισμα, λέξη – κλειδί, χαρακτηριστικό και στόχος. Από εδώ και στο εξής στην παρούσα διπλωματική εργασία θα αναφερόμαστε στην έννοια αυτή με τον όρο λέξη – κλειδί, καθώς οι κριτικές είναι κείμενα με λέξεις – κλειδιά γνωρίσματα.

2.2.5.1 Σε επίπεδο οντότητας

Η επίλυση του προβλήματος σε αυτό το επίπεδο έχει ως στόχο τον εντοπισμό της οντότητας (αντικείμενο αξιολόγησης) για την οποία μιλάει το κείμενο. Στα κείμενα κριτικών που επικεντρωνόμαστε και μελετάμε σε αυτή τη διπλωματική εργασία, όπου η οντότητα είναι το προϊόν που σχολιάζεται και θεωρείται γνωστό άρα δεν χρειάζεται να εντοπιστεί. Έτσι, δεν θα μας απασχολήσει περαιτέρω η Ανάλυση Συναισθήματος σε επίπεδο οντότητας.

2.2.5.2 Σε επίπεδο εγγράφου

Ο στόχος σε αυτό το επίπεδο είναι να ταξινομηθεί ένα ολόκληρο έγγραφο σύμφωνα με το εάν εκφράζει θετική ή αρνητική ψυχολογία. Στηρίζεται στο ότι κάθε έγγραφο εκφράζει απόψεις για μία ενιαία οντότητα (π.χ. προϊόν, υπηρεσία, άτομο, θέμα, γεγονός κ.α.), για το λόγο αυτό δεν μπορεί να εφαρμοστεί σε έγγραφα τα οποία συγκρίνουν πολλαπλές οντότητες. Το πρόβλημα της ταξινόμησης αυτού του επιπέδου όταν εστιάζει σε κριτικές προϊόντων, μπορεί να χωριστεί (1) στην πρόβλεψη κατηγορικών κλάσεων (ή κατηγοριοποίηση), αλλά και (2) στην πρόβλεψη βαθμολογίας αξιολόγησης (Rating Score) του προϊόντος μέσω ανάλυση παλινδρόμησης (Regression Analysis) καθώς κάθε διακριτή τιμή αντιπροσωπεύει διαφορετικό βαθμό αξιολόγησης (π.χ. 0 έως 5 αστεράκια), όπως στο Pang κ.α. (2005), το οποίο είναι συνεχές σύνολο.

Οι περισσότερο χρησιμοποιούμενες τεχνικές επίλυσης με Μηχανική Μάθηση και των δύο παραπάνω υπο - προβλημάτων πρόβλεψης χρησιμοποιούν μεθόδους εποπτευόμενης μάθησης (Supervised Learning) για τις οποίες γίνεται αναλυτικότερη αναφορά παρακάτω. Οι πιο σύγχρονες μέθοδοι προτείνουν cross - domain ανάλυση, οι οποίες διαμορφώνονται σύμφωνα με τον τομέα από όπου προέρχονται τα δεδομένα προς ταξινόμηση και cross - language ανάλυση που διαμορφώνονται σύμφωνα με τη γλώσσα των δεδομένων αντίστοιχα. Ως επακόλουθο της προαναφερθείσας επεξήγησης, ούτε αυτό το επίπεδο εξετάζεται βαθύτερα σε αυτή τη διπλωματική εργασία καθώς ο στόχος μας είναι η εξόρυξη γνώμης για το κάθε ξεχωριστό χαρακτηριστικό γνώρισμα του προϊόντος που περιέχει η εκάστοτε κριτική, μη λαμβάνοντας υπόψη τον απώτερο σκοπό μας που είναι η βαθμολόγηση της οντότητας «προϊόν».

2.2.5.3 Σε επίπεδο πρότασης

Ο στόχος σε αυτό το επίπεδο είναι να καθοριστεί εάν κάθε πρόταση εκφράζει θετική, αρνητική, ουδέτερη (συμπεριλαμβάνει και την μη ύπαρξη γνώμης) άποψη ή και περισσότερες. Ωστόσο, η γνώση μόνο για το εάν μια πρόταση έχει θετικό ή αρνητικό προσανατολισμό και όχι οι λέξεις – κλειδιά της, δεν είναι πολύ χρήσιμη, παρά μόνο στις περιπτώσεις όπου γνωρίζουμε το συναίσθημα των λέξεων - κλειδιών αλλά και ποιες λέξεις -

κλειδιά αναφέρονται στην πρόταση και ως επακόλουθο αναγνωρίζουμε το συναίσθημα ολόκληρης της πρότασης. Η βασική διαφορά της με την ανάλυση του συναισθήματος σε επίπεδο εγγράφου έγκειται στο ότι η πρόταση εκφράζει ένα μόνο συναίσθημα (πράγμα όχι πάντα αληθές όπως θα δούμε παρακάτω) ενώ το έγγραφο περισσότερα.

Η ανάλυση αυτού του επίπεδου είναι στενά συνδεδεμένη με την **ταξινόμηση υποκειμενικότητας** (Subjectivity Classification) και επιχειρεί να διακρίνει τις αντικειμενικές από τις υποκειμενικές προτάσεις, όπως οι Hatzivassiloglou και Wiebe (2000), Pang κ.α. (2002), Riloff και Wiebe (2003), Yu και Hatzivassiloglou (2003), Wiebe κ.α. (2004), Wilson κ.α. (2006), Riloff κ.α. (2006). Συνεπώς υπάρχει πληθώρα δυνατών τρόπων προσέγγισης του προβλήματος της κατηγοριοποίησης υποκειμενικότητας. Οι Hatzivassiloglou και Wiebe (2000) διαθέτοντας ένα σύνολο από δείγματα προτάσεων επισημασμένα ως υποκειμενικά ή αντικειμενικά κατηγοριοποιούν ένα άγνωστο δείγμα μετρώντας τη μέση ομοιότητά του με τα γνωστά δείγματα, βάσει των κοινών τους λεκτικών γνωρισμάτων. Οι Riloff κ.α. (2006) από την άλλη, εξάγουν με διαδικασίες μάθησης ένα σύνολο από χαρακτηριστικά εκφραστικά μοτίβα που παρουσιάζουν υψηλή στατιστική συσχέτιση με τα αντικειμενικά ή τα υποκειμενικά δείγματα αντίστοιχα. Οι Pang κ.α. (2002) εκμεταλλεύονται τις σχέσεις γειτόνευσης των προτάσεων με βάση την υπόθεση ότι γειτονικές προτάσεις συχνά έχουν τον ίδιο υποκειμενικό χαρακτήρα. Τα αρχικά δείγματα στα οποία στηρίζονται όλες οι παραπάνω τεχνικές μπορούν να συλλεχθούν με μη εποπτευόμενες τεχνικές, εκμεταλλευόμενοι εκ των προτέρων γνωστές ενδείξεις υποκειμενικότητας (π.χ. παρουσία ικανού αριθμού συναισθηματικά φορτισμένων όρων) και αντικειμενικότητας (π.χ. παθητική σύνταξη).

Η διάκριση των αντικειμενικών από τις υποκειμενικές προτάσεις ενέχει έναν ορισμένο βαθμό σχετικότητας καθώς τα χαρακτηριστικά που καθιστούν μια πρόταση αντικειμενική ή υποκειμενική είναι δύσκολο να προσδιοριστούν, ειδικά σε κείμενα κριτικών όπου συχνά κάνουν την εμφάνισή τους προτάσεις αντικειμενικές ως προς τη μορφή αλλά εν δυνάμει υποκειμενικές ως προς τη διάθεση. Συνεπώς η ταξινόμηση υποκειμενικότητας δεν αντιστοιχεί πάντα με το συναίσθημα, καθώς πολλές αντικειμενικές προτάσεις μπορεί να υποδηλώνουν γνώμη. Για παράδειγμα η πρόταση «*2 ημέρες αφ' ότου το αγόρασα, σταμάτησε να λειτουργεί*» αν και περιγράφει ένα γεγονός εκφράζει δυσαρέσκεια. Επομένως, θα ήταν προτιμότερο σε αυτό το σημείο η ταξινόμηση να έγκειται στο εάν η πρόταση εκφράζει γνώμη ή όχι και όχι εάν είναι αντικειμενική ή υποκειμενική, ώστε εάν εκφράζει να ακολουθήσει η ταξινόμησή της σε θετική ή αρνητική. Άλλωστε έρευνες όπως των Koppel και Schler (2006) έχουν δείξει ότι οι ταξινομητές βελτιώνουν τις προβλέψεις τους συμπεριλαμβάνοντας την ουδέτερη κλάση (μη ύπαρξη γνώμης).

Η ανάλυση στο επίπεδο πρότασης μπορεί να γίνει ακόμη πιο προκλητική διαδικασία, όπως επισημαίνει και ο Liu (2012), εάν ληφθούν υπόψιν υποθετικές, σαρκαστικές, συγκριτικές

προτάσεις και παρομοιώσεις οι οποίες δεν εκφράζουν ξεκάθαρο συναίσθημα αλλά το υπονοούν. Οι υποθετικές προτάσεις περιγράφουν κυρίως συνέπειες υποθετικών καταστάσεων, ενώ στις σαρκαστικές οι συγγραφείς γράφουν το αντίθετο από αυτό που εννοούν. Οι παρομοιώσεις και οι συγκριτικές προτάσεις δε, μεταφέρουν έμμεσες γνώμες και συνήθως δεν είναι εφικτό να τις αποτιμήσουμε μόνο με βάση τη γλωσσική πληροφορία που είναι ενσωματωμένη στην πρόταση. Έτσι εύκολα γίνεται κατανοητή η δυσκολία διαχείρισης όλων αυτών των προτάσεων.

Συμπερασματικά και εστιάζοντας στο πρόβλημα της παρούσας διπλωματικής εργασίας, η Ανάλυση Συναισθήματος σε επίπεδο προτάσεων έρχεται πιο κοντά στο συναίσθημα που επιχειρούμε να εξάγουμε για τα χαρακτηριστικά γνωρίσματα του προϊόντος από την κάθε κριτική, έχει πολλές αδυναμίες όταν εφαρμόζεται, όπως η έλλειψη γνώσης για τον προσανατολισμό του συναισθήματος για τις μεμονωμένες λέξεις - κλειδιά της πρότασης. Για παράδειγμα, όταν εξετάζεται ποιά είναι τα χαρακτηριστικά γνωρίσματα του προϊόντος που αναφέρονται σε μια πρόταση κριτικής του και τι άποψη εκφράζεται για το καθένα, καθώς μια πρόταση μπορεί να συνθέτει διαφορετικά συναισθήματα εάν περιέχει πολλές λέξεις - κλειδιά. Επιπρόσθετα, δύσκολα μπορεί να αντιμετωπίσει απόψεις σε προτάσεις που εμπεριέχουν σύγκριση (π.χ. «*Το iPad είναι καλύτερο του Samsung Galaxy Tab*»), καθώς η επίλυση στο πρόβλημα εδώ δεν αρκεί απλά να ταξινομήσει την πρόταση σε θετική, αρνητική ή ουδέτερη.

2.2.5.4 Σε επίπεδο λέξης

Η ανάλυση αυτού του επιπέδου εστιάζει απευθείας στην ίδια την λέξη – κλειδί της πρότασης, αυτό που μας ενδιαφέρει να εξετάσουμε. Βασίζεται στην ιδέα ότι η γνωμοδότηση αποτελείται από ένα συναίσθημα και μια λέξη – κλειδί, ή στόχος γνώμης (Opinion Target) όπως συχνά συναντάμε τον όρο, και επικεντρώνεται στο να ανακαλύψει τα συναισθήματα των ξεχωριστών αυτών στόχων. Για παράδειγμα, στην πρόταση *«Αν και η ανάλυση της οθόνης είναι εξαιρετική, η διάρκεια της μπαταρίας του είναι απογοητευτική»* οι στόχοι γνώμης είναι η *οθόνη* και η *μπαταρία* και εκφράζονται με διαφορετικό συναίσθημα. Ακόμη κι αν θεωρήσουμε ότι κάθε έγγραφο αξιολογεί μια μεμονωμένη οντότητα, δεν μπορούμε να εγγυηθούμε ότι η γενική άποψη για την οντότητα έγκειται και στις εκάστοτε λέξεις – κλειδιά που εμπεριέχει η κριτική της. Για παράδειγμα, μια αρνητική γνώμη για ένα συγκεκριμένο laptop μέσω μιας κριτικής δεν σημαίνει ότι αυτός που υπέβαλλε την κριτική είναι αρνητικός για κάθε χαρακτηριστικό γνώρισμα του συγκεκριμένου laptop. Η εξέταση θα πρέπει να μεταφερθεί σε επίπεδο λέξης, η οποία υλοποιείται με την τεχνική Ανάλυσης Συναισθήματος βασισμένη στις λέξεις - κλειδιά (Aspect - Based Sentiment Analysis) ή αλλιώς, με ορολογία Μηχανικής Μάθησης, εξόρυξη γνώμης βασισμένη στα χαρακτηριστικά (Feature - Based Opinion Mining) όπως εισήγαν τον όρο οι Hu και Liu (2004), η οποία είναι μια από τις

τεχνικές που μας απασχολεί σε μεγάλο βαθμό στην παρούσα διπλωματική εργασία, εξετάζοντας πληθώρα εργασιών που την χρησιμοποιούν, καθώς ο αρχικός μας στόχος είναι η εξόρυξη γνώμης για τα χαρακτηριστικά γνωρίσματα του προϊόντος στη κάθε κριτική, ο οποίος προσεγγίζεται με την ανάλυση του κειμένου σε αυτό το επίπεδο.

2.2.6 Σύνολα δεδομένων για Ανάλυση Συναισθήματος

Η επίλυση στο πρόβλημα της Ανάλυσης Συναισθήματος με Μηχανική Μάθηση (εποπτευόμενη ή μη) απαιτεί μεγάλο όγκο δεδομένων ώστε να επιτευχθούν όσο το δυνατό καλύτερα αποτελέσματα από τους αλγορίθμους, επομένως η διάθεση έτοιμων συνόλων δεδομένων (datasets) παρέχει σημαντική διευκόλυνση στο έργο τους. Υπάρχει πλέον διαθέσιμη πληθώρα συνόλων δεδομένων ανά τομέα, η χρήση των οποίων μειώνει σε μεγάλο βαθμό τον χρόνο προετοιμασίας ενός συστήματος Ανάλυσης Συναισθήματος αφού παρακάμπτεται η χρονοβόρα και επίπονη διαδικασία δημιουργίας του απαραίτητου συνόλου δεδομένων για την εκπαίδευση των μοντέλων Μηχανικής Μάθησης που θα χρησιμοποιηθούν. Παράλληλα δε, η κοινή χρήση τους από πολλούς ερευνητές σε διάφορους τομείς εφαρμογής οδηγεί στην συνεχή βελτίωσή τους, συνεπώς προτείνονται. Μία συνοπτική αναφορά σε μερικά από τα πιο διαδεδομένα που έχουν χρησιμοποιηθεί μέχρι σήμερα γίνεται κάτωθεν:

- **Pang & Lee dataset**, μια συλλογή 1.000 αρνητικών και 1.000 θετικών κριτικών ταινιών η οποία δημιουργήθηκε από τους Pang και Lee (2004) [29].
- **Multi-domain sentiment dataset**, μια συλλογή που περιέχει κριτικές προϊόντων που έχουν ληφθεί από το Amazon.com από διάφορους τομείς προϊόντων, το πλήθος των οποίων ποικίλει ανά κατηγορία. Οι κριτικές περιέχουν αξιολογήσεις αστεριών (1 έως 5 αστέρια) που μπορούν να μετατραπούν σε δυαδικές ετικέτες αν χρειαστεί. Δημιουργήθηκε από τους Blitzer κ.α. (2007) [30].
- **IMDB11 dataset**, μια μεγάλη συλλογή 50.000 κριτικών ταινιών η οποία περιέχει ένα σύνολο εκπαίδευσης 25.000 επισημασμένων κριτικών και ένα σύνολο δοκιμών 25.000 επισημασμένων κριτικών, με 12.500 θετικές και 12.500 αρνητικές κριτικές στο κάθε σύνολο. Δημιουργήθηκε από τους Maas κ.α. (2011) [31].
- **MPQA Opinion Corpus**, μια συλλογή από ειδησεογραφικά άρθρα ποικίλων πηγών που έχουν σχολιαστεί ως προς την άποψη, τις πεποιθήσεις, το συναίσθημα και τις εικασίες που εκφράζουν. Περιγράφεται στο Deng και Wiebe (2015) [32].

[29] Διαθέσιμη από: http://www.cs.cornell.edu/people/pabo/movie-review-data/
[30] Διαθέσιμη από: http://www.cs.jhu.edu/~mdredze/datasets/sentiment/
[31] Διαθέσιμη από: http://ai.stanford.edu/~amaas/data/sentiment/
[32] Διαθέσιμη από: http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/

- **International Survey on Emotion Antecedents and Reactions (ISEAR) corpus**, είναι μια συλλογή εκθέσεων φοιτητών σχετικά με καταστάσεις στις οποίες οι ερωτηθέντες αισθάνθηκαν κάποιο από τα επτά βασικά συναισθήματα: «χαρά», «θλίψη», «θυμός», «αηδία», «φόβος», «αηδία», «ντροπή» και «ενοχή». Οι απαντήσεις περιλαμβάνουν περιγραφές του τρόπου με τον οποίο εξέτασαν την κατάσταση και πώς αντιδρούν, όπως εξηγείται στο Scherer και Wallbott (1994) [33].
- **EmotiBlog corpus**, μια συλλογή από αναρτήσεις σε ιστολόγια (blogs) που δημιουργήθηκαν και σχολιάστηκαν για την ανίχνευση υποκειμενικών λέξεων, φράσεων και προτάσεων στα νέα κείμενα που γεννήθηκαν με το Web 2.0, από τους Boldrini κ.α. (2011).
- **Affective Text Corpus** μια συλλογή που αποτελείται από τίτλους ειδήσεων που προέρχονται από μεγάλες εφημερίδες όπως οι New York Times, το CNN και το BBC News, καθώς και από την μηχανή αναζήτησης εφημερίδων Google News. Περιέχει δύο σύνολα δεδομένων, το πρώτο με 1.000 τίτλους για δοκιμές και το δεύτερο με 200 τίτλους για ανάπτυξη, ο καθένας από τους οποίους είναι επισημασμένος με ένα από τα έξι συναισθήματα «χαρά», «θλίψη», «έκπληξη», «θυμός», «αηδία», «φόβος» και τον προσανατολισμό (σθένος) της πολικότητας. Χρησιμοποιήθηκε στο SemEval 2007 στο Task 14: Affective Text και αναπτύχθηκε από τους Strapparava και Mihalcea (2007) [34].

2.2.7 Ανάλυση Δεδομένων

Τα δεδομένα που αναλύονται για το συναίσθημά τους στη τρέχουσα διπλωματική εργασία είναι **κριτικές** που δημοσιεύονται σε ηλεκτρονικές πλατφόρμες για κάποιο προϊόν. Ανάλογα με το στόχο που εξυπηρετούν κάθε φορά, μπορούμε να μιλήσουμε για καλοπροαίρετες και για κακοπροαίρετες κριτικές. Μια κριτική μπορεί να συνιστά αποδοχή και επιδοκιμασία του αντικειμένου αξιολόγησης (θετική κριτική), οπότε χρησιμοποιεί ως μέσο τον έπαινο, ή να αποτελεί άρνηση, στηλίτευσή του (αρνητική κριτική), οπότε το μέσο της είναι ο στιγματισμός και η αποδοκιμασία.

Οι απόψεις των κειμένων των κριτικών μπορούν επίσης να χωριστούν, σύμφωνα με τους Jindal και Liu (2006b), σε κανονικές (regular), οι οποίες εκφράζουν ένα συναίσθημα για μια συγκεκριμένη οντότητα και συγκριτικές (comparative), οι οποίες εκφράζουν τη σχέση ομοιότητας ή διαφορετικότητας δύο ή περισσότερων οντοτήτων συνήθως με τη συγκριτική ή υπερθετική μορφή επιθέτων ή επιρρημάτων. Μερικές προτεινόμενες τεχνικές για τον εντοπισμό της **σύγκρισης**, σε επίπεδο προτάσεων είναι η εξόρυξη κανόνων με επισήμανση

[33] Διαθέσιμη από: http://www.affective-sciences.org/home/research/materials-and-online-research/research-material/

[34] Διαθέσιμη από: http://web.eecs.umich.edu/~mihalcea/downloads.html#affective

προτάσεων (Label Sequential Rule Mining), ενώ σε επίπεδο λέξης η χρήση των αλγορίθμων της Μέγιστης Εντροπίας (Maximum Entropy) και των Διανυσματικών Μηχανών Υποστήριξης (Support Vector Machines - SVM), επισημαίνοντας ότι και οι τρεις είναι μέθοδοι εποπτευόμενης μάθησης.

Μια ακόμη διάκριση των απόψεων με βάση τον τρόπο που εκφράζονται μέσα στο κείμενο, όπως εντοπίζεται στο Liu (2012), είναι η εξής:

- άμεση (explicit): μια υποκειμενική δήλωση που δίνει μια κανονική ή συγκριτική γνώμη όπως αναφέρουν οι Zhang και Liu (2011b), εκφράζοντας ξεκάθαρα θετικό ή αρνητικό συναίσθημα. Για παράδειγμα, «*Το HP laptop που αγόρασα είναι το καλύτερο όλων*».
- έμμεση (implicit): μια αντικειμενική δήλωση που υποδηλώνει μια κανονική ή συγκριτική γνώμη εκφράζοντας συνήθως ένα επιθυμητό ή ανεπιθύμητο γεγονός όπως αναφέρουν οι Greene και Resnik (2009), υπονοώντας το συναίσθημα. Για παράδειγμα, στην πρόταση «*Αυτό το tablet είναι ακριβό*», η λέξη «*ακριβό*» είναι η λέξη που υποδηλώνει τόσο το συναίσθημα όσο και το χαρακτηριστικό γνώρισμα «*τιμή*».

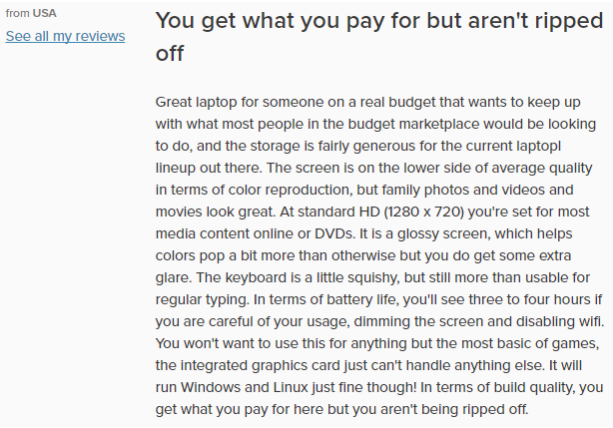
Οι κριτικές είναι πιθανό να εντοπιστούν με τις παρακάτω μορφές χωρίς απαραίτητα να συναντώνται μεμονωμένες και ανεξάρτητες, αλλά μπορεί να συνδυάζονται, ακόμη και να εντοπίζονται όλες μαζί στο εκάστοτε αντικείμενο αξιολόγησης (οντότητα / προϊόν):

- Βαθμολογία ή «αστεράκια» (Rating) με τα οποία έχει χαρακτηρίσει το προϊόν (ή μεμονωμένο χαρακτηριστικό γνώρισμά του) κάθε σχολιαστής και είναι σε κανονικές τιμές (Ordinal). Συνήθως οι τιμές τους είναι από το 1 εως και το 5 ακέραιοι (integer) και ο μέγιστος αριθμός καθορίζεται από τις απαιτήσεις του προβλήματος. Ένα παράδειγμα επεξήγησης της βαθμολογίας 5 επιπέδων φαίνεται στον Πίνακας 1.

Πίνακας 2: Σημασία Επιπέδων Βαθμολογίας (rating)

Επίπεδα Βαθμολογίας	**Επεξήγηση**
★	Το απεχθάνομαι
★★	Δεν μου αρέσει
★★★	Είναι εντάξει
★★★★	Μου αρέσει
★★★★★	Το αγαπώ

- Ελεύθερο κείμενο. Είναι η πιο διαδεδομένη μορφή συγγραφής κριτικών που συναντάμε ηλεκτρονικά, δίνοντας στον σχολιαστή την δυνατότητα να γράψει ελεύθερα την γνώμη του για το εκάστοτε προϊόν (ή μεμονωμένο χαρακτηριστικό γνώρισμά του), χωρίς καμία άλλη απαίτηση. Ένα παράδειγμα κριτικής ελεύθερου κειμένου από την ιστοσελίδα www.viewpoints.com φαίνεται στην Εικόνα 7.

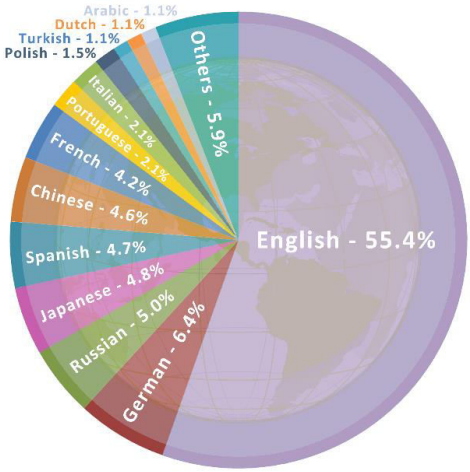

from USA
See all my reviews

You get what you pay for but aren't ripped off

Great laptop for someone on a real budget that wants to keep up with what most people in the budget marketplace would be looking to do, and the storage is fairly generous for the current laptopl lineup out there. The screen is on the lower side of average quality in terms of color reproduction, but family photos and videos and movies look great. At standard HD (1280 x 720) you're set for most media content online or DVDs. It is a glossy screen, which helps colors pop a bit more than otherwise but you do get some extra glare. The keyboard is a little squishy, but still more than usable for regular typing. In terms of battery life, you'll see three to four hours if you are careful of your usage, dimming the screen and disabling wifi. You won't want to use this for anything but the most basic of games, the integrated graphics card just can't handle anything else. It will run Windows and Linux just fine though! In terms of build quality, you get what you pay for here but you aren't being ripped off.

Εικόνα 7: Παράδειγμα κριτικής ελεύθερου κειμένου

Τα ελεύθερα κείμενα των κριτικών που εξετάζουμε επιλέχθηκε να είναι γραμμένα στην **αγγλική γλώσσα**, δεδομένου ότι είναι η πιο ευρέως χρησιμοποιούμενη γλώσσα παγκοσμίως και τήνει να γίνει η κοινή γλώσσα επικοινωνίας ανά τον κόσμο, κυρίως μέσω της εξάπλωσης του διαδικτύου, όπως ξεκάθαρα φαίνεται και στην Εικόνα 8. Επιπρόσθετα με την επιλογή αυτή, τα τυχόν συστήματα που θα αναπτυχθούν με τις μεθόδους που αναλύουμε και προτείνουμε για την αγγλική γλώσσα, θα χαρακτηρίζονται με μεγαλύτερο βαθμό γενίκευσης. Ένας ακόμη λόγος αυτής της απόφασης άρα και απόρριψης της εναλλακτικής μας επιλογής για την ανάλυση της ελληνικής γλώσσας είναι η εκθετική δυσκολία και αντιξοότητα της ικανοποιητικής διαχείρισης της απαιτητικής γραμματικής και συντακτικού της ελληνική γλώσσας. Όπως διαπιστώθηκε από την έρευνα μας πάνω στο κομμάτι της Ανάλυσης Συναισθήματος, η διαχείριση της ελληνικής γλώσσας είναι ακόμη σε πολύ πρώιμο στάδιο και θα ήταν ακόμη πιο δύσκολη και χρονοβόρα διαδικασία η έρευνα ανάλυσης αυτής από ότι της αγγλικής γλώσσας. Αυτό το προκλητικό πρόβλημα, αν και μας κίνησε το ενδιαφέρον δεν μπορούσε να συμπεριληφθεί στα πλαίσια αυτής της διπλωματικής εργασίας λόγω της μεγάλης του έκτασης, έτσι το αφήνουμε για μελλοντική και μεμονωμένη έρευνα.



Εικόνα 8: Ποσοστό χρήσης διαφόρων γλωσσών σε περιεχόμενο που βρίσκεται στους 1.000.000 πιο κορυφαίους ιστότοπους παγκοσμίως [10]

2.3 Ανάλυση Συναισθήματος με Μηχανική Μάθηση

2.3.1 Εξόρυξη Γνώμης

Όπως έχουμε ήδη προαναφέρει, εάν το πρόβλημα της Ανάλυσης Συναισθήματος προσεγγιστεί με μεθόδους Μηχανικής Μάθησης τότε μιλάμε για εξόρυξη γνώμης (ή κατηγοριοποίηση γνώμης ή ταξινόμηση κειμένου). Ο στόχος της εξόρυξης γνώμης είναι να κατατάξει αυτόματα ένα έγγραφο ή απόσπασμα εγγράφου σε μία κλάση συναισθήματος, ανάλογα με το είδος της γνώμης που εκφράζει (αναγνώριση στοιχείων υποκειμενικότητας). Ωστόσο, δεν είναι όλες οι γνώμες το ίδιο έντονες, και σε αρκετές περιπτώσεις έχει νόημα να υιοθετηθεί ένα σχήμα κατηγοριοποίησης που θα τις διακρίνει όχι μόνο ως προς τον προσανατολισμό τους, δηλαδή σε θετικές και αρνητικές, αλλά και ως προς την ένταση, δηλαδή ως ισχυρές ή ασθενείς.

Κατά την διαδικασία της εξόρυξης γνώμης σε ένα απόσπασμα κειμένου κριτικής, ως σημείο αφετηρίας μπορεί να χαρακτηριστεί η αναγνώριση των υποκειμενικών λέξεων της πρότασής της. Κάθε τέτοια λέξη χαρακτηρίζεται από ένα πρότερο συναίσθημα (prior sentiment) που σχετίζεται με το εννοιολογικό περιεχόμενό της. Η πληροφορία του πρότερου συναισθήματος αυτών των λέξεων συνήθως αντλείται μέσα από λεξικά γνώμης, είτε χρησιμοποιώντας έτοιμα προκατασκευασμένα, είτε ιδιοπαραγόμενο μέσα από διαδικασίες μάθησης, εποπτευόμενες ή μη, αξιοποιώντας είτε σύνολα δεδομένων με κείμενα (corpus), είτε γλωσσολογικές βάσεις δεδομένων (dictionaries / linguistic databases), είτε συνδυαστικά και τα δύο. Ακόμα όμως και αν εξασφαλιστεί η διαθεσιμότητα ενός κατάλληλου λεξικού γνώμης, υπάρχουν πολλές ακόμη προκλήσεις που πρέπει να αντιμετωπιστούν προκειμένου να μεταβούμε από το πρότερο συναίσθημα των επιμέρους λέξεων - κλειδιών της πρότασης, στο συνολικό της κριτικής. Οι προκλήσεις στην ανάλυση στο επίπεδο πρότασης, όπως έχουμε προαναφέρει, έγκειται στο εάν ληφθούν υπόψιν υποθετικές, σαρκαστικές, παρομοιώσεις και συγκριτικές προτάσεις οι οποίες δεν εκφράζουν ξεκάθαρο συναίσθημα αλλά το υπονοούν.

Το πρόβλημα της Ανάλυσης Συναισθήματος με Μηχανική Μάθηση για κάθε μορφής κριτική, μπορεί να προσεγγιστεί:

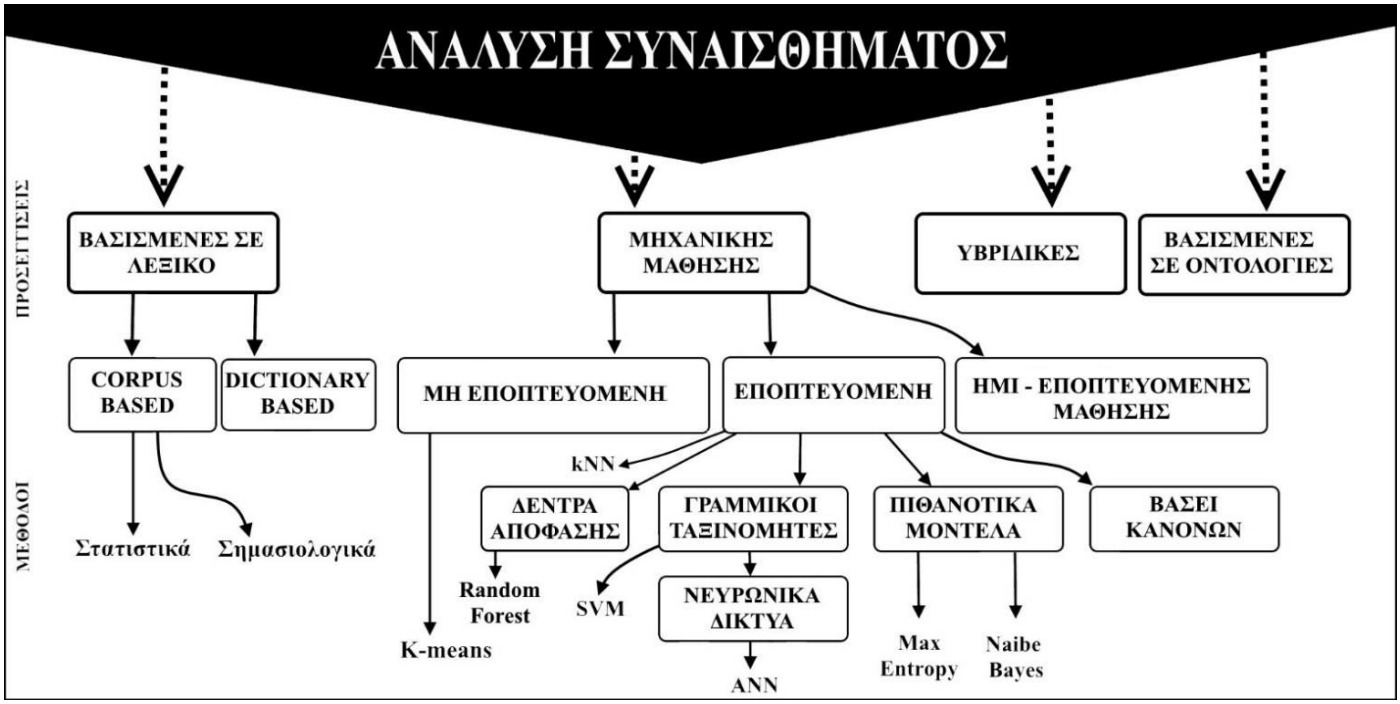
1. με **εποπτευόμενες μεθόδους,** που είναι δύσκολα προσαρμόσιμες και εντοπίζονται κυρίως σε αναλύσεις σε επίπεδο εγγράφου και οντότητας. Δεν έχουν την απαίτηση κάποιου λεξικού γνώμης και αποδίδουν υψηλή ακρίβεια στην ταξινόμηση. Εντούτοις, τις περισσότερες φορές είναι εξαρτημένες με τον τομέα στον οποίο έχουν εκπαιδευτεί και χωλαίνουν σε νέους άγνωστους τομείς.
2. με ευέλικτες μεθόδους **μη εποπτευόμενης μάθησης**, οι οποίες δεν απαιτούν επισημασμένα δεδομένα καθώς η διαδικασία μάθησης λείπει. Στην κατηγορία αυτή

μπορούν να ενταχθούν και τα μοντέλα που βασίζονται στην **χρήση λεξικών** γνώμης (Lexicon-based), αλλά δεν πρέπει να ταυτιστούν με την Μηχανική Μάθηση, καθώς δεν ανήκουν στον τομέα αυτό. Αυτού του είδους οι προσεγγίσεις απαιτούν ισχυρές γλωσσικές πηγές που αποδίδουν μια βαθμολογία πολικότητας σε κάθε όρο που είναι δύσκολα διαθέσιμες.

3. με μεθόδους **βασισμένες σε κανόνες** (Rule-based), η αποτελεσματικότητα και η ακρίβεια των οποίων εξαρτώνται από τους κανόνες που θα καθορίσουν. Εντοπίζονται πιο αποδοτικοί σε ακρίβεια στις αναλύσεις επιπέδου οντότητας.
4. με μεθόδους **βασισμένες στις οντολογίες** (Ontology-based), όπου δημιουργείται ένα μοντέλο το οποίο απεικονίζει τις λέξεις - κλειδιά ως οντολογίες με κλάση, υποκλάση, αντικείμενο και ιδιότητες αντικειμένου.
5. με **υβριδικές** μεθόδους, που είναι ένας συνδυασμός των παραπάνω μεθόδων, εκμεταλλευόμενες τα καλύτερα χαρακτηριστικά κάθε επιμέρους αλγορίθμου, που όπως είναι αναμενόμενο αποδίδουν καλύτερα καθώς είναι από τις πιο σύγχρονες στη βιβλιογραφία και έχουν τις περισσότερες αναφορές [35].

Υποστηρίζοντας το ρητό «*Μια εικόνα είναι ίση με χίλιες λέξεις*», οι προσεγγίσεις που ειπώθηκαν παραπάνω, με μια διαφοροποιημένη κατηγοριοποίηση και συμπεριλαμβανομένων και των μεθόδων που χρησιμοποιεί η κάθε μία για το έργο που έχει τεθεί να επιλύσει πάνω στην Ανάλυση Συναισθήματος, απεικονίζονται συνολικά στην Εικόνα 9 με σκοπό την συλλογική τους κατανόηση.

[35] Οι αναφορές (citations) σε μια δημοσιευμένη ή μη πηγή δηλώνει μια εγγραφή στο τμήμα βιβλιογραφικών αναφορών του έργου με σκοπό να αναγνωρίσει τη συνάφεια των έργων άλλων με το θέμα της συζήτησης στο σημείο όπου εμφανίζεται η παραπομπή. Συνήθως, όσο μεγαλύτερος είναι ο αριθμός αναφορών μιας δημοσίευσης τόσο πιο αξιόπιστη θεωρείται.



Εικόνα 9: Προσεγγίσεις & Μέθοδοι Ανάλυσης Συναισθήματος

2.3.2 Αλγόριθμοι Μηχανικής Μάθησης για Ανάλυση Συναισθήματος

Ορισμένοι ενδεικτικοί αλγόριθμοι Μηχανικής Μάθησης που εντοπίστηκαν να χρησιμοποιούνται για την Ανάλυση Συναισθήματος, σε κείμενα κριτικών αναφέρονται παρακάτω, καλύπτοντας έτσι το βασικό θεωρητικό υπόβαθρο, χωρίς εκτενής εμβάθυνση στον τρόπο λειτουργίας τους, απλά γνωστοποιώντας τις επιλογές που υπάρχουν για την επίλυση του εν λόγω προβλήματος.

2.3.2.1 Εποπτευόμενης μάθησης

- **Γλωσσικά Μοντέλα (Language Modeling)**: μια παραγωγική μέθοδος που αναθέτει σε κάθε λέξη των προτάσεων την πιθανότητα αυτή να εμφανιστεί σε ένα κείμενο της κατηγορίας των κειμένων για την οποία έχει εκπαιδευτεί το μοντέλο χρησιμοποιώντας τα σύνολα δεδομένων εκπαίδευσης. Συνεπώς καταλαβαίνουμε ότι η μέθοδος αυτή είναι εποπτευόμενης μάθησης. Στην n - gram (n - διαδοχικές λέξεις) γλωσσική μοντελοποίηση, η πιθανότητα της λέξης υπολογίζεται από το γινόμενο των εξαρτώμενων πιθανοτήτων των n - grams, των οποίων οι πιθανότητες εξαρτώνται από τις προηγούμενες *n-1* λέξεις, όπως επισημαίνουν οι Cui κ.α. (2006). Οι Pang κ.α. (2002) και οι Andreevskaia και Bergler (2008) που επίσης πειραματίζονται με μοντέλα εκπαίδευσης βασισμένα σε n – grams, παρατηρούν ότι τα n - gram μοντέλα διογκώνουν αχρείαστα τον χώρο χαρακτηριστικών του προβλήματος και μειώνουν την ακρίβεια.
- **Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machine - SVM)** : μια μέθοδος εποπτευόμενης μάθησης που χρησιμοποιείται για κατηγοριοποίηση τόσο γραμμικών όσο

και μη γραμμικών δεδομένων σε ένα χώρο πολλών διαστάσεων. Στη βασική του μορφή προβλέπει για κάθε νέο στοιχείο εισόδου, σε ποια από τις δυο προκαθορισμένες κλάσεις θα τοποθετηθεί, ενώ με τη μέθοδο του πυρήνα (Kernel Method) μπορεί να εύκολα να επεκταθεί στην ταξινόμηση σε περισσότερες των δύο κλάσεων, μετασχηματίζοντας τα μη γραμμικά δεδομένα σε ένα χώρο περισσότερων διαστάσεων στον οποίο είναι γραμμικά διαχωρίσιμα. Είναι ένας μη - πιθανοτικός ταξινομητής (δηλαδή δεν υπολογίζει πιθανότητες) και αυτό ίσως είναι και το μόνο μειονέκτημά του, το οποίο ξεπερνιέται από τα πολύ σημαντικά θετικά του στοιχεία. Οι SVM είναι αποτελεσματικές σε χώρους πολλών διαστάσεων ακόμη και όταν ο αριθμός των χαρακτηριστικών είναι μεγαλύτερος του αριθμού των δειγμάτων, αποφεύγοντας το πρόβλημα της υπερ – προσαρμογής (Overfitting) [36]. Επιπρόσθετα, η χρήση του υποσύνολο των παραδειγμάτων εκπαίδευσης για την κατασκευή του συνόλου δοκιμών / απόφασης μειώνει κατά πολύ την χρήση απαιτούμενης μνήμης και συνεπώς τις κάνουν αποδοτικότερες.

Σε μια πιθανή περίπτωση ταξινόμησης με SVM σε δύο κλάσεις το σενάριο έχει ως εξής:
το σύνολο των στοιχείων αποτελείται από δύο υποσύνολα, έστω k και m, και το αποτέλεσμα της συνάρτησης είναι $+1$ ή -1 (για $y_i = +1$ ή $y_i = -1$) ανάλογα το υποσύνολο όπου ανήκει το δοθέν στοιχείο x_i. Τα δύο αυτά υποσύνολα ονομάζονται κλάσεις και η τιμές $+1$ και -1 είναι η «ετικέτα» της κλάσης. Τα στοιχεία x_i και οι αντίστοιχες τιμές τους, y_i, αποτελούν το σύνολο εκπαίδευσης (training set). Τα στοιχεία x_i ονομάζονται πρότυπα εκπαίδευσης (training patterns) ενώ οι τιμές y_i που αντιστοιχούν σε αυτά, στόχοι εκπαίδευσης (training targets).

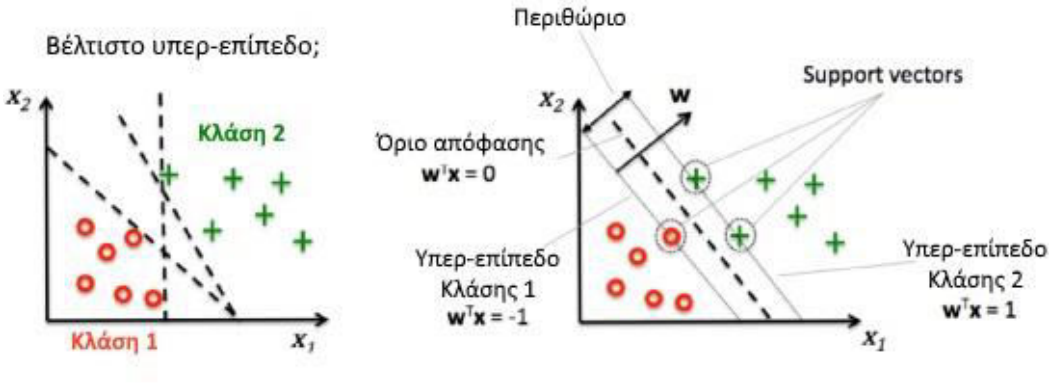
Η βασική ιδέα του αλγόριθμου SVM που στηρίζεται στη θεωρία της Στατιστικής Μάθησης (Statistical Learning Theory), είναι να δημιουργήσει ένα υπερ - επίπεδο διαχωρισμού με τέτοιο τρόπο ώστε να βρίσκεται στη μεγαλύτερη δυνατή απόσταση από τα κοντινότερα σημεία εκπαίδευσης των δύο κλάσεων, όπως φαίνεται και στην Εικόνα 10. Στις περιπτώσεις όπου οι ταξινομητές δε διαχωρίζουν απλά τα δεδομένα, αλλά επιχειρούν να τα διαχωρίσουν με τέτοιο τρόπο ώστε να συμπεριφέρονται καλύτερα σε νέα δεδομένα βασίζεται στην αρχή της απόδοσης γενίκευσης (Generalization Performance). Τα πλησιέστερα σημεία στο υπερ - επίπεδο, από τα οποία η απόσταση θέλουμε να είναι μέγιστη ονομάζονται διανύσματα υποστήριξης (Support Vectors). Παρότι η εκπαίδευση μπορεί να είναι αργή, η ακρίβεια είναι αρκετά υψηλή χάρη στην ικανότητα μοντελοποίησης σύνθετων, μη γραμμικών ορίων απόφασης. Εστιάζοντας στο πρόβλημα της Ανάλυσης Συναισθήματος σε κριτικές προϊόντων που επιδιώκουμε να

[36] Ως υπερ - προσαρμογή (Overfitting) χαρακτηρίζεται η παραγωγή μιας ανάλυσης που αντιστοιχεί σε ένα συγκεκριμένο σύνολο δεδομένων και μπορεί να αποτύχει να προσαρμόσει πρόσθετα δεδομένα ή να προβλέψει αξιόπιστα μελλοντικούς στόχους. Ένα υπερ - προσαρμοσμένο μοντέλο είναι ένα στατιστικό μοντέλο που περιέχει περισσότερες παραμέτρους από αυτές που μπορούν να δικαιολογήσουν τα δεδομένα [11].

επιλύσουμε στην διπλωματική αυτή εργασία, οι SVM μπορούν να χρησιμοποιηθούν για την ταξινόμηση του συναισθήματος των χαρακτηριστικών γνωρισμάτων του προϊόντος ή μιας πρότασης της κριτικής ή ολόκληρης της κριτικής (ανάλογα το επίπεδο ανάλυσης) στις προκαθορισμένες κλάσεις «θετικό», «αρνητικό» ή ακόμη και «ουδέτερο» και περισσότερων, ανάλογα τις απαιτήσεις ταξινόμησης.

Στην παρακάτω Εικόνα 10 διακρίνεται η γραφική μοντελοποίηση του βασικού γραμμικού αλγορίθμου, με σκοπό η οπτική αναπαράστασή του να βοηθήσει στην καλύτερη κατανόηση της λειτουργίας του. Τα βήματα του βασικού αλγόριθμου SVM έχουν ως εξής:

1. Αναπαριστά τα αρχικά δεδομένα ως σημεία σε έναν χώρο υψηλών διαστάσεων, όπου είναι ο χώρος των χαρακτηριστικών, με τέτοιο τρόπο ώστε οι διακριτές κλάσεις να είναι με όσο το δυνατόν μεγαλύτερο κενό διαχωρισμένες.
2. Αναζητεί για το βέλτιστο γραμμικά διαχωριζόμενο υπερ - επίπεδο που διαχωρίζει τις δύο κλάσεις, δηλαδή αυτού που ελαχιστοποιεί το σφάλμα κατηγοριοποίησης στα άγνωστα δεδομένα, επιλέγοντας αυτό με το μέγιστο περιθώριο ανάμεσα στις δύο κλάσεις. Όταν η συνάρτηση SVM θα συγκλίνει πάντα ντετερμινιστικά στο ίδιο μοναδικό ελάχιστο σημαίνει ότι έχει εντοπίσει την βέλτιστη λύση, το υπερ - επίπεδο, δηλαδή, που μπορεί να διαχωρίσει τα δεδομένα στο χώρο των χαρακτηριστικών και ονομάζεται Maximum Marginal Hyperplane (MMH). Στις περιπτώσεις που οι κλάσεις είναι μη διακριτές, ο αλγόριθμος SVM ψάχνει ένα υπερ - επίπεδο που μεγιστοποιεί το περιθώριο και ταυτόχρονα ελαχιστοποιεί τα σφάλματα κατηγοριοποίησης.
3. Αναθέτει τα νέα στοιχεία στην κατάλληλη κλάση.



Εικόνα 10: Υλοποίηση αλγορίθμου SVM, Raschka (2015)

- **Naive Bayes**: από τις δημοφιλέστερες μεθόδους εποπτευόμενης μάθησης που χρησιμοποιείται στην ταξινόμηση κειμένου με στόχο να ταξινομήσει ένα νέο – άγνωστο στοιχείο σε μια από τις προκαθορισμένες κλάσεις με την προϋπόθεση ότι έχει εκπαιδευτεί

ώστε να αρχικοποιηθούν οι παράμετροί του. Οι ταξινομητές Naive Bayes είναι μια οικογένεια πιθανοτικών ταξινομητών, εξαιρετικά επεκτάσιμων, που βασίζονται στην εφαρμογή του Μπεϊσιανού θεωρήματος (Bayesian Decision Theory) που στηρίζεται στην αφελή (naive) υπόθεση ανεξαρτησίας μεταξύ των χαρακτηριστικών που κάνει, η οποία είναι ότι η παρουσία (ή μη) ενός χαρακτηριστικού σε μια κλάση είναι ανεξάρτητη από την παρουσία (ή μη) ενός άλλου χαρακτηριστικού (Class Conditional Independence). Αν και υπάρχουν περιπτώσεις που αυτό πράγματι συμβαίνει, δεν ισχύει πάντα, για το λόγο αυτό χαρακτηρίζεται και ως υπόθεση.

Το μοντέλο Naive Bayes είναι απλό και εύκολο στην υλοποίησή του καθώς δεν χρειάζονται πολύπλοκες επαναληπτικές εκτιμήσεις των παραμέτρων του, συμβάν που το καθιστά ιδιαίτερα χρήσιμο για μεγάλα σύνολα δεδομένων. Οι απαιτήσεις του ως προς τη CPU, την μνήμη αλλά και των δεδομένων του συνόλου εκπαίδευσης είναι μικρές, ενώ ο χρόνος εκπαίδευσής του είναι σημαντικά μικρότερος σε σχέση με άλλες μεθόδους. Αν και παρουσιάζει ανταγωνιστικά αποτελέσματα που του επιτρέπουν να χρησιμοποιείται ευρέως σε μεγάλο αριθμό μεθόδων ταξινόμησης, θεωρείται «*κακός*» εκτιμητής καθώς συχνά υπερεκτιμά τις πιθανότητες εξόδου.

Ο βασικός αλγόριθμος Naive Bayes έχει ως εξής:

Έστω ένα νέο στοιχείο $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ του οποίου η κλάση αναζητείται με βάση το διάνυσμα χαρακτηριστικών του. Υπολογίζονται για το νέο αυτό στοιχείο οι πιθανότητες να ανήκει σε κάθε μία από τις προκαθορισμένες κλάσεις και ταξινομείται τελικά στην κλάση για την οποία η πιθανότητα είναι η μεγαλύτερη. Η πιθανότητα κάποιο στοιχείο $\boldsymbol{x}$ να ανήκει στην κλάση C_i $i = 1, 2, \ldots, k$ δηλώνεται ως η δεσμευμένη πιθανότητα, γνωστή και ως εκ των υστέρων πιθανότητα (Posterior Probability)

$$P(C_i \mid \mathbf{x}) = P(C_i \mid x_1, x_2, \ldots, x_n)$$

οπότε η απόφαση του ταξινομητή $\hat{y}$ για το στοιχείο $\boldsymbol{x}$ είναι

$$\hat{y} = arg\ max_{\ i \in \{1,2,\ldots,k\}} P(C_i \mid \mathbf{x})$$

Η πιθανότητα $P(c_j|\boldsymbol{x})$, εφαρμόζοντας το θεώρημα του Bayes, υπολογίζεται ως εξής:

$$P(c_j|\boldsymbol{x}) = P(c_j, \boldsymbol{x})P(\boldsymbol{x}) = P(\boldsymbol{x}|c_j)P(c_j)P(\boldsymbol{x})$$

Όπου:

- $P(c_j)$ είναι η προγενέστερη πιθανότητα (Prior Probability) της κλάσης j
- $P(\boldsymbol{x}|c_j)$ είναι η πιθανότητα του στοιχείου $\boldsymbol{x}$ δεδομένης της κλάσης c_j (Class Conditional Probability Density Function)

Στο σημείο αυτό έρχεται να εφαρμοστεί και η υπόθεση της ανεξαρτησίας μεταξύ των χαρακτηριστικών δεδομένης της κλάσης c_j, οπότε η πιθανότητα $P(\boldsymbol{x}|c_j)$ υπολογίζεται ως το γινόμενο των επιμέρους $P(x_i|c_j)$

$$P(\boldsymbol{x}|c_j) = \prod_{i=1}^{n} P\left(x_i \middle| c_j\right)$$

Για την απόφαση του ταξινομητή αρκεί ο υπολογισμός των πιθανοτήτων (c_j) και $(x_i|c_j)$. Οι πιθανότητες αυτές εκτιμώνται κάνοντας χρήση της εκτίμησης Μέγιστης Πιθανοφάνειας (Maximum Likelihood Estimation – MLE) πάνω στο σύνολο εκπαίδευσης. Σύμφωνα με την MLE, οι παράμετροι ενός στατιστικού μοντέλου επιλέγονται έτσι ώστε να συμφωνούν με τα δεδομένα που έχει στη διάθεσή του. Έτσι, η πιθανότητα $P(c_j)$ υπολογίζεται ως το ποσοστό των στοιχείων στο σύνολο εκπαίδευσης που ανήκουν στη κλάση c_j και η πιθανότητα $P(\boldsymbol{x}|c_j)$ υπολογίζεται από τις επιμέρους πιθανότητες $P(x_i|c_j)$ οι οποίες εκτιμώνται ίσες με τις αντίστοιχες συχνότητες των χαρακτηριστικών στο ίδιο σύνολο εκπαίδευσης.

$$P(C_j) = \frac{\pi\lambda\acute{\eta}\theta o\varsigma\ \delta\varepsilon\delta o\mu\acute{\varepsilon}\nu\omega\nu\ \sigma\tau\eta\nu\ \kappa\lambda\acute{\alpha}\sigma\eta\ C_j}{\sigma\upsilon\nu o\lambda\iota\kappa\acute{o}\ \pi\lambda\acute{\eta}\theta o\varsigma\ \delta\varepsilon\delta o\mu\acute{\varepsilon}\nu\omega\nu}$$

Υπάρχουν διάφορες παραλλαγές του αλγορίθμου Naive Bayes. Η διαφορά τους έγκειται στην υπόθεση που κάνουν σχετικά με την κατανομή $P(x_i|c_j)$. Ορισμένες εκδόσεις του Naive Bayes είναι οι εξής:

- Gaussian Naive Bayes.

  Εδώ γίνεται η υπόθεση ότι τα δεδομένα ακολουθούν κανονική και Γκαουσιανή κατανομή (Gaussian Distribution). Δηλαδή:

  $$P(x_i|c_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}\right)$$

  όπου η μέση τιμή μ_j και η τυπική απόκλιση ή διασπορά σ_j υπολογίζονται μέσω της εκτίμησης Μέγιστης Πιθανοφάνειας (MLE) και ορίζουν επαρκώς τις συναρτήσεις πυκνότητας πιθανότητας. Ο αλγόριθμος αυτός συνήθως δεν βρίσκει εφαρμογή σε εργασίες Επεξεργασίας Φυσικής Γλώσσας (NLP).

- Multinomial Naive Bayes.

  Η έκδοση αυτή υλοποιεί τον αλγόριθμο Naive Bayes για πολυωνυμικά κατανεμημένα δεδομένα [37] και χρησιμοποιείται στην ταξινόμηση κειμένου και ειδικότερα στην τεχνική αναπαράστασης κειμένου σε διάνυσμα Bag-of-Words (BoW). Εδώ τα δεδομένα αναπαρίστανται συνήθως ως μετρήσεις μεμονωμένων λέξεων ή n - grams ανάλογα με τη θεώρηση. Η κατανομή παραμετροποιείται από τα διανύσματα $\boldsymbol{\theta}c_j=(\theta_{c_j1},\dots,\theta_{c_jn})^T$, όπου ο αριθμός n των χαρακτηριστικών για την ταξινόμηση κειμένου ισούται με το μέγεθος του λεξικού και θ_{c_j1} είναι η πιθανότητα $P(x_i|c_j)$ του

[37] Η πολυωνυμική κατανομή (Multinomial Distribution) χρησιμοποιείται για την εύρεση της πιθανότητας να προβλεφθεί σωστά μια σειρά επαναλήψεων ανεξάρτητων τυχαίων μεταβλητών, το καθένα εκ των οποίων έχει τη δική του γνωστή πιθανότητα να συμβεί [13].

στοιχείου i να εμφανιστεί σε ένα δείγμα της κλάσης c_j. Τα στοιχεία του διανύσματος θ_{c_j} υπολογίζονται μέσω μίας εξομαλυμένης εκδοχής MLE ως εξής:

$$\theta_{c_j i} = \frac{N_{c_j i} + a}{N_{c_j} + an}$$

όπου $N_{c_j 1}$ είναι ο αριθμός των φορών που το χαρακτηριστικό i εμφανίζεται στα δείγματα της κλάσης c_j στο σύνολο εκπαίδευσης D και N_{c_j} είναι το συνολικό πλήθος των χαρακτηριστικών για τη κλάση c_j. Η παράμετρος ομαλοποίησης a εισάγεται για την αντιμετώπιση χαρακτηριστικών που δεν εμφανίζονται καθόλου στο σύνολο εκπαίδευσης και εμποδίζει τη διάδοση μηδενικών πιθανοτήτων στους υπολογισμούς. Αν αντί να μετράμε όλες τις εμφανίσεις ενός n - gram στο κείμενο, τις μετράμε μόνο μία φορά, τότε προκύπτει η δυαδικοποιημένη (Binarized) εκδοχή του Multinomial Naive Bayes που ονομάζεται και Boolean Multinomial Naive Bayes. Συνήθως ο Multinomial Naive Bayes χρησιμοποιείται όταν οι πολλαπλές εμφανίσεις των λέξεων είναι σημαντικές στο πρόβλημα ταξινόμησης όπως στην ταξινόμηση με βάση το θέμα (Topic Classification). Ο Boolean Multinomial Naive Bayes χρησιμοποιείται όταν οι συχνότητες των λέξεων δεν παίζουν σημαντικό ρόλο στην ταξινόμηση, όπως στην Ανάλυση Συναισθήματος, όπου δεν ενδιαφέρει ο αριθμός εμφάνισης της λέξης αλλά περισσότερο το γεγονός ότι εμφανίστηκε.

- Bernoulli Naive Bayes.

  Εδώ τα δεδομένα ακολουθούν κατανομή Bernoulli [38] και συνεπώς χειρίζονται δυαδικά χαρακτηριστικά, δηλαδή κάθε χαρακτηριστικό x_j μπορεί να πάρει την τιμή *0* ή *1*. Η διαφορά του από τον Boolean Naive Bayes είναι ότι λαμβάνει υπόψη τους όρους που δεν εμφανίζονται στο κείμενο, οι οποίοι παραγοντοποιούνται όταν υπολογίζονται οι δεσμευμένες πιθανότητες και άρα η απουσία των όρων συνυπολογίζεται. Η πιθανότητα $P(x_i|c_j)$ υπολογίζεται ως:

  $$P(x_i|c_j) = P(i|c_j)x_i + (1 - P(i|c_j))(1 - x_i)$$

  Μπορεί να χρησιμοποιηθεί όταν στο πρόβλημα η απουσία κάποιας συγκεκριμένης λέξης παίζει ρόλο. Για παράδειγμα, ο Bernoulli Naive Bayes βρίσκει εφαρμογή συνήθως στην ανίχνευση ανεπιθύμητης αλληλογραφία (Spam Detection) με πολύ καλά αποτελέσματα.

- **Μέγιστης Εντροπίας (Maximum Entropy)**: ένας πιθανοτικός ταξινομητής του οποίου οι πιθανότητες εξόδου υπολογίζονται κάνοντας χρήση μιας λογιστικής συνάρτησης,

[38] Η κατανομή Bernoulli είναι η κατανομή πιθανότητας μιας τυχαίας μεταβλητής η οποία παίρνει την τιμή *1* με πιθανότητα p και την τιμή *0* με πιθανότητα $q = 1 - p$. Περιγράφει κάθε μεμονωμένο τυχαίο πείραμα με δυο πιθανά αποτελέσματα (επιτυχία - αποτυχία) και πιθανότητα επιτυχίας p. [14]

μέλος της οικογένειας σιγμοειδών συναρτήσεων [39], για το λόγο αυτό ονομάζεται και αλγόριθμος λογιστικής παλινδρόμησης (Logistic Regression) (γνωστή και ως παλινδρόμηση με σιγμοειδή συνάρτηση) αν και δεν πρέπει να μας μπερδεύει όρος καθώς στοχεύει στην ταξινόμηση και όχι στην παλινδρόμηση. Αποτελεί την γενίκευση του αλγορίθμου Naive Bayes, δηλαδή της απλής γραμμικής παλινδρόμησης για την περίπτωση όπου η εξαρτημένη μεταβλητή *Y* είναι δίτιμη (τιμή *0*: όταν απουσιάζει το χαρακτηριστικό - τιμή *1*: όταν υπάρχει το χαρακτηριστικό) και βρίσκει εφαρμογή κυρίως σε προβλήματα ταξινόμησης κειμένου. Όπως φανερώνει το όνομα του μοντέλου, βασίζεται στην αρχή της Μέγιστης Εντροπίας [40] πράγμα που καθιστά την κατανομή των δεδομένων να είναι όσο το δυνατόν πιο ομοιόμορφη, καθώς οι μόνες υποθέσεις που γίνονται για την επιλογή του μοντέλου είναι οι περιορισμοί που επιβάλλονται από το σύνολο εκπαίδευσης. Ο χρήστης είναι αυτός που καθορίζει ποιους συνδυασμούς από ετικέτες και ποια χαρακτηριστικά (Features) πρέπει να έχουν τις δικές τους παραμέτρους, καθώς η χρήση μιας παραμέτρου μπορεί να συσχετίσει ένα χαρακτηριστικό με πολλές ετικέτες ή να συσχετιστούν πολλά χαρακτηριστικά με μια ετικέτα. Ο ταξινομητής Μέγιστης Εντροπίας εφόσον, όπως προαναφέραμε, δεν κάνει κάποια υπόθεση ανεξαρτησίας μεταξύ των χαρακτηριστικών, αδυνατεί να υπολογίσει άμεσα όλες τις εξαρτήσεις για όλους τους συνδυασμούς από χαρακτηριστικά. Έτσι, για να βρεθεί το σύνολο των παραμέτρων που θα μεγιστοποιούν τη συνολική πιθανοφάνεια (Likelihood) του συνόλου εκπαίδευσης και γενικότερα την απόδοσή του, χρησιμοποιεί επαναληπτικές τεχνικές βελτιστοποίησης (Iterative Optimization Techniques). Αρχικά αρχικοποιούνται οι παράμετροι του μοντέλου με τυχαίες τιμές και επαναληπτικά ανανεώνονται ώστε να έρθουν πιο κοντά στις βέλτιστες τιμές. Αν και οι μέθοδοι βελτιστοποίησης εξασφαλίζουν ότι οι παράμετροι θα φτάσουν όντως στις βέλτιστες τιμές, δεν είναι δυνατόν να καθοριστεί ο χρόνος που χρειάζεται για την επίτευξη αυτού, με συνέπεια η διαδικασία της εκπαίδευσης συνήθως να διαρκεί αρκετό χρόνο.

Συμπερασματικά και συνοπτικά ο αλγόριθμος έχει ως εξής:

Για κάθε λέξη *w* και class $c \in C$, καθορίζουμε ένα κοινό χαρακτηριστικό $f(w, c) = N$ όπου *N* είναι το πλήθος των *w* που εμφανίζονται μέσα στο κείμενο στην κλάση *c* ή σαν λογική τιμή (boolean), για παρουσία ή απουσία της λέξης. Στη συνέχεια, μέσω επαναληπτικής βελτιστοποίησης, εκχωρείται ένα βάρος σε κάθε κοινό χαρακτηριστικό ούτως ώστε να μεγιστοποιηθεί η λογαριθμική πιθανοφάνεια των δεδομένων εκπαίδευσης.

[39] Η σιγμοειδής συνάρτηση (Sigmoid Function) είναι μια μαθηματική συνάρτηση η οποία έχει μορφή *S* και ονομάζεται και ως σιγμοειδές καμπύλη [15].

[40] Η αρχή της Μέγιστης Εντροπίας του Jayne (1957) είναι μια μέθοδος στατιστικής παρεμβολής, όταν η πληροφορία μας για ένα πρόβλημα υπάρχει με την μορφή μέσων τιμών [16].

Η πιθανότητα της κλάσης c δοθέντος του κειμένου d και των βαρών λ καθορίζεται ως εξής:

$$P(c|d,\lambda) = \frac{exp(\sum_i \lambda_i f_i(c,d))}{\sum_{c' \in C} exp(\sum_i \lambda_i f_i(c',d))}$$

- **Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks - ANNs)**: είναι υπολογιστικά μοντέλα εποπτευόμενης κυρίως μάθησης που εμπνέονται από τις βιολογικές διαδικασίες μάθησης του ανθρώπινου εγκεφάλου. Η μη γραμμικότητα, η παραλληλία, το μεγάλο πλήθος νευρώνων, οι πολύπλοκες διασυνδέσεις κ.α. είναι όλα χαρακτηριστικά των νευρωνικών δικτύων, βιολογικών και τεχνητών. Χρησιμοποιούνται σε μια πληθώρα εφαρμογών με πολύ υψηλές επιδόσεις, όπως η υπολογιστική όραση (Computer Vision) και η αναγνώριση φωνής (Speech Recognition) όπου μέσω της βαθιάς μάθησης (Deep Learning), δηλαδή την εκπαίδευση νευρωνικών δικτύων με πολλά κρυφά στρώματα, επιτυγχάνονται κορυφαία αποτελέσματα.

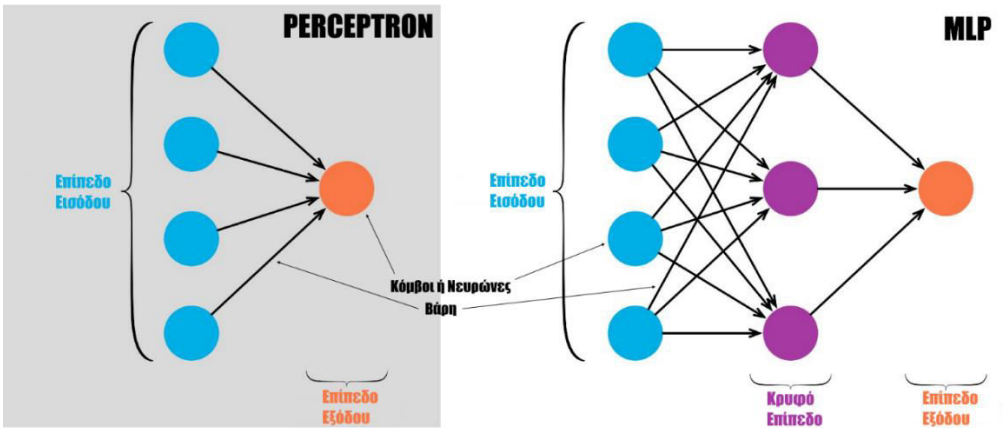
  Τα Τεχνητά Νευρωνικά Δίκτυα (ANN) απαρτίζονται από κόμβους, τους λεγόμενους νευρώνες (neurons), οι οποίοι συνδέονται μεταξύ τους δημιουργώντας ένα δίκτυο. Κάθε διασύνδεση νευρώνων στο δίκτυο χαρακτηρίζεται από κάποιο βάρος w_i (όπως φαίνεται στην Εικόνα 11), το οποίο παραμετροποιείται κατά τη φάση εκπαίδευσης, πράγμα που συνιστάται για όλα τα βάρη του δικτύου, ελαχιστοποιώντας μια συνάρτηση κόστους.

  Το βασικό στοιχείο του μοντέλου είναι ο τεχνητός νευρώνας, η απλούστερη μορφή του οποίου ονομάζεται Perceptron, ένα νευρωνικό δίκτυο δύο επιπέδων για την ταξινόμηση γραμμικά διαχωρίσιμων δεδομένων. Σε μη γραμμικά διαχωρίσιμα προβλήματα, ο αλγόριθμος δεν τερματίζει ποτέ. Για την επίλυση προβλημάτων που δεν είναι γραμμικά διαχωρίσιμα χρησιμοποιούνται νευρωνικά δίκτυα πολλών επιπέδων που καλούνται πολυστρωματικά Perceptrons (MultiLayer Perceptrons – MLP) τα οποία αποτελούν την γενίκευση του Perceptron και χρησιμοποιούν κρυφά επίπεδα (Hidden Layers), ανάμεσα σε αυτά της εισόδου και της εξόδου, για να απεικονίσουν δεδομένα σε χώρους υψηλότερης διάστασης. Έτσι ένα δίκτυο MLP αποτελείται από το επίπεδο εισόδου, το οποίο απλά στέλνει τα σήματα εισόδου σε όλους τους νευρώνες του κρυφού επιπέδου, ένα ή περισσότερα κρυφά επίπεδα μη γραμμικών νευρώνων και το επίπεδο εξόδου, το οποίο αποτελείται από γραμμικούς ή μη γραμμικούς νευρώνες (όπως φαίνεται στην Εικόνα 11).

  Στα MLP δίκτυα αφού στηθεί η αρχιτεκτονική του δικτύου επιλέγοντας τον αριθμό των επιπέδων και των νευρώνων ανά επίπεδο αλλά και τις συναρτήσεις ενεργοποίησης σε κάθε επίπεδο ακολουθεί η εκπαίδευσή τους. Η φάση της εκπαίδευσης ενός νευρωνικού δικτύου στοχεύει στις σωστές εξόδους για κάθε είσοδο σύμφωνα με τα επιλεγμένα βάρη για τους νευρώνες. Η κατάλληλη επιλογή γίνεται με βάση την ελαχιστοποίηση μιας συνάρτησης κόστους με έναν αλγόριθμο που ονομάζεται Οπισθοδιάδοση

(Backpropagation), έναν αποτελεσματικό τρόπο εκπαίδευσης δικτύων πολλών επιπέδων. Υπάρχουν δύο είδη εκπαίδευσης, ανάλογα με τη συνάρτηση που ελαχιστοποιείται και το πότε γίνονται οι ανανεώσεις των βαρών: (1) η on - line μάθηση και (2) η μαζική μάθηση (Batch Learning). Η on - line μάθηση είναι απλή στην υλοποίηση και συγκλίνει γρηγορότερα, ενώ η μαζική μάθηση αν και παραλληλοποιείται, απαιτεί περισσότερο χώρο αποθήκευσης και χρόνο.

Σύμφωνα με το θεώρημα του Καθολικού Προσεγγιστή (Universal Approximator) ένα μόνο κρυφό επίπεδο μη γραμμικών νευρώνων είναι αρκετό για την προσέγγιση οποιασδήποτε συνεχούς συνάρτησης. Η πολυπλοκότητα και οι απαιτήσεις της εκάστοτε εφαρμογής είναι αυτά που καθορίζουν τον αριθμό των κρυφών νευρώνων που θα χρησιμοποιηθούν στο εκάστοτε πρόβλημα.



Εικόνα 11:Αρχιτεκτονική Δικτύου Perceptron πολλών επιπέδων (MultiLayer Perceptron)

- **k- Πλησιέστερου Γείτονα (k- Nearest Neighbors - kNN)**: Στην περίπτωση αυτή ο ταξινομητής βασίζει την απόφαση του σε τοπολογικά κριτήρια, αντί πιθανοτικών που έχουμε εξετάσει εως τώρα. Αποτελεί έναν από τους πιο απλούς αλγορίθμους Μηχανικής Μάθησης και βασίζεται στην έννοια της εγγύτητας. Ταξινομεί σημεία του χώρου χαρακτηριστικών στην κλάση που είναι η πιο κοινή μεταξύ των k πλησιέστερων σημείων εκπαίδευσης. Η παράμετρος k καθορίζεται από τον χρήστη και επιλέγεται συνήθως μέσω πειραμάτων. Συνήθως μεγαλύτερες τιμές k μπορούν να βελτιώσουν το αποτέλεσμα της ταξινόμησης αλλά και να συμπεριλάβουν στα υποψήφια απομακρυσμένα δείγματα (Outliers) με αρνητικές βέβαια συνέπειες. Στην περίπτωση ταξινόμησης δύο κλάσεων αποφεύγονται άρτιες τιμές k για να μην υπάρχει ισοπαλία στην ψήφο της κλάσης και με την ίδια λογική, στη γενική περίπτωση K κλάσεων αποφεύγονται πολλαπλάσια του K δηλαδή $k=K$, $k=2K$, $k=3K$ κ.λπ.

  Ως μετρική απόστασης μπορεί να χρησιμοποιηθεί οποιαδήποτε μαθηματικά θεμελιωμένη απόσταση σημείων σε n - διάστατο χώρο για τον υπολογισμό της ομοιότητας ή ανομοιότητας μεταξύ ζευγών δεδομένων. Οι συνηθέστερες επιλογές είναι:

- η απόσταση Minkowski (Minkowski Distance)

$$D(x_i, x_j) = \left(\sum_{k=1}^{d} |x_{i,k} - x_{j,k}|^p \right)^{1/p}$$

- η Ευκλείδεια απόσταση (Euclidean Distance), η πιο δημοφιλής ειδική περίπτωση της απόστασης Minkowski χαρακτηρίζεται ως ένα μέτρο απόστασης μεταξύ δύο σημείων στον επίπεδο *n* - διάστατο χώρο κάνοντας επανειλημμένη χρήση του Πυθαγόρειου θεωρήματος [41],

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2}$$

- η απόσταση Mahalanobis (Mahalanobis Distance), ένα μέτρο απόστασης μεταξύ ενός σημείου *P* και μιας κατανομής *D*, που εισήχθη από τον P. C. Mahalanobis το 1936 και αποτελεί μια πολυδιάστατη γενίκευση της ιδέας της μέτρησης πόσων τυπικών αποκλίσεων μακριά είναι το *P* από τη μέση της *D*. Αυτή η απόσταση είναι μηδέν εάν το *P* είναι στη μέση του *D* και αυξάνεται καθώς το *P* απομακρύνεται από το μέσο. Διαφέρει από την Ευκλείδεια απόσταση στο ότι λαμβάνει υπόψη τη συσχέτιση μεταξύ των δεδομένων και δεν επηρεάζεται από το πλάτος των δεδομένων [17].

$$D(x_i, x_j) = \sqrt{(x_i - x_j) S^{-1} (x_i - x_j)^T}$$

όπου S^{-1}: πίνακας συνδιασποράς [42].

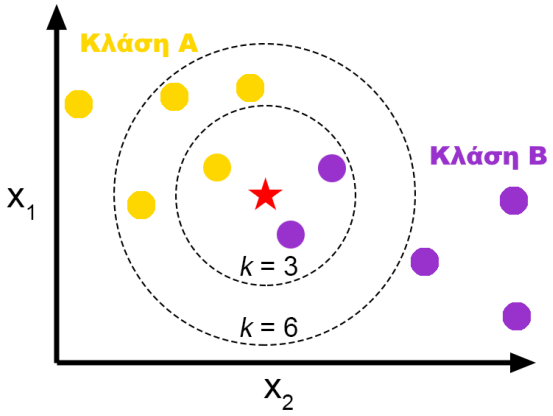
- η απόσταση Hamming (Hamming Distance) υπολογίζει την απόσταση μεταξύ δύο συμβολοσειρών ίσου μήκους όπου ορίζεται ο αριθμός θέσεων στις οποίες τα αντίστοιχα σύμβολα είναι διαφορετικά. Μετρά τον αριθμό των λαθών που μετέτρεψαν την μία συμβολοσειρά στην άλλη. Για παράδειγμα, η απόσταση Hamming μεταξύ 10**1**1**1**01 και 10**0**1**0**01 είναι 2. [18]

Ο αλγόριθμος του Πλησιέστερου Γείτονα (kNN) είναι πολύ απλός στη σύλληψη και γρήγορος στην υλοποίησή του, αλλά υπολογιστικά ακριβός για μεγάλα σύνολα δεδομένων εκπαίδευσης καθώς για κάθε νέο στοιχείο απαιτείται υπολογισμός των αποστάσεων του από όλα τα σημεία εκπαίδευσης. Επηρεάζεται αρνητικά από κλάσεις δεδομένων που δεν είναι ισορροπημένες, καθώς αν μία κλάση περιλαμβάνει πολύ περισσότερα δεδομένα από τις υπόλοιπες είναι πιθανότερο να υπερισχύσει σε μία

[41] Το Πυθαγόρειου θεώρημα είναι η σχέση της ευκλείδειας γεωμετρίας ανάμεσα στις πλευρές ενός ορθογώνιου τριγώνου: *γ2 + β2 = α2* όπου *β* και *γ* τα μήκη των δύο κάθετων πλευρών και *α* το μήκος της υποτείνουσας.

[42] Ο πίνακας Συνδιασποράς (Covariance matrix) είναι ένας πίνακας του οποίου το στοιχείο στη θέση *i, j* είναι η συνδιακύμανση μεταξύ των *i* και *j* στοιχείων ενός τυχαίου διανύσματος [19].

διαδικασία ψήφου. Η βασική του διαφορά από τους υπόλοιπους αλγορίθμους είναι ότι στην ουσία δεν περιλαμβάνει φάση εκπαίδευσης.

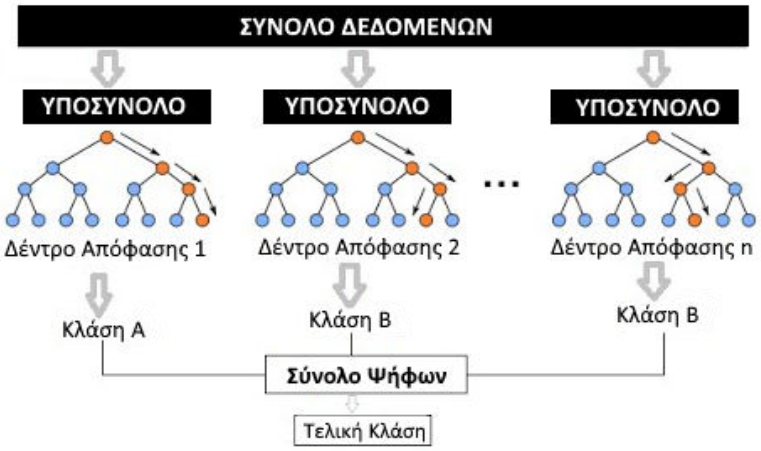


Εικόνα 12: Ταξινόμηση με k- Nearest Neighbors [21]

- **Δέντρα απόφασης (Random forest)**: Ο ταξινομητής αυτός επιλέγεται λόγω της υψηλής απόδοσής του σε ένα μεμονωμένο δέντρο απόφασης όσον αφορά την ακρίβεια. Είναι ουσιαστικά μια μέθοδος που βασίζεται στην εμφωλίαση [43]. Εκτός αυτών, μειώνει τη διακύμανση και βοηθά στην αποφυγή της υπερφόρτωσης (overfitting).

  Ο ταξινομητής λειτουργεί ως εξής: Δεδομένου ενός συνόλου D, ο ταξινομητής δημιουργεί αρχικά δείγματα k του D, με καθένα από τα δείγματα να δηλώνεται ως D_i. Κάθε D_i έχει τον ίδιο αριθμό πλειάδων με τον D, ο οποίος δειγματίζεται με αντικατάσταση από το D. Η δειγματοληψία με αντικατάσταση σημαίνει ότι μερικές από τις αρχικές πλειάδες του D δεν μπορούν να συμπεριληφθούν στο D_i, ενώ άλλες μπορεί να εμφανιστούν περισσότερες από μία φορές. Ο ταξινομητής κατασκευάζει τότε ένα δέντρο απόφασης με βάση κάθε D_i. Ως αποτέλεσμα, σχηματίζεται ένα «δάσος» (forest) που αποτελείται από k δέντρα αποφάσεων. Για να ταξινομήσει μια άγνωστη πλειάδα x, κάθε δέντρο επιστρέφει την πρόβλεψη της ταξινόμησής του ως μία ψήφο. Η τελική απόφαση της τάξης του x αποδίδεται σε αυτή που έχει τις περισσότερες ψήφους, όπως αναφέρεται στο Fang και Zhan (2015). Η Εικόνα 13 απεικονίζει μια τυπική διαδικασία ταξινόμησης με χρήση του αλγορίθμου Random forest.

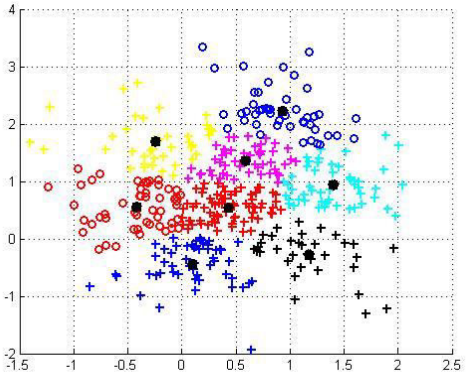
[43] Η εμφωλίαση (Bootstrap Aggregating ή Bagging) είναι ένας αλγόριθμος που έχει σχεδιαστεί για να βελτιώνει τη σταθερότητα και την ακρίβεια των αλγορίθμων Μηχανικής Μάθησης που χρησιμοποιούνται στη στατιστική ταξινόμηση και παλινδρόμηση [20].



Εικόνα 13:Ταξινόμηση με Random forest

2.3.2.2 Μη εποπτευόμενης μάθησης

- **K-means**: ένας δημοφιλής, απλός αλγόριθμος συσταδοποίησης ο οποίος στοχεύει στον διαχωρισμό στοιχείων σε συστάδες οι οποίες παρουσιάζονται γύρω από ένα σημείο το οποίο αποκαλούμε κέντρο. Κάθε στοιχείο ανήκει στην συστάδα με το κοντινότερο σε αυτό κέντρο, όπως φαίνεται και στην Εικόνα 14.



Εικόνα 14: Συσταδοποίηση με K- means [22]

2.3.2.3 Σύγκριση Αλγορίθμων Μηχανικής Μάθησης για Ανάλυση Συναισθήματος

Έπειτα της περιγραφής των διαθέσιμων μεθόδων Μηχανικής Μάθησης που χρησιμοποιούνται κατά περίπτωση για την επίλυση στα εκάστοτε προβλήματα της Ανάλυσης Συναισθήματος (π.χ. Ταξινόμηση Υποκειμενικότητας, Ταξινόμηση Πολικότητας Συναισθήματος κ.λπ.) δημιουργήσαμε τον παρακάτω Πίνακα 2 παράθεσης / σύγκρισης των συχνότερα χρησιμοποιούμενων αλγορίθμων για Ανάλυση Συναισθήματος, για να γίνουν ακόμη πιο ξεκάθαρες οι διαφορές τους. Όπως εύλογα μπορεί κάποιος να παρατηρήσει, οι αλγόριθμοι Μηχανικής Μάθησης είναι κυρίως εποπτευόμενες, καθώς όπως έχουμε προαναφέρει η Ανάλυση Συναισθήματος θεωρείται πρόβλημα εποπτευόμενης μάθησης.

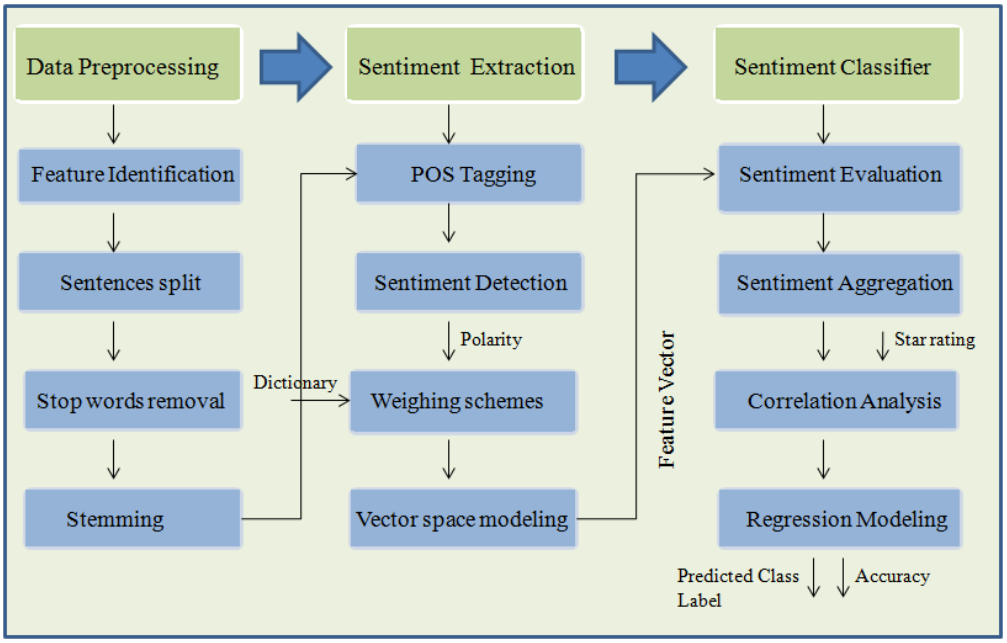
Παρατηρώντας τους μαζικά και σημειώνοντας ότι σε αυτό το σημείο να συνυπολογίζεται ο τομέας εφαρμογής τους, που θεωρείται ισχυρός παράγοντας αποδοτικότητας, διαπιστώνουμε ότι τα Δέντρα Αποφάσεων και τα Νευρωνικά Δίκτυα (ANN) είναι αδιαμφισβήτητα από τις καλύτερες επιλογές αλγορίθμων, ενώ ο kNN είναι ο πιο απαιτητικός αλλά με πλεονέκτημα την ικανοποιητική διαχείριση μεγάλου αριθμού δεδομένων. Επιπρόσθετα, παρατηρούμε ότι ο Naïve Bayes είναι από τους απλούστερους, αποδοτικότερους και πιο επεκτάσιμους αλγορίθμους όταν οι απαιτήσεις σε μνήμη και υπολογιστική ισχύ είναι μικρές, αλλά μειονεκτεί και στο ότι υποθέτει την ανεξαρτησία μεταξύ των γλωσσικών χαρακτηριστικών και συχνά υπερεκτιμά τις πιθανότητες εξόδου. Τέλος, οι SVM και Μέγιστης Εντροπίας κατατάσσονται στην ίδια πάνω – κάτω κλίμακα απόδοσης, με διαφορετικούς φυσικά τομείς εφαρμογής ο καθένας, με τον Μέγιστης Εντροπίας να χαρακτηρίζεται δυσκολότερα διαχειρίσιμος, ενώ ο SVM απαιτεί μεγάλο σύνολο εκπαιδευτικών δεδομένων γεγονός πολύ κουραστικό και χρονοβόρο, με πλεονέκτημα την αποφυγή της υπερφόρτωσης.

Πίνακας 3: Συχνότεροι αλγόριθμοι για Ανάλυση Συναισθήματος

	Naïve Bayes	**SVM**	**Max Entropy**	**ANN**	**Random forest**	**kNN**
Βασισμένος σε	Θεώρημα Bayes	Διανυσματικές Αποστάσεις	Σιγμοειδής Συνάρτηση, Γενίκευση Bayes	Deep Learning, Βιολογική μάθηση εγκεφάλου	Συγχώνευση Δέντρων Απόφασης	Πλησιέστερου γείτονα (απόσταση)
Απλότητα	Πολύ απλός	Μέτριος	Δύσκολος	Μέτρια	Απλός	Απλά
Απόδοση	★★★★	★★★★	★★★	★★★★★	★★★★★	★
Απαιτήσεις Μνήμης						
Ακρίβεια	★★★	★★★	★★★★	★★★★	★★★★★	★★★
Χρόνο που απαιτεί						
Ενδεικτικές εφαρμογές	Ανίχνευση ανεπιθύμητων μηνυμάτων, Ταξινόμηση εγγράφων, Ανάλυση Συναισθήματος	Αναγνώριση χειρόγραφου, Βιοπληροφορική, Κατηγοριοποίηση κειμένου, Ανίχνευση προσώπου	Δοκιμές διάγνωσης στην ιατρική, Φυσική	Υπολογιστική Όραση, Αναγνώριση Φωνής	Βιομετρική, Ηλεκτρονικό εμπόριο, Φάρμακα, Τράπεζες, Χρηματιστήριο	Συστήματα συστάσεων, Αναζήτηση έννοιας

2.4 *Ανάλυση Συναισθήματος βασισμένη στις λέξεις - κλειδιά*

Η Ανάλυση Συναισθήματος σε επίπεδο εγγράφου και πρότασης είναι χρήσιμη σε πολλές εφαρμογές εάν υποθέσουμε ότι το κείμενο που αναλύεται φέρει άποψη για ένα χαρακτηριστικό γνώρισμα μιας οντότητας. Η πλειοψηφία των κειμένων όμως, περιέχουν τόσο θετικές όσο και αρνητικές γνώμες για περισσότερα του ενός χαρακτηριστικά γνωρίσματα της οντότητας και όχι μια ξεκάθαρη άποψη για μια ενιαία οντότητα, έτσι για να εντοπίσουμε αυτές τις κρυμμένες πληροφορίες, θα πρέπει να μεταφερθούμε στην Ανάλυση Συναισθήματος βασισμένη σε λέξεις – κλειδιά (Aspect-based Sentiment Analysis - **ABSA**) [44] (ή feature-based) και σε επίπεδο λέξης. Ο στόχος της είναι αρχικά να εντοπίσει τα χαρακτηριστικά γνωρίσματα της οντότητας και στη συνέχεια να αποδώσει σε αυτά το συναίσθημα που εκφράζουν όπως επισημαίνουν οι Laskari και Sanampudi (2016). Στην περίπτωσή μας, εστιάζουμε στα χαρακτηριστικά γνωρίσματα του προϊόντος για τα οποία εκφράζεται συναίσθημα μέσω μιας κριτικής και ποιο είναι αυτό (π.χ. *θετική γνώμη για την κάμερα του νέου iPad*). Μια τυπική διαδικασία εντοπισμού αυτών φαίνεται στην Εικόνα 15.



Εικόνα 15: Διαδικασία Ανάλυσης Συναισθήματος βασισμένη σε λέξεις - κλειδιά, Abirami και Askarunisa (2016)

Για την επίτευξη του παραπάνω στόχου με μεθόδους Μηχανικής Μάθησης απαιτείται η αναπαράσταση του κειμένου σε διάνυσμα χαρακτηριστικών *x* ώστε να υλοποιηθούν τεχνικές προ - επεξεργασίας δεδομένων, να γίνει εξόρυξη των γνωρισμάτων και ο εντοπισμός του συναισθηματικού τους προσανατολισμού με τη δημιουργία των αντίστοιχων λεξικών, όπως παρουσιάζουμε αναλυτικότερα στα επόμενα κεφάλαια. Εδώ αξίζει να αναφέρουμε ότι η

[44] Για λόγους συντομίας θα χρησιμοποιούμε στο υπόλοιπο του κειμένου της διπλωματικής εργασίας τον όρο Ανάλυση Συναισθήματος βασισμένη σε λέξεις – κλειδιά ως ABSA από τα αρχικά της αγγλικής ορολογίας Aspect-based Sentiment Analysis, όπως είναι και ευρέως χρησιμοποιούμενος.

διαδικασία υλοποίησης της ABSA στηρίζεται στην παραδοχή ότι κάθε κριτική μιλάει για ένα γνωστό προϊόν καθώς έχει εξαχθεί από τη σελίδα του συγκεκριμένου προϊόντος, άρα δεν χρειάζεται να γίνει ανίχνευσή του. Για το λόγο αυτό ασχολούμαστε μόνο με την εξαγωγή των χαρακτηριστικών γνωρισμάτων του συγκεκριμένου προϊόντος στα οποία αναφέρεται ο σχολιαστής. Η ABSA βρίσκει εφαρμογή σε μεγάλη ποικιλία δεδομένων όπως κριτικές ταινιών, ταξιδιών και καταστημάτων εστίασης, προϊόντων τεχνολογίας και υπηρεσίες.

2.4.1 Κείμενο σε διάνυσμα

Οι αλγόριθμοι Μηχανικής Μάθησης που εφαρμόζονται για την Ανάλυση Συναισθήματος στα κείμενα των κριτικών, απαιτούν αναπαράσταση του κειμένου σε διάνυσμα, ώστε να είναι σε θέση να πραγματοποιήσουν πράξεις μεταξύ των λέξεων αλλά και να είναι επιτεύξιμη και η σύγκρισή τους. Το ζητούμενο είναι η αντιστοίχιση κάθε κειμένου κριτικής σε μία διανυσματική αναπαράσταση $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ όπου n είναι η διάσταση του χώρου χαρακτηριστικών. Αυτή η διαδικασία εντοπίζεται και με τον όρο διανυσματοποίηση (Vectorization) και μπορεί να πραγματοποιηθεί με τα μοντέλα αναπαράστασης που παρουσιάζονται παρακάτω. Η παρουσίασή τους γίνεται συνοπτικά καθώς στόχος μας είναι να γίνουν γνωστές στον αναγνώστη οι επιλογές που υπάρχουν για την αναπαράσταση των δεδομένων στο διανυσματικό χώρο, αναφέροντας τα μειονεκτήματα και τις εφαρμογές τους, ώστε να είναι σε θέση να ξεχωρίσει το κατάλληλο για κάθε περίπτωση και εάν χρειαστεί να ανατρέξει στην αντίστοιχη πηγή για την μαθηματική υλοποίησή του.

2.4.1.1 Bag of Words – BoW

Η απλούστερη αναπαράσταση του κειμένου ως σύνολο από λέξεις / φράσεις (όρους) στον διανυσματικό χώρο. Όπως δηλώνει και το όνομά της, κάθε κείμενο αντιμετωπίζεται σαν ένα «σακίδιο» ανεξάρτητων λέξεων ή φράσεων, όπου αυτό πρακτικά σημαίνει ότι δεν λαμβάνεται υπόψιν η σειρά των λέξεων / φράσεων αλλά εστιάζει στην παρουσία και την συχνότητα παρουσίας αυτών στο κείμενο. Άμεσα, λοιπόν, γίνεται αντιληπτή η αδυναμία του μοντέλου, εφόσον η σειρά των λέξεων / φράσεων παίζει καθοριστικό ρόλο, ιδιαίτερα στην Ανάλυση Συναισθήματος. Το πιο χαρακτηριστικό παράδειγμα είναι η παρουσία της άρνησης σε κάποιο σημείο του κειμένου, το οποίο αλλάζει εντελώς το νόημά του, αν και έχει την ίδια αναπαράσταση, όπως εύκολα διακρίνουμε στο παράδειγμα του Πίνακα 3.

Πίνακας 4: Αναπαράσταση BoW

Πραγματικό Κείμενο	**Αναπαράσταση BoW**
«Η ανάλυση της οθόνης μου άρεσε, αλλά δεν το προτείνω»	*{Η, ανάλυση, της, οθόνης, μου, άρεσε,,, αλλά, δεν, το, προτείνω}*
«Η ανάλυση της οθόνης δεν μου άρεσε, αλλά το προτείνω»	*{Η, ανάλυση, της, οθόνης, δεν, μου, άρεσε,,, αλλά, το, προτείνω}*

Ένα επιπλέον μειονέκτημα του μοντέλου είναι ότι ο διανυσματικός χώρος που δημιουργεί, δεν παρέχει πληροφορίες για τις σημασιολογικές σχέσεις μεταξύ των λέξεων / φράσεων και έτσι δεν εγγυάται η αναπαράσταση των κειμένων με παρόμοιο νόημα σε κοντινές θέσεις του χώρου χαρακτηριστικών. Ωστόσο, λόγω της απλότητάς του είναι ιδιαίτερα δημοφιλής, εξασφαλίζοντας ικανοποιητικά αποτελέσματα σε προβλήματα ταξινόμησης κειμένου και θεματικών ενοτήτων.

Πιο αναλυτικά, η διαδικασία αναπαράστασης του μοντέλου BoW ξεκινά με την δημιουργία ενός λεξιλογίου (vocabulary) που περιέχει όλες τις λέξεις ή φράσεις (n - grams) που εμφανίζονται στα κείμενα. Ακολουθεί η αναπαράσταση κάθε κειμένου σε διάνυσμα, όπου κάθε χαρακτηριστικό του αντιστοιχίζεται σε κάθε λέξη του κειμένου, έτσι διαπιστώνουμε ότι η διανυσματική αναπαράσταση έχει διάσταση ίση με το μέγεθος του λεξικού. Οι τιμές των χαρακτηριστικών μπορεί να είναι *1* ή *0* για παρουσία ή απουσία αντίστοιχα, της λέξης στο κείμενο που καλείται Term Occurrence ή ο αντίστοιχος φυσικός αριθμός που χαρακτηρίζει την συχνότητα παρουσίας της λέξης / φράσης στο κείμενο και καλείται Term Frequency.

Η BoW αναπαράσταση μπορεί να αντιμετωπίσει τις λέξεις των κειμένων σαν unigrams ($n = 1$), δηλαδή μεμονωμένα την κάθε λέξη αλλά και σαν bigrams ($n = 2$), δηλαδή δύο διαδοχικές λέξεις ή και n - grams (n διαδοχικές λέξεις), για αυτό και αναφερόμαστε σε αυτές σαν λέξεις / φράσεις ή συνολικά όροι. Η επιλογή του n εξαρτάται από την εφαρμογή και στόχο έχει την βελτίωση του ταξινομητή. Συνήθως προτιμώνται μικρά n ώστε να αποφευχθεί η αύξηση των χαρακτηριστικών x στο διάνυσμα, που σημαίνει αύξηση των διαστάσεων και μείωση της απόδοσης, αν και η επιλογή μεγάλων n ($n >= 3$) εντοπίζει πιο ακριβείς και ντετερμινιστικές (προ - καθορισμένες) εκφράσεις όπως υποστηρίζουν οι Cui κ.α. (2006).

Για παράδειγμα, έχοντας τα δύο κείμενα από κριτικές προϊόντων τεχνολογίας:

Κείμενο 1: *είναι πολύ καλό tablet με πολύ καλές επιδόσεις!*

Κείμενο 2: *η διάρκεια ζωής της μπαταρίας του tablet είναι πολύ απογοητευτική …*

Για $n = 1$:

Λεξικό$_{n=1}$: [*'είναι', 'πολύ', 'καλό', 'tablet', 'με', 'καλές', 'επιδόσεις', '!', 'η', 'διάρκεια', 'ζωής', 'της', 'μπαταρίας', 'του', 'απογοητευτική', '.'*]

Term occurrence (Κείμενο 1): [1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0]

Term occurrence (Κείμενο 2): [1,1,0,1,0,0,0,0,0,1,1,1,1,1,1,1,1]

Term frequency (Κείμενο 1): [1,2,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0]

Term frequency (Κείμενο 2): [1,1,0,1,0,0,0,0,0,1,1,1,1,1,1,1,3]

Για $n = 2$:

Λεξικό$_{n=2}$: [*'είναι πολύ', 'πολύ καλό', 'καλό tablet', 'tablet με', 'με πολύ', 'πολύ καλές', 'καλές επιδόσεις', 'επιδόσεις!', 'η διάρκεια', 'διάρκεια ζωής', 'ζωής της', 'της μπαταρίας', 'μπαταρίας του', 'του tablet', 'tablet είναι', 'πολύ απογοητευτική', 'απογοητευτική.'*]

Term occurrence (Κείμενο 1): [1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0]

Term occurrence (Κείμενο 2): [1,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1]

Term frequency (Κείμενο 1): [1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0]

Term frequency (Κείμενο 2): [1,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1]

2.4.1.2 Word Vectors

Προσμετρώντας τα μειονεκτήματα του μοντέλου BoW, που εξετάσαμε προηγουμένως, οδηγούμαστε στις αναπαραστάσεις Word Vectors που υπόσχονται να ανακαλύπτουν τόσο σημασιολογικές όσο και συντακτικές σχέσεις μεταξύ των λέξεων στο κείμενο προς ανάλυση. Εδώ κάθε λέξη αντιστοιχίζεται σε ένα πυκνό διάνυσμα ενός χώρου μικρής διάστασης με πραγματικές τιμές. Για την δημιουργία των διανυσματικών αναπαραστάσεων χρησιμοποιούνται κυρίως μη εποπτευόμενης μάθησης τεχνικές μείωσης διάστασης σε πίνακες συν - εμφάνισης (co - occurrence), οι οποίοι περιέχουν, συν τοις άλλοις, πληροφορίες για την συνύπαρξη των λέξεων στο κείμενο. Στην απλούστερη περίπτωση, είναι τετραγωνικοί και συμμετρικοί και κάθε γραμμή τους όπως και κάθε στήλη τους αντιστοιχούν σε έναν όρο από το λεξικό. Στην μέθοδο αυτή αναφερόμαστε μόνο συνοπτικά για να γίνει μία σύνδεση των word vectors με τη μέθοδο Bag-of-Words και να αποσαφηνιστεί η ιδέα της απεικόνισης λέξεων σε διανυσματικούς χώρους με τη χρήση στατιστικών μετρήσεων για την γειτνίασή τους, καθώς δεν είναι ένα από τα πεδία που χρειάζεται εδώ να αναλύσουμε περαιτέρω.

Η αναπαράσταση των Word Vectors επιτυγχάνεται με νευρωνικά γλωσσικά μοντέλα (Neural Language Models - NLM) τα οποία αντιμετωπίζουν το ζήτημα της έλλειψης δεδομένων στις λέξεις μέσω της παραμετροποίησής τους ως διανυσμάτα και τη χρήση τους ως είσοδο σε ένα νευρωνικό δίκτυο, όπως τα Word2vec και GloVe τα οποία αποσαφηνίζονται παρακάτω. Οι παράμετροι διδάσκονται ως μέρος της εκπαιδευτικής διαδικασίας.

1. Το **Word2vec** μοντέλο χρησιμοποιείται για την απεικόνιση λέξεων σε διανυσματικούς χώρους σχετικά μικρών (50 ως 300) διαστάσεων, δημιουργώντας σχέσεις μεταξύ των λέξεων χωρίς τη βοήθεια εξωτερικών επισημειωτών, άρα υπόκεινται στην μη εποπτευόμενη μάθηση. Τα συστατικά των διανυσμάτων αντιπροσωπεύουν τα βάρη ή τη σημασία της κάθε λέξης στο κείμενο. Είναι ικανό να ανακαλύπτει πολύπλοκες σημασιολογικές και συντακτικές σχέσεις μεταξύ των λέξεων και να τις μετατρέπει σε γραμμικές ιδιότητες του διανυσματικού χώρου. Γενικά πρόκειται για ένα window-based γλωσσικό μοντέλο πρόβλεψης που βασίζεται στην αρχιτεκτονική ενός απλού νευρωνικού δικτύου.

Προτάθηκε από τους Mikolov κ.α. (2013) και σε συνεργασία με την Google έδωσαν αρχικά δύο διαφορετικές υλοποιήσεις του για συντακτικά προβλήματα που αναφέρονται παρακάτω, αλλά κατά το πέρασμα των χρόνων, εντοπίζεται να βελτιώνεται με τις νέες εκδόσεις του σε ποικιλία γλωσσών προγραμματισμού.

- την Continuous Bag-of-Words (CBOW) που προβλέπει την κεντρική λέξη, δεδομένων των λέξεων γύρω από αυτή στο διανυσματικό χώρο,
- την Skip - Gram (SG) που προβλέπει τις τριγύρω λέξεις δεδομένης της λέξης στο κέντρο, όπως αναφέρουν οι Ling κ.α. (2015).

Ένα παραγόμενο διάνυσμα του μοντέλου θα μπορούσε να ήταν το εξής:

$$vec(\text{'Αθήνα'}) = vec(\text{'Ρώμη'}) - vec(\text{'Ιταλία'}) + vec(\text{'Ελλάδα'})$$

2. Ένα υβριδικό μοντέλο, συνδυασμός των δύο τεχνικών εξαγωγής διανυσματικών αναπαραστάσεων, Word2vec και BoW, είναι το **GloVe** (Global Vectors), που δημιουργεί διανυσματικές αναπαραστάσεις λέξεων ενσωματώνοντας συνολική (Global) πληροφορία. Το μοντέλο αυτό χρησιμοποιεί τον πίνακα συν - εμφάνισης του συνόλου δεδομένων αλλά δεν εφαρμόζει κάποια μέθοδο μείωσης της διάστασης. Αντίθετα χρησιμοποιεί βαθμίδα κατάβασης (Gradient Descent) με σκοπό την ελαχιστοποίηση της συνάρτησης κόστους.

2.4.1.3 Διανυσματικές Αναπαραστάσεις Κειμένου

Οι Le και Mikolov (2014) προτείνουν ένα νευρωνικό μοντέλο για την εκπαίδευση διανυσματικών αναπαραστάσεων κειμένου που αποτελεί άμεση γενίκευση και επέκταση του μοντέλου Word2vec. Ονομάζουν το μοντέλο Paragraph Vector (PV) ωστόσο συχνά το ίδιο μοντέλο αναφέρεται με το όνομα **Doc2Vec**, ονομασία που καταδεικνύει τη στενή σχέση του με το μοντέλο Word2vec αλλά και τη δυνατότητα του μοντέλου να παράγει αναπαραστάσεις για οποιασδήποτε μορφής έγγραφα, από φράσεις και προτάσεις μέχρι παραγράφους και ολόκληρα κείμενα, που είναι και η βασική διαφορά του με το Word2vec.

2.4.2 Προεπεξεργασία δεδομένων

Έπειτα της αναπαράστασης του κειμένου προς ανάλυση σε διάνυσμα, εντοπίζεται η αναγκαιότητα να μειωθούν οι διαστάσεις του, ώστε να γίνει ευκολότερη η εργασία των ταξινομητών που θα χρησιμοποιηθούν έπειτα στις μεθόδους Μηχανικής Μάθησης. Η προεπεξεργασία των δεδομένων είναι μια από τις ενέργειες που συμβάλλει στην βελτίωση της ακρίβειας των ταξινομητών, μετατρέποντας σε μια «φιλικότερη» μορφή το αδόμητο κείμενο που δέχονται ως είσοδο, σύμφωνα με τους Haddi κ.α. (2013). Συνεπώς καταλαβαίνουμε την σημαντικότητά της και την ανάγκη να γίνει με προσοχή. Η μετατροπή αυτή εξαλείφει την άχρηστη πληροφορία ή «θόρυβος» όπως συχνά συναντάται, από τα κείμενα και πιο συγκεκριμένα στην ABSA μέσω της μεθόδου της ευρετηρίασης

(**tokenization**) επιτυγχάνεται ο διαχωρισμός του κειμένου μόνο σε τμήματα που αποφέρουν νόημα (tokens). Τα tokens συμβαίνει να έχουν διαφορετικό μήκος ανάλογα με τις ανάγκες του κάθε προβλήματος και την εφαρμογή ορισμένων κανόνων. Μπορεί να τα συναντήσουμε ως λέξη, αριθμό, σημείο στίξης, έννοιες, φράσεις κ.α. Το μικρότερο τμήμα κειμένου με νόημα είναι μια μεμονωμένη λέξη και με την μέθοδο της ευρετηρίασης όλες οι μορφές αυτής της λέξης, που ονομάζονται λέξημα (lexeme), θεωρούνται ένα token.

Η προεπεξεργασία δεδομένων ανήκει στο πεδίο Επεξεργασίας Φυσικής Γλώσσας (NLP) και εύκολα διαπιστώνουμε ότι θα υπάρχει διαφορετική αντιμετώπιση σε κείμενα που μιλάνε για μαθηματικά / φυσική / χημεία, άρθρα σε εφημερίδες / blogs, κείμενα email / chat, κείμενα σε κάποια γλώσσα προγραμματισμού κ.α. Για παράδειγμα, στον τομέα της Εξόρυξης Πληροφορίας τα σημεία στίξης δεν παίζουν κανένα ρόλο στα κείμενα αναζήτησης και θα πρέπει να εξαλείφονται καθώς για τον διαχωρισμό των λέξεων χρησιμοποιείται το κενό (space). Ακόμη ένα κατατοπιστικό παράδειγμα είναι σε κείμενα γραμμένα σε HTML ή XML (γλώσσες προγραμματισμού) όπου εξετάζονται για την εξόρυξη συναισθήματος, θα πρέπει να αφαιρούνται οι ετικέτες (tags) διότι σχετίζονται μόνο με τη μορφοποίηση του κειμένου και δεν αποφέρουν καμία λειτουργικότητα στο νόημά του.

Συνεπώς, η υλοποίηση της ευρετηρίασης δεν περιλαμβάνει συγκεκριμένες λειτουργίες και δεν υπάρχει προκαθορισμένη διαδικασία αλλά υπάρχει πληθώρα αλγορίθμων που διαχωρίζουν τα συστατικά του κειμένου κατά περίπτωση και σύμφωνα με τον τομέα εξέτασης των κειμένων και εντοπίζει τα κατάλληλα token. Για παράδειγμα, εάν θέλουμε να εκτελέσουμε αφαίρεση μια λίστας stopwords, θα χωρίσουμε το κείμενο σε tokens μια λέξης και θα την εφαρμόσουμε πάνω σε αυτά και όχι σε ολόκληρο το κείμενο, καθώς η λίστα των stopwords αποτελείται (συνήθως) από μεμονωμένες λέξεις.

Κάτωθεν παρατίθενται οι βασικότεροι αλγόριθμοι που χρησιμοποιούνται ευρύτερα στην προεπεξεργασία των κειμένων κριτικών, πάντα με τον ίδιο σκοπό, την βελτίωση της ακρίβειας των ταξινομητών.

- **Part-of-speech Tagging (POS Tag)**: μια διαδικασία η οποία ανήκει κι αυτή στο πεδίο Επεξεργασίας Φυσικής Γλώσσας (NLP) και χαρακτηρίζει τι μέρος του λόγου (γραμματική χρήση) είναι η κάθε λέξη μέσα στο κείμενο (π.χ. ρήμα, επίθετο, επίρρημα κ.λπ.). Μπορεί να αναπαρασταθεί με ποικίλους τρόπους προσαρμοζόμενη και αυτή στις ανάγκες κάθε προβλήματος και τομέα εφαρμογής. Ένα παράδειγμα αναπαράστασης θα μπορούσε να είναι προσθέτοντας μπροστά από κάθε ρήμα το αρχικό αναγνωριστικό μέσα σε ετικέτες π.χ. *<V>* από το *verb*. Ωστόσο, το αποτέλεσμα έπειτα της υλοποίησης του αλγορίθμου, θα είναι πάντα το ίδιο, καθώς σε αυτό το σημείο αυτό που ενδιαφέρει είναι να διατηρηθούν στο λεξιλόγιο μόνο όροι χαρακτηρισμένοι ως ένα συγκεκριμένο μέρος του λόγου. Για παράδειγμα, έχει αποδειχθεί ότι τα επίθετα είναι σημαντικοί δείκτες

απόψεων, έτσι θα διαχειριστούν διαφορετικά σε εξεταζόμενα κείμενα κριτικές χρηστών από ότι σε tweets, όπου τα κείμενα λόγω της μικρής τους έκτασης δεν είναι περιγραφικά αλλά εστιάζουν στο θέμα και συνεπώς δεν χρησιμοποιούν πολλά επίθετα.

Το POS Tagging στηρίχθηκε στη μετρική PMI, η οποία περιγράφεται στο Κεφάλαιο 2.4.3, καθώς ο εντοπισμός επιτυγχάνεται με την μέτρηση της εξάρτησης των γειτονικών λέξεων μέσα σε προτάσεις ή απλά φράσεις. Αξίζει να σημειωθεί ότι μια λέξη μπορεί να έχει περισσότερους του ενός χαρακτηρισμούς, ανάλογα το περιεχόμενο του κειμένου. Οι αλγόριθμοι που καθορίζουν τον χαρακτηρισμό της λέξης μπορεί να λειτουργούν αυτοματοποιημένα με τη χρήση Μέγιστης Εντροπίας ή Δέντρων Αποφάσεων σε συνδυασμό με την στατιστική και ονομάζονται στοχαστικοί (stochastic), επιτυγχάνοντας υψηλό ποσοστό ακρίβειας. Η εναλλακτική περίπτωση είναι χειροκίνητα, με την χρήση κανόνων, με τους λεγόμενους rule - based αλγόριθμους οι οποίοι αν και έχουν χαμηλότερο ποσοστό ακρίβειας, απαιτούν μικρότερο όγκο δεδομένων, βελτιώνονται, εφαρμόζονται ευκολότερα και αποδίδουν και με ένα μικρό σύνολο κανόνων.

Το πιο διαδεδομένο POS Tagging της αγγλικής γλώσσας, όπως αναφέρει ο Liu (2012), είναι του Penn Treebank [45] και οι χαρακτηρισμοί του φαίνονται στον Πίνακας 4.

Πίνακας 5: Επεξήγηση Penn Treebank Part-Of-Speech (POS) tags

TAG	Περιγραφή
CC	Coordinating Conjunction: Συνδετικοί Συντονιστές (π.χ. and, but)
CD	CarDinal number: Απόλυτος αριθμός (π.χ. 1, two, thousand)
DT	DeTerminer: Προσδιοριστής, λέξη πριν από ένα ουσιαστικό που δείχνει που αναφέρεται το ουσιαστικό (π.χ. my)
EX	EXistential there: Ύπαρξη λέξης
FW	Foreign Word: Ξένη λέξη
IN	Preposition or subordinating conjunction: Πρόθεση ή υποδεέστερος σύνδεσμος (π.χ. on, before)
JJ	Adjective: Επίθετο (π.χ. fast)
JJR	Comparative adjective: Συγκριτικό επίθετο (π.χ. faster)
JJS	Superlative adjective: Υπερθετικό επίθετο (π.χ. fastest)
LS	LiSt item marker: Δείκτης στοιχείου λίστας
MD	Modal: Βοηθητικό ρήματος (π.χ. can, should)
NN	Singular Noun: Ουσιαστικό ενικού αριθμού (π.χ. laptop)
NNS	Plural Noun: Ουσιαστικό πληθυντικού αριθμού (π.χ. laptops)
NNP	Proper singular noun: Κύριο όνομα ενικού αριθμού που αναφέρεται σε μια μοναδική οντότητα (π.χ. John, Samsung)
NNPS	Proper plural noun: Κύριο όνομα πληθυντικού αριθμού (π.χ. Kennedys)
PDT	PreDeTerminer: Λέξη πριν τον προσδιοριστή για περισσότερες πληροφορίες (π.χ. all, both)

[45] Διαθέσιμο από: http://www.cis.upenn.edu/~treebank/home.html

POS	Possessive ending: Κτητική κατάληξη (π.χ. Companies' workers)
PRPS	Possessive pronoun: Κτητική αντωνυμία (π.χ. mine, you)
RB	Adverb: Επίρρημα (π.χ. early)
RBR	Comparative adverb: Συγκριτικό επίρρημα (π.χ. earlier)
RBS	Superlative adverb: Υπερθετικό επίρρημα (π.χ. earliest)
RP	Particle: Πρόθεση που αλλάζει ένα ρήμα (π.χ. away, off)
SYM	Symbol (π.χ. ✓)
TO	To
UH	Interjection: Επιφώνημα (π.χ. oops, yeah)
VB	Base form of VerB: Βάση μορφή ρήματος (π.χ. play)
VBD	Past tense of VerB: Αόριστος χρόνος ρήματος (π.χ. played)
VBG	Gerund or present participle of VerB: Γερούνδιος ή ενεστώτας ρήματος (π.χ. playing)
VBN	Past participle of VerB: Παθητική μετοχή ρήματος (π.χ. have played)
VBP	Non-3rd person singular of VerB: Μη 3ου ενικού πρόσωπο ρήματος (π.χ. play)
VBZ	3rd person singular present of VerB: 3ο ενικό πρόσωπο ρήματος (π.χ. play)
WDT	Wh-determiner: Προσδιοριστής λέξης που ξεκινάει με wh (π.χ. what, which)
WP	Wh-pronoun: Αντωνυμία λέξης που ξεκινάει με wh (π.χ. who)
WPS	Possessive wh-pronoun: Κτητική αντωνυμία (π.χ. whose)

- Η αντικατάσταση ή μαρκάρισμα (Tagging) ορισμένων ακολουθιών χαρακτήρων που εκφράζουν μια έννοια λεγόμενα σαν **emoticons**, τα οποία εντοπίζονται συχνά σε κείμενα κριτικών. Ο τρόπος παρουσίασης ενός emoticon δεν είναι μοναδικός, υπάρχουν παραλλαγές που περιλαμβάνουν κενά, γράμματα, επιπλέον σημεία στίξης, επαναλήψεις γραμμάτων κ.α. Για παράδειγμα το emoticon *«:D»* μπορεί να εμφανιστεί με τις επιπλέον μορφές: *«:d», «: D», «:ddd», «:DDD»* κ.α. Συνεπώς ο εντοπισμός όλων των δυνατών τρόπων εμφάνισης ενός emoticon είναι αρκετά δύσκολος εώς αδύνατος, ενώ παράλληλα υπάρχει και το πρόβλημα των περιπτώσεων όπου η ακολουθία των χαρακτήρων ταιριάζει σε κάποιο emoticon αλλά στην πράξη δεν είναι, και απλά τυχαίνει αυτή η ακολουθία χαρακτήρων στο κείμενο. Για παράδειγμα, στην ακολουθία των χαρακτήρων *«:Dell»* εντοπίζεται το emoticon *«:D»* το οποίο σημαίνει «*πολύ χαρούμενος*» αλλά στην πραγματικότητα δεν θέλει το κείμενο να περάσει αυτή την έννοια. Στον Πίνακας 5 εντοπίζονται ορισμένα από τα συχνότερα χρησιμοποιούμενα emoticons με την αντίστοιχη σημασία τους.

Πίνακας 6: Συχνότερα χρησιμοποιούμενα emoticons

EMOTICON						**ΣΗΜΑΣΙΑ**
:D	:-D	=D	=-D			Πολύ χαρούμενος
:)	:-)	:]	=]	=)	(:	Χαρούμενος
;)	;-)					Κλείνει το μάτι

:(	:-(	:|	:-|	=(	=|		Λυπημένος
:'(	='(						Κλαίει
<3							Καρδιά (ερωτευμένος, αρέσει)
:O	:o	:-O	:-o	:0	:-0		Άφωνος, έκπληκτος
:P	:p	:d	:-P	:-p	:-d		Βγάζει γλώσσα, κοροϊδεύει

- Η αφαίρεση των **Stopwords** τα οποία περιλαμβάνουν πολύ συχνά χρησιμοποιούμενες λέξεις, άρθρα, επιρρήματα, προθέσεις, αριθμούς κ.α. όπως «*and*», «*this*», «*a/an*» κ.λπ. Η αφαίρεσή τους, λοιπόν, είναι μια απαραίτητη διαδικασία εάν στόχος είναι να μειωθεί ο όγκος των δεδομένων και να βελτιωθεί η απόδοση του συστήματος εξόρυξης συναισθήματος, καθώς όχι μόνο δεν βοηθάνε στην εξόρυξη γνώσης αλλά δυσκολεύουν ακόμη περισσότερο την διαδικασία των αλγορίθμων Μηχανικής Μάθησης. Εύκολα γίνεται αντιληπτό όμως, ότι όσο χρήσιμη κι αν είναι η αφαίρεσή τους, τόσο δύσκολη είναι η δημιουργία μιας κατάλληλης λίστας Stopwords. Δεν υπάρχει ιδανική λίστα που μπορεί να χρησιμοποιείται πάντα κι από όλους, καθώς όπως έχουμε αναφερθεί επανειλημμένως, η Ανάλυση Συναισθήματος σε κάθε τομέα εξετάζεται με διαφορετικά κριτήρια και εδώ εντοπίζεται η ανάγκη διατήρησης διαφορετικών λέξεων. Για παράδειγμα, σε εξεταζόμενα κείμενα στον προγραμματισμό υπολογιστών και πιο συγκεκριμένα σε HTML σελίδες η διατήρηση των ετικετών (tags) στα κείμενα είναι απαραίτητη διότι έχουν μεγάλη σημασία, ενώ ο εντοπισμός ετικετών σε άλλο τομέα εξέτασης θα θεωρούνταν άχρηστη πληροφορία. Επομένως, η διαδικασία δημιουργίας της καταλληλότερης λίστας είναι μια επίπονη διαδικασία αλλά απαραίτητη.

  Υπάρχουν αρκετές έτοιμες λίστες με προκαθορισμένες λέξεις, συνήθως για κάθε ξεχωριστή γλώσσα, που μπορούν να χρησιμοποιηθούν σαν το εφαλτήριο και να γίνουν οι κατάλληλες προσθήκες πάνω στις ανάγκες του εκάστοτε προβλήματος. Οι λίστες αυτές μπορούν να εμπλουτιστούν είτε με λέξεις με μεγάλη συχνότητα εμφάνισης (Term Frequence High), είτε με λέξεις που εμφανίζονται μόνο μία φορά στο κείμενο (Term Frequence - TF1), είτε με λέξεις που έχουν χαμηλό βάρος σημασίας (Inverse Document Frequency - IDF), η επιλογή εξαρτάται ανά περίπτωση. Για παράδειγμα, λέξεις που εμφανίζονται πολύ συχνά σε κείμενα κριτικών είναι αξιοσημείωτες και χρήζουν προσοχής καθώς γίνεται συχνή αναφορά, ενώ αντίθετα λέξεις που εμφανίζονται μια φορά σε κάθε κριτική είναι προτιμότερο να εξαλείφονται γιατί αναμένεται να μην επηρεάζουν το νόημα.
- **Κανονικοποίηση**: η μείωση κάθε λέξης με οποιαδήποτε μορφή στη μικρότερή της μορφή, θεωρώντας ότι δεν ενδιαφέρει η γραμματική αξία της λέξης. Στον τομέα της προ - επεξεργασίας των δεδομένων υπάρχουν οι μέθοδοι λημματοποίηση (lemmatization) και αποκοπή (stemming) με κοινό στόχο την μείωση του όγκου των κειμένων προς ανάλυση.

Κρατώντας μόνο το λήμμα ή το πρόθεμα των λέξεων στο λεξικό εξαλείφεται μεγάλος όγκος επαναλαμβανόμενων λέξεων που προέρχονται από την ίδια βασική λέξη, πράγμα που βοηθάει πολύ την ακρίβεια των ταξινομητών. Οι διαδικασίες που ακολουθούν οι αλγόριθμοι για να επιτύχουν τον κοινό αυτό στόχο είναι διαφορετικές.

- Η λημματοποίηση προσπαθεί να βρει τη βασική κλιτική μορφή της λέξης η οποία ονομάζεται λήμμα συνήθως μέσα από κάποιο λεξικό, δηλαδή εξαλείφει τον γραμματικό χρόνο της λέξης και όλες τις κλιτικές μορφές. Για παράδειγμα, οι λέξεις «*plays*», «*played*», «*playing*» είναι μορφές του ίδιου λήμματος «*play*».
- Η αποκοπή διατηρεί την «ρίζα» της λέξης (παράγωγο) που ονομάζεται πρόθεμα, αποκόπτοντας τις καταλήξεις από τις διάφορες μορφές που μπορούμε να την συναντήσουμε. Για παράδειγμα, οι λέξεις «*uncontrollably*», «*controlled*», «*controls*», «*uncontrollable*» είναι μορφές του ίδιου προθέματος «*control*». Η διαδικασία αυτή αν και υλοποιείται ευκολότερα και είναι γρηγορότερη από την λημματοποίηση, είναι πιο επιρρεπείς σε σφάλματα (Over - stemming, Under - stemming) διότι δεν χρησιμοποιεί γνώση περιεχομένου αλλά λειτουργεί ανεξάρτητα για κάθε νέα λέξη. Για το λόγο αυτό δεν μπορεί να διακρίνει λέξεις που μοιάζουν μεταξύ τους αλλά έχουν διαφορετικό νόημα. Συνεπώς, υπάρχουν διάφορα είδη αλγορίθμων αποκοπής (stemmers) όπως με Πίνακα Αναζήτησης (Table Look Up Approach), σύμφωνα με τον αριθμό των χαρακτήρων που ακολουθούν μια λέξη σε ένα κείμενο (Successor Variety) και σύμφωνα με τους *N* χαρακτήρες που είναι ο ένας δίπλα στον άλλον (N - Gram), δηλαδή κυρίως rule - based προσεγγίσεις. Ο πιο συνηθισμένος αλγόριθμος αποκοπής (stemmer) της αγγλικής γλώσσας και αυτός που έχει αποδειχθεί επανειλημμένα ότι είναι εμπειρικά πολύ αποτελεσματικός, είναι ο αλγόριθμος του Porter (1980) [46]. Εναλλακτικά, οι αλγόριθμοι που υπάρχουν είναι ο παλαιότερος One - pass Lovins Stemmer [47] του Lovins (1968) και ο νεότερος Paice/Husk Stemmer [48] του Paice (1990).
- Μια ακόμη μέθοδος κανονικοποίησης είναι η μετατροπή όλων των γραμμάτων σε πεζά με στόχο ίδιες λέξεις απλά γραμμένες διαφορετικά, με έναν τουλάχιστον κεφαλαίο χαρακτήρα, να θεωρούνται μία. Για παράδειγμα οι λέξεις «*design*» και «*Design*» και «*DESIGN*» μετατρέπονται στην λέξη «*design*» και θεωρούνται ένα token, καθώς έχουν την ίδια σημασιολογική έννοια.

- Η χρήση μιας κοινής κωδικοποίησης (π.χ. UTF-8) σε όλα τα κείμενα είναι μια χρήσιμη λειτουργία της ευρετηρίασης, διότι μετατρέπει ειδικούς χαρακτήρες που πιθανώς να

[46] Διαθέσιμο από: https://tartarus.org/martin/PorterStemmer/

[47] Διαθέσιμες 2 υλοποιήσεις από: http://snowball.tartarus.org/algorithms/lovins/stemmer.html, https://www.cs.waikato.ac.nz/~eibe/stemmers/

[48] Διαθέσιμο από: http://www.comp.lancs.ac.uk/computing/research/stemming/

εντοπισθούν στα κείμενα κριτικών (που κατά κανόνα συλλέγονται ηλεκτρονικά) σε μια πρότυπη κωδικοποίηση ώστε να αναγνωριστούν κατάλληλα. Ενδεικτικά παραδείγματα:

& ➔ & , < ➔ < , > ➔ >

- Η αφαίρεση ή αντικατάσταση των σημείων στίξης (μεμονωμένων και επαναλαμβανόμενων) στις περιπτώσεις ανάλυσης συναισθήματος σε επίπεδο λέξης, όπου έχει προηγηθεί ο διαχωρισμός των κειμένων σε προτάσεις οπότε δεν χρειάζονται πλέον σαν πληροφορία, εκτός των περιπτώσεων που λαμβάνονται υπόψη στον εντοπισμό του συναισθήματος. Για παράδειγμα, τα 3 θαυμαστικά στη σειρά *(!!!)* θα μπορούσαν να εκφράζουν ενθουσιασμό, κατά συνέπεια σε αυτή τη περίπτωση δεν θα πρέπει να αγνοηθούν, απεναντίας θα πρέπει να ληφθούν σοβαρά υπόψη (μεγάλη βαρύτητα), επομένως κι αυτή η μέθοδος επιλέγεται ανά περίπτωση.
- Μαρκάρισμα λέξεων που θέλουν να δώσουν έμφαση στο λόγο, όπως επιμηκυμένες λέξεις (Elongated Words), δηλαδή λέξεις με επαναλαμβανόμενους χαρακτήρες και λέξεις γραμμένες αποκλειστικά με κεφαλαία γράμματα, οι οποίες μπορούν να βοηθήσουν στον εντοπισμό της έντασης (Intensity) του συναισθήματος. Για παράδειγμα, η λέξη «*very*» γραμμένη ως «*VERY*» ή «*veeeeery*» ή «*veryyyyyy*» θέλει να εκφράσει μεγάλο ενθουσιασμό άρα θα πρέπει να θεωρείται με μεγαλύτερη ένταση από ότι η λέξη «*very*».
- Η αφαίρεση επιθέτων και επιρρημάτων συγκριτικού και υπερθετικού βαθμού, διατηρώντας μόνο αυτά θετικού βαθμού. Πρέπει να δοθεί προσοχή στον εντοπισμό τόσο της μονολεκτικής (με τις καταλήξεις *–er* και *–est*), όσο και της περιφραστικής (με τα επιρρήματα more/less και most/least) μορφής των βαθμών σύγκρισης των επιθέτων και επιρρημάτων.
- Μαρκάρισμα αρνητικών λέξεων, μια εξίσου σημαντική διαδικασία καθώς οι λέξεις αυτές αποδίδουν εντελώς αντίθετο νόημα στο υπό εξέταση κείμενο.

Τα tokens που προκύπτουν έπειτα των επιλεγμένων μεθόδων προ - επεξεργασίας στα αρχικά δεδομένα, αποτελούν την αρχική μορφή του λεξιλογίου της συλλογής που θα χρησιμοποιηθεί σαν είσοδος στους αλγορίθμους Μηχανικής Μάθησης για τον εντοπισμό του συναισθήματος. Το λεξιλόγιο αυτό μπορεί κατόπιν να εμπλουτιστεί με επιπλέον πληροφορίες όπως η ύπαρξη ή όχι του token ή τον αριθμό εμφάνισής του σε κάθε πρόταση, κριτική ή / και προϊόν ακόμη και συνολικά στη συλλογή.

2.4.3 Εξαγωγή λέξεων – κλειδιών

Η διαδικασία εξαγωγής λέξεων – κλειδιών από τα προ - επεξεργασμένα πλέον δεδομένα για μια οντότητα μπορεί να έχει τις παρακάτω προσεγγίσεις, όπως εντοπίζονται στην βιβλιογραφία του συγκεκριμένου ερευνητικού πεδίου.

2.4.3.1 Βάσει συχνότητας εμφάνισης

Στηριζόμενη στην παραδοχή ότι αυτές εκφράζονται, κατά ένα αξιοσέβαστο ποσοστό (60% - 70%), όπως υποστηρίζουν οι Liu κ.α. (2007) μέσα από τα συχνότερα χρησιμοποιούμενα ουσιαστικά της πρότασης. Αυτά μπορεί να αποτελούνται και από περισσότερες της μια λέξης (multi - word aspect term) (π.χ. *φωτεινότητα οθόνης*), πράγμα που καθιστά την διαδικασία ακόμη δυσκολότερη. Ένας απλός αλλά αποτελεσματικός τρόπος για την ανίχνευση αυτών των σύνθετων ουσιαστικών είναι να οριστεί ένα ελάχιστο κατώφλι στο ποσοστό συν - εμφάνισης (co - occurence) των πιθανών λέξεων - κλειδιών στις προτάσεις ώστε να θεωρούνται σαν μια ενιαία λέξη - κλειδί. Για τον υπολογισμό της συν - εμφάνισης μπορούν να χρησιμοποιηθούν:

- το μέτρο PMI (Pointwise Mutual Information) το οποίο μετρά το βαθμό στατιστικής εξάρτησης δύο όρων

$$PMI(x_1, x_2) = log_2(\frac{P(x_1 \wedge x_2)}{P(x_1)\,P(x_2)})$$

  όπου $P(x_1 \wedge x_2)$ είναι η πραγματική πιθανότητα συν - εμφάνισης των όρων x_1 και x_2, και $P(x_1)\,P(x_2)$ η πιθανότητα συν-εμφάνισης των δύο όρων εάν είναι στατιστικά ανεξάρτητες.

- η απόσταση βάσει Λανθάνουσας Σημασιολογικής Ανάλυσης (Latent Semantic Analysis - LSA) που συλλαμβάνει το νόημα των λέξεων με στατιστικούς υπολογισμούς σε μια συλλογή κειμένων, από τους Turney και Littman (2003). Η βασική ιδέα είναι να βρεθεί ο λόγος των κείμενων στα οποία εμφανίζεται μία συγκεκριμένη λέξη προς αυτά που δεν εμφανίζονται, παρέχοντας ένα σύνολο αμοιβαίων περιορισμών που καθορίζουν την ομοιότητα της σημασίας των λέξεων ή συνόλων αυτών. Η ικανότητα της LSA έχει εδραιωθεί με ποικίλους τρόπους και στην απεικόνιση της ανθρώπινης γνώσης. Για παράδειγμα στην απόκτηση γνώσης που συμπίπτει με αυτή των ανθρώπων σε ένα συνηθισμένο λεξιλόγιο, όπως και στην μίμηση της ανθρώπινης συμπεριφοράς στην σύνταξη των λέξεων.

  Επομένως, ο αλγόριθμος της LSA χαρακτηρίζεται ως μια αυτοματοποιημένη στατιστική τεχνική για την εύρεση συσχετίσεων μεταξύ λέξεων, όπως αναφέρουν οι Landauer κ.α. (1998). Χρησιμοποιεί προκατασκευασμένα λεξικά γνώμης, βάσεις πληροφοριών, σημασιολογικά δίκτυα, γραμματικές, συντακτικούς διαπερατές ή μορφολογίες και παίρνει σαν είσοδο μόνο απλό κείμενο χωρισμένο σε λέξεις, η καθεμία ορισμένη μοναδικά και χωρισμένες σε κομμάτια προτάσεων και παραγράφων. Τα βήματα του αλγορίθμου LSA έχουν ως εξής:

  1. αντιστοίχιση κειμένου σε έναν πίνακα. Κάθε γραμμή δηλώνει μια μοναδική λέξη και κάθε στήλη ένα κείμενο. Κάθε κελί περιέχει την πιθανότητα με την οποία κάθε λέξη εμφανίζεται σε ένα κείμενο.

2. μεταμόρφωση των κελιών, στην οποία κάθε συχνότητα κελιού ζυγίζεται από μια συνάρτηση που αντιστοιχεί στην σημαντικότητα της λέξης στο συγκεκριμένο κείμενο και στον βαθμό που η λέξη αυτή παρέχει πληροφορία.
3. εφαρμόζει μια αποσύνθεση μοναδικής τιμής (Singular Value Decomposition - SVD) στον πίνακα, μια μορφή ανάλυσης παραγόντων, σύμφωνα με την οποία ένας ορθογώνιος πίνακας μετατρέπεται σε γινόμενο τριών άλλων πινάκων. Ο πρώτος πίνακας περιγράφει τις οντότητες γραμμών σαν διανύσματα παραγόμενα από έναν ορθοκανονικό παράγοντα τιμών. Ο δεύτερος περιγράφει τις οντότητες στήλης με τον ίδιο τρόπο και ο τρίτος είναι ένας διαγώνιος πίνακας που περιέχει αύξουσες τιμές, έτσι ώστε όταν και οι τρεις πίνακες πολλαπλασιαστούν, ξαναδημιουργείται ο πρώτος πίνακας. Υπάρχει μαθηματική απόδειξη ότι κάθε πίνακας μπορεί να αποσυνδεθεί τέλεια, χρησιμοποιώντας όχι περισσότερους παράγοντες από την μικρότερη διάσταση του αρχικού πίνακα. Όταν λιγότεροι από τον απαραίτητο αριθμό παραγόντων χρησιμοποιούνται, ο ανακατασκευασμένος πίνακας είναι καλύτερα καθορισμένος. Μπορεί να μειωθεί η διάσταση της λύσης απλά διαγράφοντας συντελεστές του διαγώνιου πίνακα, συνήθως αρχίζοντας από τον μικρότερο. Ο τελικός αυτός πίνακας που δημιουργείται είναι ο πίνακας συν – εμφάνισης.

Σαν τελικό συμπέρασμα στην επιλογή της μεθόδου εύρεσης της συν – εμφάνισης και σύμφωνα με τον Turney (2001) το PMI σε σύγκριση με το LSA είναι προτιμότερη μέθοδος.

Μερικά μοντέλα εύρεσης των συχνότερα χρησιμοποιούμενων ουσιαστικών σαν λέξεις / φράσεις - κλειδιά των προτάσεων κριτικών είναι των: Hu και Liu (2004), Popescu και Etzioni (2005), Blair-Goldensohn κ.α. (2008), Ku κ.α. (2006), Moghaddam και Ester (2010), Scaffidi κ.α. (2007), Zhu κ.α. (2009), Long κ.α. (2010).

2.4.3.2 Βάσει συντακτικού

Αξιοποιώντας τη σχέση των λέξεων που εκφράζουν γνώμη και των λέξεων στόχους (αυτές για τις οποίες εκφράζεται η γνώμη), καθώς συνήθως οι λέξεις γνώμης είναι γνωστές μέσα από γλωσσικούς πόρους, όπως λεξικά συνωνύμων (περιγράφονται στο Κεφάλαιο 2.4.4.3.1.1) και σε συνδυασμό με ορισμένους κανόνες που μπορούν να χρησιμοποιηθούν για να εξάγουν επιπλέον λέξεις στόχους. Για παράδειγμα, σαν υποψήφια λέξη - κλειδί θα μπορούσε να θεωρηθεί το κοντινότερο ουσιαστικό σε κάποιο επίθετο χαρακτηρισμένου ως λέξη γνώμης σύμφωνα με τον κανόνα ότι τα επίθετα πριν από τα ουσιαστικά σε μία πρόταση είναι αυτά που δίνουν το συναίσθημα για το ουσιαστικό. Χαρακτηριστικά παραδείγματα τέτοιων μεθόδων είναι των Hu και Liu (2004), Blair-Goldensohn κ.α. (2008), Zhuang κ.α. (2006), Somasundaran και Wiebe (2009), Kobayashi κ.α. (2006), Qiu κ.α. (2011), Wu κ.α. (2009), Kessler και Nicolov (2009), Kamps κ.α. (2004), Takamura κ.α. (2005).

2.4.3.3 Βάσει προγενέστερης γνώσης

Αξιοποιώντας προγενέστερη γνώση (βάσει γνώσεων) στην διαδικασία ανακάλυψης των λέξεων – κλειδιών. Για παράδειγμα, οι Fahrni και Klenner (2008) αξιοποιούν την ιεραρχία κατηγοριών της Wikipedia [49] ως βοήθημα στην αναγνώριση λέξεων – κλειδιών, ενώ οι Thet κ.α. (2008) αναπτύσσουν εξειδικευμένους κανόνες αναγνώρισης εκφράσεων ως λέξεις – κλειδιά στο πεδίο των κριτικών ταινιών, που περιλαμβάνουν, μεταξύ άλλων, αναγνώριση ονοματικών οντοτήτων (NER) [50] και επίλυση συναναφορών [51].

2.4.3.4 Βάσει εποπτευόμενης Μηχανικής Μάθησης

Χρησιμοποιώντας εποπτευόμενη μάθηση η οποία προϋποθέτει την χειροκίνητη επισήμανση των κειμένων για την εκπαίδευση του μοντέλου, με τις επικρατέστερες μεθόδους να στηρίζονται στην διαδοχική μάθηση (sequential learning), όπως οι Hidden Markov Models (HMM) στο Rabiner (1989) και Conditional Random Fields (CRF) των Lafferty κ.α. (2001), αλλά και σε μεθόδους ταξινόμησης βασισμένες σε δέντρα αποφάσεων (Decision Trees) όπως οι Kobayashi κ.α. (2007) , καθώς και Διανυσματικές Μηχανές Υποστήριξης (SVM) όπως οι Yu κ.α. (2011).

2.4.3.5 Βάσει θεματικής μοντελοποίησης

Η θεματική μοντελοποίηση (topic modeling) είναι μια μέθοδος που σαν έξοδο έχει ένα σύνολο κλάσεων (topics), όπου κάθε κλάση σχηματίζει μία θεματική ενότητα. Η έξοδος αυτή μπορεί να επιτευχθεί είτε με μέθοδο μη εποπτευόμενης μάθησης μέσω μιας κατανομής πιθανοτήτων των λέξεων του εγγράφου, είτε βασιζόμενη σε μια ομάδα δυαδικών ταξινομητών, ένας για κάθε κλάση όπου η τελική απόφαση συναρμολογείται από τις αποφάσεις των επιμέρους ταξινομητών. Οι θεματικές ενότητες που προκύπτουν δεν εμφανίζονται πάντα σαν λέξεις – κλειδιά μέσα στη πρόταση (π.χ. ανίχνευση κατηγορίας «*κόστος*» μέσω του χαρακτηριστικού «*τιμή*» που αναφέρεται σε μια κριτική για το προϊόν «*iPad Tab*») και γενικότερα παρατηρούνται πιο αραιές από ότι οι λέξεις – κλειδιά όπως παρατηρούν οι Laskari και Sanampudi (2016). Η θεματική μοντελοποίηση όχι μόνο ανακαλύπτει λέξεις – κλειδιά αλλά και συνώνυμες αυτών.

[49] Η Wikipedia είναι μια δωρεάν ηλεκτρονική εγκυκλοπαίδεια.

[50] Η αναγνώριση ονοματικών οντοτήτων (Named Entity Recognition - NER) είναι μια υποδιεργασία της εξαγωγής πληροφορίας που επιδιώκει να εντοπίσει και να κατατάξει το όνομα οντότητας σε προκαθορισμένες κατηγορίες, όπως τα ονόματα προσώπων, οργανώσεων, τις εκφράσεις της εποχής, νομισματικές αξίες, ποσοστά, κ.λπ.

[51] Η επίλυση συναναφορών (coreference resolution) είναι η διαδικασία εντοπισμού όλων των εκφράσεων που αναφέρονται στην ίδια οντότητα σε ένα κείμενο [12].

Σε αυτό το σημείο αξίζει να σημειωθεί ότι οι λέξεις – κλειδιά που συγχωνεύονται σε μία θεματική ενότητα δεν είναι απαραίτητο να είναι κοντινά συνώνυμες (π.χ. οι λέξεις – κλειδιά «*σχεδίαση*», «*χρώμα*», «*αίσθηση*», «*υλικό*» σε μια κατηγορία). Πολλά συνώνυμα εξαρτώνται από τον τομέα στον οποίο αναφέρονται, όπως υποστηρίζουν και οι Liu κ.α. (2005). Για παράδειγμα, οι λέξεις – κλειδιά «*ταινία*» και «*εικόνα*» θεωρούνται συνώνυμες σε κριτικές ταινιών, αλλά όχι σε κριτικές για φωτογραφικές μηχανές καθώς σε αυτό το τομέα η λέξη – κλειδί «*εικόνα*» είναι πιο πιθανό να είναι συνώνυμη με την λέξη – κλειδί «*φωτογραφία*», ενώ η λέξη – κλειδί «*ταινία*» με τη λέξη – κλειδί «*βίντεο*». Μια εξίσου σημαντική παρατήρηση του Liu (2012), είναι ότι αρκετές λέξεις – κλειδιά που συναθροίζονται σε θεματική ενότητα δεν είναι ούτε γενικά, ούτε συγκεκριμένου τομέα συνώνυμες, αλλά ακόμη και αντώνυμες (π.χ. οι λέξεις – κλειδιά «*ακριβό*» και «*φθηνό*» μπορεί να ανήκουν στην ίδια θεματική ενότητα «*κόστος*»). Συνεπώς, εστιάζοντας στην θεματική μοντελοποίηση, μια προσέγγιση για πολλαπλές διασπορές θα ήταν προτιμότερη από τις σύνηθες προσεγγίσεις που βασίζονται στις συνώνυμες λέξεις – κλειδιά των λεξικών όπως υποστηρίζουν οι Pavlopoulos και Androutsopoulos (2014).

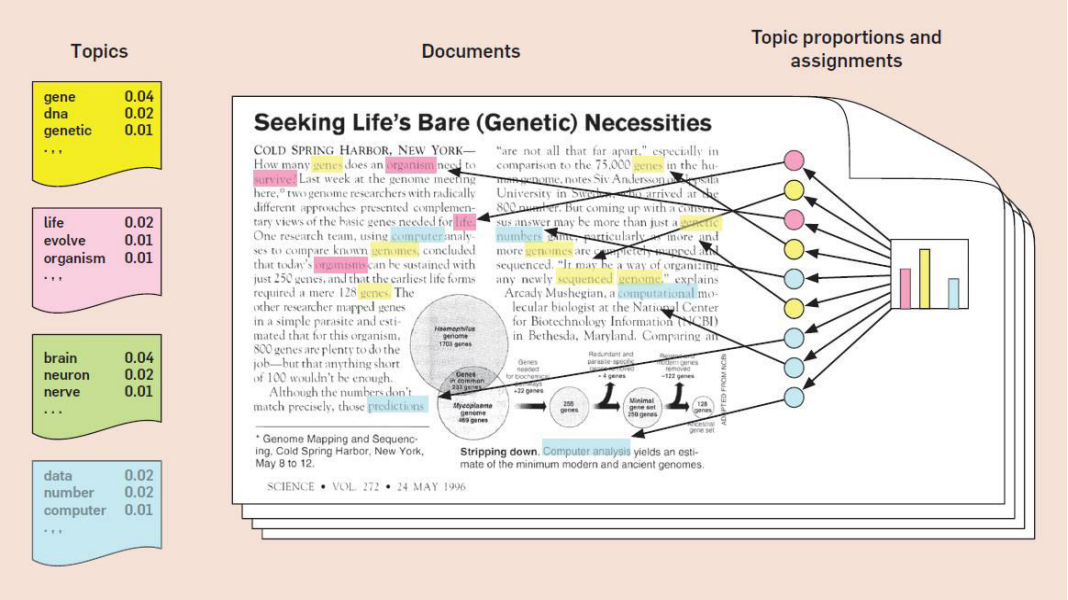
Ορισμένες τεχνικές επίλυσης στο εν λόγω πρόβλημα των **συνωνύμων** που έχουν ως στόχο την κατηγοριοποίηση των λέξεων – κλειδιών σε κατηγορίες από όπου προκύπτουν οι θεματικές ενότητες είναι: Carenini κ.α. (2005), Yu κ.α. (2011), Zhai κ.α. (2010), Guo κ.α. (2009), Andrzejewski κ.α. (2009), Mukherjee και Liu (2012).

Τεχνικά, τα μοντέλα θεματικών ενοτήτων είναι ένας τύπος γραφικών μοντέλων που βασίζονται σε Μπεϊσιανά (Bayesian) δίκτυα και παρόλο που χρησιμοποιούνται για τη μοντελοποίηση και την εξαγωγή θεματικών ενοτήτων από συλλογές κειμένων, μπορούν να επεκταθούν ώστε να μοντελοποιήσουν ταυτόχρονα και άλλα είδη πληροφοριών όπως το συναίσθημα. Εκεί, μπορεί να σχεδιαστεί ένα κοινό μοντέλο το οποίο θα μοντελοποιεί ταυτόχρονα τόσο τις λέξεις γνώμης όσο και τις λέξεις στόχους, λόγω της παρατήρησης ότι κάθε γνώμη έχει στόχο αλλά και να ομαδοποιήσει συνώνυμες λέξεις στόχους.

Αν και η θεματική μοντελοποίηση έχει ικανοποιητικά αποτελέσματα, δεν προτείνεται για εφαρμογές Ανάλυσης Συναισθήματος διότι απαιτεί μεγάλο αριθμό δεδομένων και είναι πολύ χρονοβόρα διαδικασία. Πιο συγκεκριμένα, λόγω χρήσης των παραγωγικών μοντέλων (generative models) ως επίλυση στο πρόβλημα της εξαγωγής λέξεων – κλειδιών, η τεχνική αυτή δεν προτιμάται, διότι τα θέματα ή θεματικές ενότητες (topics) έχουν καθολικό χαρακτήρα και χαρακτηρίζουν ολόκληρα τα έγγραφα ενώ οι λέξεις – κλειδιά χαρακτηρίζουν μεμονωμένες προτάσεις ή φράσεις, έτσι ακυρώνεται η βασική παραδοχή των παραγωγικών μοντέλων ότι ένα έγγραφο αποτελεί μείγμα ενός σχετικά μικρού αριθμού θεμάτων.

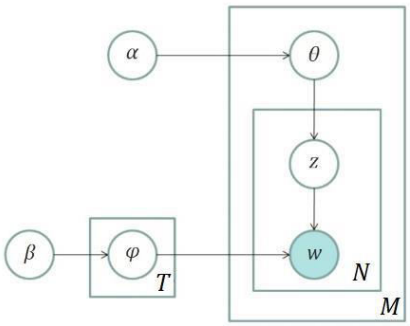
Παρά όλα αυτά, παρακάτω αναφέρονται οι δημοφιλέστερες τεχνικές υλοποίησης θεματικής μοντελοποίησης, καθώς είναι μία ευρέως χρησιμοποιούμενη και δημοφιλής μέθοδος:

- **PLSA** (Probabilistic Latent Semantic Analysis - Πιθανοτική Λανθάνουσα Σημασιολογική Ανάλυση), του Hofmann (1999) γνωστή και ως PLSI (Probabilistic Latent Semantic Indexing - Πιθανοτική Λανθάνουσα Σημασιολογική Ευρετηρίαση) στην Ανάκτηση Πληροφοριών, μια στατιστική τεχνική για την ανάλυση συν-εμφανιζόμενων λέξεων – κλειδιών. Στην πράξη, δημιουργεί μια χαμηλής διάστασης αναπαράσταση των υποψήφιων λέξεων – κλειδιών όσον αφορά τη συγγένειά τους με άλλες κρυφές λέξεις – κλειδιά, όπως και στην Λανθάνουσα Σημασιολογική Ανάλυση (LSA), από την οποία εξελίχθηκε το PLSA. Σε σύγκριση με την τυπική LSA η οποία προέρχεται από την Γραμμική Άλγεβρα και μειώνει τους πίνακες συν - εμφάνισης μέσω αποσύνθεσης μιας μοναδικής τιμής, η PLSA βασίζεται σε ένα μείγμα αποσύνθεσης που προέρχεται από ένα μοντέλο λανθάνουσας κλάσης, το οποίο οδηγεί σε μια πιο προσεγμένη προσέγγιση με σταθερή βάση στη στατιστική.
- **LDA** (Latent Dirichlet Allocation - Λανθάνουσα κατανομή Dirichlet), των Blei κ.α. (2003) Griffiths και Steyvers (2003) και Griffiths και Steyvers (2007), ένα μοντέλο που παράγει πιθανότητες για συλλογές διακριτών δεδομένων όπως κείμενα κριτικών, τα οποία μοντελοποιούνται ως μια κατανομή θεματικών ενοτήτων και κάθε θεματική ενότητα με τη σειρά της, ως κατανομή λέξεων. Οι θεματικές ενότητες θεωρούνται κρυφές (hidden) ή αλλιώς λανθάνουσες (latent) και οι πιο συχνά εμφανιζόμενες λέξεις σε μια θεματική ενότητα είναι και αυτές που την αντιπροσωπεύουν. Διαθέτει πολλές παραλλαγές η κάθε μια υλοποιημένη για την επίλυση διαφορετικού προβλήματος προσαρμοζόμενη στις εκάστοτε διαθέσιμες συλλογές δεδομένων. Το κύριο πλεονέκτημα του μοντέλου, άλλωστε, είναι το υψηλό ποσοστό προσαρμοστικότητάς του. Όσον αφορά την επίλυση του δικού μας προβλήματος, το LDA προτιμάται εξαιτίας του Μπεϊσιανού (Bayesian) χαρακτήρα του που αποφεύγει το πρόβλημα της υπερ - προσαρμογής (overfitting) των αποτελεσμάτων στα δεδομένα εκπαίδευσης. Ωστόσο, η εφαρμογή του μοντέλου για την ανακάλυψη λέξεων - κλειδιών σε κριτικές αντιμετωπίζει αρκετές δυσκολίες καθώς είναι σχεδιασμένο να αναγνωρίζει καθολικά θέματα που καλύπτουν ένα ολόκληρο έγγραφο, σε αντίθεση με τις λέξεις – κλειδιά, το επίπεδο ανάλυσης όπου εστιάζουμε, που έχουν τοπικό χαρακτήρα και αναφέρονται σε συγκεκριμένες προτάσεις.



Εικόνα 16: Επεξήγηση της διαίσθησης πίσω από το LDA, Blei (2012)

Η επεξήγηση της διαίσθησης πίσω από το LDA σύμφωνα με έναν από τους δημιουργούς του, Blei (2012), φαίνεται στην παραπάνω Εικόνα 16 ενώ η γραφική αναπαράσταση του μοντέλου LDA φαίνεται στην Εικόνα 17 με μορφή plate notation [52].



Εικόνα 17: Γραφική αναπαράσταση LDA, Blei κ.α. (2003)

όπου ο σκιασμένος κόμβος είναι η παρατηρούμενη μεταβλητή και οι μη-σκιασμένοι είναι οι λανθάνουσες μεταβλητές. Τα βέλη αναπαριστούν τις εξαρτήσεις και τα ορθογώνια τις διαδικασίες επαναλαμβανόμενης δειγματοληψίας. Το M αντιπροσωπεύει το πλήθος των κειμένων στη συλλογή, το N το πλήθος των λέξεων σε κάθε κείμενο και το T το πλήθος των θεματικών ενοτήτων. Η τιμή z αντιστοιχεί στη θεματική ενότητα από την οποία εξάγεται μια συγκεκριμένη λέξη w. Οι ανά-κείμενο πολυωνυμικές κατανομές θεματικών ενοτήτων δίνονται από το θ, ενώ το φ δίνει τις ανά- θεματική ενότητα πολυωνυμικές κατανομές λέξεων. Προγενέστερες (prior) κατανομές Dirichlet τοποθετούνται πάνω από αυτές τις κατανομές.

[52] Η plate σημειογραφία (notation) είναι μια μέθοδος αντιπροσώπευσης των μεταβλητών που επαναλαμβάνονται σε ένα γραφικό μοντέλο [24].

Το μοντέλο LDA ακολουθώντας παραδοχές όπως το ότι κάθε κείμενο μπορεί να αποτελείται από πολλαπλές θεματικές ενότητες και ότι κάθε θεματική ενότητα μπορεί να αποτελείται από πολλαπλές λέξεις, δημιουργούνται ανάγκες μοντελοποίησης αυτών των συσχετίσεων και αυτές καλύπτονται από τις κατανομές Dirichlet που είναι στη φύση τους να αντιμετωπίζουν τέτοιες πολλαπλότητες, η Dirichlet θεωρείται μια πολύ - μεταβλητή κατανομή. Αυτές οι Dirichlet κατανομές παραμετροποιούνται από τα α και β αντίστοιχα και δεν είναι εμφανή από τα δεδομένα, για το λόγο αυτό αποκαλούνται και λανθάνουσες (latent) ή κρυφές (hidden). Όσο πιο χαμηλές οι τιμές των α και β τόσο πιο λίγες οι θεματικές ενότητες ανά κείμενο και οι λέξεις ανά θεματική ενότητα αντίστοιχα.

Πιο συγκεκριμένα για *N* αριθμό θεματικών ενοτήτων έχουμε τη σχέση:

$$p(w) = \int_{\theta} \left(\prod_{n=1}^{N} \sum_{z_{n=1}}^{k} p(w_n | z_n; \beta) p(z_n \theta) \right) p(\theta; \alpha) d\theta$$

Όπου α είναι η προγενέστερη παράμετρος κατανομής ανά λέξη και η β είναι η προγενέστερη Dirichlet παράμετρος κατανομής ανά θεματική ενότητα.

Τα βήματα του αλγορίθμου LDA είναι τα εξής:

1. Επιλέγονται οι τιμές για τις υπερ - παραμέτρους α, β και για το πλήθος των θεματικών ενοτήτων *T*. Υπάρχουν δε αλγόριθμοι οι οποίοι μπορούν και εκτιμούν τις παραμέτρους α και β κατά το στάδιο εκπαίδευσης του μοντέλου. Οι τιμές των α και β βασίζονται στο *T* καθώς και στο μέγεθος του λεξικού. Ιδανικές συνήθως επιλογές αυτών είναι $\alpha = 50/T$ και $\beta = 0.01$, στα Griffiths και Steyvers (2003) και Griffiths και Steyvers (2007).
2. Επιλέγεται ο αριθμός των λέξεων *N* για κάθε κείμενο.
3. Για κάθε λέξη των κειμένων γίνεται δειγματοληψία *z* από $\theta(j)$, όπου *j* είναι ο δείκτης του τρέχοντος κειμένου και δειγματοληψία w από $\varphi(z)$.
4. Υπολογίζεται η πιθανότητα $P(z|w)$ δεδομένων των α, β και *T* για κάθε λέξη των κειμένων, καθώς οι πιο αντιπροσωπευτικές λέξεις των θεματικών ενοτήτων καθορίζονται από τις υψηλότερες πιθανότητες αυτών. Ο υπολογισμός αυτός θεωρείται ένα πρόβλημα δύσκολα επιλύσιμο για αυτό προτείνονται οι τεχνικές μεταβολικής προσέγγισης EM (Expectation-Maximization) των Blei κ.α. (2003) και η δειγματοληψία Gibbs στο Griffiths και Steyvers (2004).
5. Εκτιμώνται οι κατανομές φ και θ για κάθε θεματική ενότητα και κάθε κείμενο. Οι κατανομές θ των θεματικών ενοτήτων είναι αυτές που διαμορφώνουν τη βάση της μεθόδου συσταδοποίησης των κειμένων σε θεματικές ενότητες με LDA.

Ένα παράδειγμα χρήσης του LDA μοντέλου ίσως βοηθήσει στην καλύτερη κατανόησή του. Αν υποθέσουμε ότι έχουμε τις τρεις προτάσεις:

1. *I like to* ***eat fish*** *and* ***vegetables***.
2. Cats ***eat fish***.
3. Fish are pets.

Παρατηρούμε ότι οι έντονες (bold) λέξεις μιλούν για *τρόφιμα* ενώ οι υπογραμμισμένες για *ζώα*. Το LDA είναι σε θέση να ταξινομήσει τις λέξεις των προτάσεων σε θέματα, έτσι οι προτάσεις αυτές σύμφωνα με το LDA, περιέχουν τα ακόλουθα θέματα:

Πρόταση 1: 100% τρόφιμα
Πρόταση 2: 66.7% ζώα, 33.3% τρόφιμα
Πρόταση 3: 100% ζώα

Και κάθε θέμα περιέχει τις ακόλουθες λέξεις των προτάσεων:

Θέμα "*τρόφιμα"*: 40% eat, 40% fish, 20% vegetables
Θέμα "*ζώα"*: 80% fish, 20% cats

Οι θεματικές ενότητες στα μοντέλα θεματικών ενοτήτων είναι οι λέξεις στόχοι στα πλαίσια της Ανάλυσης Συναισθήματος, επομένως, η θεματική μοντελοποίηση μπορεί να εφαρμοστεί και για την εξαγωγή αυτών των λέξεων στόχων. Ωστόσο, οι θεματικές ενότητες μπορεί να καλύπτουν τόσο τις λέξεις στόχους όσο και τις λέξεις γνώμης. Για την ανάλυση του συναισθήματος, αυτές πρέπει να διαχωριστούν. Αυτό, όπως προαναφερθήκαμε, μπορεί να επιτευχθεί επεκτείνοντας το βασικό μοντέλο (π.χ. LDA) με μοντέλα αναγνώρισης συναισθήματος.

2.4.3.6 Βάσει συσταδοποίησης

Εσκεμμένα τελευταία μέθοδος διότι σπανιότερα χρησιμοποιούμενη, η συσταδοποίηση μέσω του αλγορίθμου k- means, ο οποίος διαχωρίζει *n* αντικείμενα σε *k* συστάδες όπως περιεγράφηκε παραπάνω, άρα επιτυγχάνει τον τελικό στόχο που είναι η ανάθεση θεματικής ενότητας στο κείμενο. Προϋποθέτει την μετατροπή των δεδομένων σε διανύσματα, με tf – idf (term frequency – inverse document frequency) τιμές που δείχνουν την συχνότητα εμφάνισης κάθε λέξης σε ολόκληρη τη συλλογή κειμένων, άρα αντικατοπτρίζουν την σημαντικότητα της λέξης στη συλλογή, αλλά και τον αριθμό των κέντρων οπότε και συστάδων που δημιουργούνται γύρω από αυτά.

2.4.4 Ανίχνευση συναισθήματος λέξεων – κλειδιών

Για να επιτευχθεί ο τελικός στόχος εντοπισμού του συνολικού συναισθήματος του κειμένου της κριτικής και κατά επέκταση του προϊόντος, είναι απαραίτητη η ανίχνευση του συναισθήματος των εκάστοτε λέξεων – κλειδιών (ή με όρους Μηχανικής Μάθησης, η πολικότητα των χαρακτηριστικών του διανύσματος), που εξήχθησαν στα προηγούμενα βήματα. Στη διαδικασία αυτή θα πρέπει να λαμβάνεται υπόψιν ότι οι φράσεις που εκφράζουν το συναίσθημα μπορεί να είναι άμεσες αλλά και έμμεσες, όπως έχουμε ήδη αναφέρει. Ο

εντοπισμός του συναισθήματος στις λέξεις - κλειδιά προσεγγίζεται τόσο με τεχνικές εποπτευόμενης μάθησης, όσο και με τεχνικές βασισμένες σε λεξικό γνώμης που αν και δεν εντάσσονται στο πεδίο την Μηχανικής Μάθησης εντοπίζεται να χρησιμοποιούνται σε μεγάλο βαθμό όσον αφορά την επίτευξη του στόχου μας για το λόγο αυτό αναλύονται παρακάτω στο Κεφάλαιο 2.4.4.3.

2.4.4.1 Με μεθόδους εποπτευόμενης Μηχανικής Μάθησης

Η εποπτευόμενη Μηχανική Μάθηση συνήθως επιλέγεται για ταξινόμηση συναισθήματος σε επίπεδο εγγράφων, καθώς τα έγγραφα περιέχουν περισσότερες λέξεις - κλειδιά από ότι πιο εστιασμένα μια πρόταση. Οι τεχνικές εποπτευόμενης μάθησης εξαρτώνται από τα δεδομένα εκπαίδευσης, έτσι εκπαιδεύονται για συγκεκριμένο τομέα και δεν αποδίδουν το ίδιο καλά και σε άλλους άγνωστους τομείς, συνεπώς δεν χρησιμοποιούνται σε μεγάλο φάσμα τομέων. Καθώς η ταξινόμηση συναισθήματος είναι κατά κύριο λόγο ένα πρόβλημα ταξινόμησης κειμένου σε δύο κλάσεις, θετική και αρνητική, οποιαδήποτε μέθοδος εποπτευόμενης μάθησης μπορεί να εφαρμοστεί.

Το Pang κ.α. (2002) ήταν το πρώτο άρθρο που υιοθέτησε αυτήν την προσέγγιση για να ταξινομήσει τις κριτικές ταινιών σε δύο κατηγορίες, θετικές και αρνητικές. Απέδειξαν ότι χρησιμοποιώντας unigrams (n - grams μεγέθους 1) ως χαρακτηριστικά γνωρίσματα η ταξινόμηση απέδωσε αρκετά καλά εκτελώντας είτε Naive Bayes είτε SVM, αν και πειραματίστηκαν και με άλλες επιλογές χαρακτηριστικών τις οποίες δεν προτείνουν. Ορισμένες επιπλέον τεχνικές υλοποίησης εποπτευόμενης μάθησης που χρησιμοποιούνται για ανίχνευση του συναισθήματος των λέξεων – κλειδιών στις κριτικές, είναι: των Wei και Gulla (2010) προτείνοντας ένα ιεραρχικό μοντέλο ταξινόμησης, χρησιμοποιώντας ανάλυση εξάρτησης των λέξεων – κλειδιών στο Jiang κ.α. (2011) και παρομοίως οι Boiy και Moens (2009) και οι Ding κ.α. (2009) και Ganapathibhotla και Liu (2008) που αντιμετωπίζουν τις συγκριτικές προτάσεις.

2.4.4.2 Με μεθόδους μη εποπτευόμενης Μηχανικής Μάθησης

Εν τούτοις, δεδομένου ότι οι λέξεις γνώμης είναι συχνά ο κυρίαρχος παράγοντας για την ταξινόμηση του συναισθήματος στα κείμενα, δεν είναι δύσκολο να φανταστούμε ότι οι λέξεις και φράσεις που χαρακτηρίζουν το συναίσθημα στις προτάσεις μπορούν να χρησιμοποιηθούν για την ταξινόμηση του συναισθήματος με μη εποπτευόμενο τρόπο. Οι τεχνικές μη εποπτευόμενης μάθησης είναι χρήσιμες για συστήματα ABSA που προορίζονται να χρησιμοποιηθούν σε τομείς με ελάχιστες αλλαγές, όπως υποστηρίζουν οι Pavlopoulos και Androutsopoulos (2014). Μια υλοποίηση είναι η μέθοδος του Turney (2002) που ταξινομεί βάσει συντακτικών προτύπων έχοντας εκπονήσει POS tagging.

2.4.4.3 Με χρήση λεξικού

Επιπρόσθετα, υπάρχουν οι προσεγγίσεις βασισμένες σε λεξικό, οι οποίες για την απόδοση του συναισθηματικού προσανατολισμού, χρησιμοποιούν προκατασκευασμένα λεξικά συναισθήματος (sentiment lexicons) (ή λεξικά γνώμης) που περιέχουν τις λέξεις γνώμης του κειμένου, κανόνες σύνθετων εκφράσεων των απόψεων και το δέντρο ανάλυσης. Πολλοί συσχετίζουν αυτού του είδους τις προσεγγίσεις με την μη εποπτευόμενη Μηχανική Μάθηση αλλά δεν ανήκει στο πεδίο της, μολαταύτα λόγω της μεγάλης τους απήχησης και συνεισφοράς στο συγκεκριμένο πρόβλημα εξετάζονται και παρατίθενται. Μερικές ενδεικτικές σχετικές εργασίες που προσεγγίζουν το πρόβλημα με χρήση λεξικού είναι: των Ding κ.α. (2008) που αντιμετωπίζουν συγκριτικές προτάσεις, των Hu και Liu (2004) που συνοψίζουν τα αποτελέσματα των εκφραζόμενων συναισθημάτων όλων των λέξεων σε μια πρόταση για να την χαρακτηρίσουν, των Kim και Hovy (2004) που πολλαπλασιάζουν τα αποτελέσματα των συναισθημάτων των λέξεων και των Taboada κ.α. (2011) που ενσωματώνουν την εντατικοποίηση[53] και διαχειρίζονται την άρνηση για να υπολογίσουν το βαθμό του αισθήματος για κάθε έγγραφο.

Μπορούν να διαχωριστούν σε:

- Dictionary-based λεξικά, τα οποία προϋποθέτουν την παρουσία λέξεων γνώμης σηματοδότησης και κάποιου είδος μηχανισμό μέτρησης για την πρόβλεψη του συναισθήματος. Οι προσεγγίσεις που τα χρησιμοποιούν είναι αρκετά απλές, γρήγορες και ημι - αυτοματοποιημένες αλλά έχουν χαμηλή ακρίβεια. Ενδεικτικές προσεγγίσεις που χρησιμοποιούν Dictionary-based λεξικά είναι των Hu και Liu (2004) και Andreevskaia και Bergler (2008).
- Corpus-based λεξικά, τα οποία παρέχουν πρόσβαση όχι μόνο στις ετικέτες συναισθήματος αλλά σε ένα πλαίσιο το οποίο μπορεί να χρησιμοποιηθεί προς όφελός του αλγορίθμου Μηχανικής Μάθησης. Μπορεί να είναι μια προσέγγιση βασισμένη σε κανόνες ή ανάλυση φυσικής γλώσσας, ή ακόμα και ένας συνδυασμός αυτών. Παρέχουν υψηλή αυτοματοποίηση και ευκολότερη προσαρμογή σε νέες γλώσσες. Επιπλέον, φέρουν και κάποια εξειδίκευση τομέα, που μπορεί να ενημερώσει τον αλγόριθμό για την ποικιλία των ετικετών συναισθήματος για μια λέξη ανάλογα με το πλαίσιο / τομέα. Τα Corpus-based λεξικά είναι είτε στατιστικά που απαιτούν μεγάλο αριθμό δεδομένων για καλύτερη κάλυψη, είτε σημασιολογικά τα οποία δεν είναι ανεξάρτητα τομέα και συμφραζομένων. Μερικές προσεγγίσεις που χρησιμοποιούν Corpus -based λεξικά είναι των Hatzivassiloglou και McKeown (1997), Turney (2002), Yu και Hatzivassiloglou (2003), Ding κ.α. (2008).

[53] Η εντατικοποίηση είναι η ενίσχυση ή η εξομάλυνση της έντασης του συναισθήματος.

- Χειροκίνητα κατασκευασμένα λεξικά γνώμης, τα οποία τείνουν να είναι πιο ακριβή αλλά δεν παρέχουν κανέναν αυτοματισμό και φυσικά είναι μια επίπονη διεργασία που απαιτεί πολύ χρόνο για να στηθεί.

Η προσέγγιση εδώ, ενέχει στην απόδοση μιας βαθμολογίας για κάθε λέξη γνώμης, που εκφράζει το κατά πόσο το νόημά της ταιριάζει σε κάποια προκαθορισμένη κατηγορία συναισθήματος. Η συνηθέστερη κατηγοριοποίηση συναισθήματος ενέχει δύο κλάσεις, την «θετικό συναίσθημα» και την «αρνητικό συναίσθημα». Ωστόσο, εντοπίζονται και λεξικά με περισσότερες και πιο εξειδικευμένες κατηγορίες συναισθημάτων, όπως «χαρά», «ενθουσιασμός», «θυμός», «λύπη» κ.α. Κάθε λέξη προς ανάλυση αναζητείται στο λεξικό γνώμης σημειώνοντας την βαθμολογία της και το συνολικό συναίσθημα του κειμένου προσδιορίζεται από το άθροισμα των βαθμολογιών των επιμέρους λέξεων γνώμης, με την βοήθεια κατωφλίων (thresholds) στις περιπτώσεις πολυεπίπεδης συναισθηματικής κατάταξης. Όπως καταλαβαίνουμε από την παραπάνω περιγραφή, αυτού του είδους η προσέγγιση επίλυσης χρησιμοποιείται ως επί το πλείστον σε προβλήματα της Ανάλυσης Συναισθήματος σε επίπεδο πρότασης ή λέξης. Οι μέθοδοι βασισμένες σε λεξικό, έχει αποδειχθεί ότι αποδίδουν καλύτερα σε μεγαλύτερο αριθμό τομέων, χρησιμοποιώντας ένα λεξικό γνώμης που περιέχει μια λίστα από εκφραστικές λέξεις, σύνθετες εκφράσεις, ιδιωματισμούς, κανόνες απόψεων κ.α.

2.4.4.3.1.1 Λεξικά Γνώμης

Για την χρήση της συγκεκριμένης προσέγγισης στο πρόβλημα της Ανάλυσης Συναισθήματος επιτάσσεται η χρήση ενός λεξικού γνώμης (ή συναισθήματος), έτσι εδώ είναι σκόπιμο να αναφερθούμε σε χαρακτηριστικά παραδείγματα τέτοιων λεξικών και λεξιλογικών βάσεων της αγγλικής γλώσσας που μας απασχολεί και σε αυτή την διπλωματική εργασία. Παραθέτονται σύμφωνα με την ημερομηνία δημοσίευσής τους καθώς η υλοποίηση των περισσότερων στηρίχθηκε στα προηγούμενα. Εντούτοις μπορούν να χρησιμοποιηθούν σαν βάση για την παραγωγή δικών μας λεξικών γνώμης, με την αποδοχή της πιο επίπονης και χρονοβόρας διαδικασίας, αλλά με το πλεονέκτημα να είναι προσαρμοσμένο ακριβώς στις ανάγκες του προβλήματος άρα η μέθοδος που θα το χρησιμοποιήσει να αποφέρει βελτιωμένα αποτελέσματα.

- Το **General Inquirer** Lexicon του Harvard των Stone κ.α. (1966), είναι ένα λεξικό που συνδέει συντακτικές, σημασιολογικές και ρεαλιστικές πληροφορίες σε λέξεις με ετικέτα για το τι μέρος του λόγου είναι σε μορφή υπολογιστικού φύλλου, η οποία είναι η πιο εύκολη μορφή για χρήση με τις περισσότερες υπολογιστικές εφαρμογές.
- Το **Wordnet** αποτελεί το πιο διαδεδομένο εργαλείο στη γλωσσική έρευνα και την εξόρυξη συναισθήματος. Είναι μια λεξικογραφική βάση δεδομένων για την αγγλική

γλώσσα, διαθέσιμη κάτω από ελεύθερη άδεια χρήσης, το οποίο δημιουργήθηκε στο Πανεπιστήμιο Princeton το 1985. Είναι οργανωμένο ως ένα σημασιολογικό δίκτυο λεξικογραφικών λημμάτων που αναπαριστά σχέσεις νοηματικής συνάφειας μεταξύ δύο λημμάτων, όπως υπερνυμία (*is - a*), ολονυμία (*has - a*), μερωνυμία (*part - of*), αντωνυμία (*opposite - of*) κ.α. Επιπρόσθετα, παρέχει ερμηνείες, όπως ένα λεξικό αλλά και λίστες συνωνύμων - αντωνύμων, όπως ένα thesaurus [54]. Αντίθετα με τα συνήθη λεξικά, που είναι οργανωμένα κατά όρους με πολλαπλές ερμηνείες, στο Wordnet η βασική μονάδα οργάνωσης είναι η ερμηνεία (Gloss), η οποία μπορεί να αντιστοιχεί σε πολλαπλούς όρους. Η συλλογή των όρων που εκφράζουν μια συγκεκριμένη ερμηνεία ονομάζεται σύνολο συνωνύμων (Synset). Ένας όρος ανήκει σε τόσα synsets όσες και οι ερμηνείες του και όλοι οι όροι που απαρτίζουν ένα synset ανήκουν σε ένα από τα τέσσερα μέρη του λόγου (ρήμα, ουσιαστικό, επίθετο, επίρρημα) που απαρτίζουν το λεξιλόγιο ανοιχτής κλάσης. Η πιο πρόσφατη έκδοσή του (WordNet 3.1, 2012) περιλαμβάνει 155.287 λέξεις οργανωμένες σε 117.659 synsets και 206.941 σημασιολογικές σχέσεις μεταξύ synsets και λέξεις.

- Η **Linguistic Inquiry & Word Counts** (LIWC) (2001) [55], είναι μια ιδιόκτητη βάση δεδομένων που αποτελείται από πολλές κατηγοριοποιημένες κανονικές εκφράσεις και είναι επί πληρωμή. Οι ταξινομήσεις του συσχετίζονται σε μεγάλο βαθμό με το General Inquirer λεξικό του Harvard.
- Το **Opinion Lexicon**, των Hu και Liu (2004) περιέχει περίπου 6.800 λέξεις (2006 θετικές και 4783 αρνητικές) και ορισμένες αξιοσημείωτες ιδιότητές του είναι ότι συμπεριλαμβάνει κακή ορθογραφία, μορφολογικές παραλλαγές και αργκό. Εφαρμόζουν τις σχέσεις αντωνυμίας – συνωνυμίας του WordNet σε ένα αρχικό σύνολο επιθέτων «σπόρους» (seeds) όπως τα «good» και «bad» για την συναισθηματική αξιολόγηση των λέξεων. Το τελικό αποτέλεσμα είναι μια λίστα θετικών και μια λίστα αρνητικών λέξεων χωρίς καμία άλλη πληροφορία.
- Το **WordNet Affect**, των Strapparava και Valitutti (2004) χρησιμοποιεί μεγαλύτερο αρχικό σύνολο λέξεων «σπόρους», ταξινομημένων σύμφωνα με πέντε κατηγορίες συναισθήματος, «χαρά» «λύπη» «φόβος» «έκπληξη» «αηδία», που θεωρούνται ως οι βασικές που μπορούν να καλύψουν όλα τα συναισθήματα. Έπειτα επεκτείνεται σύμφωνα με το σημασιολογικού γράφο του WordNet, που ουσιαστικά αποτελείται από λέξεις που έχουν συναισθηματική έννοια, χαρακτηρισμένη με μια ετικέτα (a-label). Για παράδειγμα, η ετικέτα «*sensation*» (αίσθηση) δίνει την σημασιολογική έννοια του ουσιαστικού «*coldness*» (ψυχρότητα) με το ρήμα «*feel*» (αισθάνομαι).

[54] Ένα thesaurus (θησαυρός) είναι ένα λεξικό που απαριθμεί λέξεις που ομαδοποιούνται σύμφωνα με την ομοιότητα του νοήματός τους (π.χ. συνώνυμα - αντώνυμα) [23].

[55] Διαθέσιμη από: http://liwc.wpengine.com

- Το **MPQA (Multi-Perspective Question Answering) Subjectivity** Lexicon [56] από τους Wilson κ.α. (2005) είναι κατά κύριο λόγο μια λίστα ενδείξεων υποκειμενικότητας και χρησιμοποιείται από το σύστημα OpinionFinder που επεξεργάζεται τα κείμενα και εντοπίζει αυτόματα υποκειμενικές προτάσεις καθώς και διάφορες πτυχές της υποκειμενικότητας μέσα σε προτάσεις, συμπεριλαμβανομένων παραγόντων που είναι πηγές γνώμης, άμεσες υποκειμενικές εκφράσεις και γεγονότα ομιλίας και εκφράσεις συναισθημάτων.
- Το **SentiWordNet** των Esuli και Sebastiani (2006) αποδίδει θετικές και αρνητικές πραγματικές εκτιμήσεις αισθημάτων στα synsets του WordNet. Η βασική ιδέα πίσω από την δημιουργία του είναι ότι λέξεις με παρόμοιο χαρακτηρισμό στο WordNet τείνουν να έχουν παρόμοια συναισθηματική πολικότητα. Εξ' αυτού, στο SentiWordNet οι λέξεις «σπόροι» επεκτείνονται αξιοποιώντας την ομοιότητα χαρακτηρισμών του WordNet. Σε κάθε λέξη έχει αποδοθεί μια βαθμολογία ως προς την θετικότητα (*PosScore*) και την αρνητικότητά της (*NegScore*), ενώ η βαθμολογία της αντικειμενικότητας υπολογίζεται από τον τύπο: *ObjScore = 1 - (PosScore + NegScore)*. Πέραν των βαθμολογιών, δίνεται και η ερμηνεία κάθε λέξης καθώς προσδιορίζεται και το synset στο οποίο ανήκει.
- **SenticNet**, των Cambria κ.α. (2014) [57], είναι μια λεξική πηγή για την ανάλυση των συναισθημάτων σε επίπεδο ιδεών. Επικαλείται ως Sentic Computing, ένα καινοτόμο πολυεπιστημονικό υπόδειγμα για την Ανάλυση Συναισθήματος. Σε σχέση με τα προαναφερθέντα λεξικά γνώμης, το SenticNet είναι σε θέση να συσχετίσει και να περιπλέξει την πολικότητα και τις συναισθηματικές πληροφορίες μαζί. Κάθε όρος αναπαρίστανται από την ένταση τεσσάρων βασικών συναισθηματικών διαστάσεων (ευαισθησία, ικανότητα, προσοχή, ευχαρίστηση). Υιοθετεί ένα κομμάτι της ιεραρχίας του WordNet-Affect όπως η επίτευξη του στόχου, καθώς παρέχει βαθμολογίες αισθήματος (στο διάστημα *-1* και *1*) για 14.000 έννοιες της κοινής λογικής. Το συναίσθημα που μεταφέρεται από κάθε όρο είναι ορισμένο με βάση την ένταση δεκαέξι βασικών συναισθημάτων, που ορίζονται σε ένα μοντέλο που ονομάζεται Κλεψύδρα Συναισθημάτων (Hourglass of Emotions).
- Η δημιουργία του λεξικού **MicroWNOp** των Cerini κ.α. (2007) στηρίχθηκε σε ένα σύνολο 100 λέξεων «σπόροι» για κάθε μια από τις θετικές, αρνητικές και ουδέτερες κατηγορίες συναισθήματος, που προήλθαν από το έτοιμο λεξικό General Inquirer Lexicon του Harvard. Το λεξικό τους επεκτάθηκε με την προσθήκη όλων των 1.105 synsets του WordNet, που περιείχαν τις συγκεκριμένες λέξεις και χωρίστηκε σε τρία τμήματα (Common, Group1, Group2). Κάθε γραμμή στα επιμέρους τμήματα αντιστοιχεί

[56] Διαθέσιμη από: http://mpqa.cs.pitt.edu/lexicons

[57] Διαθέσιμη από: http://sentic.net/api/

σε ένα synset και περιλαμβάνει: βαθμολογίες για την θετικότητα και την αρνητικότητα του υποσυνόλου και το σύνολο των λέξεων που ανήκουν στο συγκεκριμένο υποσύνολο. Κάθε λέξη χαρακτηρίζεται για το τι μέρος του λόγου είναι και από την ερμηνεία της, με αναφορά στο WordNet. Ένα παράδειγμα γραμμής ενός από τα τμήματα του λεξικού MicroWNOp είναι:

Τμήμα Common		
Positive Score	**Negative Score**	**Synset**
1	0	real#a#7 true#a#2

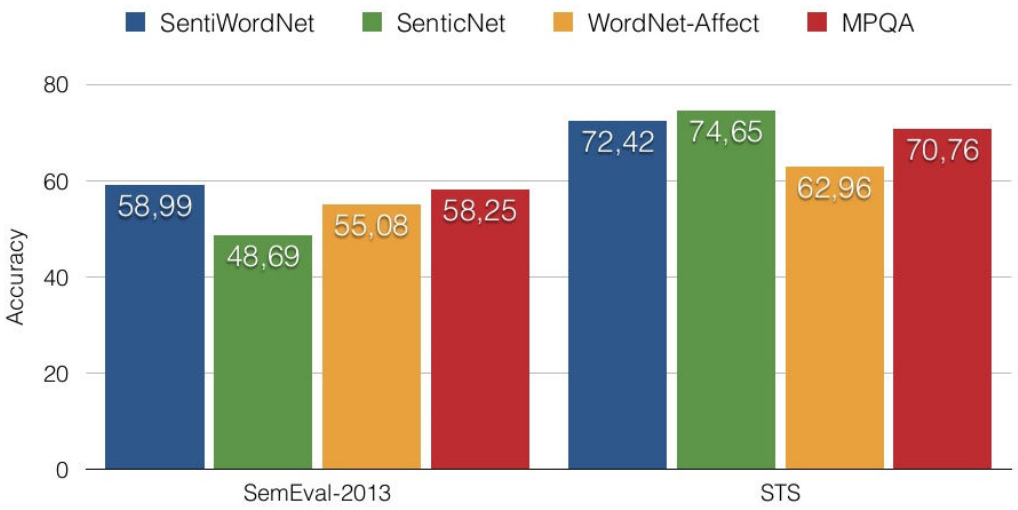
2.4.4.3.1.2 Σύγκριση Λεξικών Γνώμης

Όλα τα παραπάνω λεξικά παρέχουν βασικές ταξινομήσεις πολικότητας. Τα υποκείμενα λεξικά τους είναι διαφορετικά, επομένως είναι δύσκολο να τα συγκρίνουμε ολοκληρωτικά, αλλά μπορούμε να δούμε πόσο συχνά διαφωνούν μεταξύ τους επειδή παρέχουν τιμές αντίθετης πολικότητας για μια δεδομένη λέξη, όπως εξετάζει ο Potts (2011). Ο Πίνακας 6 αναφέρει τα αποτελέσματα αυτών των συγκρίσεων καθώς διακρίνονται τα ποσοστά διαφωνίας μεταξύ ορισμένων από αυτά.

Πίνακας 7: Ποσοστά διαφωνίας μεταξύ λεξικών συναισθημάτων Potts (2011)

	MPQA	**Opinion Lexicon**	**Inquirer**	**SentiWordNet**	**LIWC**
MPQA	–	33/5402 (0.6%)	49/2867 (2%)	1127/4214 (27%)	12/363 (3%)
Opinion Lexicon		–	32/2411 (1%)	1004/3994 (25%)	9/403 (2%)
Inquirer			–	520/2306 (23%)	1/204 (0.5%)
SentiWordNet				–	174/694 (25%)
LIWC					–

Επιπροσθέτως, σύμφωνα με την έρευνα των Musto κ.α. (2014) πάνω στα τέσσερα από τα προαναφερθέντα λεξικά γνώμης MPQA (8.222 κάλυψη όρων), SentiWordNet (117.659 κάλυψη όρων), SenticNet (14.000 κάλυψη όρων), WordNet Affect (200 κάλυψη όρων), καταλήγουν στο συμπέρασμα ότι τα MPQA και SentiWordNet υπερέχουν των άλλων. Τα πειράματά τους εκτελέστηκαν με δύο σύνολα δεδομένων από tweets, τα SemEval 2013 έκτασης 14.435 κειμένων και STS Dataset έκτασης 1.600.00 κειμένων. Τα αποτελέσματά τους φαίνονται στη παρακάτω Εικόνα 18. Συμπερασματικά, το SentiWordNet και το MPQA είναι αυτά που είχαν την υψηλότερη ακρίβεια και στα δύο σύνολα δεδομένων όπου εξετάστηκαν στη συγκεκριμένη έρευνα.



Εικόνα 18: Αποτελέσματα σύγκρισης λεξικών γνώμης των Musto κ.α. (2014)

2.4.4.3.1.3 Αιτίες απόρριψης μεθόδων βασισμένων σε λεξικά

Αν και οι προσεγγίσεις βασισμένες σε λεξικό είναι εύκολα κατανοητές και υλοποιήσιμες και μπορούν να αποφύγουν πολλά θέματα, όπως στα Ding κ.α. (2008), Hu και Liu (2004), υπάρχουν αρκετοί λόγοι που δεν προτείνονται για Ανάλυση Συναισθήματος και πιο συγκεκριμένα στην ταξινόμηση κειμένου. Τα αποτελέσματά τους είναι πολύ πιθανό να απέχουν κατά πολύ από την πραγματική ανθρώπινη ερμηνεία, λόγω της έλλειψης χρήσης της αλληλεπίδρασης των λέξεων που θα βοηθούσε στην κατανόηση του συνολικού νοήματος και την αποφυγή της αστοχίας. Μερικοί λόγοι αστοχίας άρα και απόρριψης είναι:

- Μια λέξη γνώμης μπορεί να έχει αντίθετους προσανατολισμούς σε διαφορετικούς τομείς εφαρμογής (αμφισημία). Για παράδειγμα στις προτάσεις:

  "This camera sucks."
  "This vacuum cleaner sucks very well."

  η ίδια λέξη *sucks*, στην πρώτη πρόταση έχει αρνητικό προσανατολισμό, εννοώντας ότι η κάμερα είναι χάλια, ενώ στην δεύτερη θετικό προσανατολισμό, εννοώντας ότι η ηλεκτρική σκούπα ρουφάει πολύ καλά.
- Μια πρόταση αν και περιέχει λέξεις γνώμης μπορεί να μην εκφράζει κανένα συναίσθημα. Όπως στη πρόταση:

  "If I could find a good camera in the shop, I will buy it."

  υπάρχει η λέξη γνώμης *good* αλλά εκφράζεται είτε θετικό είτε αρνητικό συναίσθημα. Αυτό το φαινόμενο παρατηρείται συχνά σε διάφορους τύπους προτάσεων, όπως ερωτήσεις και προτάσεις καταστάσεων, αλλά όχι σε όλες τις προτάσεις αυτών των ειδών.
- Δύσκολα αντιμετωπίζονται σαρκαστικές και ειρωνικές προτάσεις με ή χωρίς λέξεις γνώμης καθώς είναι δύσκολο να εντοπιστεί το πραγματικό συναίσθημα της πρότασης. Για παράδειγμα η ειρωνική πρόταση:

"What a great laptop! It stopped working in two days."

- Πολλές προτάσεις χωρίς την παρουσία λέξεων γνώμης μπορεί να υποδηλώνουν γνώμη. Οι περισσότερες από αυτές τις προτάσεις είναι αντικειμενικές προτάσεις που χρησιμοποιούνται για να εκφράσουν κάποιες πραγματικές πληροφορίες. Όπως η πρόταση:

"This app uses a lot of battery."

- Αγνοείται η άρνηση. Η ύπαρξη μίας λέξης άρνησης επηρεάζει σαφώς το νόημα της λέξης ή των λέξεων που ακολουθούν αντιστρέφοντας τη πολικότητα. Για παράδειγμα, οι λέξεις *not, never* κ.α.
- Αγνοούνται λέξεις μεταβολής έντασης που αυξάνουν (intensifiers) ή μειώνουν (downtoners) την ένταση της επόμενης λέξης. Παραδείγματα τέτοιων λέξεων είναι *very, really, more, less* κ.α.
- Αγνοείται η σειρά των λέξεων που εμφανίζονται στο κείμενο η οποία μπορεί να αντιστρέψει τη πολικότητα μίας πρότασης. Για παράδειγμα οι παρακάτω προτάσεις:

"That's not true, I'm a fan of this tablet."
"That's true, I'm not a fan of this tablet."

χρησιμοποιούν το ίδιο σύνολο λέξεων αλλά έχουν αντίθετο νόημα.

- Αγνοείται η σειρά των φράσεων που εμφανίζονται στο κείμενο η οποία πολλές φορές προσδιορίζει το συνολικό συναίσθημα καθότι το συναίσθημα που υπερισχύει στον αναγνώστη συνήθως είναι της τελευταίας φράσης ή εξαρτάται από την κρίση του. Για παράδειγμα η πρόταση:

"*Nicely filmed and well-acted, but falls short in narrations*."

αφήνει συνήθως αρνητική εντύπωση διότι η αρνητική φράση είναι στο τέλος.

- Αγνοούνται οι ιδιωματισμοί που είναι εκφράσεις που προσδίδουν ένα ιδιαίτερο νόημα. Παράδειγμα αποτελεί η φράση "*once in a blue moon*", που μεταφράζεται σαν πολύ σπάνια.
- Αγνοούνται οι εναντιωματικοί σύνδεσμοι σε μία πρόταση, οι οποίοι συνδέουν δύο φράσεις αντίθετης πολικότητας. Παραδείγματα τέτοιων είναι οι λέξεις *but, although, however* κ.ά.
- Πολλές φορές σε κείμενα κριτικών γίνεται αναφορά σε παραπάνω από μία οντότητες ή ακόμα και σε διαφορετικά χαρακτηριστικά της ίδιας οντότητας. Σε αυτές τις περιπτώσεις πρώτον, δεν ενδιαφέρει η ταξινόμηση του κειμένου συνολικά σε θετική ή αρνητική άποψη, αλλά εξετάζεται το κείμενο κυρίως σε επίπεδο λέξης και δεύτερον όταν η οντότητα δεν είναι ξεκάθαρη μπορεί να αποδοθεί λανθασμένο συναίσθημα που αναφέρεται σε δευτερεύουσα οντότητα. Όπως για παράδειγμα στην πρόταση:

"Nokia is better than Sony."

όπου δεν είναι ξεκάθαρο για το ποια είναι η οντότητα που αναλύεται.

2.4.4.3.1.4 Αντιμετώπιση προβλήματος χρήσης λεξικών

Λαμβάνοντας υπόψιν τους παραπάνω λόγους αστοχίας των προσεγγίσεων της Ανάλυσης Συναισθήματος βασισμένες σε λεξικό γνώμης, γίνονται συνεχώς προσπάθειες αντιμετώπισής τους με σκοπό πάντα την βελτίωση τους. Μια από τις σημαντικότερες προσπάθειας έγκειται στην επέκταση του λεξικού γνώμης με τους ιδιωματισμούς της, διαφορετικών για κάθε ξεχωριστή γλώσσα. Ακόμη, για την αντιμετώπιση της άρνησης προτείνεται η (όχι πάντα) ικανοποιητική λύση αλλαγής του προσήμου της συναισθηματικής βαθμολογίας των λέξεων που έπονται μιας γνωστής λέξης άρνησης. Εκτός ετούτου, η ύπαρξη των ενισχυτών έντασης ή εντατικοποιητών (intensifiers) στο κείμενο, οι οποίοι μπορούν να χωριστούν σε ενισχυτές (amplifiers) και εξομαλυντές (downtoners), πολλές φορές μπορεί να αξιοποιηθεί αυξάνοντας ή μειώνοντας την βαθμολογία της λέξης που συνοδεύουν, όπως προτείνουν οι Taboada κ.α. (2011). Όσον αφορά τον εντοπισμό της οντότητας στην οποία αναφέρεται κάθε κείμενο, γίνεται χρήση σημασιολογικών ρόλων που εξετάζουν τις σχέσεις των οντοτήτων με το κύριο ρήμα της πρότασης, διαδικασία που δεν χρειάζεται να υλοποιηθεί στις κριτικές των προϊόντων που εξετάζουμε καθώς είναι εκ των προτέρων γνωστή η οντότητα «προϊόν» που σχολιάζεται. Σε γενικά πλαίσια, προτείνεται η δημιουργία λεξικών γνώμης πιο εξειδικευμένων στον εκάστοτε θεματικό τομέα ανάλυσης των κειμένων, όσο πιο χρονοβόρα διαδικασία κι αν είναι, προτιμάται διότι έχει αποδειχθεί μεγαλύτερη ακρίβεια αποτελεσμάτων. Τέλος η ειρωνεία είναι από τις δυσκολότερα διαχειρίσιμες περιπτώσεις για την οποία δεν έχουν εντοπιστεί αξιόπιστες και υποσχόμενες λύσεις, άρα χρήζουν περαιτέρω έρευνας.

2.5 Αξιολόγηση

Τα μέτρα αξιολόγησης που υπάρχουν στη βιβλιογραφία του ερευνητικού πεδίου της εξαγωγής λέξεων – κλειδιών και του εντοπισμού του συναισθήματος που εκφράζουν και συνήθως χρησιμοποιούνται, βασίζονται κυρίως στην κατανόηση και τη μέτρηση της συνάφειας. Σε γενικά πλαίσια η απόδοση ενός συστήματος Ανάλυσης Συναισθήματος κρίνεται από τη συμφωνία που παρουσιάζουν τα αποτελέσματα της ταξινόμησής του με την αντίστοιχη ανθρώπινη κρίση. Δεδομένου ότι η ανθρώπινη κρίση διαφέρει ανά άνθρωπο κρίνεται πολλές φορές απαραίτητη η εκτίμηση του βαθμού συμφωνίας των κριτών (inter-rater agreement). Επιγραμματικά αναφέρουμε ότι για τον υπολογισμό της συμφωνίας μπορούν να χρησιμοποιηθούν οι στατιστικοί δείκτες: Joint-probability of Agreement, Cohen's kappa, Fleiss' kappa, Inter-rater Correlation, Concordance Correlation Coefficient και Intra-class

Correlation. Έπειτα της εν λόγω αποσαφήνισης είμαστε σε θέση να αξιολογήσουμε το σύστημα Ανάλυσης Συναισθήματος αυτό καθαυτό.

Δεδομένων των:

- positive (*P*): ο αριθμός των πραγματικών θετικών στοιχείων στα δεδομένα
- negatives (*N*): ο αριθμός των πραγματικών αρνητικών στοιχείων στα δεδομένα
- true positive (*TP*): ο αριθμός των σχετικών στοιχείων που ανακτήθηκαν σωστά από τα δεδομένα (σωστή ανάκτηση)
- true negative (*TN*): ο αριθμός των μη σχετικών στοιχείων που ανακτήθηκαν σωστά από τα δεδομένα (σωστή απόρριψη)
- false positive (*FP*): ο αριθμός των σχετικών στοιχείων που δεν ανακτήθηκαν από τα δεδομένα (λάθος ανάκτηση)
- false negative (*FN*): ο αριθμός των μη σχετικών στοιχείων που δεν ανακτήθηκαν από τα δεδομένα (λάθος απόρριψη)

και με τη βοήθεια ενός Πίνακα Σύγχυσης [58] γίνεται η απεικόνιση των αποτελεσμάτων της ταξινόμησης του συστήματος. Παρατηρώντας λοιπόν τον Πίνακας 6 προκύπτει ότι το άθροισμα των στοιχείων της διαγωνίου αποτελούν τον αριθμό των σωστών προβλέψεων, ενώ το άθροισμα των υπολοίπων στοιχείων αποτελεί τον αριθμό των λάθος προβλέψεων.

Πίνακας 8: Πίνακας Σύγχυσης

		Προβλεπόμενη κλάση			
		Κλάση 1	Κλάση 2	…	Κλάση n
Πραγματική κλάση	Κλάση 1	TP_{11}	FN_{12}	…	FN_{1n}
	Κλάση 2	FN_{21}	TP_{22}	…	FN_{2n}
	…	…	…	…	…
	Κλάση m	FN_{m1}	FN_{m2}	…	TP_{mn}

Έτσι, προκύπτουν τα εξής μέτρα για την αξιολόγηση της ταξινόμησης:

1. **precision**: ο λόγος των σχετικών στοιχείων προς των ανακτημένων στοιχείων

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

2. **recall**: ο λόγος των σχετικών στοιχείων που έχουν ανακτηθεί στο σύνολο των σχετικών στοιχείων

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

3. ο αρμονικός μέσος των precision και recall, το **F- measure** (ή *F1 score*)

[58] Ο Πίνακας Σύγχυσης (Confusion Matrix) είναι μια συγκεκριμένη διάταξη πίνακα που επιτρέπει την απεικόνιση της απόδοσης ενός αλγορίθμου.

$$F = \frac{2 * P * R}{P + R}$$

4. Ορισμένες άλλες τεχνικές εστιασμένες σε πιο συγκεκριμένους τομείς που εντοπίζεται να χρησιμοποιούνται όπως από τους Socher κ.α. (2008), Yu κ.α. (2011), Zhao κ.α. (2010) και Pavlopoulos και Androutsopoulos (2014), επιγραμματικά είναι το Weighted Precision, το Weighted recall και το Normalized Discounted Cumulative Gain (**nDCG**), ένα μέτρο της ποιότητας κατάταξης για την αξιολόγηση κάθε λεξικού των διαφορετικών στοιχείων που περιέχουν.

$$nDCG = \frac{1}{Z}\sum_{i=1}^{m}\frac{2^{t(i)} - 1}{\log_2(1 + i)}$$

όπου *Z* είναι ένας συντελεστής κανονικοποίησης για να εξασφαλιστεί ότι μια άριστη κατάταξη παίρνει τιμή *nDCG = 1* και *t(i)* είναι μια συνάρτηση ανταμοιβής για ένα στοιχείο που τοποθετείται στη θέση *i* του λεξικού που έχει επιστραφεί.

5. Το **accuracy**, ένας σταθμισμένος μέσος όρος του precision και του αντίστροφου (inverse) precision. Χαρακτηρίζεται ως μια περιγραφή των σφαλμάτων των συστημάτων και ως μέτρο στατιστικής προκατάληψης, καθώς προκαλεί διαφορά μεταξύ του αποτελέσματος και μιας «*true*» τιμής.

$$ACC = \frac{TP + TN}{P + N}$$

Αν και είναι ένα καλά εδραιωμένο μέτρο, είναι γνωστό ότι όταν εφαρμόζεται σε διαστρεβλωμένες (ακανόνιστες) κατανομές μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα. Θα ήταν προτιμότερο να αποφεύγεται στις περιπτώσεις εξαγωγής των χαρακτηριστικών γνωρισμάτων των προϊόντων σε κριτικές και να χρησιμοποιείται το precision αντί αυτού. Εδώ καλό είναι να σημειωθεί ότι η έννοια και η χρήση του precision διαφέρει από τομέα σε τομέα εφαρμογής αλλά και άλλους κλάδους της επιστήμης και της τεχνολογίας.

6. Το **error rate** (λόγος σφάλματος), που εκφράζει το ποσοστό των εσφαλμένων ταξινομήσεων του συστήματος και συχνά χρησιμοποιείται αντί του accuracy.

$$error\ rate = 1 - accuracy$$

Μεταξύ των άλλων θα πρέπει να σημειωθεί ότι τα μέτρα precision, recall και F-measure δεν λαμβάνουν υπόψη ότι η θετική κλάση είναι πιο κοντά στην ουδέτερη από ό, τι στην αρνητική. Για παράδειγμα, η ταξινόμηση ενός θετικού όρου ως αρνητικό θα πρέπει να «τιμωρείται» περισσότερο από την ταξινόμησή του ως ουδέτερο.

Αξιοσημείωτο επίσης είναι ότι για μεθόδους που επιχειρούν να ταξινομήσουν τις λέξεις – κλειδιά μιας οντότητας (π.χ. χαρακτηριστικά γνωρίσματα ενός προϊόντος) σύμφωνα με την γνώμη που εκφράζουν (από τις πιο θετικές ως τις πιο αρνητικές), τα μέτρα αξιολόγησης που

ως επί το πλείστον προτείνονται στην βιβλιογραφία είναι τα βασισμένα στην βαθμολογία / κατάταξη (Ranking - based) (π.χ. αστεράκια), όπως επισημαίνουν οι Pavlopoulos και Androutsopoulos (2014).

3

Σχετικές Εργασίες / Δημοσιεύσεις

3.1 Παρουσίαση Σχετικών Εργασιών

Παρακάτω παρουσιάζεται η τρέχουσα γενική εικόνα των βασικότερων σχετικών δημοσιεύσεων / εργασιών για την επίλυση του προβλήματος της Ανάλυσης Συναισθήματος σε κριτικές προϊόντων σε επίπεδο προτάσεων που μας απασχολεί στην προκειμένη διπλωματική εργασία, ταξινομημένες με την ημερομηνία δημοσίευσής τους, ώστε να παρατηρηθούν και οι εξελίξεις στην επίλυση του εν προκειμένω προβλήματος κατά το πέρασμα των χρόνων. Οι συγγραφείς κάθε δημοσίευσης ξεχωριστά προτείνουν το δικό τους μοντέλο επίλυσης του προβλήματος, παρουσιάζοντας τον τρόπο υλοποίησής του και τα αποτελέσματα αξιολόγησής του. Ακολουθεί παρουσίαση κάθε μιας μεμονωμένα αναφέροντας αξιοσημείωτα στοιχεία της και οι τέσσερις πίνακες σύγκρισής τους, ταξινομημένων σύμφωνα με τις τεχνικές προσεγγίσεις τους. Για περισσότερες πληροφορίες για την εκάστοτε δημοσίευση, παρακαλούμε ανατρέξτε στην Βιβλιογραφία.

1. **Hatzivassiloglou και McKeown (1997)**: πρότειναν την πρώτη μέθοδο για τον προσδιορισμό των προσανατολισμών των επιθέτων ανιχνεύοντας ζεύγη λέξεων που συνδέονται με συζεύξεις σε ένα σύνολο δεδομένων. Για παράδειγμα, στην πρόταση, "*Αυτό το αυτοκίνητο είναι όμορφο και ευρύχωρο*", αν γνωρίζουμε ότι το "*όμορφο*" είναι θετικό, μπορούμε να συμπεράνουμε ότι και το "*ευρύχωρο*" είναι θετικό. Η υποκείμενη διαίσθηση είναι ότι οι προσανατολισμοί των συνδυασμένων επίθετων υπόκεινται σε ορισμένους γλωσσικούς περιορισμούς. Ένα λογαριθμικό γραμμικό μοντέλο

παλινδρόμησης χρησιμοποιεί αυτούς τους περιορισμούς για να προβλέψει αν τα επίθετα είναι συνδυασμένα με ίδιους ή διαφορετικούς προσανατολισμούς, επιτυγχάνοντας 82% ακρίβεια όταν καθεμία θεωρείται ανεξάρτητη. Συνδυάζοντας τους περιορισμούς σε πολλά επίθετα, διαχωρίζουν τα επίθετα σε ομάδες διαφορετικών προσανατολισμών με έναν αλγόριθμο ομαδοποίησης και τέλος, τα επίθετα επισημαίνονται σαν θετικά ή αρνητικά. Οι αξιολογήσεις τους υποδεικνύουν υψηλά επίπεδα απόδοσης με τιμή precision να είναι μεγαλύτερη από 90% για τα επίθετα. Η αδυναμία της μεθόδου τους είναι ότι καθώς βασίζεται στις σχέσεις σύνδεσης, δεν είναι σε θέση να εξάγει επίθετα που δεν συνδέονται.

2. **Turney (2002)**: εκτελεί την ταξινόμηση βάσει ορισμένων συντακτικών προτύπων που είναι πιθανό να χρησιμοποιηθούν για να εκφράσουν γνώμη. Τα συντακτικά αυτά πρότυπα εκπονούνται με βάση τις ετικέτες στις λέξεις για το τι μέρος του λόγου είναι μέσα στη πρόταση (με POS tagging). Συνεπώς καταλαβαίνουμε ότι ο αλγόριθμός του είναι μη εποπτευόμενης μάθησης και αποτελείται από τρία στάδια. Στο πρώτο, εξάγονται δύο συνεχόμενες λέξεις που πληρούν κάποιο από τα πρότυπα του Πίνακα 8 σύμφωνα με οποία συνηθίζεται να εκφράζεται γνώμη.

Πίνακας 9: Πρότυπα POS tags για την εξαγωγή φράσεων δύο λέξεων στο Turney (2002)

	ΠΡΩΤΗ ΛΕΞΗ	ΔΕΥΤΕΡΗ ΛΕΞΗ	ΤΡΙΤΗ ΛΕΞΗ (ΔΕΝ ΕΞΑΓΕΤΑΙ)
1	JJ	NN ή NNS	οτιδήποτε
2	RB, RBR ή RBS	JJ	όχι NN ούτε NNS
3	JJ	JJ	όχι NN ούτε NNS
4	NN ή NNS	JJ	όχι NN ούτε NNS
5	RB, RBR ή RBS	VB, VBD, VBN ή VBG	οτιδήποτε

Τα ουσιαστικά και τα ρήματα δρουν σαν τα συμφραζόμενα, καθώς αυτά συνήθως εκφράζουν διάφορα συναισθήματα. Στο δεύτερο στάδιο, υπολογίζεται ο σημασιολογικός προσανατολισμός των εξαγόμενων φράσεων μέσω του μέτρου PMI, βάση της σχέσης της φράσης με την θετική λέξη αναφοράς «*excellent*» (εξαιρετικό) και της αρνητικής «*poor*» (κακό)

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

Οι πιθανότητες υπολογίζονται με την εκτέλεση ερωτημάτων στη μηχανή αναζήτησης AltaVista, συλλέγοντας τον αριθμό των hits. Για κάθε ερώτημα, μια μηχανή αναζήτησης συνήθως δίνει τον αριθμό των σχετικών εγγράφων στο ερώτημα, που είναι ο αριθμός των hits. Έτσι, με την αναζήτηση των δύο όρων μαζί και ξεχωριστά υπολογίζονται οι πιθανότητες. Ο Turney (2002) χρησιμοποιεί την μηχανή αναζήτησης AltaVista επειδή έχει την λειτουργία NEAR η οποία περιορίζει την αναζήτηση στα έγγραφα που περιέχουν τις λέξεις, μέσα σε δέκα λέξεις η μια με την άλλη με οποιαδήποτε σειρά. Στο τρίτο

στάδιο, δεδομένης μιας κριτικής, ο αλγόριθμος υπολογίζει το μέσο όρο του σημασιολογικού προσανατολισμού όλων των φράσεων της κριτικής και την ταξινομεί ως θετική εάν ο μέσος όρος είναι θετικός ή διαφορετικά ως αρνητική. Ο αλγόριθμός του επιτυγχάνει μέση ακρίβεια 74%, στην αξιολόγηση 410 κριτικών, δείγματα από τέσσερις διαφορετικούς τομείς αξιολόγησης (κριτικές αυτοκινήτων, τραπεζών, ταινιών, και ταξιδιωτικών προορισμών). Η ακρίβεια κυμαίνεται από 84% για τις κριτικές σε αυτοκίνητα έως 66% για κριτικές ταινιών.

3. **Pang κ.α. (2002)**: χρησιμοποίησαν κριτικές ταινιών ως δεδομένα και τους αλγορίθμους Μηχανικής Μάθησης Naive Bayes, Μέγιστης Εντροπίας και SVM για την ταξινόμησή τους σε προτεινόμενες ή όχι (thumbs up - thumbs down), αλλά δεν απέδωσαν τόσο καλά στην ταξινόμηση συναισθημάτων, όσο στα παραδοσιακά προβλήματα ταξινόμησης. Επομένως οι συγγραφείς οδηγήθηκαν σε δοκιμή νέων μεθόδων και στο Pang κ.α. (2005) εξελίσσουν την προηγούμενη εργασία τους προσπαθώντας να εκτιμήσουν την αξιολόγηση των ταινιών σε ένα σύστημα βαθμολόγησης (π.χ. 1 μέχρι 5 αστέρια). Έτσι προκύπτουν παραπάνω των δύο κλάσεων προς ταξινόμηση σε σχέση με την προηγούμενη εργασία τους. Η συγκεκριμένη εργασία είναι μια πιο ενδιαφέρουσα προσέγγιση για ένα σύστημα ταξινόμησης κειμένου σε πολλές κλάσεις διότι υπάρχει διαφορετικός βαθμός ομοιότητας μεταξύ των κλάσεων. Παράδειγμα η κλάση «3 αστέρια» είναι πιο κοντά στην κλάση «*4 αστέρια*» από ότι στην κλάση «*1 αστέρι*».

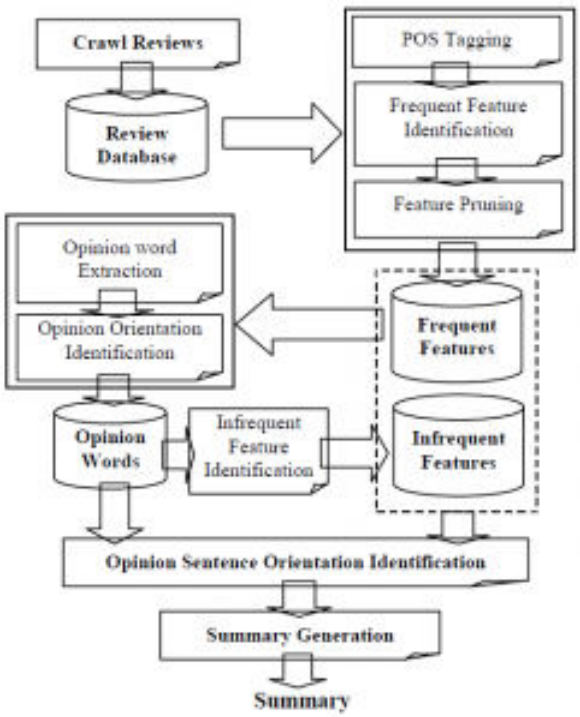
   Αρχικά αξιολογούν την ανθρώπινη απόδοση στο έργο δίνοντάς σε ανθρώπους να βαθμολογήσουν τα κείμενα των κριτικών τους, με σκοπό να δείξουν ότι αν και οι βαθμολογίες των ανθρώπων διαφέρουν μεταξύ τους είναι κοντά στις ίδιες κλάσεις. Για παράδειγμα αν μια ταινία βαθμολογηθεί με 4 αστέρια από κάποιον είναι πιο πιθανό και οι υπόλοιποι να την βαθμολογήσουν με 3, 4 ή 5 αστέρια παρά με 1 που είναι πιο μακριά από το 4. Τέλος, εφαρμόζουν τους τρείς αλγορίθμους ένας – εναντίον - όλων (one-vs-all), παλινδρόμηση (regression) και μετρική σήμανση (metric labeling) οι οποίοι όλοι βασίζονται στο SVM με σκοπό και εδώ να διασφαλιστεί ότι παρόμοια αντικείμενα θα λάβουν παρόμοιες ετικέτες και δείχνουν ότι ο αλγόριθμος της μετρικής σήμανσης μπορεί να προσφέρει σημαντικές βελτιώσεις σε εκδοχές πολλαπλών κλάσεων και παλινδρόμησης των SVM, όταν χρησιμοποιούν ένα νέο μέτρο ομοιότητας κατάλληλο για το πρόβλημα.

4. **Hu και Liu (2004)**: χρησιμοποιούν μια τεχνική εξόρυξης κανόνων συσχέτισης, εξάγοντας τα πιο συχνά εμφανιζόμενα ουσιαστικά μιας πρότασης ως λέξεις - κλειδιά. Αφού έχει συλλεχθεί ένας ικανοποιητικός αριθμός κριτικών, εξάγουν όλα τα ουσιαστικά (μεμονωμένες λέξεις ή φράσεις με μήκος το πολύ τρεις λέξεις) από τις κριτικές κάθε

συνόλου δεδομένων με POS tagging χρησιμοποιώντας το εργαλείο NLProcessor [59] που η έξοδός του είναι XML δεδομένα και τα θεωρούν ως υποψήφιες λέξεις – κλειδιά. Στη συνέχεια δημιουργούν ένα αρχείο (λεξικό), όπου κάθε γραμμή του αντιπροσωπεύει κάθε πρόταση της κριτικής και περιέχει τα ουσιαστικά του προηγούμενου βήματος αφαιρώντας ορισμένες πολύ χρησιμοποιούμενες λέξεις (stop words) και τις καταλήξεις τους (stemming), υλοποιώντας και ορθογραφικό έλεγχο. Κάνοντας την παραδοχή ότι όταν οι χρήστες γράφουν στις κριτικές για τα χαρακτηριστικά γνωρίσματα του προϊόντος χρησιμοποιούν παρεμφερές λεξιλόγιο, τα ουσιαστικά που εμφανίζονται πιο συχνά στην πρόταση της κριτικής θεωρούνται και τα χαρακτηριστικά γνωρίσματα του προϊόντος τα οποία σχολιάζει ο εκάστοτε χρήστης. Πραγματοποιούν έλεγχο, βασιζόμενοι στον αλγόριθμο Apriori, για το αν εμφανίζεται με την ίδια σειρά και πόσες φορές το κάθε ζευγάρι ή τριπλέτα ουσιαστικών (n - grams) στο πλήθος των προτάσεων (γραμμών λεξικού) και εάν ξεπερνά το ελάχιστο κατώφλι του 1% (minimum support) του συνόλου των προτάσεων τότε αποτελεί υποψήφια λέξη – κλειδί. Ο αλγόριθμος προχωρά με δύο βήματα κλαδέματος με σκοπό να διορθώσει τον εαυτό του, εντοπίζοντας λέξεις – κλειδιά που αποτελούνται από πολλές ή μία λέξεις. Πρώτον, ελέγχουν εάν η απόσταση μεταξύ των λέξεων που αποτελούν την κάθε υποψήφια λέξη – κλειδί είναι μικρότερη από τρία, σε πάνω από μια πρόταση των κριτικών τότε θεωρούνται σαν μια ενιαία υποψήφια λέξη – κλειδί αποτελούμενη από πολλές λέξεις. Για παράδειγμα, η φράση – κλειδί «διάρκεια ζωής οθόνης» εμφανίζεται στην πρόταση «*η διάρκεια ζωής του είναι πολύ χειρότερη από την ανάλυση οθόνης*», δεν προσμετράται σαν υποψήφια γιατί η απόσταση των λέξεων της είναι μεγαλύτερη των τριών λέξεων. Δεύτερον, ελέγχουν εάν η κάθε υποψήφια φράση – κλειδί περιέχεται σε πάνω από τρεις προτάσεις χωρίς να αποτελεί μέρος κάποιας από τις υπόλοιπες φράσεις – κλειδιά ώστε να αποτελέσει μια μεμονωμένη υποψήφια φράση – κλειδί. Για παράδειγμα, αν η υποψήφια λέξη – κλειδί «*επεξεργαστής*» εμφανίζεται σε 10 προτάσεις και η φράση – κλειδί «*ταχύτητα επεξεργαστή*», που προέκυψε από το προηγούμενο βήμα, σε τέσσερα, τότε σημαίνει ότι η υποψήφια λέξη – κλειδί «*επεξεργαστής*» εμφανίζεται σε έξι προτάσεις μεμονωμένα άρα αποτελεί και αυτή λέξη – κλειδί. Στη συνέχεια, δημιουργούν ένα αρχείο επιθέτων (λεξικό γνώμης) αποτελούμενο από τα πιο κοντινά επίθετα των φράσεων – κλειδιών που εξήχθησαν στα προηγούμενα βήματα, στηριζόμενοι στην υπόθεση ότι μέσω αυτών εκφράζεται η γνώμη για το κάθε χαρακτηριστικό γνώρισμα του προϊόντος στο οποίο αναφέρεται η πρόταση. Ο προσανατολισμός του συναισθήματος των επιθέτων αυτών εντοπίζεται μέσω των συνωνύμων του λεξικού WordNet. Για παράδειγμα, στην πρόταση «*Το iPad βγάζει υπέροχες φωτογραφίες*» το επίθετο «*υπέροχες*» μπαίνει στο αρχείο με τις λέξεις γνώμης.

[59] Διαθέσιμο από: http://www.infogistics.com/textanalysis.html

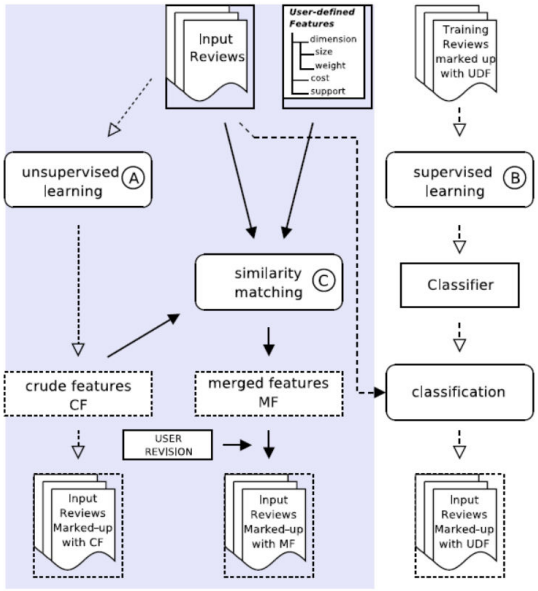
Επιπρόσθετα, μέσω αυτού του ελέγχου είναι πιθανό να εντοπιστούν και επιπλέον φράσεις – κλειδιά ελέγχοντας τα κοντινά ουσιαστικά των συγκεκριμένων επιθέτων στις προτάσεις που δεν περιέχουν υποψήφια φράση – κλειδί, μέσω της παραδοχής ότι με τις ίδιες λέξεις γνώμης οι σχολιαστές αναφέρονται σε πολλά χαρακτηριστικά γνωρίσματα του προϊόντος, για παράδειγμα «*εξαιρετική ανάλυση οθόνης*», «*εξαιρετική σχεδίαση*». Τέλος, για κάθε υποψήφια φράση – κλειδί, όλες οι προτάσεις που την αναφέρουν ομαδοποιούνται σε θετικές και αρνητικές. Η πολικότητα της πρότασης εντοπίζεται ανεξάρτητα από την πολικότητα της φράσης – κλειδί μέσα στην πρόταση. Οι φράσεις – κλειδιά ταξινομούνται σύμφωνα με τη συχνότητά τους στις κριτικές με τις μακρύτερες να παρουσιάζονται πριν από αυτές της μιας λέξης. Η τελική επίδοση του αλγορίθμου τους ήταν, precision: 72% και recall: 80% στον εντοπισμό των λέξεων – κλειδιών, precision: 64.2% και recall: 69.3% στον εντοπισμό των προτάσεων γνώμης και accuracy: 84.2% στην ανίχνευση προσανατολισμού των προτάσεων. Η αρχιτεκτονική του συστήματός τους φαίνεται και στην παρακάτω Εικόνα 19.



Εικόνα 19: Αρχιτεκτονική συστήματος των Hu και Liu (2004)

5. **Kamps κ.α. (2004)**: επωφελούνται από το λεξικό WordNet για να δημιουργήσουν ένα δίκτυο συνωνύμων, συνδέοντας ζεύγη συνώνυμων λέξεων και καθορίζουν τον σημασιολογικό προσανατολισμό μιας λέξης από τις πιο σύντομες διαδρομές στο δίκτυο σε δύο βασικές λέξεις «*καλό*» και «*κακό*» που επιλέχθηκαν ως εκπρόσωποι των θετικών και αρνητικών προσανατολισμών. Η βαθμολογία που πέτυχαν για τον αξιολογικό παράγοντα είναι 68.19%, για τον παράγοντα ισχύος είναι 71.36% και για τον παράγοντα δραστηριότητας είναι 61.85%.

6. **Carenini κ.α. (2005):** η μέθοδος τους βασίστηκε σε μετρήσεις ομοιότητας που καθορίστηκαν χρησιμοποιώντας την ομοιότητα συμβολοσειρών, τα συνώνυμα και τις λεκτικές αποστάσεις χρησιμοποιώντας και αυτοί το λεξικό WordNet. Η μέθοδος προϋποθέτει την ταξινόμηση των λέξεων – κλειδιών που θα δοθούν και συνδυάζει κάθε λέξη γνώμης που ανακαλύφθηκε σε έναν κόμβο της ταξινόμησης βασιζόμενη στις ομοιότητες. Τα πειράματα τους στηρίχθηκαν σε κριτικές ψηφιακών φωτογραφικών μηχανών και DVD με πολλά υποσχόμενα αποτελέσματα τόσο στην ακρίβειας όσο και στη μείωση του σημασιολογικού πλεονασμού των ακατέργαστων χαρακτηριστικών. Στην παρακάτω Εικόνα 20 διακρίνεται η προσέγγισή τους στην διαδικασία εξαγωγής των λέξεων – κλειδιών.

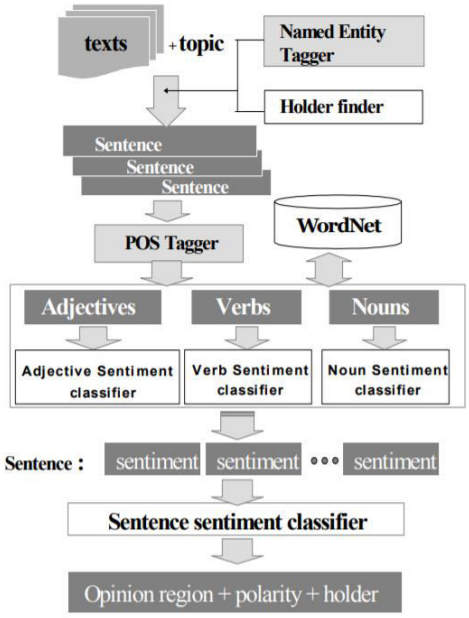


Εικόνα 20: Προσεγγίσεις στην εξαγωγή λέξεων - κλειδιών, Carenini κ.α. (2005)

7. **Popescu και Etzioni (2005)**: αντιμετωπίζουν τα πιο συχνά εμφανιζόμενα ουσιαστικά μιας πρότασης ως υποψήφιες λέξεις – κλειδιά (*F*) και αναγνωρίζουν το χαρακτηριστικό γνώρισμα του προϊόντος (*P*) στο οποίο αναφέρεται κάθε λέξη - κλειδί μετρώντας τη συνύπαρξή τους σε φράσεις που περιέχουν μοτίβα μερωνυμικών σχέσεων (π.χ. "*F του P*", "*P έχει F*"), χρησιμοποιώντας το μέτρο PMI. Το σύστημά τους (ονόματι OPINE) είναι μη εποπτευόμενης μάθησης και χρησιμοποιώντας τη τεχνική χαλαρωτικής σήμανσης (tagging) για τον προσδιορισμό του σημασιολογικού προσανατολισμού των δυνητικών λέξεων γνώμης επιτυγχάνει 22% υψηλότερο precision και 3% χαμηλότερο recall σε σύγκριση με παλαιότερα συστήματα.
8. **Wilson κ.α. (2005)**: προσδιορίζουν το συναίσθημα του σχολιαστή σε επίπεδο προτάσεων με λεξικό γνώμης, μία προσέγγιση εποπτευόμενης μάθησης χωρισμένη σε δύο βήματα. Ο στόχος τους ήταν να ταξινομήσουν το συμφραζόμενο συναίσθημα των προτάσεων που περιέχουν περιπτώσεις ενδείξεων υποκειμενικότητας στο λεξικό υποκειμενικότητας.

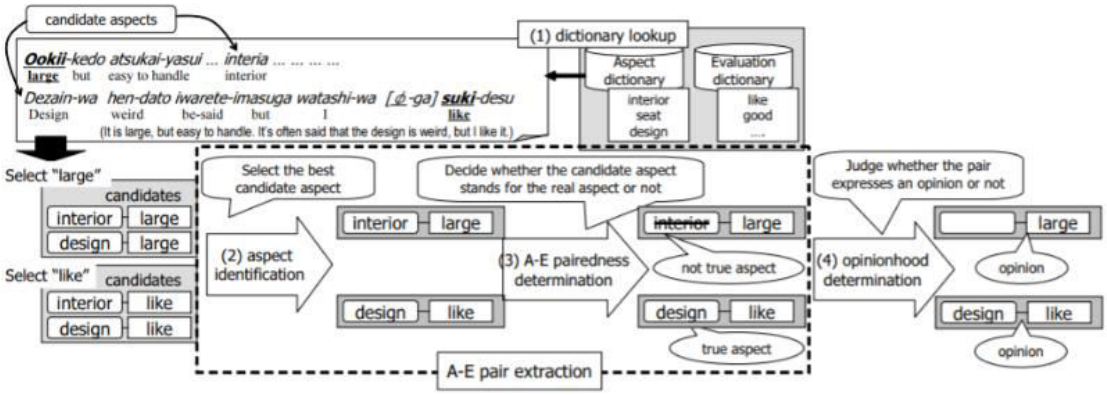
Αρχικά ελέγχουν αν η πρόταση εκφράζει συναίσθημα και εάν εντοπιστεί προσδιορίζεται η πολικότητά της μέσω λεξικού 8000 υποκειμενικών όρων βασισμένο στο λεξικό των Riloff και Wiebe (2003). Η πολικότητα της πρότασης μπορεί να είναι θετική (π.χ. *εξαιρετικό tablet*), αρνητική (π.χ. *απογοητευτικό laptop*), θετική και αρνητική μαζί (π.χ. *ο επεξεργαστής του έχει σπουδαίες επιδόσεις αλλά η μπαταρία του δεν κρατάει ούτε ώρα*) ή ουδέτερη (π.χ. *η κατανάλωση της μπαταρίας δεν είχε καμία διαφορά μεταξύ των δύο tablet*). Σύμφωνα με τα αποτελέσματά τους πέτυχαν σε διάφορα σύνολα δεδομένων 75% ποσοστό ακρίβειας (accuracy) στο πρώτο βήμα και εως 65% στο δεύτερο, ικανοποιητικά αλλά όχι αξιόπιστα ποσοστά. Ωστόσο, ο αλγόριθμός τους προτείνεται και προτιμάται για την εξόρυξη γνώμης με μεθόδους Μηχανικής Μάθησης.

9. **Kim και Hovy (2004, 2006)**: προτείνουν ένα μοντέλο που εξάγει ως λέξεις - κλειδιά λεκτικά σύνολα που συγγενεύουν συντακτικά με υποκειμενικά ρήματα ή επίθετα χρησιμοποιώντας σημασιολογική επισήμανση ρόλων. Η μέθοδος ξεκινάει με τρεις κατηγορίες θετικών, αρνητικών και ουδέτερων λέξεων, βρίσκει τα συνώνυμά τους (μέσω του WordNet) και στη συνέχεια χρησιμοποιεί μια Bayesian φόρμουλα για να υπολογίσει την εγγύτητα κάθε λέξης - κλειδί σε κάθε κατηγορία (θετική, αρνητική και ουδέτερη) ώστε να προσδιορίσει την πιο πιθανή κλάση της. Πειραματίζονται σε δύο συλλογές δεδομένων από δύο διαφορετικούς τομείς ώστε να δοκιμάσουν την προσέγγισή τους σε δύο διαφορετικές καταστάσεις κριτικών προϊόντων και εστιατορίων. Η πρώτη ήταν από κριτικές που συλλέχθηκαν από το epinions.com. Οι 3.241 κριτικές προϊόντων με 115.029 φράσεις ήταν από συσκευές αναπαραγωγής mp3 κατασκευασμένες από διάφορους κατασκευαστές, ενώ οι 7.524 κριτικές εστιατορίων με 194.393 φράσεις για διάφορους τύπους εστιατορίων. Ο μέσος όρος των προτάσεων σε μία κριτική προϊόντος ήταν 35 και 25 αντίστοιχα σε μία κριτική εστιατορίου. Η δεύτερη ήταν από δεδομένα που συλλέχθηκαν από το complaints.com με 59 αναφορές παραπόνων σχετικά με συσκευές αναπαραγωγής mp3 και 322 σχόλια για εστιατόρια. Τα πειραματικά τους αποτελέσματα επιτυγχάνουν 66% precision και 76% recall. Στην Εικόνα 21 παρουσιάζεται η αρχιτεκτονική του συστήματός τους.



Εικόνα 21: Αρχιτεκτονική συστήματος Kim και Hovy (2004)

10. **Kobayashi κ.α. (2006)**: απευθύνονται στο έργο της εξαγωγής γνώμης υποθέτοντας ότι μια γνώμη μπορεί να εκπροσωπείται ως μια τριπλέτα (θέμα, χαρακτηριστικό, αξιολόγηση) και προτείνουν μια υπολογιστική μέθοδο για την εξαγωγή τέτοιων τριπλετών από κριτικές. Σε αυτή τη μέθοδο, το κύριο έργο αποσυντίθεται στην διαδικασία εξόρυξης ζευγών χαρακτηριστικό – αξιολόγηση από ένα δεδομένο απόσπασμα κειμένου και την διαδικασία κρίσης αν εκφράζεται η άποψη του συγγραφέα. Απορρίπτουν την χρήση λεξικών γνώμης για την ανίχνευση υποψηφίων ζευγών διότι εξαρτάται σε μεγάλο βαθμό από τον τομέα. Εφαρμόζουν τεχνικές Μηχανικής Μάθησης και στις δύο υποτιμήσεις. Η διαδικασία εξαγωγής της γνώμης που ακολουθούν φαίνεται και στην Εικόνα 22. Χρησιμοποίησαν σαν σύνολο δεδομένων 288 κριτικές αυτοκινήτων με 4.442 φράσεις γραμμένες στα Ιαπωνικά καθώς δημιούργησαν λεξικά γνώμης τόσο για την εξαγωγή των ζευγών όσο και για την αξιολόγηση των υποψηφίων. Η τιμή του precision που επιτυγχάνουν στην εξαγωγή λέξεων - κλειδιών είναι 78% και στην εξαγωγή



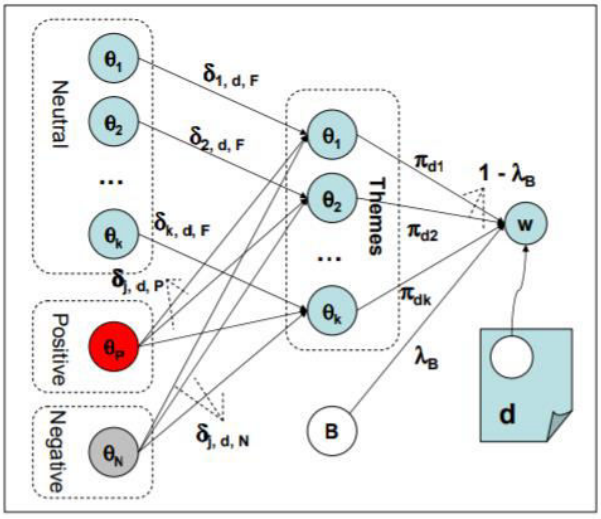
ζευγών είναι 61.7%. Στον προσδιορισμό της γνώμης όταν έπειται της διαδικασίας εξαγωγής λέξεων - κλειδιών επιτυγχάνει τα καλύτερα αποτελέσματα precision με 82.2%.

Εικόνα 22: Διαδικασία εξαγωγής γνώμης στο Kobayashi κ.α. (2006)

11. **Kobayashi κ.α. (2007)** : χρησιμοποιούν έναν αναλυτή εξάρτησης για τον εντοπισμό των σχέσεων εξάρτησης των λέξεων - κλειδιών με τις λέξεις γνώμης. Πρώτα βρίσκουν ζεύγη υποψήφιων λέξεων - κλειδιών και λέξεων γνώμης χρησιμοποιώντας ένα δέντρο εξάρτησης και στη συνέχεια χρησιμοποιούν μια δενδροειδής μέθοδο ταξινόμησης για να μάθει ο αλγόριθμός τους και να ταξινομήσει τα υποψήφια ζεύγη σαν μια λέξη - κλειδί. Οι λέξεις - κλειδιά εξάγονται, εν κατακλείδι, από τα υψηλότερα βαθμολογημένα ζεύγη χρησιμοποιώντας, μεταξύ άλλων, στοιχεία από τα συμφραζόμενα (contextual clues) και στοιχεία στατιστικής συνύπαρξης (common co-occurrence statistics). Όσον αφορά την αξιολόγηση, διαπιστώνουν ότι τα μοντέλα που χρησιμοποιούν τις ενδείξεις συμφραζομένων δείχνουν σχεδόν 10% βελτίωση τόσο στην ακρίβεια (precision) όσο και στην ανάκληση (recall) σε σχέση με παλαιότερα συστήματα. Αυτό υποδηλώνει ότι η

μέθοδος που βασίζεται στη Μηχανική Μάθηση πλεονεκτεί έναντι της προσέγγισης με βάση τα πρότυπα. Παρόμοια αποτελέσματα παρατηρούνται στην εξαγωγή σχέσεων εξάρτησης. Χρησιμοποίησαν σαν σύνολο δεδομένων 116 ηλεκτρονικά άρθρα εστιατορίων γραμμένα στα Ιαπωνικά, αναρτημένα στην κατηγορία "gourmet" του ιστολογίου blog.livedoor.com.

12. **Snyder και Barzilay (2007)** : αντιμετωπίζουν την ανάλυση πολλαπλών απόψεων που υπάρχουν σε κείμενα κριτικών και σχετίζονται μεταξύ τους και η ταξινόμησή τους έγκειται σε μία κλίμακα πολλαπλών κλάσεων. Για παράδειγμα, η κριτική ενός εστιατορίου μπορεί να περιλαμβάνει κριτικές για το φαγητό, την ατμόσφαιρα και την εξυπηρέτηση. Ο στόχος τους ήταν η παραγωγή μιας σειράς από αριθμητικά δεδομένα που το καθένα αντιστοιχεί σε μία άποψη. Ο αλγόριθμος που χρησιμοποιείται μαθαίνει να ταξινομεί τις ανεξάρτητες κριτικές και προβλέπει τις βαθμολογίες τους, αναλύοντας τις σχέσεις ανάμεσα στις κριτικές για το αν συμφωνούν ή διαφωνούν και αποδεικνύουν ότι το μοντέλο τους για τις κριτικές που συμφωνούν είναι πιο ακριβές με ποσοστό 67%, σε σχέση με τα ανεξάρτητα μοντέλα με ποσοστό 58%.
13. **Mei κ.α. (2007)**: δημιούργησαν ένα κοινό μοντέλο βασισμένο σε ένα μοντέλο θεματικών ενοτήτων, ένα μοντέλο θετικού αισθήματος και ένα μοντέλο αρνητικού αισθήματος που εκπαιδεύτηκε με ορισμένα εξωτερικά δεδομένα εκπαίδευσης. Το μοντέλο τους βασίστηκε στο pLSA και η διαδικασία δημιουργίας αυτού φαίνεται στην Εικόνα 23.

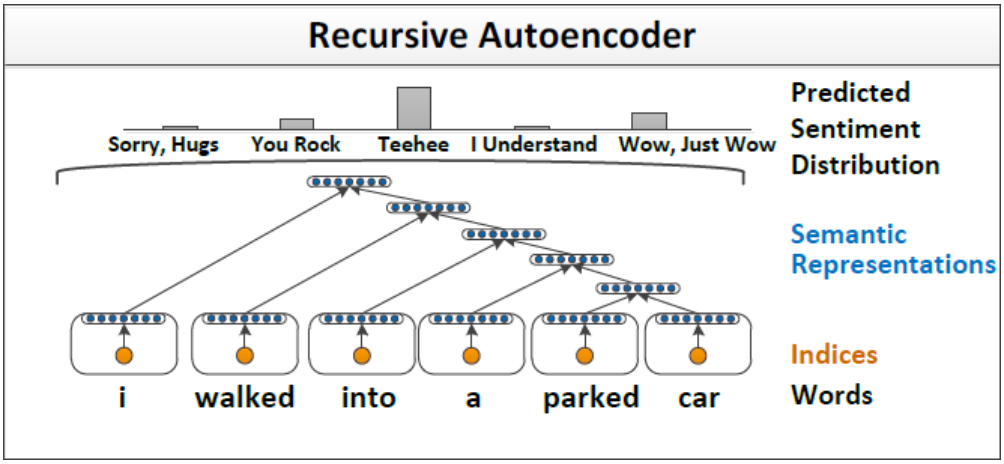


Εικόνα 23: Η διαδικασία δημιουργίας του μοντέλου στο Mei κ.α. (2007)

14. **Scaffidi κ.α. (2007)**: εξερευνούν την υπόθεση ότι οι λέξεις - κλειδιά που αντιπροσωπεύουν τα χαρακτηριστικά γνωρίσματα των προϊόντων στις προτάσεις εμφανίζονται πιο συχνά σε κείμενα κριτικών από ότι σε γενικού σκοπού αγγλικά κείμενα. Το σύστημα που δημιουργούν ονομάζεται Red Opal και η αξιολόγησή τους στην ακρίβεια (precision) εξαγωγής χαρακτηριστικών είναι 88%, ο χρόνος εκτέλεσης εξαγωγής χαρακτηριστικών είναι ανάλογος με τον αριθμό των κριτικών, ο αλγόριθμος

βαθμολόγησης επιτυγχάνει ακρίβεια (precision) 80% για τα προϊόντα με υψηλή βαθμολογία ενώ εξοικονομείται χρόνος στους τελικούς χρήστες εφόσον το σύστημά τους μειώνει το χρόνο για να βρεί τα αντικείμενα από 10 - 15 λεπτά σε 3 λεπτά.

15. **Socher κ.α. (2008)**: Μία ενδιαφέρουσα προσέγγιση στο πρόβλημα της αυτόματης εκτίμηση της διάθεσης κειμένων γίνεται στο Stanford University, όπου οι ερευνητές εισάγουν ένα μοντέλο ημι – εποπτευόμενης Μηχανικής Μάθησης, που βασίζεται σε αναδρομικούς αυτοσυσχετιστές (recursive autoencoders - RAE) για την πρόβλεψη του συναισθήματος με κατανομή σε επίπεδο προτάσεων, μαθαίνοντας να αναπαριστά προτάσεις ως διάνυσμα. Στην πρόβλεψη του συναισθήματος αυτή η προσέγγιση έχει καλύτερα αποτελέσματα από άλλες μεθόδους, που δεν κάνουν χρήση προκαθορισμένων λεξικών ή κανόνων αλλαγής πολικότητας σε ευρέως χρησιμοποιημένες συλλογές όπως οι κριτικές ταινιών. Το συγκεκριμένο μοντέλο αξιολογήθηκε, εκτός των άλλων, στην ικανότητα να προβλέψει κατανομές συναισθημάτων σε μία νέα συλλογή δεδομένων που αποτελείται από προσωπικές ιστορίες χρηστών σχολιασμένες με πολλαπλές ετικέτες συναισθημάτων και περιγράφουν διάφορες συναισθηματικές αντιδράσεις. Ο αλγόριθμός τους μπορεί και προβλέπει με μεγαλύτερη ακρίβεια την κατανομή αυτών των συναισθημάτων σε σχέση με άλλους αλγόριθμούς. Η εργασία τους εκμεταλλεύεται την ιεραρχική δομή των λέξεων μιας πρότασης και την χρησιμοποιεί για να κατανοήσει την συναισθηματική σημασιολογία τους. Επιπλέον, μπορεί να εκπαιδευτεί για οποιαδήποτε δεδομένα και δεν απαιτεί ειδικά λεξικά γνώμης, χαρακτηρίζοντάς το με μεγαλύτερο βαθμό γενίκευσης. Τέλος, η πρόβλεψη έγκειται σε μια πολυδιάστατη κατανομή από συναισθηματικές καταστάσεις και όχι στην συνηθέστερη δισδιάστατη κατανομή θετικό / αρνητικό συναίσθημα. Στην Εικόνα 24 απεικονίζεται το μοντέλο τους. Οι αναδρομικοί αυτοσυσχετιστές χρησιμοποιούν ως είσοδο συνεχή διανύσματα λέξεων και το μοντέλο μαθαίνει να αναπαριστά διανύσματα προτάσεων, καθώς γίνεται εμφανής και η ιεραρχική δομή τους. Στην συνέχεια το μοντέλο μαθαίνει να κατανέμει σε μία σειρά συναισθημάτων για κάθε κόμβο.

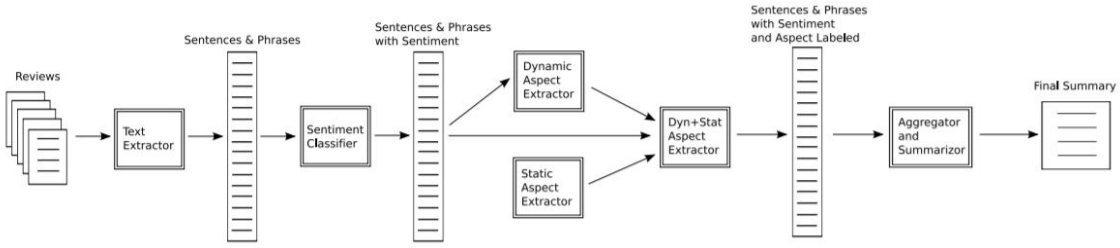


Εικόνα 24: Αναπαράσταση των recursive autoencoders στο Socher κ.α. (2008)

16. **Stoyanov και Cardie (2008):** οι συγγραφείς αντιμετώπισαν την εξόρυξη των στόχων ως πρόβλημα επίλυσης συσχέτισης θεματικών ενοτήτων (topic coreference). Αρχικά προσδιορίζουν το εύρος του θέματος, στη συνέχεια εκτελούν κατά ζεύγη ταξινόμηση των συναφών γνωμών ως προς το αν υπάρχει συσχέτιση θεματικής ενότητας ή όχι και ομαδοποιούν τις απόψεις σύμφωνα με τα αποτελέσματα. Πρότειναν να εκπαιδεύεται ένας ταξινομητής που θα κρίνει αν δύο λέξεις γνώμης μιλούν για τον ίδιο στόχο, πράγμα που δείχνει ότι η προσέγγισή τους είναι εποπτευόμενης μάθησης και δείχνει να υπερβαίνει σημαντικά πολλές μη τετριμμένες βασικές γραμμές σύμφωνα με τα τρία μέτρα αξιολόγησης συσχέτισης *B-CUBED* (B^3), *α* και *CEAF (Constrained EntityAlignment F-Measure)*. Αξιολογούν την προσέγγισή τους χρησιμοποιώντας ένα υπάρχον σύνολο δεδομένων, το MPQA των Deng και Wiebe (2015), επεκτείνοντάς το με χειροκίνητους σχολιασμούς που κωδικοποιούν πληροφορίες για την θεματική ενότητα με ποσοστά επιτυχίας για την συσχέτιση με τα μέτρα που προαναφέρθηκαν: B^3=64%, α=54%, *CEAF*=69% για όλες τις γνώμες.
17. **Titov και McDonald (2008)**: προτείνουν μια επέκταση του κλασσικού μοντέλου LDA, την οποία ονομάζουν multi - grain LDA (MG-LDA), υιοθετώντας έναν περιορισμό κυλιόμενου παραθύρου (κατώφλι) με μήκος λίγων προτάσεων επί των αναλυόμενων κειμένων. Το μοντέλο τους αποκτά την ικανότητα να διακρίνει τοπικές εκτός από καθολικές θεματικές ενότητες, επιλύοντας το πρόβλημα του LDA. Στις καθολικές θεματικές ενότητες αντιστοιχούν τα χαρακτηριστικά γνωρίσματα που είναι δύσκολο να εντοπιστούν, όπως ο κατασκευαστής του προϊόντος, ενώ στις τοπικές θεματικές ενότητες τα εύκολα εξαγόμενα χαρακτηριστικά γνωρίσματα, όπως η ταχύτητα του επεξεργαστή. Εδώ, κάθε λέξη - κλειδί που ανακαλύπτεται είναι μια πολυωνυμική κατανομή αυτών, καθώς διαφορετικές λέξεις - κλειδιά που εκφράζουν τις ίδιες ή σχετικές λέξεις - κλειδιά ομαδοποιούνται αυτόματα μαζί με την ίδια λέξη – κλειδί. Ωστόσο, αυτό το μοντέλο είναι αρκετά πολύπλοκο και δεν διαχωρίζει λέξεις – κλειδιά στόχους και λέξεις γνώμης. Το

συγκεκριμένο μοντέλο έδειξε ότι τα μοντέλα θεματικών ενοτήτων όπως το LDA, ενδέχεται να μην είναι κατάλληλα για την ανίχνευση λέξεων – κλειδιών από προτάσεις. Οι κριτικές που αναλύουν μιλάνε για εστιατόρια και τα ποσοστά που έπιασαν στο precision στην ανάλυση συναισθήματος ήταν 75.8% για τον εξεταζόμενο τομέα των υπηρεσιών, 85.5% για την τοποθεσία και 75% για τα δωμάτια, ενώ στη λογική παλινδρόμηση 80.8%, 94% και 88.3% αντίστοιχα.

18. **Blair-Goldensohn κ.α. (2008)**: Εξάγουν τις λέξεις – κλειδιά στηριζόμενοι στην προσέγγιση των συχνότερα χρησιμοποιούμενων ουσιαστικών σε προτάσεις που φέρουν συναίσθημα ή που ακολουθούν κάποιο συντακτικό πρότυπο συναισθήματος. Αφαιρούν τα stopwords και τις λέξεις – κλειδιά με μικρή συχνότητα εμφάνισης. Στη συνέχεια, χρησιμοποιώντας ένα χειροκίνητα κατασκευασμένο λεξικό γνώμης από γνωστές λέξεις αρνητικής και θετικής πολικότητας μέσω του Wordnet, απορρίπτουν τις λέξεις – κλειδιά που το άθροισμα των βαρών των λέξεων του λεξικού που συνυπάρχουν στα συντακτικά πρότυπα είναι μικρότερο από ένα κατώφλι που ορίζεται. Για να εντοπιστούν λέξεις – κλειδιά που αναφέρονται σε κάποιο ευρύτερο χαρακτηριστικό, χρησιμοποιήθηκε ο ταξινομητής Maximum Entropy, ο οποίος εκπαιδεύτηκε σε δύο διαφορετικούς τομείς τα χαρακτηριστικά των οποίων είχαν οριστεί εκ των προτέρων ως οι κλάσεις του ταξινομητή. Ως εκ τούτου, χρησιμοποιούν εποπτευόμενη μάθηση. Για παράδειγμα, με τον τρόπο αυτό εντοπίζονται τα χαρακτηριστικά γνωρίσματα «*φωτεινότητα*», «*χρώματα*» του ευρύτερου χαρακτηριστικού γνωρίσματος «*οθόνη*». Ο αλγόριθμός τους επιτυγχάνει μέσο precision 83.1% στη θετική κατηγορία και 84.4% στην αρνητική στις κριτικές που οι χρήστες είχαν αφήσει βαθμολογία και average precision 80.3% στη θετική κατηγορία και 71.3% στην αρνητική στις κριτικές χωρίς καμία επιπλέον πληροφορία από τους χρήστες. Η επισκόπηση του συστήματός τους φαίνεται και στην κάτωθεν Εικόνα 25. Τα κουτιά με διπλό περιθώριο αντιπροσωπεύουν στοιχεία του συστήματος και τα κουτιά μονού περιθωρίου, αρχεία κείμενων.



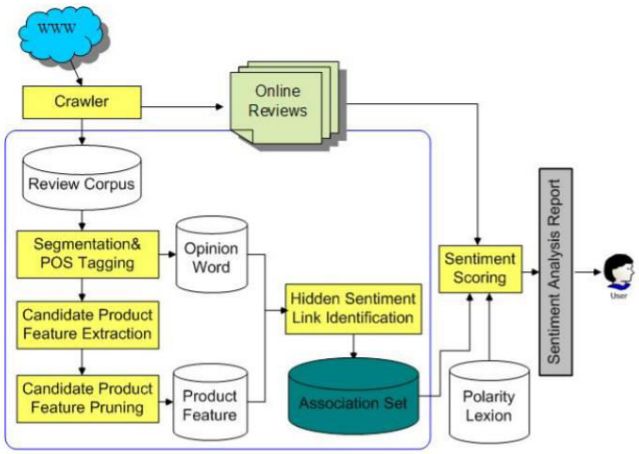
Εικόνα 25: Επισκόπηση Συστήματος Blair-Goldensohn κ.α. (2008)

19. **Ding κ.α. (2008)** : προτείνουν μια προσέγγιση εντοπισμού του σημασιολογικού προσανατολισμού των απόψεων στηριζόμενη σε λεξικό γνώμης που χρησιμοποιεί εξωτερικές ενδείξεις και γλωσσικές συμβάσεις της φυσικής γλώσσας. Με αυτή τους τη μέθοδο δείχνουν ότι πολλές λέξεις, ακόμη και του ίδιου τομέα μπορεί να έχουν

διαφορετικό προσανατολισμό ανάλογα την χρήση τους. Για παράδειγμα, η λέξη «*υψηλή*» στην πρόταση «*Η διάρκεια ζωής της μπαταρίας είναι υψηλή*» έχει θετικό προσανατολισμό ενώ στην πρόταση «*Η οθόνη εκπέμπει υψηλή ακτινοβολία*» αρνητικό. Χρησιμοποιούν γλωσσολογικά πρότυπα για τον ταυτόχρονο εντοπισμό των λέξεων που φέρουν συναίσθημα και το ίδιο το συναίσθημα αλλά και τις λέξεις - κλειδιά. Σε αυτή τη προσέγγιση διαχειρίζονται εύκολα πολλαπλές λέξεις γνώμης που συγκρούονται μέσα στην ίδια πρόταση καθώς και λέξεις γνώμης που εξαρτώνται από το περιεχόμενου του τομέα στο οποίο αναφέρονται. Χρησιμοποιούν σαν σύνολο δεδομένων κριτικές προϊόντων και στηριζόμενοι στις λίστες γνώμης (Opinion Lexicon) των Hu και Liu (2004) που χρησιμοποίησαν σαν βάση το λεξικό WordNet, επιτυγχάνουν πολλά υποσχόμενα αποτελέσματα της τάξης των precision: 92%, recall: 91% και F-Score: 91%.

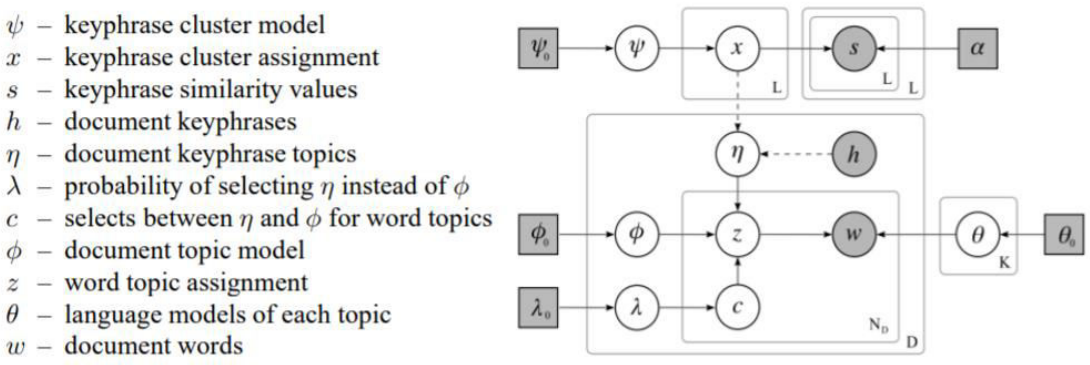
20. **Su κ.α. (2008)**: αντιμετωπίζουν τις έμμεσα εκφραζόμενες γνώμες στις κριτικές. Προτείνουν μια μέθοδο συσταδοποίησης για να χαρτογραφήσουν τις έμμεσες εκφράσεις των κριτικών, οι οποίες θεωρήθηκαν ως λέξεις γνώμης στις αντίστοιχες άμεσες εκφράσεις. Η μέθοδος εκμεταλλεύεται τη σχέση αμοιβαίας ενίσχυσης μεταξύ μιας άμεσης έκφρασης και μιας λέξης γνώμης που σχηματίζει ένα συνυπάρχον ζεύγος σε μια πρόταση το οποίο μπορεί να υποδεικνύει ότι η λέξη γνώμης περιγράφει την έκφραση ή ότι η έκφραση συνδέεται με την λέξη γνώμης. Ο αλγόριθμος μοντελοποιείται επαναλαμβάνοντας κατά συστάδες το σύνολο των άμεσων εκφράσεων και το σύνολο των λέξεων γνώμης ξεχωριστά. Σε κάθε επανάληψη, πριν από τη συσταδοποίηση κάθε συνόλου, τα αποτελέσματα συσταδοποίησης του άλλου συνόλου χρησιμοποιούνται για να ενημερώσουν την ομοιόμορφη ομοιότητα του συνόλου. Η ομοιότητα ανά ζεύγος σε ένα σύνολο καθορίζεται από ένα γραμμικό συνδυασμό ομοιότητας εντός συνόλου και ομοιότητας μεταξύ των συνόλων. Η εσωτερική ομοιότητα δύο στοιχείων είναι η παραδοσιακή ομοιότητα. Η ομοιόμορφη ομοιότητα δύο στοιχείων υπολογίζεται βάσει του βαθμού συσχετισμού μεταξύ εκφράσεων και λέξεων γνώμης. Η συσχέτιση (ή η σχέση αμοιβαίας ενίσχυσης) μοντελοποιείται χρησιμοποιώντας ένα διμερές γράφημα. Μια έκφραση και μια λέξη γνώμης συνδέονται εάν έχουν συνυπάρξει σε μια πρόταση. Οι δεσμοί σταθμίζονται επίσης με βάση τη συχνότητα συν-εμφάνισης. Μετά την επαναληπτική συσταδοποίηση, οι ισχυρότεροι δεσμοί *n* μεταξύ των εκφράσεων και των ομάδων λέξεων γνώμης σχηματίζουν τη χαρτογράφηση. Πειραματίζονται με κριτικές προϊόντων γραμμένες στα Κινέζικα και χρησιμοποιούν ένα έτοιμο WordNet-like λεξικό, το Chinese Concept Dictionary (CCD) για τη λήψη πληροφοριών της σημασιολογικής κλάσης για κάθε στοιχείο. Το precision που αποκτάνε με την προσέγγιση αμοιβαίας ενίσχυσης είναι 81,9%, σχεδόν 13 μονάδες υψηλότερη από την προσέγγιση γειτνίασης. Τα αποτελέσματα τους, όπως επισημαίνουν εξαρτώνται σε μεγάλο βαθμό από τον ορισμό

της υποκειμενικότητας που χρησιμοποιείται στον σχολιασμό των κειμένων. Η αρχιτεκτονική του συστήματός τους φαίνεται και στην Εικόνα 26.



Εικόνα 26: Αρχιτεκτονική συστήματος Su κ.α. (2008)

21. **Branavan κ.α. (2008)**: προτείνουν μια μέθοδο η οποία χρησιμοποιεί τις περιγραφές των λέξεων – κλειδιών ως τις βασικές φράσεις, σε μια μορφή κριτικών που περιλαμβάνει τα πλεονεκτήματα και μειονεκτήματα του προϊόντος, για να βοηθήσουν στην εύρεση των λέξεων – κλειδιών στο ελεύθερο κείμενο της κριτικής που είναι υπό εξέταση. Το μοντέλο τους αποτελείται από δύο μέρη. Το πρώτο μέρος συγκεντρώνει αυτές τις φράσεις από τα πλεονεκτήματα και τα μειονεκτήματα, σε κατηγορίες με βάση την ομοιότητα. Το δεύτερο μέρος δημιουργεί ένα μοντέλο θεματικών ενοτήτων με LDA, μοντελοποιώντας τις θεματικές ενότητες ή τις λέξεις - κλειδιά του κειμένου, ταυτόχρονα με το πρώτο μέρος. Τα δύο μέρη είναι υλοποιημένα με βάση την ιδέα ότι το μοντέλο αποτρέπει την ανάθεση των κρυφών θεματικών ενοτήτων στο κείμενο της κριτικής να είναι παρόμοιες με τις θεματικές ενότητες που αντιπροσωπεύουν τις βασικές φράσεις στα πλεονεκτήματα και μειονεκτήματα της κριτικής, αλλά επιτρέπει και ορισμένες λέξεις - κλειδιά στο έγγραφο να αντληθούν από άλλες θεματικές ενότητες που δεν αντιπροσωπεύονται από τις βασικές φράσεις. Αυτή η ευελιξία στη σύζευξη επιτρέπει στο μοντέλο να μάθει αποτελεσματικά με την παρουσία ελλιπών βασικών φράσεων, ενώ παράλληλα ενθαρρύνει τη συσταδοποίηση των βασικών φράσεων ώστε να συμπίπτει με τις θεματικές ενότητες του κείμενου της κριτικής. Παρατηρούν ότι η κοινή εξαγωγή συμπερασμάτων παράγει καλύτερες συστάδες από τη χρήση μόνο λέξεων - κλειδιών. Ωστόσο, αυτή η προσέγγιση εξακολουθεί να μην διαχωρίζει λέξεις - κλειδιά στόχους και λέξεις γνώμης. Στηρίζονται σε κριτικές κινητών τηλεφώνων με αποτελέσματα recall: 88%, precision: 58% και f-score: 70% και εστιατορίων με τα αντίστοιχα αποτελέσματα recall: 90%, precision: 60% και f-score: 75%. Το βασικό τους μοντέλο μπορεί να αποδοθεί και από την παρακάτω Εικόνα 27 σε μορφή plate notation.

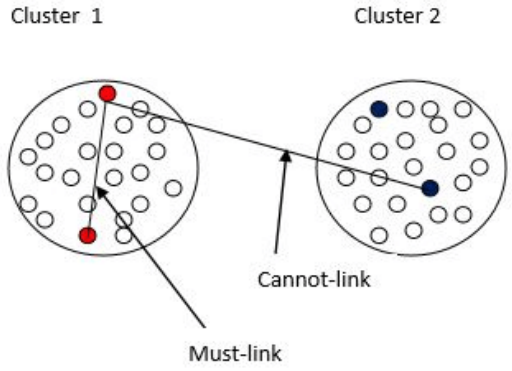


Εικόνα 27: Διάγραμμα μοντέλου Branavan κ.α. (2008)

22. **Guo κ.α. (2009)**: εδώ παρουσιάστηκε μια μέθοδος που ονομάζεται πολύ - επίπεδη λανθάνουσα σημασιολογική συσχέτιση (latent semantic association - LaSA). Στο πρώτο επίπεδο, όλες οι λέξεις / φράσεις – κλειδιά ομαδοποιούνται σε ένα σύνολο θεματικών ενοτήτων με τη χρήση του αλγορίθμου LDA και τα αποτελέσματα χρησιμοποιούνται για την κατασκευή δομών λανθάνουσας θεματικής ενότητας για λέξεις – κλειδιά εκφράσεις, αποτελούμενες δηλαδή από περισσότερες της μιας λέξης. Για παράδειγμα, έχουμε τέσσερις εκφράσεις: "*φωτογραφίες ημέρας*", "*ημερήσια φωτογραφία*", "*ημερήσιες φωτογραφίες*" και "*φωτογραφία κατά τη διάρκεια της ημέρας*". Εάν ο LDA ομαδοποιήσει τις μεμονωμένες λέξεις "*ημέρα*" και "*ημερήσια*" στο θέμα1, και " *φωτογραφία* " και "*φωτογραφίες*" στο θέμα2, το σύστημα θα ομαδοποιήσει και τις τέσσερις εκφράσεις σε μία ομάδα, το οποίο αποκαλούμε "θέμα1 – θέμα2" και ονομάζεται δομή λανθάνουσας θεματικής ενότητας. Στο δεύτερο επίπεδο, οι λέξεις – κλειδιά εκφράσεις ομαδοποιούνται ξανά από τον LDA αλλά σύμφωνα με τις δομές που έχουν παραχθεί στο επίπεδο 1 και τις τριγύρω λέξεις. Ακολουθώντας το παραπάνω παράδειγμα, οι "*φωτογραφίες ημέρας*", "*ημερήσια φωτογραφία*", "*ημερήσιες φωτογραφίες*" και "*ημερήσια φωτογραφία*" στο θέμα "θέμα1 – θέμα2" σε συνδυασμό με τις τριγύρω λέξεις αποτελούν ένα έγγραφο. Το LDA τρέχει σε αυτά τα έγγραφα για να παράγει το τελικό αποτέλεσμα. Αξιολογούν την προτεινόμενη προσέγγισή τους (που ονομάζουν Multi-LaSA) χρησιμοποιώντας κριτικές τριών τομέων ηλεκτρονικών προϊόντων. Τα δύο σύνολα δεδομένων είναι αγγλικές κριτικές στους τομείς ψηφιακών φωτογραφικών μηχανών και φορητών υπολογιστών και το δεύτερο είναι κινεζικές κριτικές στον τομέα των κινητών τηλεφώνων. Η μέθοδός τους όπως επισημαίνουν, καταργεί αποτελεσματικά τους άκυρους υποψήφιους όρους με την επαναλαμβανόμενη επαλήθευση και εντοπίζει χαρακτηριστικά γνωρίσματα του προϊόντος χαμηλότερης συχνότητας εμφάνισης. Το Multi-LaSA επιτυγχάνει accuracy 81%, 85% και 82% στις φωτογραφικές μηχανές, φορητούς υπολογιστές και κινητά τηλέφωνα, αντίστοιχα. Σε σύγκριση με τον K-means αυξάνει την ακρίβεια κατά 10,33%, 14,08% και 29,90%, αντίστοιχα ενώ σε σύγκριση με τη μέθοδο LDA, βελτιώνει σημαντικά την ακρίβεια με 7.41%, 12.50% και 25.29% αντίστοιχα. Επιπλέον, δείχνουν

ότι η μέθοδός τους είναι ανεξάρτητη γλώσσας και τομέα (language & domain independent). Στην επόμενη έκδοσή τους χρησιμοποίησαν μια παρόμοια ιδέα για την ομαδοποίηση των λέξεων – κλειδιών από διαφορετικές γλώσσες σε κατηγορίες, οι οποίες μπορούν να χρησιμοποιηθούν για να συγκρίνουν απόψεις με διαφορετικές απόψεις από διαφορετικές χώρες.

23. **Andrzejewski κ.α. (2009)**: εδώ χρησιμοποιούνται γνώσεις ή περιορισμοί τομέα για να καθοδηγήσουν τη θεματική μοντελοποίηση στη δημιουργία καλύτερων θεματικών ενοτήτων. Οι περιορισμοί έχουν είτε τη μορφή Must - Links όπου στην ομαδοποίηση δύο στοιχεία πρέπει να βρίσκονται στην ίδια κλάση είτε τη μορφή Cannot - Links όπου δύο στοιχεία δεν μπορούν να βρίσκονται στην ίδια κλάση, όπως φαίνεται και στην Εικόνα 28. Ωστόσο, η μέθοδος μπορεί να οδηγήσει σε μια εκθετική ανάπτυξη στην κωδικοποίηση των Cannot-Links περιορισμών και αυτός ο μεγάλος αριθμός περιορισμών δημιουργεί τη δυσκολία διαχείρισής τους. Αποδεικνύουν ότι οι θεματικές ενότητες που προκύπτουν όχι μόνο ενσωματώνουν με επιτυχία τις συγκεκριμένες γνώσεις του τομέα, αλλά γενικεύουν πέρα από αυτόν, συμπεριλαμβάνοντας / εξαιρώντας άλλες σχετικές. Η συλλογή δεδομένων που χρησιμοποίησαν για τα πειράματα αποτελείται από 18.193 αποσπάσματα που επιλέχθηκαν από τη βάση δεδομένων MEDLINE για τη σχέση τους με γονίδια ζύμης.



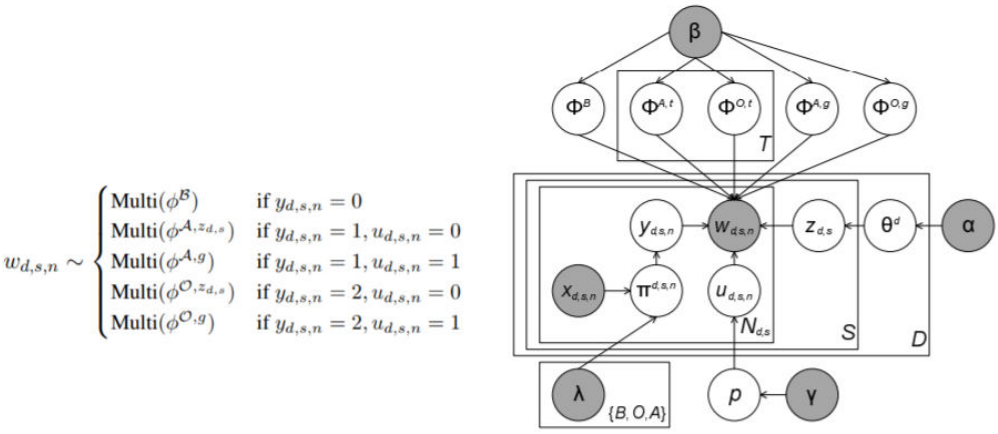
Εικόνα 28: Περιορισμοί must-link & cannot-link των Andrzejewski κ.α. (2009) [26]

24. **Brody και Elhadad (2010)**: χρησιμοποιούν το LDA σε επίπεδο προτάσεων. Πιο συγκεκριμένα, προτείνουν να εντοπίζονται αρχικά οι λέξεις - κλειδιά χρησιμοποιώντας θεματική μοντελοποίηση και στη συνέχεια να προσδιορίζεται το συναίσθημα των λέξεων λαμβάνοντας υπόψιν μόνο τα επίθετα της πρότασης. Η προσέγγιση αυτή στηρίζεται στην φύση των κειμένων προς ανάλυση, καθώς όταν το μοντέλο LDA χρησιμοποιείται σε επίπεδο κειμένων, τα χαρακτηριστικά γνωρίσματα που εξάγει συνήθως είναι αυτά που δεν μπορούν να εντοπιστούν (π.χ. κατασκευαστής) ενώ εάν χρησιμοποιηθεί σε επίπεδο προτάσεων παράγει τις θεματικές ενότητες των επιθυμητών χαρακτηριστικών. Παρουσιάζουν ένα απλό και ευέλικτο, σε σχέση με τον τομέα και τη γλώσσα, σύστημα μη εποπτευόμενης μάθησης για την εξαγωγή γνώμης λαμβάνοντας υπόψη την επίδραση

του τομέα στην πολικότητα των συναισθημάτων, ένα ζήτημα που αγνοήθηκε σε μεγάλο βαθμό στην προηγούμενη βιβλιογραφία. Δείχνουν την αποτελεσματικότητά του και στις δύο περιπτώσεις, όπου επιτυγχάνει παρόμοια αποτελέσματα για πιο πολύπλοκες μεθόδους ημι-εποπτευόμενης μάθησης που περιορίζονται από την εξάρτησή τους με τον χειροκίνητο σχολιασμό. Για την σύγκριση των βαθμολογίων (rankings [60]) χρησιμοποιήσαν τα μέτρα αξιολόγησης Kendall's coefficient (τ_k) με μέσο όρο 0.36 για τα αυτόματα δεδομένα και 0.45 για τα χειροκίνητα επισημασμένα ενώ Kendall's distance (D_k) με μέσο όρο 0.32 για τα αυτόματα δεδομένα και 0.27 για τα χειροκίνητα επισημασμένα δεδομένα.

25. **Zhao κ.α. (2010)**: πρότειναν το υβριδικό μοντέλο MaxEnt-LDA (έναν συνδυασμό της Μέγιστης Εντροπίας και του LDA) για να ανακαλύψουν από κοινού / ταυτόχρονα / παράλληλα τόσο τις λέξεις - κλειδιά στόχους όσο και τις ειδικές λέξεις γνώμης και αξιοποιώντας τα συντακτικά χαρακτηριστικά να διαχωρίσουν τις λέξεις - κλειδιά στόχους από τις λέξεις γνώμης με το μειονέκτημα ότι δεν είναι προσαρμόσιμα σε όλους τους τομείς. Για την εκπαίδευση του μοντέλου χρησιμοποιήθηκαν κριτικές για εστιατόρια με 46 προτάσεις, για κινητά τηλέφωνα με 125 προτάσεις και για DVD players με 180 προτάσεις. Αξιολόγησαν το μοντέλο τους MaxEnt-LDA χρησιμοποιώντας δύο μεγάλα σύνολα δεδομένων 1.644.923 κριτικών για εστιατόρια και 1.097.739 κριτικών για ξενοδοχεία, αφαιρώντας stopwords και υλοποιώντας τον Stanford POS Tagger. Υπολογίζοντας το μέτρο rank precision εντοπίζουν καλύτερα αποτελέσματα του αλγορίθμου MaxEnt-LDA στο *p@5*=82.5% σε σχέση με προηγούμενες εργασίες και ξεχωριστά για κάθε τομέα, υπολόγισαν το μέτρο nDCG και βρήκαν ότι ο MaxEnt-LDA αποδίδει καλύτερα στον τομέα των εστιατορίων με *nDCG@10*=89.7% σε σχέση με το *nDCG@10*=78.1% προηγούμενων προσεγγίσεων και στον τομέα των ξενοδοχείων με *nDCG@5*=82% σε σχέση με το *nDCG@10*=78.2% προηγούμενων προσεγγίσεων. Αυτό που αξίζει να σημειωθεί για το μοντέλο τους είναι μπορεί να αποδώσει καλά και με σχετικά μικρό σύνολο δεδομένων εκπαίδευσης και από διαφορετικό τομέα. Στην Εικόνα 29 διακρίνεται σε μορφή plate notation το μοντέλο τους.

[60] Η διαφορά των rating και ranking έγκειται στο ότι στο rating δεν υπάρχει η έννοια της σύγκρισης.



Εικόνα 29: Διάγραμμα μοντέλου των Zhao κ.α. (2010)

26. **Wei και Gulla (2010)**: προτείνουν μια νέα προσέγγιση ονόματι HL-SOT για την επισήμανση των χαρακτηριστικών γνωρισμάτων ενός προϊόντος και τα σχετικά τους συναισθήματα σε κριτικές προϊόντων με μία διαδικασία ιεραρχικής μάθησης (Hierarchical Learning - HL) ορισμένης με ένα δέντρο οντολογίας συναισθήματος (Sentiment Ontology Tree - SOT). Επισημαίνουν ότι αν και η προσέγγισή τους είναι κυρίως για την ανάλυση των συναισθημάτων πάνω σε κριτικές προϊόντων, μπορεί εύκολα να γενικευτεί στην επισήμανση περισσότερων του ενός τομέα προϊόντων. Το σύνολο δεδομένων τους αποτελείται από 1.446 κριτικές πελατών στις ψηφιακές φωτογραφικές μηχανές σύμφωνα με τις οποίες δημιούργησαν χειροκίνητα ένα δέντρο οντολογίας (SOT) για το προϊόν των ψηφιακών φωτογραφικών μηχανών αποτελούμενος από 105 κόμβους που περιλαμβάνουν 35 κόμβους χωρίς φύλλα (non-leaf) που αντιπροσωπεύουν τα χαρακτηριστικά γνωρίσματα της ψηφιακής φωτογραφικής μηχανής και 70 κόμβους με φύλλα (leaf) που αντιπροσωπεύουν τα συναφή συναισθήματα με το εκάστοτε χαρακτηριστικό γνώρισμα κόμβος. Δεδομένου ότι η προτεινόμενη προσέγγιση HL-SOT είναι ιεραρχική διαδικασία ταξινόμησης, χρησιμοποιούν τρεις κλασσικές λειτουργίες απώλειας για τη μέτρηση της απόδοσης ταξινόμησης, One-error Loss (O-Loss), the Symmetric Loss (S-Loss) και Hierarchical Loss (H-Loss) στα οποία η μικρότερη αξία απώλειας σημαίνει και καλύτερη απόδοση. Επιτυγχάνουν τα καλύτερα αποτελέσματα με διαστασιμότητα [61]: 220 με O-Loss=0.84, S-Loss=2.28, H-Loss=1.02. Έτσι καταλήγουν στο συμπέρασμα ότι ο αντίκτυπος της διαστασιμότητας δείχνει ότι η ευρετηρίαση περισσότερων όρων βελτιώνει την ακρίβεια της προτεινόμενης προσέγγισης HL-SOT αλλά μειώνοντας σημαντικά την υπολογιστική αποδοτικότητα.
27. **Jakob και Gurevych (2010)**: σε αυτή την εργασία οι συγγραφείς επικεντρώνονται στον εντοπισμό γνώμης των στόχων ως μέρος της εξαγωγής συναισθήματος. Μοντελοποιούν

[61] Η διαστασιμότητα (dimensionality) είναι η διάσταση ενός διανυσματικού χώρου, το πλήθος των διαστάσεών του [25].

το πρόβλημα ως πληροφορία εξόρυξης, την οποία αντιμετωπίζουν με βάση Conditional Random Fields (CRF). Ως βασική γραμμή χρησιμοποιούν τον εποπτευόμενης μάθησης αλγόριθμο από τους Zhuang κ.α. (2006) και αξιολογούν εκτενώς τους αλγορίθμους τους σε σύνολα δεδομένων από τέσσερις διαφορετικούς τομείς σχολιασμένα με συγκεκριμένες περιπτώσεις στόχων γνώμης σε επίπεδο προτάσεων. Διερευνούν την προσέγγισή τους που βασίζεται στα CRF η οποία βελτιώνει την απόδοση από 0.077, 0.126, 0.071 και 0.178 όσον αφορά το F-Measure στην εξαγωγή στους τέσσερις τομείς αντίστοιχα.

28. **Zhai κ.α. (2010)**: μια μέθοδος ημι - εποπτευόμενης μάθησης για την ομαδοποίηση των λέξεων – κλειδιών σε κατηγορίες που έχει προ-καθορίσει ο χρήστης. Αρχικά ο χρήστης επισημαίνει έναν μικρό αριθμό λέξεων – κλειδιών για κάθε κατηγορία που αντικατοπτρίζει τις ανάγκες του και το σύστημα, στη συνέχεια, αποδίδει στις υπόλοιπες λέξεις – κλειδιά τις κατάλληλες κατηγορίες χρησιμοποιώντας μια μέθοδο ημι - εποπτευόμενης μάθησης που λειτουργεί με επισημασμένα και μη παραδείγματα. Η μέθοδος που χρησιμοποιείται είναι η Expectation - Maximization (EM) καθώς αποδεικνύεται αποτελεσματική και επιτρέπει την εύκολη χρήση προηγούμενης γνώσης (prior knowledge) (ή αλλιώς παραδοχές). Εδώ χρησιμοποιείται με την παραδοχή ότι οι λέξεις – κλειδιά εκφράσεις ή φράσεις – κλειδιά που μοιράζονται κοινές λέξεις (π.χ. "*διάρκεια ζωής μπαταρίας*" και "*ισχύς μπαταρίας*") ή είναι συνώνυμες σε ένα λεξικό (π.χ. "*εμφάνιση*" και "*εικόνα*") είναι πιθανό να ανήκουν στην ίδια κατηγορία. Για την αποδοχή της γενικότητας των προτεινόμενων μεθόδων τους πειραματίστηκαν χρησιμοποιώντας κριτικές προϊόντων homecinema, ασφάλειες, στρώματα, αυτοκίνητα και ηλεκτρικές σκούπες. Τα αποτελέσματά τους εξακριβώνουν ότι η χρήση της προηγούμενης γνώσης βοηθά τον EM να παράγει καλύτερα αποτελέσματα ταξινόμησης. Με τον προτεινόμενο αλγόριθμο Soft Constrained Expectation Maximizer (SC-EM) πέτυχαν accuracy: 81%, purity: 82%, entropy:69%.

    Στην επόμενη έκδοσή τους οι Zhai κ.α. (2011) προτείνουν μια διαφορετική αλλά ευρετική προσέγγιση, το λεγόμενο Constrained - LDA, όπου χρησιμοποιήθηκαν οι ίδιοι χαλαροί περιορισμοί (ή παραδοχές) των κοινών λέξεων / φράσεων – κλειδιών και της λεξικής ομοιότητας με τον αλγόριθμο EM, αλλά αποκλείστηκε η ανάγκη να ζητηθεί από τον χρήστη η παροχή επισήμανσης των αρχικών λέξεων – κλειδιών για κάθε κατηγορία. Χρησιμοποιεί τους περιορισμούς στη δειγματοληψία Gibbs των Griffiths και Steyvers (2004) για να προκαλέσουν τον υπολογισμό της δεσμευμένης πιθανότητας ανάθεσης θεματικής ενότητας σε μια λέξη – κλειδί.

29. **Zhang και Liu (2011)**: εντοπίζουν τις λέξεις γνώμης σύμφωνα με τον τομέα. Τα ουσιαστικά και φράσεις ουσιαστικών που υποδεικνύουν το χαρακτηριστικό γνώρισμα του προϊόντος εντοπίζονται χρησιμοποιώντας ένα μοντέλο εξόρυξης γνώμης βασισμένο στις λέξεις – κλειδιά (ABSA). Το συναίσθημα για κάθε χαρακτηριστικό ουσιαστικό

καθορίζεται στο βήμα υποψήφιας αναγνώρισης όπου παράγονται ένα λεξικό υποψηφίων χαρακτηριστικών γνωρισμάτων με θετικές απόψεις και ένα λεξικό υποψηφίων χαρακτηριστικών γνωρισμάτων αρνητικών απόψεων. Το χαρακτηριστικό γνώρισμα του προϊόντος μεταβάλλεται άμεσα σε θετικές και αρνητικές λέξεις γνώμης στο βήμα κλαδέματος. Το λεξικό γνώμης που ακολουθούν οι Ding κ.α. (2008) χρησιμοποιήθηκε για να προσδιορίσει την πολικότητα της γνώμης για κάθε χαρακτηριστικό γνώρισμα προϊόντος στη πρόταση. Για μια φράση *s* που περιέχει ένα χαρακτηριστικό γνώρισμα του προϊόντος *f*, οι λέξεις της φράσης σε πρώτη φάση προσδιορίζονται με την ταύτιση των λέξεις στο λεξικό της γνώμης. Προσδιορίζεται ένας βαθμός προσανατολισμού για το *f* και ο σημασιολογικός προσανατολισμός της θετικής λέξεων έχει βαθμολογία *+1* και μια αρνητική λέξη έχει βαθμολογία *-1*. Κατά την άθροιση όλων των βαθμολογιών, εάν η τελική βαθμολογία είναι θετική, τότε η γνώμη για το χαρακτηριστικό γνώρισμα στο *s* είναι θετική, αλλιώς είναι αρνητική. Πραγματοποίησαν πειράματα χρησιμοποιώντας τέσσερα διαφορετικά σύνολα δεδομένων κριτικών προϊόντων για φάρμακα και στρώματα από μια εμπορική εταιρεία που τις παρείχε και για ραδιόφωνα και δρομολογητές που εξόρυξαν οι ίδιοι. Αξίζει να σημειωθεί ότι εντοπίζουν ουσιαστικά που φέρουν άποψη σε συγκεκριμένο τομέα και τέτοια ουσιαστικά δεν εμφανίζονται σε κανένα λεξικό γενικής γνώμης. Υιοθετούν την ακρίβεια της κατάταξης (rank precision *precision@N*) σαν μετρική αξιολόγησης που δίνει το ποσοστό των σωστών χαρακτηριστικών του ουσιαστικού που φέρουν άποψη στην θέση τάξης *N* και πετυχαίνουν τα καλύτερα αποτελέσματα για *N=10, Precision@10*: 60% στα στρώματα, 60% στα φάρμακα, 50% στους δρομολογητές, 40% στα ραδιόφωνα. Η προτεινόμενη μέθοδός τους καθορίζει την πολικότητα των λέξεων – κλειδιών όχι μόνο από λέξεις γνώμης που τροποποιούν τις λέξεις στόχους, αλλά και από το περιβάλλον που την περιβάλλει.

30. **Yu κ.α. (2011)**: μια πιο εξελιγμένη μέθοδος στηριζόμενη κι αυτή σε κριτικές προϊόντων για την τελική ταξινόμηση σε κατηγορίες η οποία χρησιμοποιεί ένα σύνολο μέτρων απόστασης συνδυασμένων με μια στρατηγική βελτιστοποίησης. Η προσέγγισή τους αποτελείται από τέσσερα συστατικά, αρχικά αποκτούν την ιεραρχία, ακολουθεί ο προσδιορισμός των λέξεων – κλειδιών, η σημασιολογική μάθηση και η δημιουργία της ιεραρχίας των λέξεων – κλειδιών. Το σύνολο δεδομένων που χρησιμοποιήθηκε περιέχει κριτικές από καταναλωτές σε έντεκα δημοφιλή προϊόντα σε τέσσερις τομείς που εξήχθησαν από διάφορα διαδεδομένα φόρουμ και ιστότοπους όπως το cnet.com, viewpoints.com, reevoo.com και gsmarena.com. Εξετάζουν την αποτελεσματικότητα της προσέγγισής τους στην αναγνώριση των λέξεων – κλειδιών, στη δημιουργία της ιεραρχίας και στην εξαγωγή μη ξεκάθαρων λέξεων – κλειδιών. Αρχικά, διαπιστώνουν ότι οι μέθοδοι ομαδοποίησης και οι μέθοδοι βασισμένες σε πρότυπα παρουσιάζουν χαμηλή απόδοση. Ξεπερνώντας τις επιδόσεις της μεθόδου των Su κ.α. (2008) κατά περισσότερο

από 9,18% στο F1, δείχνουν ότι η προσέγγισή τους μπορεί να προσδιορίζει μη ξεκάθαρα χαρακτηριστικά γνωρίσματα με την αξιοποίηση υποκείμενων ενώσεων μεταξύ των λέξεων γνώμης και κάθε λέξεων στόχους της ιεραρχίας. Για την ταξινόμηση του συναισθήματος με τον ταξινομητή SVM επιτυγχάνουν μέσο F1: 78.7% στα έντεκα προϊόντα.

31. **Lakkaraju κ.α. (2011)**: μια σειρά κοινών μοντέλων προτάθηκε εδώ με βάση το σύνθετο μοντέλο θεματικών ενοτήτων του HMM-LDA (Hidden Markov Model - Latent Dirichlet Allocation), το οποίο λαμβάνει υπόψιν τόσο μια ακολουθία λέξεων όσο και μεμονωμένη λέξη (n - grams). Είναι κατά κύριο λόγο μία μη εποπτευόμενη προσέγγιση που αντιμετωπίζει το καθήκον της αναλύσεως συναισθημάτων βασισμένης στις λέξεις - κλειδιά (ABSA) ως διαδικασία δύο σταδίων. Έτσι, τα μοντέλα τους, FACTS (FACeT & Sentiment extraction), CFACTS (Coherence based FACTS), CFACTS-R (CFACTS με Rating) και FACTS-R (FACTS με Rating), μπορούν να συλλάβουν τόσο τις συντακτικές δομές όσο και τις σημασιολογικές εξαρτήσεις, καθώς είναι σε θέση να ανακαλύψουν κρυμμένες λέξεις - κλειδιά και τις αντίστοιχες αξιολογήσεις τους (rating). Εκτέλεσαν τα πειράματά τους με το σύνολο δεδομένων κριτικές προϊόντων του amazon.com στις κατηγορίες ψηφιακές φωτογραφικές μηχανές (61.482), φορητούς υπολογιστές (10.011), κινητά τηλέφωνα (6.348), τηλεοράσεις LCD (2.346) και εκτυπωτές (2.397). Στον Πίνακα 9 φαίνεται η ακρίβεια που πετυχαίνουν τα μοντέλα τους για τα δύο στάδια ανάλυσης. Το JST (joint sentiment topic) αναφέρει την βασική (baseline) ακρίβεια των εργασιών της ανάλυσης του συναισθήματος που υπάρχουν.

Πίνακας 10: Accuracy των μοντέλων των Lakkaraju κ.α. (2011) στα 2 στάδια ανάλυσης

1. Ανίχνευση Συναισθήματος σε επίπεδο λέξης και πρότασης			**2. Ανίχνευση πολικότητας**	
Model	Word accuracy	Sentence accuracy	accuracy 2 κλάσεων	accuracy 5 κλάσεων
CFACTS-R	77.68%	80.54%	83.98%	**77.87%**
CFACTS	**78.22%**	**81.28%**	**84.52%**	75.02%
FACTS-R	72.02%	72.25%	78.02%	71.76%
FACTS	72.98%	75.72%	78.19%	70.28%
JST	73.14%	76.18%	78.38%	69.39%

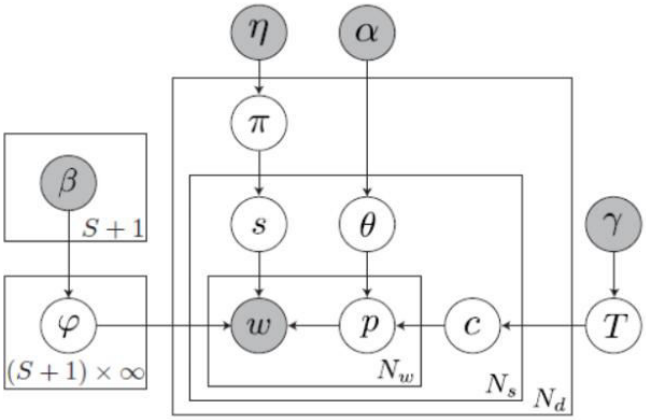
32. **Qiu κ.α. (2011)**: χρησιμοποιούν ημι - εποπτευόμενες μεθόδους για την ανάλυση του προβλήματος εξόρυξης γνώμης, όπως επέκταση λεξικών γνώμης και εξαγωγή λέξεων γνώμης. Η μεθοδολογία τους ορίζει ένα σύνολο κανόνων στο συντακτικό δέντρο κάθε πρότασης, που τους επιτρέπει να αναγνωρίζουν λέξεις - κλειδιά που γειτονεύουν με υποκειμενικούς όρους και αντιστρόφως. Ονομάζουν το μοντέλο τους διπλή διάδοση (Double Propagation) καθώς διαδίδει πληροφορίες μεταξύ λέξεων – κλειδιών στόχους αλλά και στόχων γνώμης. Ένα βασικό πλεονέκτημα της προτεινόμενης μεθόδου είναι ότι

χρειάζεται μόνο ένα αρχικό λεξικό γνώμης για να ξεκινήσει η διαδικασία εκκίνησης. Κατά την αξιολόγηση, συγκρίνουν την προτεινόμενη μέθοδό τους Prop-dep (που χρησιμοποιεί διπλή διάδοση) με πολλές άλλες παρόμοιας εργασίας, χρησιμοποιώντας την συλλογή κριτικών προϊόντων των Hu και Liu (2004) για φωτογραφικές μηχανές, DVD player, mp3 player, κινητά και τα αποτελέσματα του F-score δείχνουν ότι η προσέγγισή τους ξεπερνά των Hu και Liu (2004), Popescu και Etzioni (2005), PLSA, CRF και CRF-D κατά 10%, 4%, 32%, 43% και 32% αντίστοιχα, γεγονός που υποδεικνύει ότι οι κανόνες που ορίζονται βάσει των σχέσεων είναι αποτελεσματικοί και η ιδέα διάδοσης (propagation) είναι ισχυρή.

33. **Mukherjee και Liu (2012)**: ένα ημι - εποπτευόμενο κοινό μοντέλο, το οποίο επιτρέπει στον χρήστη να παρέχει ορισμένες λέξεις - κλειδιά για κάποιες θεματικές ενότητες προκειμένου να παράγει τις κατανομές των λέξεων – κλειδιών σε θεματικές ενότητες που να ανταποκρίνονται στις ανάγκες του χρήστη. Το μοντέλο επίσης διαχωρίζει τις λέξεις γνώμης από τις λέξεις – κλειδιά στόχους. Χρησιμοποίησαν 101.234 κριτικές ξενοδοχείων από το tripadvisor.com με 692.783 προτάσεις σαν σύνολο δεδομένων, από το οποίο αφαίρεσαν τα σημεία στίξης, τα stopwords και λέξεις που εμφανίζονταν λιγότερο από 5 φορές στο κείμενο. Το προτεινόμενο μοντέλο τους το ονομάζουν ME-SAS (Maximum-Entropy Seeded Aspect & Sentiment) και το εκπαίδευσαν με τον αλγόριθμο Μέγιστης Εντροπίας, ορίζοντας την παράμετρό του λ χρησιμοποιώντας το λεξικό γνώμης (Opinion Lexicon) των Hu και Liu (2004) και χωρίς καμία επιπλέον επισημασμένη πληροφορία από το χρήστη. Χρησιμοποιούν σαν μετρική αξιολόγησης την ακρίβεια της κατάταξης (rank precision *precision@N*) επιτυγχάνοντας τα καλύτερα αποτελέσματα 88% με τον αλγόριθμο ME-SAS για *p@10* σε σύγκριση με τους τέσσερις αλγορίθμους ME-LDA (Maximum Entropy LDA), DF-LDA (Dirichlet Forest LDA) των Andrzejewski κ.α. (2009), DF-LDA-Relaxed, SAS (Seeded Aspect & Sentiment) σε κάθε θεματική ενότητα.
34. **Kim κ.α. (2013)**: προτείνουν ένα μοντέλο ιεραρχικής συναισθηματικής διάστασης που ονομάζουν HASM (Hierarchical Aspect Sentiment Model) για την ανακάλυψη μιας ιεραρχικής θεματικής ενότητας συναισθημάτων βασισμένων σε λέξεις – κλειδιά από μη επισημασμένες κριτικές. Στο HASM, ολόκληρη η δομή είναι ένα δέντρο και κάθε μεμονωμένος κόμβος είναι ένα δέντρο δύο επιπέδων, του οποίου η ρίζα αντιπροσωπεύει μια λέξη - κλειδί και οι κόμβοι παιδιά αντιπροσωπεύουν την πολικότητα των συναισθημάτων που συνδέονται με αυτή. Κάθε πολικότητα συναισθημάτων ή λέξη - κλειδί μοντελοποιείται ως κατανομή λέξεων. Για την αυτόματη εξαγωγή τόσο της δομής όσο και των παραμέτρων του δέντρου, χρησιμοποιήθηκε ένα Bayesian μη παραμετρικό μοντέλο. Η γραφική αναπαράσταση του HASM διακρίνεται σε μορφή plate notation στην κάτωθεν Εικόνα 30. Χαρακτηρίζεται ως ευέλικτο καθώς μπορεί να ανακαλύψει θεματικές ενότητες με περισσότερα από δύο συναισθήματα, δυνατότητα χρήσιμη στην διαδικασία

ανάλυσης διάθεσης. Τα πειράματά τους σε σύνολα δεδομένων κριτικές για laptops και ψηφιακές φωτογραφικές μηχανές από το amazon.com [62], δείχνουν ότι το HASM επιτυγχάνει καλύτερη ακρίβεια (accuracy) ταξινόμησης, της τάξεως του 85% στο μικρό σύνολο δεδομένων με 1 (ισχυρά αρνητικό) ή 5 (ισχυρά θετικό) «αστεράκια» και 76% στο μεγάλο σύνολο δεδομένων με 1 ή 2 (αρνητικό) και 4 ή 5 (θετικό) «αστεράκια», σε επίπεδο προτάσεων σε σχέση με προγενέστερα ιεραρχικά μοντέλα θεματικών ενοτήτων.

HYPERPARAMETERS	
α	Uniform Dirichlet prior for θ
β	Non-uniform Dirichlet prior for β
γ	rCRP prior
η	Beta prior for π
LATENT & VISIBLE VARIABLES	
s	Sentence sentiment
p	Word subjectivity
θ	In-sentence subjectivity distribution
c	Aspect for each sentence
π	In-document sentiment distribution
Φ	Aspect-sentiment node
φ	Aspect and sentiment-polar topic
T	rCRP tree
w	Word
PLATE NOTATIONS	
S	Number of sentiment polarities
N_d	Number of documents
N_s	Number of sentences for each document
N_w	Number of tokens for each sentence
SAMPLING NOTATIONS	
m_n	# of sentences assigned to aspect node n
M_n	# of sentences assigned to aspect node n and its all subnodes, including leaf
β_s	$= \sum_w \beta_{s,w}$
$n_{k,s}^{w,(v)}$	# of words having v^{th} vocabulary assigned to aspect k and sentiment s
$n_d^{s,(k)}$	# of k-polar sentences in document d
$n_{d,i}^{p,(k)}$	# of k-subjective words in sentence i of d
n_{-i}	Counter variable excluding index i



Εικόνα 30: Γραφική Αναπαράσταση HASM, Kim κ.α. (2013)

35. **Toh και Wang (2014)**: οι συγγραφείς εδώ μοντελοποίησαν την εξαγωγή των λέξεων – κλειδιών ως διαδοχική διεργασία τόσο επισήμανσης όσο και εξαγωγής λέξεων – κλειδιών που χρησιμοποιούν για την εκπαίδευση CRF (Conditional Random Fields). Εκτός από τα κοινά χαρακτηριστικά που χρησιμοποιούνται στα παραδοσιακά συστήματα αναγνώρισης ονοματικών οντοτήτων (NER), χρησιμοποιούν και εκτεταμένους εξωτερικούς πόρους για την κατασκευή διαφόρων λιστών ονομάτων και ομάδων λέξεων, γεγονός που όπως αποδεικνύουν βελτιώνει την απόδοση της εξαγωγής. Χρησιμοποίησαν μη επισημασμένα δεδομένα από το Multi-Domain Sentiment Dataset των Blitzer κ.α. (2007) από το amazon.com και το Yelp Phoenix Academic Dataset με κριτικές χρηστών και μεταχειρίστηκαν κριτικές εστιατορίων και laptop για την εκπαίδευση του συστήματός τους που ονομάζουν DLIREC. Τα υψηλότερα αποτελέσματα τους στην εξαγωγή λέξεων – κλειδιών στα εστιατόρια ήταν precision: 85%, recall: 82.7%, F1: 84% ενώ στα laptop ήταν precision: 81.9%, recall: 67.1%, F1: 73.7%.
36. **Pavlopoulos και Androutsopoulos (2014)**: μία μη εποπτευόμενης μάθησης επέκταση του μοντέλου των Hu και Liu (2004) για την εξαγωγή λέξεων – κλειδιών με χρήση

[62] Διαθέσιμα από: http://uilab.kaist.ac.kr/research/WSDM11/

Word2Vec προτείνεται εδώ, προσθέτοντας ένα επιπλέον βήμα κλαδέματος το οποίο χρησιμοποιεί διανυσματικές αναπαραστάσεις συνεχούς διαστήματος των λέξεων – κλειδιών, όπως στο Word2vec μοντέλο των Mikolov κ.α. (2013). Οι διανυσματικές αυτές αναπαραστάσεις μπορούν να παραχθούν εκπαιδεύοντας ένα γλωσσικό μοντέλο είτε για να προβλέπει την επόμενη λέξη - κλειδί, είτε για την πρόβλεψη της τρέχουσας λέξης - κλειδί έχοντας σαν δεδομένο τις γύρω λέξεις– κλειδιά. Σε κάθε περίπτωση, κάθε λέξη – κλειδί του λεξικού αναπαρίσταται ως ένα πυκνό διάνυσμα ενός συνεχούς διανυσματικού χώρου και οι διανυσματικές λέξεις αντιμετωπίζονται ως λανθάνουσες μεταβλητές που πρέπει να μάθουν κατά την εκπαίδευση. Χρησιμοποιήθηκε η αγγλική Wikipedia ως δεδομένα εκπαίδευσης (training dataset) του μοντέλου για την απόκτηση των διανυσματικών λέξεων, καθώς τα διανύσματα των φράσεων για τις υποψήφιες λέξεις – κλειδιά πολλών λέξεων, παράγονται με παρόμοιο τρόπο. Δημιουργούν ένα σύστημα με τρεις υπο - διεργασίες. Πρώτον, γίνεται εξαγωγή των λέξεων – κλειδιών ενός ή πολλών λέξεων τα οποία αντιστοιχούν στα χαρακτηριστικά γνωρίσματα του εκάστοτε προϊόντος που σχολιάζεται (aspect term extraction). Δεύτερον, ομαδοποιούνται όμοιες λέξεις – κλειδιά που περιγράφουν το ίδιο αντικείμενο (π.χ. «*τιμή*», «*κόστος*») (aspect term aggregation). Τρίτον, υπολογίζεται το συναίσθημα ανά λέξη – κλειδί ή κλάση, που εντοπίστηκε στο προηγούμενο βήμα (aspect term polarity estimation). Η διαφοροποίηση αυτής της προσέγγισης εντοπίζεται στο τρίτο αυτό βήμα όπου πρώτα υπολογίζεται το συναίσθημα όλης της πρότασης με δύο SVM ταξινομητές (ένας για ανίχνευση συναισθήματος και ένας για ανίχνευση πολικότητας συναισθήματος) και έπειτα χαρακτηρίζει με αυτό όλες τις λέξεις - κλειδιά της πρότασης αυτής, αντιμετωπίζοντας βέβαια πρόβλημα με τις προτάσεις που περιέχουν διαφορετικές πολικότητες. Χρησιμοποιούν τρία διαφορετικά σύνολα δεδομένων με κριτικές εστιατορίων με 3.710 αγγλικές προτάσεις, 30 ξενοδοχείων με 3.600 αγγλικές προτάσεις και 394 φορητών υπολογιστών με 3.085 αγγλικές προτάσεις. Πειραματίζονται με αυτά, με τις τέσσερις διαφορετικές μεθόδους FREQ (συχνότερα χρησιμοποιούμενα ουσιαστικά), FREQ+W2V (συχνότερα χρησιμοποιούμενα ουσιαστικά με Word2vec), H&L (μέθοδος των Hu και Liu), H&L+W2V (μέθοδος των Hu και Liu με Word2vec). Αποδεικνύουν ότι η H&L+W2V, στην οποία συμπεριλαμβάνουν ένα επιπλέον βήμα κλάδευσης που χρησιμοποιεί συνεχείς αναπαραστάσεις διανυσματικών διαστημάτων των λέξεων, με Average Weighted Precision (AWP) 66.8% στα εστιατόρια, 53.3% στα ξενοδοχεία, 38.9% στα laptop, αποδίδει καλύτερα αποδεικνύοντας πως το σύστημά τους έχει μεγάλη ικανότητα γενίκευσης. Τέλος δημοσιοποιούν τρία νέα σύνολα δεδομένων και προτείνουν τρεις νέους τρόπους αξιολόγησης Weighted Precision, Weighted Recall και Average Weighted Precision (AWP) υποστηρίζοντας ότι είναι καταλληλότερες για εργασίες εξόρυξης θεματικών ενοτήτων.

37. **Alghunaim κ.α. (2015)**: διερευνούν την αποτελεσματικότητα των Word Vector αναπαραστάσεων στο πρόβλημα της ABSA. Συγκεκριμένα, στοχεύουν σε τρείς υπο-διεργασίες: την εξόρυξη των λέξεων - κλειδιών, την ανίχνευση κατηγορίας των λέξεων-κλειδιών και την πρόβλεψη του συναισθήματός τους. Η αποτελεσματικότητα εξετάζεται σε διαφορετικά δεδομένα κειμένου από κριτικές εστιατορίων του Yelp και αποσπάσματα ειδήσεων του GoogleNews και αξιολογούν την ποιότητα των διανυσμάτων που παράγουν τα οποία εξαρτώνται από το συγκεκριμένο τομέα. Χρησιμοποιούν το μοντέλο skip-gram τουWord2Vec των Mikolov κ.α. (2013) για να υπολογίσουν τις αναπαραστάσεις των διανυσμάτων λέξεων. Τελικά, επιτυγχάνουν F1: 79,91% στην εξόρυξη των λέξεων - κλειδιών, F1: 86,75% στην ανίχνευση κατηγοριών και ακρίβεια: 72,39% στην πρόβλεψη συναισθημάτων.
38. **Wang και Liu (2015)**: ένα μοντέλο βαθιάς μάθησης (Deep Learning) που είχε από τα καλύτερα αποτελέσματα στον ετήσιο διαγωνισμό SemEval [63] του 2015. Χρησιμοποιούν βαθιά νευρωνικά δίκτυα και για τις δύο διαδικασίες εξόρυξης των λέξεων – κλειδιών και του συναισθήματος και προτείνουν μια νέα μέθοδο για τον συνδυασμό της συντακτικής δομής και των Συνελικτικών Νευρωνικών Δικτύων (CNN) ώστε να ταιριάζουν άμεσα οι λέξεις - κλειδιά με τις αντίστοιχες πολικότητές τους. Χρησιμοποίησαν δύο σύνολα δεδομένων εκπαίδευσης περίπου 550 κριτικών φορητών υπολογιστών και εστιατορίων, επισημασμένα με τις αντίστοιχες θεματικές ενότητες και πολικότητές τους επιτυγχάνοντας στο πρώτο τους πείραμα εντοπισμού θεματικής ενότητας και οντότητας precision=52% recall=50% F1=51%, στο δεύτερο πείραμα εντοπισμού πολικότητας accuracy=78% και στο τρίτο πείραμα για την εξάρτηση του τομέα accuracy=80%.

3.2 Σύγκριση Σχετικών Εργασιών

Οι σχετικές εργασίες / δημοσιεύσεις που παρουσιάστηκαν λεκτικά άνωθεν μπορούν να διαχωριστούν στους αντίστοιχους Πίνακες 10 και 11, σύμφωνα με το υπο – πρόβλημα που εντοπίζεται να αναλαμβάνουν να επιλύσουν, όπως αναφέρουμε τις δύο διεργασίες της εξαγωγής λέξεων - κλειδιών και ανίχνευσης του συναισθήματός τους της ABSA. Στον Πίνακα 12 παρουσιάζονται αυτές που προσεγγίζουν και τις δύο διεργασίες μαζί.

Πίνακας 11: Σχετικές εργασίες Εξαγωγής λέξεων - κλειδιών

ΒΑΣΗ ΣΥΝΤΑΚΤΙΚΟΥ	ΒΑΣΗ ΣΥΧΝΟΤΗΤΑΣ ΕΜΦΑΝΙΣΗΣ	ΕΠΟΠΤΕΥΟΜΕΝΗΣ ΜΑΘΗΣΗΣ	ΜΗ ΕΠΟΠΤΕΥΟΜΕΝΗΣ ΜΑΘΗΣΗΣ	ΥΒΡΙΔΙΚΑ
Zhao κ.α. (2010)	Hu και Liu (2004)	Jakob και Gurevych (2010)	Titov και McDonald (2008)	Popescu και Etzioni (2005)

[63] Το SemEval (Semantic Evaluation) είναι ένα ετήσιο workshop αξιολογήσεων υπολογιστικών συστημάτων σημασιολογικής ανάλυσης.

Qiu κ.α. (2011)	Long κ.α. (2010)		Moghaddam και Ester (2010)	Blair-Goldensohn κ.α. (2008)
Zhang κ.α. (2010)	Liu κ.α. (2005)		Lakkaraju κ.α. (2011)	Yu κ.α. (2011)
	Scaffidi κ.α. (2007)			

Πίνακας 12: Σχετικές εργασίες Ανάλυσης Συναισθήματος

ΒΑΣΗ ΛΕΞΙΚΟΥ	ΕΠΟΠΤΕΥΟΜΕΝΗΣ ΜΑΘΗΣΗΣ	ΜΗ ΕΠΟΠΤΕΥΟΜΕΝΗΣ ΜΑΘΗΣΗΣ
Hu και Liu (2004)	Blair-Goldensohn κ.α. (2008)	Popescu και Etzioni (2005)
Moghaddam και Ester (2010)	Yu κ.α. (2011)	
Zhu κ.α. (2009)	Titov και McDonald (2008)	
	Pang κ.α. (2002)	

Πίνακας 13: Σχετικές εργασίες Εξαγωγής λέξεων - κλειδιών & Ανάλυσης Συναισθήματος

ΒΑΣΗ ΣΥΝΤΑΚΤΙΚΟΥ	ΕΠΟΠΤΕΥΟΜΕΝΗΣ ΜΑΘΗΣΗΣ	ΜΗ ΕΠΟΠΤΕΥΟΜΕΝΗΣ ΜΑΘΗΣΗΣ	ΥΒΡΙΔΙΚΑ
Zhuang κ.α. (2006)	Kobayashi κ.α. (2006)	Mei κ.α. (2007)	Zhao κ.α. (2010)
	Wang και Liu (2015)	Titov και McDonald (2008)	Mukherjee και Liu (2012)
		Moghaddam και Ester (2010)	
		Yu κ.α. (2011)	
		Alghunaim κ.α. (2015)	
		Scaffidi κ.α. (2007)	
		Wilson κ.α. (2005)	

Επιπροσθέτως οι σχετικές εργασίες, ομαδοποιούνται μαζικά, ταξινομημένες σύμφωνα με την χρονιά δημοσίευσής τους, στους τέσσερις Πίνακες 13, 14, 15, 16. Η ομαδοποίηση αυτή έγινε σύμφωνα με τις αντίστοιχες κατηγορίες προσεγγίσεων της Ανάλυσης Συναισθήματος με Μηχανική Μάθηση (εποπτευόμενης μάθησης, μη εποπτευόμενης μάθησης, ημι – εποπτευόμενης μάθησης) και προσεγγίσεις με λεξικά όπως έχουμε ήδη εξετάσει, επιδεικνύοντας τις βασικές διαφορές τους, με στόχο την πιο ολοκληρωμένη και σφαιρική εικόνα της βιβλιογραφίας του τομέα (state-of-the-art) που εξετάζεται στην παρούσα διπλωματική εργασία. Όπως αντιλαμβανόμαστε δεν χρησιμοποιούν τις ίδιες μετρικές αξιολόγησης για την αποτελεσματικότητα του εκάστοτε μοντέλου που προτείνουν, εν τούτοις δεν είναι εφικτή μια ξεκάθαρη σύγκρισή τους. Αντ' αυτού σαν μέτρο σύγκρισης θα μπορούσε να θεωρηθεί ο αριθμός των αναφορών (citations) που ενέχει η κάθε δημοσίευση, παράγοντας που τις περισσότερες φορές εκδηλώνει το κύρος και την αξιοπιστία της, αλλά δεν ισχύει πάντα.

Πίνακας 14:Σχετικές Εργασίες Εποπτευόμενης Μηχανικής Μάθησης

Δημοσίευση	Citations	Έτοιμο Λεξικό	Αλγόριθμος / Μέθοδος	Dataset	Απόδοση
Pang κ.α. (2002)	7.347	-	Naive Bayes, Maximum Entropy, SVM	Κριτικές Ταινιών	SVM είχε καλύτερη επίδοση με accuracy 81.5%
Wilson κ.α. (2005)	2.661	General	BoosTexter	Επισημασμένες	accuracy 75% στην εξαγωγή λέξεων

		Inquirer Harvard Lexicon, MPQA Opinion Corpus	AdaBoost αλγόριθμος	κριτικές προϊόντων	– κλειδιών, accuracy 65% στην Ανάλυση Συναισθήματος
Kobayashi κ.α. (2007)	178	-	Co-occurrence, Contextual clues	Ιαπωνικά άρθρα εστιατορίων από το blog.livedoor.com	10% βελτίωση precision & recall από παρόμοιες εργασίες
Snyder και Barzilay (2007)	330	-	PRanking	Κριτικές εστιατορίων από το www.we8there.com.	accuracy: 67%
Mei κ.α. (2007)	777	-	pLSA, HMM	Weblog άρθρα	αποτελεσματικό για ανάλυση συναισθήματος σε θεματικές ενότητες
Scaffidi κ.α. (2007)	253	-	Information Retrieval, TFIDF για την ταξινόμηση αποτελεσμάτων αναζήτησης	Κριτικές προϊόντων με rating	precision: 88% στην εξαγωγής λέξεων – κλειδιών, precision: 80% στην βαθμολόγηση λέξεων – κλειδιών
Stoyanov και Cardie (2008)	146	MPQA Corpus	Topic coreference	MPQA Topic Corpus	Μέτρα αξιολόγησης συσχέτισης: B^3=64%, α=54%, *CEAF*=69%
Blair-Goldensohn κ.α. (2008),	342	Wordnet	Maximum Entropy	Κριτικές εστιατορίων, ξενοδοχείων	precision: 83.1% στη θετική κατηγορία και 84.4% στην αρνητική
Wei και Gulla (2010)	121	-	Hierarchical Learning, Sentiment Ontology Tree	Κριτικές πελατών στις ψηφιακές φωτογραφικές μηχανές	Με dimensionality: 220 O-Loss=0.84, S-Loss=2.28, H-Loss=1.02
Yu κ.α. (2011)	58	MPQA, WordNet, Open Directory Project (ODP), Wikipedia	PMI, SVM	Κριτικές από προϊόντα σε 4 τομείς από cnet.com, viewpoints.com, reevoo.com και gsmarena.com	Ταξινόμηση συναισθήματος με SVM F1: 78.7%
Jiang κ.α. (2011)	648	-	SVM	Tweets	Accuracy: 68.2%
Alghunaim κ.α. (2015)	11	-	Skip-gram τουWord2Vec	Κριτικές εστιατορίων του Yelp, άρθρα GoogleNews	F1: 79,91% στην εξόρυξη των λέξεων - κλειδιών, F1: 86,75% στην ανίχνευση κατηγοριών, accuracy : 72,39% στην πρόβλεψη συναισθημάτων
Wang και Liu (2015)	9	-	Deep learning, CNN	Κριτικές εστιατορίων, ξενοδοχείων, laptops του SemEval	εντοπισμός θεματικής ενότητας / οντότητας precision=52%, recall=50%, F1=51%, εντοπισμός πολικότητας accuracy=78%, ανεξαρτήτου τομέα δεδομένα accuracy=80%.
Jakob και Gurevych (2010)	338	MPQA	CRF	Κριτικές σε κάμερες, ταινίες,υπηρεσίες web, αυτοκίνητα	βελτιώνει την απόδοση από 0.077, 0.126, 0.071 και 0.178 όσον αφορά το F-Measure

Πίνακας 15: Σχετικές Εργασίες Μη Εποπτευόμενης Μηχανικής Μάθησης

Δημοσίευση	Citations	Έτοιμο Λεξικό	Αλγόριθμος / Μέθοδος	Dataset	Απόδοση
Hatzivassiloglou και McKeown	2.108	-	log-linear regression model,	Wall Street Journal corpus (1987)	82% των όμοιων επιθέτων έχουν παρόμοια πολικότητα.

(1997)			non-hierarchical clustering		Precision: 90%
Turney (2002)	5.163	-	συντακτικά πρότυπα, POS tagging, PMI	Κριτικές αυτοκινήτων, τραπεζών, ταινιών και ταξιδιωτικών προορισμών	65,83% στις ταινίες. 84% στα αυτοκίνητα. Accuracy:74%
Popescu και Etzioni (2005)	2.081		συνύπαρξη λέξεων με PMI, POS tagging	Κριτικές προϊόντων	precision: 22% υψηλότερο από παλαιότερες εργασίες
Kobayashi κ.α. (2006)	34	-	Dependency parser, τριπλέτες	Ιαπωνικές κριτικές αυτοκινήτων	precision εξαγωγής λέξεων – κλειδιών 78%, εξαγωγή ζευγών 61.7%, προσδιορισμό της γνώμης 82.2%
Titov και McDonald (2008)	673	-	MG-LDA	Κριτικές εστιατορίων, ξενοδοχείων, mp3 players	precision: 75.8% στις κριτικές υπηρεσιών, 85.5% στις κριτικές τοποθεσίας, 75% στις κριτικές δωματίων logistic regression: 80.8% στις κριτικές υπηρεσιών, 94% στις κριτικές τοποθεσίας, 88.3% στις κριτικές δωματίων
Su κ.α. (2008)	223	WordNet-like lexicon, CCD	clustering, K-Means, mutual reinforcement	Κινέζικες κριτικές προϊόντων	precision: 81.9%
Branavan κ.α. (2008)	111	-	clustering, hierarchical Bayesian model, LDA, cosine similarity	Κριτικές κινητών, εστιατορίων	στις κριτικές κινητών τηλεφώνων recall: 88%, precision: 58%, f-score: 70% στις κριτικές εστιατορίων recall: 90%, precision: 60%, f-score: 75%
Guo κ.α. (2009)	119	-	LDA, latent semantic association (LaSA) model	Κινέζικες κριτικές κινητών τηλεφώνων, Αγγλικές κριτικές φωτογραφικών μηχανών και laptops	accuracy: 81% στις φωτογραφικές μηχανές, 85% στους φορητούς υπολογιστές 82% στα κινητά τηλέφωνα
Andrzejewski κ.α. (2009)	302	-	θεματική μοντελοποίηση, Dirichlet trees, Must-Links / Cannot-Links	MEDLINE database	γενικεύουν πέραν των γνώσεων που καθορίζονται από το χρήστη
Brody και Elhadad (2010)	398	-	θεματική μοντελοποίηση με LDA	κριτικές εστιατορίων	average Kendall coefficient: 0.36 για τα αυτόματα δεδομένα - 0.45 για τα χειροκίνητα επισημασμένα, average Kendall's distance: 0.32 για τα αυτόματα δεδομένα - 0.27 για τα χειροκίνητα επισημασμένα
Zhang και Liu (2011)	146	Λεξικό γνώμης των Ding κ.α. (2008)	feature-based opinion mining model	κριτικές προϊόντων για φάρμακα, στρώματα, ραδιόφωνα, δρομολογητές	rank precision@10: 60% στις κριτικές στρωμάτων, 60% στις κριτικές φαρμάκων, 50% στις κριτικές δρομολογητών, 40% στις κριτικές ραδιοφώνων
Lakkaraju κ.α. (2011)	70	SentiWordNet	HMM-LDA	κριτικές προϊόντων του amazon.com για ψηφιακές φωτογραφικές μηχανές, φορητούς υπολογιστές, κινητά τηλέφωνα,	CFACTS: accuracy 78.22% σε επίπεδο λέξης, accuracy 81.28% σε επίπεδο πρότασης, accuracy 84.52% σε 2 κλάσεις, CFACTS-R: 77.87%

				τηλεοράσεις LCD, εκτυπωτές	
Kim κ.α. (2013)	80	-	Bayesian nonparametric model	κριτικές laptops και ψηφιακές φωτογραφικές μηχανές από το amazon.com	accuracy: 85% και 76% στα 2 datasets
Pavlopoulos και Androutsopoulos (2014)	4	Αγγλική Wikipedia	Word2vec, SVM	Κριτικές εστιατορίων, ξενοδοχείων, laptops	H&L+W2V με average weighted precision 66.8% στα εστιατόρια, 53.3% στα ξενοδοχεία, 38.9% στα laptop

Πίνακας 16: Σχετικές Εργασίες Ημι - Εποπτευόμενης Μηχανικής Μάθησης

Δημοσίευση	Citations	Έτοιμο Λεξικό	Αλγόριθμος / Μέθοδος	Dataset	Απόδοση
Socher κ.α. (2008)	810	MPQA	Recursive Autoencoders (RAE)	Κριτικές ταινιών, MPQA opinions	Accuracy: 77.7% στις κριτικές ταινιών, 86.4% στο MPQA opinions
Qiu κ.α. (2011)	637	Opinion Lexicon των Hu και Liu (2004)	Double Propagation	Hu και Liu datasets: κριτικές προϊόντων για φωτογραφικές μηχανές, DVD player, mp3 player, κινητά	F-score ξεπερνά των Hu και Liu (2004), Popescu και Etzioni (2005), PLSA, CRF και CRF-D κατά 10%, 4%, 32%, 43% και 32% αντίστοιχα
Zhai κ.α. (2011)	103	-	Expectation - Maximization (EM), Constrained - LDA, naïve Bayesian classification	Κριτικές homecinema, ασφάλειες, στρώματα, αυτοκίνητα, ηλεκτρικές σκούπες	Με SC-EM accuracy: 81%, purity: 82%, entropy:69%.
Mukherjee και Liu (2012)	233	Opinion Lexicon των Hu και Liu (2004)	Maximum-Entropy, LDA	Κριτικές ξενοδοχείων	precision@10: 88% με τον αλγόριθμο ME-SAS
Zhao κ.α. (2010)	306	-	Maximum-Entropy, LDA	Κριτικές εστιατορίων, ξενοδοχείων	MaxEnt-LDA: p@5=82.5%, nDCG@10=89.7% στα εστιατόρια, nDCG@5=82% στα ξενοδοχεία
Toh και Wang (2014)	59	WordNet, Opinion Lexicon των Hu και Liu (2004)	CRF, Double Propagation, K-means με word2vec	Multi-Domain Sentiment Dataset, Yelp	precision: 85%, recall: 82.7%, F1: 84% στον τομέα των εστιατορίων ενώ precision: 81.9%, recall: 67.1%, F1: 73.7% στον τομέα των laptop

Πίνακας 17: Σχετικές Εργασίες βασισμένες σε λεξικό

Δημοσίευση	Citations	Έτοιμο Λεξικό	Αλγόριθμος / Μέθοδος	Dataset	Απόδοση
Hu και Liu (2004)	5.368	WordNet	POS tagging, Apriori	κριτικές προϊόντων	accuracy: 84.2%, precision: 72%, recall: 80% στην εξαγωγή λέξεων – κλειδιών, precision: 64.2%: στην Ανάλυση Συναισθήματος, average recall: 80%, average precision: 72%,
Kamps κ.α.	775	WordNet,	graph-theoretic model	Stanford Political	παράγοντας αξιολόγησης: 68.19%,

(2004)		General Inquirer		Dictionary	παράγοντας ισχύος: 71.36%, παράγοντας δραστηριότητας: 61.85%
Kim και Hovy (2004)	1.610	WordNet	Bayesian φόρμουλα για να υπολογίσει την εγγύτητα κάθε λέξης σε κάθε κατηγορία	κριτικές προϊόντων και εστιατορίων από το epinions.com	precision: 66% recall: 76%
Carenini κ.α. (2005)	210	WordNet	μετρήσεις ομοιότητας λέξεων	κριτικές ψηφιακών φωτογραφικών μηχανών και DVD	υψηλή ακρίβεια και μείωση του σημασιολογικού πλεονασμού των ακατέργαστων λέξεων – κλειδιών
Ding κ.α. (2008)	1.15	WordNet, Opinion Lexicon των Hu και Liu (2004)	γλωσσολογικά πρότυπα	κριτικές προϊόντων	precision: 92%, recall: 91%, F-Score: 91%

Ακολουθούν μερικές αξιοσημείωτες παρατηρήσεις και συμπεράσματα που εξάγονται από τους εν λόγω πίνακες σύγκρισης:

- Το υπο – πρόβλημα της εξαγωγής των λέξεων – κλειδιών μιας πρότασης κριτικής παρατηρούμε ότι επιλύεται πολύ σπάνια με μεθόδους εποπτευόμενης Μηχανικής Μάθησης, καθώς το πρόβλημα της ανίχνευσης του συναισθήματος των λέξεων – κλειδιών της πρότασης αποφεύγεται να προσεγγίζεται με μεθόδους μη εποπτευόμενης Μηχανικής Μάθησης.
- Αξιοσημείωτη παρατήρηση είναι ότι οι σχετικές εργασίες που επιλύουν συνολικά το πρόβλημα της ABSA, δηλαδή όλα τα υπο – προβλήματά της, είναι κατά μεγαλύτερο ποσοστό με μεθόδους μη εποπτευόμενης Μηχανικής Μάθησης.
- Διαπιστώνουμε ότι όσο παλαιότερη είναι η δημοσίευση είναι πιθανότερο να έχει και περισσότερες αναφορές, πράγμα λογικό εφόσον ολοένα και περισσότερες νέες εργασίες στηρίζονται πάνω σε αυτές κατά το πέρασμα των χρόνων. Λόγω αυτού του γεγονότος δεν θα πρέπει να απορρίπτονται ή να αγνοούνται ή να λαμβάνονται λιγότερο υπόψιν πρόσφατες δημοσιεύσεις με μικρό αριθμό αναφορών, καθώς είναι πολύ πιθανό κάποια να αποτελεί αξιόλογη εργασία αλλά να μην έχει προλάβει, λόγω του μικρού χρόνου ύπαρξής της να αποκτήσει τον απαραίτητο αριθμό αναφορών που θα την καθορίσει αξιόπιστη. Έτσι, αν και θεωρείται σημαντικός παράγοντας, δεν είναι απόλυτος και πάντα αξιόπιστος. Προτείνεται να λαμβάνεται υπόψιν στις ακραίες περιπτώσεις όπου μία παλιά δημοσίευση έχει πολύ μικρό αριθμό αναφορών άρα την καθιστά αναξιόπιστη ή όταν μία νέα δημοσίευση έχει μεγάλο αριθμό αναφορών γεγονός που την καθιστά πάρα πολύ σημαντική.
- Ο μεγαλύτερος αριθμός σχετικών εργασιών χρησιμοποιεί σαν σύνολο δεδομένων (dataset) κριτικές χρηστών από το διαδίκτυο ώστε να υλοποιήσει την ανάλυση του συναισθήματος των χαρακτηριστικών γνωρισμάτων ενός προϊόντος, καθώς υπάρχει

διαφορά στον τρόπο έκφρασης της γνώμης σε κείμενα κριτικών προϊόντων από ότι σε οποιουδήποτε άλλου τύπου κείμενα.

- Ο μεγαλύτερος αριθμός σχετικών εργασιών εντοπίζεται να χρησιμοποιούν μεθόδους εποπτευόμενης Μηχανικής Μάθησης για την επίλυση του συνολικού προβλήματος.
- Μόλις το 45% των σχετικών εργασιών που εξετάζονται (18/40) χρησιμοποιούν λεξικό γνώμης για την αντιμετώπιση του προβλήματος, κάτι που αποδεικνύει για ακόμη μία φορά τέτοιου είδους προσεγγίσεις δεν θεωρούνται κατάλληλες για ανάλυση συναισθήματος σε αυτό το επίπεδο για το λόγους που εξετάσαμε στο Κεφάλαιο 2.4.4.3.1.3.
- Οι 9 στις 14 σχετικές εργασίες (64.2%) βασισμένες σε λεξικό γνώμης χρησιμοποιούν το WordNet, κάτι που το καθορίζει περισσότερο αξιόπιστο σε σύγκριση με όμοιά του.
- Μεγάλος αριθμός σχετικών εργασιών χρησιμοποιεί έτοιμα προκατασκευασμένα λεξικά γνώμης στις περιπτώσεις απόδοσης του συναισθηματικού προσανατολισμού, επιβεβαιώνοντας την εικασία μας ότι η κατασκευή ενός νέου λεξικού γνώμης από την αρχή είναι μια επίπονη και χρονοβόρα διαδικασία που αποφεύγεται, αν και μπορεί να δώσει καλύτερα αποτελέσματα, εφόσον θα είναι εστιασμένο ακριβώς στον τομέα ανάλυσης.
- Οι σχετικές εργασίες με την καλύτερη απόδοση επιτυγχάνονται με μεθόδους εποπτευόμενης Μηχανικής Μάθησης.
- Μια ακόμη παρατήρηση, όχι και τόσο ισχυρή λόγω της μη εγκυρότητάς της διότι μπορεί να αποδειχθεί λανθασμένη, αλλά άξια λόγου, είναι ότι δεν εντοπίστηκε καμία αξιόλογη σχετική εργασία κατά τα τελευταία τρία χρόνια, πράγμα αξιοπερίεργο καθώς αναμένεται με την αύξηση του περιεχομένου παραγόμενου από τους χρήστες (UGC) και την βαθιά επιρροή του στην καθημερινότητα των ανθρώπων, να γίνονται ολοένα και περισσότερες και αποδοτικότερες προσπάθειες επίλυσης του προβλήματος Ανάλυσης Συναισθήματος με σύγχρονες τεχνολογίες.
- Τέλος, θέτοντας έναν μέσο όρο της τάξεως του 70% - 80% στην απόδοση των προτεινόμενων συστημάτων και άρα επίλυσης στο πρόβλημα της Ανάλυσης Συναισθήματος σε κριτικές προϊόντων σε επίπεδο προτάσεων, η κρισιμότερη παρατήρηση που γεννάται είναι ότι εάν και δείχνουν να αποδίδουν ικανοποιητικά, υλοποιώντας διάφορες μεθόδους, καμία δεν έχει επιλύσει 100% το πρόβλημα, γεγονός μάλλον αδύνατο, λόγω της τεράστιας ιδιαιτερότητας της έκφρασης του λόγου από τους ανθρώπους και την απαιτητική κατανόηση του συντακτικού και των σημασιολογικών κανόνων της φυσικής γλώσσας από τους υπολογιστές. Η παρατήρηση αυτή μπορεί να καθοριστεί και το τελικό συμπέρασμα αυτής της διπλωματικής εργασίας.

4

Επίλογος

4.1 Σύνοψη και συμπεράσματα

Η Ανάλυση Συναισθήματος αντιμετωπίζεται ως ένα κοινό γνωστό πρόβλημα στο επιστημονικό πεδίο της Επεξεργασίας Φυσικής Γλώσσας (NLP). Στη συγκεκριμένη διπλωματική εργασία, με στόχο την αποτίμηση μιας βαθμολογίας σε προϊόντα που σχολιάζουν χρήστες στο διαδίκτυο αντιμετωπίζουμε το πρόβλημα της εξαγωγής συναισθήματος στις κριτικές αυτών των προϊόντων μέσω των κυριότερων χαρακτηριστικών γνωρισμάτων τους με μεθόδους Μηχανικής Μάθησης Επαγωγικά εστιάζουμε στον τομέα της Ανάλυσης Συναισθήματος βασισμένης σε λέξεις – κλειδιά (Aspect Based Sentiment Analysis - ABSA). Στο γενικό πλαίσιο, τα συστήματα ABSA δέχονται σαν είσοδο ένα σύνολο δεδομένων κειμένων που μιλούν για μια συγκεκριμένη οντότητα και προσπαθούν να εντοπίσουν τις λέξεις - κλειδιά της συγκεκριμένης οντότητας σε επίπεδο πρότασης και να εκτιμήσουν το μέσο συναίσθημα του κειμένου ανά λέξη - κλειδί. Δεν υπάρχει καθορισμένη διαδικασία υλοποίησης των υπο - διεργασιών της, ούτε προκαθορισμένος τρόπος αξιολόγησης, αντί αυτών, εξαρτάται από τις απαιτήσεις του εκάστοτε προβλήματος, τα διαθέσιμα δεδομένα και τον τομέα όπου εφαρμόζεται. Παρόλα αυτά, παραθέτουμε μια τυπική και γενική διαδικασία επίλυσης, σύμφωνα με όσα μελετήθηκαν και εντοπίστηκαν να χρησιμοποιούνται σε σχετικές εργασίες. Αναφερόμαστε στις δυσκολίες που αντιμετωπίζει η ανάλυση αυτού του επιπέδου καθώς και στις διαθέσιμες τεχνικές Μηχανικής Μάθησης που είναι σε θέση να τις επιλύσουν. Διακρίνεται να χρησιμοποιούνται τόσο μέθοδοι

εποπτευόμενης όσο και μη εποπτευόμενης μάθησης για την εκάστοτε διεργασία. Πιο συγκεκριμένα, έπειτα της σύγκρισης των σχετικών εργασιών που υλοποιήσαμε στα πλαίσια αυτής της διπλωματικής εργασίας, διαπιστώσαμε ότι για το πρώτο βήμα της εξαγωγής λέξεων – κλειδιών σε επίπεδο πρότασης προτιμώνται κυρίως μέθοδοι μη εποπτευόμενης Μηχανικής Μάθησης, ενώ για τον εντοπισμό του συναισθήματός τους κυρίως εποπτευόμενης, με συγκριτικά μεγαλύτερο αριθμό προσεγγίσεων που επιλύουν συνολικά το πρόβλημα να χρησιμοποιεί τις ευέλικτες, ευκολότερα προσαρμόσιμες μη εποπτευόμενες μεθόδους.

Πέραν των συμπερασμάτων που εξήχθησαν από την σύγκριση των σχετικών εργασιών στο Κεφάλαιο 3.2, μερικά ακόμη συμπεράσματα που αξίζει να αναφερθούν όσον αφορά τις μεθόδους για την επίλυση του προβλήματος της ABSA, σύμφωνα με την μελέτη των σχετικών εργασιών είναι τα εξής:

- η θεματική μοντελοποίηση κατά την διαδικασία εξαγωγής των λέξεων – κλειδιών δεν προτείνεται λόγω των υψηλών απαιτήσεων αριθμού δεδομένων και χρόνου υλοποίησης, αλλά και λόγω του καθολικού χαρακτήρα των θεματικών ενοτήτων που εντοπίζουν, πέραν του στόχου μας εντοπισμού των λέξεων – κλειδιών, γεγονός που συμφωνούν και οι Yu κ.α. (2011) επισημαίνοντας ότι οι μέθοδοι ομαδοποίησης σε θεματικές ενότητες παρουσιάζουν χαμηλή απόδοση, αλλά και οι Titov και McDonald (2008).
- οι προσεγγίσεις που βασίζονται σε λεξικά γνώμης εντοπίζεται να προτιμώνται σε μεγάλο βαθμό μολονότι δεν μεταχειρίζονται τις αλληλεπιδράσεις των λέξεων ώστε να αποφύγουν τον σπουδαίο αριθμό αστοχίας τους.
- οι προσεγγίσεις βάσει συντακτικών κανόνων / προτύπων αν και χαρακτηρίζονται αποτελεσματικοί από τους Qiu κ.α. (2011), αποδοκιμάζονται από τους Yu κ.α. (2011) και τους Zhao κ.α. (2010) που επισημαίνουν ότι είναι εξαρτημένες από τον τομέα για τον οποίο αναλύονται.
- τα μοντέλα βασισμένα σε n – grams μειώνουν την ακρίβεια όπως παρατηρούν οι Andreevskaia και Bergler (2008) καθώς και ο Pang κ.α. (2002).
- πολλές λέξεις ακόμη και του ίδιου τομέα εξέτασης μπορεί να έχουν διαφορετικό συναίσθημα ανάλογα τον τρόπο χρήσης τους, όπως εντοπίζουν οι Ding κ.α. (2008).

Δεδομένης της ανάγκης υλοποίησης συστημάτων Ανάλυσης Συναισθήματος που να βοηθούν τους χρήστες στις τελικές τους αποφάσεις και συνάμα τις εταιρείες να καταλάβουν τους καταναλωτές τους, έχουν γίνει σημαντικά βήματα βελτίωσης με το πέρασμα των χρόνων, εντούτοις απαιτείται βαθύτερη διερεύνηση για κατασκευή τόσο αυτοματοποιημένων όσο και ολοκληρωμένων και αποδοτικών συστημάτων που να αντιμετωπίζουν όλα τα υπο - προβλήματα παράλληλα, διότι οι αλληλεπιδράσεις τους μπορεί να βοηθήσουν στην επίλυση κάθε επιμέρους προβλήματος. Εντούτοις, καμία προσέγγιση δεν την επιλύει εξολοκλήρου ή έστω αρκετά ικανοποιητικά. Η σημαντική αυτή παρατήρηση, που αποδεικνύεται και από την

σύγκρισή μας, μπορεί να γράψει τον επίλογο αφήνοντας ανοιχτό το ζήτημα της Ανάλυσης Συναισθήματος σε κείμενα κριτικών που χρήζει περαιτέρω και εκτενέστερη μελέτη.

4.2 Μελλοντικές επεκτάσεις

Όπως έχουμε ήδη εξετάσει από την παραπάνω βιβλιογραφική ανασκόπηση στην Ανάλυση Συναισθήματος αλλά και πιο συγκεκριμένα στην εξαγωγή συναισθήματος κείμενων κριτικών μέσω των λέξεων – κλειδιών, φτάνουμε εύκολα στο συμπέρασμα ενός προκλητικού ερευνητικού τομέα που είναι σχεδόν απίθανο να καλυφθεί σε μία διπλωματική εργασία και φυσικά δεν παύει ποτέ να έχει νέα προβλήματα προς επίλυση. Η εξόρυξη γνώμης από τη φυσική γλώσσα του ανθρώπου είναι μία πολύ σκληρή και επαχθείς διαδικασία καθώς απαιτεί μια βαθιά κατανόηση του συντακτικού και των σημασιολογικών κανόνων που έχει η φυσική γλώσσα, σε οποιαδήποτε γλώσσα κι αν εκφράζεται. Επομένως, θεωρήσαμε σκόπιμο στο τελευταίο αυτό κεφάλαιο, να αναφέρουμε ορισμένους τομείς που θα ήταν ωφέλιμο να προσεγγιστούν μελλοντικά, πράγμα που θα οδηγήσει σε μία πιο ολοκληρωμένη βιβλιογραφική / ερευνητική εργασία πάνω στο εξεταζόμενο πρόβλημα και κάλυψη ενός μεγαλύτερου εύρους προκλήσεων πάνω στην Ανάλυση Συναισθήματος κριτικών χρηστών σε προϊόντα για τα κυριότερα χαρακτηριστικά γνωρίσματά τους με μεθόδους Μηχανικής Μάθησης.

- Η ανάλυση κειμένων κριτικών γραμμένων στα ελληνικά, κάτι ακόμη πιο προκλητικό λόγω του τεράστιου εύρους γραμματικών και συντακτικών κανόνων της ελληνικής γλώσσας.
- Η βαθύτερη μελέτη ταξινόμησης κειμένων κριτικών σε μεγαλύτερο και διαφορετικό εύρος συναισθημάτων, όχι μόνο θετικό και αρνητικό όπως συχνότερα συναντάται. Για παράδειγμα η πολικότητα θα μπορούσε να επεκταθεί σε συναισθήματα όπως θυμωμένος, χαρούμενος, λυπημένος, ικανοποιημένος κ.α. Η πιο εύκολη και άμεση προέκταση θα ήταν λαμβάνοντας υπόψη τις αντικειμενικές προτάσεις και το ουδέτερο συναίσθημα που εκφράζουν.
- Η μελέτη της ταξινόμησης συναισθημάτων μεταξύ των γλωσσών (Cross-language sentiment classification) η οποία σημαίνει την επίτευξη ταξινόμησης των συναισθημάτων των κειμένων γνώμης σε πολλές γλώσσες.
- Η αναλυτικότερη εξέταση επίλυσης του προβλήματος ταξινόμησης με παλινδρόμηση (regression) και όχι κατηγοριοποίηση (classification), όπου αντί να ταξινομούνται οι προτάσεις που περιέχουν το χαρακτηριστικό γνώρισμα υπολογίζοντας τον μέσο όρο του συναισθήματος των χαρακτηριστικών γνωρισμάτων, να επιστρέφεται μια βαθμολογία συναισθήματος για κάθε πρόταση. Συνοπτικά, αναφερόμαστε στην πρόβλεψη της

βαθμολογίας (rating) του προϊόντος ή των μεμονωμένων χαρακτηριστικών γνωρισμάτων του στις κριτικές του προϊόντος.

- Η βαθύτερη μελέτη εντοπισμού επώνυμων οντοτήτων ως λέξεις – κλειδιά.
- Η υλοποίηση Ανάλυσης Συναισθήματος σε κριτικές που απαιτούν συν τοις άλλοις εντοπισμό της οντότητας για την οποία σχολιάζουν και δεν θεωρείται δεδομένη.
- Η βαθύτερη εξέταση αντιμετώπισης της ειρωνείας κατά τον εντοπισμό του συναισθήματος σε επίπεδο πρότασης.
- Η βαθύτερη εξέταση εφαρμογής της Ανάλυσης Συναισθήματος σε άλλης μορφής κείμενα που δεν είναι κριτικές χρηστών, καθώς η διαχείρισή τους θα πρέπει να είναι διαφορετική, όπως για παράδειγμα στα tweets, όπου τα κείμενα λόγω της μικρής τους έκτασης δεν είναι περιγραφικά αλλά εστιάζουν κατευθείαν στο θέμα.
- Η υλοποίηση ενός αποδοτικού, αυτοματοποιημένου και ολοκληρωμένου συστήματος σύγκρισης προϊόντων αναλύοντας το συναίσθημα των χαρακτηριστικών γνωρισμάτων τους εκμεταλλευόμενοι τις τεχνικές Μηχανικής Μάθησης που εξετάστηκαν και αποσαφηνίστηκαν για την επίτευξη όσο το δυνατόν καλύτερων αποτελεσμάτων.

4.3 Βοηθητικοί Πίνακες

Πίνακας 18:Αντιστοίχιση αγγλικής – ελληνικής ορολογίας που χρησιμοποιείται στην διπλωματική εργασία.

Αγγλική Επίσημη Ορολογία	Ελληνική Μετάφραση	Αγγλική Επίσημη Ορολογία	Ελληνική Μετάφραση
Aspect	Λέξη-κλειδί / γνώρισμα	lexeme	λέξημα
Advertising Opinion Mining	Στοχευμένη τοποθέτηση διαφημίσεων	Logistic Regression	Λογιστική Παλινδρόμηση
Affective Computing	Συναισθηματική Υπολογιστική	Machine Learning	Μηχανική Μάθηση
Affective Science	Συναισθηματική Επιστήμη	Markov Chain Monte Carlo (MCMC)	Μαρκοβιανή Αλυσίδα
amplifiers	ενισχυτές	Maximum Entropy	Μέγιστη Εντροπία
Arousal	Διέγερση	Maximum Likelihood Estimation (MLE)	εκτίμησης μέγιστης πιθανοφάνειας
Artificial Intelligence (AI)	Τεχνητή Νοημοσύνη	metric labeling	μετρική σήμανση
Artificial Neural Networks (ANN)	Τεχνητά Νευρωνικά Δίκτυα	minimum support	ελάχιστο κατώφλι
Aspect Category Detection	Ανίχνευση θεματικών ενοτήτων	Multilayers	Πολυστρωματικά / πολυεπίπεδα
Aspect Category Polarity	Πολικότητα συναισθήματος κατηγορίας	Multinomial Distribution	πολυωνυμική κατανομή
Aspect level	επίπεδο λέξης	n - gram	n - διαδοχικές λέξεις
aspect term aggregation	ομαδοποίηση λέξεων – κλειδιών	naive	αφελή
Aspect Term Extraction (ATE)	Εξαγωγή λέξεων – κλειδιών	Named Entity Recognition	Εξαγωγή Ονοματικών Οντοτήτων
Aspect Term Polarity	Πολικότητα συναισθήματος	Natural Language Processing	Επεξεργασία φυσικής γλώσσας

Αγγλική Επίσημη Ορολογία	Ελληνική Μετάφραση
	λέξεων – κλειδιών
Aspect-based Sentiment Analysis (ABSA)	Ανάλυσης Συναισθήματος βασισμένη σε λέξεις - κλειδιά
Backpropagation	Οπισθοδιάδοση
Batch Learning	Μαζική Μάθηση
Binarized	Δυαδικοποιημένη
Blogs	Ιστολόγια
Bootstrap Aggregating / Bagging	Εμφωλίαση
Brand name	εμπορικό σήμα / εταιρική ταυτότητα
Business intelligence (BI)	Επιχειρηματική Ευφυΐα
citations	αναφορές
Class Conditional Independence	Ανεξαρτήσια κλάση υπό όρους
Classification	Ταξινόμηση
classifier	ταξινομητής
Clustering	Συσταδοποίηση
Co-occurence	Συν-εμφάνιση
Cognitive Science	Γνωσιακή Επιστήμη
comparative	συγκριτική
Computational Linguistics	Υπολογιστικής Γλωσσολογία
Computer Vision	Υπολογιστική Όραση
Confusion Matrix	Πίνακας Σύγχυσης
Content Management System (CMS)	Σύστημα Διαχείρισης Περιεχομένου
contextual clues	συμφραζόμενα
contextual sentiment	συναίσθημα βάσει συμφραζομένων
coreference resolution	επίλυση συναναφορών
Customer relationship management (CRM)	Μέσα Διαχείρισης Πελατειακών Σχέσεων
Data Compression	Συμπίεση Δεδομένων
dataset	Σύνολο δεδομένων
Deep-learning	Βαθιά Μάθηση
Dependent / Response variable	εξαρτώμενη / απόκρισης μεταβλητή
Dimensionality Reduction	Μείωση Διαστάσεων
Document level	επίπεδο εγγράφου
Dominance	Κυριαρχία
double propagation	διπλή διάδοση
downtoners	εξομαλυντές
Elongated Words	Επιμηκυμένες λέξεις
Entity	Οντότητα
error rate	λόγος σφάλματος

Αγγλική Επίσημη Ορολογία	Ελληνική Μετάφραση
(NLP)	
negative	αρνητικό
Neurons	νευρώνες
non-leaf	χωρίς φύλλα
one-vs-all	ένας – εναντίον - όλων
Opinion Mining	Εξόρυξη Γνώμης
Opinion Target	Στόχος / Λέξη Γνώμης
Ordinal values	Κανονικές τιμές
Outliers	απομακρυσμένα
Overfitting	υπερ – προσαρμογή
Pattern Recognition	Αναγνώριση Προτύπων
patterns	πρότυπα
Polarity	Πολικότητα
positive	θετικό
Posterior Probability	εκ των υστέρων πιθανότητα
precision	ακρίβεια
Prior Probability	προγενέστερη πιθανότητα
prior sentiment	πρότερο συναίσθημα
Probabilistic Latent Semantic Analysis (PLSA)	Πιθανοτική Λανθάνουσα Σημασιολογική Ανάλυση
Random forest	Δέντρα απόφασης
Ranking - based	βασισμένα στην βαθμολογία
Rating Score	Βαθμολογία αξιολόγησης
recall	ανάκληση
Recommender Systems	Συστήματα Συστάσεων
Recursive autoencoders (RAE)	αναδρομικοί αυτοσυσχετιστές
Regression Analysis	Ανάλυση παλινδρόμησης
regular	κανονική
Reinforcement	Ενισχυτική
Reviews	Κριτικές
rule - based	Βάσει κανόνων
seeds	σπόροι
Semi - Supervised	Ημι - εποπτευόμενη
Sentiment Analysis	Ανάλυση Συναισθήματος
sentiment lexicons	λεξικά συναισθήματος / γνώμης
Sentiment Ontology Tree (SOT)	Δέντρο οντολογίας συναισθήματος
Sentiment Orientation (SO)	Προσανατολισμός Συναισθήματος
Sequential Learning	Διαδοχική Μάθηση

Αγγλική Επίσημη Ορολογία	Ελληνική Μετάφραση
Euclidean Distance	Ευκλείδεια Απόσταση
Expectation Maximization (EM)	μεταβολική προσέγγιση
explicit	άμεση
Feature	Χαρακτηριστικό / γνώρισμα
Feature-based Opinion Mining	Εξόρυξη γνώμης βασισμένη σε χαρακτηριστικά
Feedback management	Διαχείριση των σχολίων
Forum	Τόποι δημόσιας συζήτησης
Generalization Performance	Απόδοσης Γενίκευσης
Generative Models	Παραγωγικά μοντέλα
Gloss	Ερμηνεία
Gradient Descent	Βαθμίδα Κατάβασης
Hidden Layers	Κρυφά επίπεδα
Hierarchical Learning (HL)	Ιεραρχική Μάθηση
implicit	έμμεση
Independent / Predictor variables	Ανεξάρτητες / ελεγχόμενες μεταβλητές
Information Retrieval	Ανάκτηση Πληροφορίας
Instant Messaging (IM)	Άμεσα μηνύματα
intensifiers	εντατικοποιητές
intensity	ένταση
inter-rater agreement	εκτίμηση του βαθμού συμφωνίας
inverse	αντίστροφο
Iterative Optimization Techniques	επαναληπτικές τεχνικές βελτιστοποίησης
Kernel Method	Μέθοδος Πυρήνα
keywords	λέξεις-κλειδιά
Label Sequential Rule Mining	Εξόρυξη κανόνων με επισήμανση προτάσεων
language & domain independent	ανεξαρτήτου γλώσσας & τομέα
Language Modeling	Γλωσσικά Μοντέλα
Latent	Λανθάνουσα
Latent Semantic Analysis (LSA)	Λανθάνουσα Σημασιολογική Ανάλυση
lemmatization	λημματοποίηση

Αγγλική Επίσημη Ορολογία	Ελληνική Μετάφραση
Sigmoid Function	Σιγμοειδής Συνάρτηση
Singular Value Decomposition (SVD)	αποσύνθεση μοναδικής τιμής
Social Media	Κοινωνικά Δίκτυα
Spam Detection	Ανίχνευση ανεπιθύμητης αλληλογραφίας
Speech Recognition	αναγνώριση φωνής
stemming	αποκοπή
Subjectivity Classification	Ταξινόμηση Υποκειμενικότητας
Supervised Learning	Εποπτευόμενη Μάθηση
Support Vector Machines (SVM)	Διανυσματικές μηχανές υποστήριξης
Synset	σύνολο συνωνύμων
Table Look Up Approach	Πίνακας Αναζήτησης
Tagging	Μαρκάρισμα
tags	ετικέτες
target	στόχος
Term Frequence	συχνότητα εμφάνισης
Testing dataset	Δεδομένα δοκιμής
threshold	κατώφλι
tokenization	ευρετηρίαση
tokens	τμήματα που αποφέρουν νόημα
Topic Classification	ταξινόμηση με βάση το θέμα
topic coreference	επίλυση συσχέτισης κατηγοριών
Topic Modeling	θεματική μοντελοποίηση
Training Dataset	Δεδομένα εκπαίδευσης
Universal Approximator	καθολικός προσεγγιστής
Unsupervised Learning	Μη εποπτευόμενη μάθηση
User Generated Content (UGC)	Περιεχόμενο παραγόμενο από χρήστες
Valence	Σθένος
Vector	Διάνυσμα
Vectorization	Διανυσματοποίηση
Vocabulary	Λεξιλόγιο
World Wide Web (www)	Παγκόσμιο Ιστό

5

Βιβλιογραφία

5.1 Δημοσιεύσεις

Abirami A. M. and Askarunisa A. *Feature Based Sentiment Analysis for Service Reviews.* J. UCS 22: 650-670. 2016.

Alghunaim A., Mohtarami M., Cyphers S. and Glass J. *A Vector Space Approach for Aspect Based Sentiment Analysis*. In Proceedings of NAACL-HLT 2015, pp. 116–122, 2015.

Andreevskaia A. and Bergler S. *When specialists and generalists work together: Overcoming domain dependence in sentiment tagging*. Proceedings of ACL-08: HLT, pp. 290–298, 2008.

Andrzejewski, David, Xiaojin Zhu, and Mark Craven. *Incorporating domain knowledge into topic modeling via Dirichlet forest priors*. In Proceedings of ICML. 2009. doi:10.1145/1553374.1553378

Blair-Goldensohn S., Kerry H., McDonald R., Neylon T., Reis G. A. and Reynar J. *Building a sentiment summarizer for local service reviews*. In Proceedings of WWW-2008 workshop on NLP in the Information Explosion Era. 2008.

Blei D. M., Ng A. Y. and Jordan M. I., *Latent dirichlet allocation*. The Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

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

Blitzer J., Dredze M., and Pereira F., *Biographies, Bollywood, Boom-boxes and Blenders: Domain adaptation for sentiment classification*. In Proceedings of ACL. 2007.

Boiy E. and Moens M. F. *A machine learning approach to sentiment analysis in multilingual Web texts*. Information Retrieval. 12(5): pp. 526–558. 2009. doi: 10.1007/s10791-008-9070-z

Boldrini E., Fernández J., Gómez J., Patricio M. B. *Sentiment Analysis and Opinion Mining: The EmotiBlog Corpus*. In Proceedings of the Lenguaje Natural. 47. 179-187. 2011.

Branavan S. R. K., Chen H., Eisenstein J. and Barzilay R.. *Learning document-level semantic properties from free-text annotations*. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (AC L-2008). 2008.

Brody S. and Elhadad N.. *An Unsupervised Aspect-Sentiment Model for Online Reviews*. In Proceedings of the 2010 Annual Conference of the North American Chapter of the AC L., 2010.

Cambria E., D. Olsher, and D. Rajagopal. *Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis*. AAAI, Quebec City, pages 1515-1521, 2014.

Carenini G., Raymond Ng and Ed Zwart. *Extracting knowledge from evaluative text*. In Proceedings of Third International Conference on Knowledge Capture (K-CA P-05). 2005. doi:10.1145/1088622.1088626

Cerini S., Compagnoni V., Demontis A., Formentelli M. and Gandini G. *Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining*. In Language Resources and Linguistic Theory: Typology, Second Language Acquisition, English Linguistics, Franco Angeli Editore, Milano, IT. 2007.

Cui, Hang, Vibhu Mittal and Mayur Datar. *Comparative experiments on sentiment classification for online product reviews*. In Proceedings of AAAI-2006. 2006.

Deng L. and Wiebe J. *MPQA 3.0: Entity/Event-Level Sentiment Corpus*. 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies, Denver, Colorado, USA. 2015.

Ding X., Liu B. and Zhang L. *Entity discovery and assignment for opinion mining applications*. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2009). 2009. doi:10.1145/1557019.1557141

Ding X., Liu B. and Philip S. Yu. *A holistic lexicon-based approach to opinion mining*. In Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008). 2008. doi: 10.1145/1341531.1341561

Esuli, A. and Sebastiani, F. *SentiWordNet: a high-coverage lexical resource for opinion mining*. Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR). 2006.

Fahrni A. and Klenner M., *Old wine or warm beer: Target-specific sentiment analysis of adjectives*, Computational Linguistics, vol. 2, no. 3, pp. 60–63, 2008.

Fang X. and Zhan J., *Sentiment analysis using product review data*. J Big Data. 2015. doi: 10.1186/s40537-015-0015-2

Ganapathibhotla M. and Liu B. *Mining opinions in comparative sentences*. In Proceedings of International Conference on Computational Linguistics (CO LING-2008). 2008. doi:10.3115/1599081.1599112

Greene S. and Resnik P. *More than words: Syntactic packaging and implicit sentiment*. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the AC L (NAAC L-2009). 2009.

Griffiths T. L. and Steyvers M. *Prediction and semantic association*. In Neural Information Processing Systems 15. 2003.

Griffiths T. L. and Steyvers M. *Finding Scientific Topics*. Proceedings of the National Academy of Sciences 101.Supplement 1: 5228-235. 2004.

Griffiths T. L. and Steyvers M., *Probabilistic topic models*. Handbook of latent semantic analysis. 427(7): pp. 424–440. 2007.

Guo H., H. Zhu, Z. Guo, X. Zhang and Z. Su. *Product feature categorization with multilevel latent semantic association*. In Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2009). 2009. doi:10.1145/1645953.1646091

Haddi Emma, Liu Xiaohui and Shi Yong. *The Role of Text Pre-processing in Sentiment Analysis*, Procedia Comput. Sci., vol. 17, pp. 26–32. 2013. doi:10.1016/j.procs.2013.05.005

Hatzivassiloglou V. and McKeown K. R. *Predicting the semantic orientation of adjectives*. In Proceedings of ACL'97, pp 174–181. 1997

Hatzivassiloglou V. and Wiebe J. *Effects of adjective orientation and gradability on sentence subjectivity*. In Proceedings of International Conference on Computational Linguistics (CO LING-2000). 2000. doi:10.3115/990820.990864

Hofmann T. *Probabilistic latent semantic indexing*. In Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI -1999). 1999.

Hu M. and Liu B. *Mining and summarizing customer reviews*. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), 2004. doi:10.1145/1014052.1014073

Jakob N. and Gurevych I., *Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields*, in Proceedings of the 2010 Conference on Empirical Meth-JOURNAL TKDE 19ods in Natural Lang. 2010.

Jiang Long, Mo Yu, Ming Zhou, Xiaohua Liu and Tiejun Zhao. *Target-dependent twitter sentiment classification*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (AC L-2011). 2011.

Jindal N. and Liu B. *Mining comparative sentences and relations*. In Proceedings of National Conf. on Artificial Intelligence (AAAI-2006). 2006b.

Kamps J., Marx M., Mokken R. J. and Rijke M. *Using wordnet to measure semantic orientation of adjectives*. In Proceedings of LREC -2004, pp 1115–1118. 2004.

Kessler J. S. and Nicolov N. *Targeting sentiment expressions through supervised ranking of linguistic configurations*. In Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM-2009). 2009.

Kim S. and Hovy E. *Determining the sentiment of opinions*. In Proceedings of International Conference on Computational Linguistics (CO LING-2004). 2004. doi:10.3115/1220355.1220555

Kim S. and Hovy E. *Automatic identification of pro and con reasons in online reviews*. In Proceedings of CO LING/AC L 2006 Main Conference Poster Sessions (AC L-2006). 2006. doi:10.3115/1273073.1273136

Kim Suin, Zhang Jianwen, Chen Zheng, Oh Alice and Liu Shixia. *A Hierarchical Aspect-Sentiment Model for Online Reviews*. Association for the Advancement of Artificial Intelligence. 2013.

Kobayashi Nozomi, Ryu Iida, Kentaro Inui and Yuji Matsumoto. *Opinion mining on the Web by extracting subject-attribute-value relations*. In Proceedings of AAAI-CAA W'06. 2006.

Kobayashi Nozomi, Kentaro Inui and Yuji Matsumoto. *Extracting aspect-evaluation and aspect-of relations in opinion mining*. In Proceedings of EMNLP'07. 2007.

Koppel M. and Schler J., *The importance of Neutral examples for learning sentiment*. Computational Intelligence. 22: 100–109. 2006. doi:10.1111/j.1467-8640.2006.00276.x

Ku Lun-Wei, Yu-Ting Liang and Hsin-Hsi Chen. *Opinion extraction, summarization and tracking in news and blog corpora*. In Proceedings of AAAI-CAA W'06. 2006.

Lafferty John, Andrew McCallum and Fernando Pereira. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In Proceedings of International Conference on Machine Learning (ICML-2001). 2001.

Lakkaraju H., C. Bhattacharyya, I. Bhattacharya and S. Merugu. *Exploiting Coherence for the Simultaneous Discovery of Latent Facets and associated Sentiments*. In Proceedings of SIAM Conference on Data Mining (SDM-2011). 2011.

Landauer T. K., Foltz P. W. and Laham D. *Introduction to Latent Semantic Analysis*. Discourse Processes, 25, pp. 259-284. 1998.

Laskari N. and Sanampudi S. *Aspect Based Sentiment Analysis Survey*. IOSR Journals. 18. pp. 2278-661. 2016. doi: 10.9790/0661-18212428.

Le Q. and Mikolov T. *Distributed Representations of Sentences and Documents*, International Conference on Machine Learning, 2014.

Leung C.W.K., and Chan S.C.F., *Sentiment Analysis of Product Reviews*. J. Wang, (Eds.), Encyclopedia of data warehousing and mining-Second Edition, Information Science Reference, 1794-1799, 2008.

Ling Wang, Dyer Chris, Black Alan W and Trancoso Isabel. *Two/too simple adaptations of word2vec for syntax problems*. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015.

Liu, Bing, Minqing Hu, and Junsheng Cheng. Opinion observer: *Analyzing and comparing opinions on the web*. In Proceedings of International Conference on World Wide Web (WWW-2005). 2005.

Liu B. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool. 2012

Liu Jing, Cao Yunbo, Lin Chin-Yew, Huang Yalou and Zhou Ming. *Low-quality product review detection in opinion summarization*. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-2007). 2007.

Long Chong, Jie Zhang and Xiaoyan Zhu. *A review selection approach for accurate feature rating estimation*. In Proceedings of Coling 2010: Poster Volume. 2010.

Lovins Julie Beth. *Development of a Stemming Algorithm*. Mechanical translation and computational linguistics, 11:22-31, 1968.

Maas L. Andrew, Daly E. Raymond, Pham T. Peter, Huang Dan, Ng Y. Andrew, and Potts Christopher. *Learning word vectors for sentiment analysis*. In Proceedings of the 49th

Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol 1, pages 142–150. 2011.

Mei Qiaozhu, Xu Ling, Matthew Wondra, Hang Su and ChengXiang Zhai. *Topic sentiment mixture: modeling facets and opinions in weblogs*. In Proceedings of International Conference on World Wide Web (WWW-2007). 2007.

Mikolov Tomas, Chen Kai, Corrado Greg and Dean Jeffrey. *Efficient Estimation of Word Representations in Vector Space*, CoRR, abs/1301.3781, 2013.

Moghaddam Samaneh and Ester Martin. *Opinion digger: an unsupervised opinion miner from unstructured product reviews*. In Proceeding of the ACM Conference on Information and Knowledge Management (CIKM-2010). 2010.

Mukherjee, Arjun and Bing Liu. *Aspect Extraction through Semi-Supervised Modeling*. In Proceedingsof 50th Anunal Meeting of Association for Computational Linguistics (AC L-2012). 2012.

Musto, C., Semeraro, G., Polignano, M.: *A Comparison of Lexicon-based Approaches for Sentiment Analysis of Microblog Posts*. In DART 2014, Information Filtering and Retrieval. Proceedings of the 8th International Workshop on Information Filtering and Retrieval, collocated with XIII AI*IA Symposium on Artificial Intelligence (AI*IA 2014), Pisa, Italy, December 10, 2014. CEUR Workshop Proceedings, 1314, 59-68, (2014).

Paice C. D., *Another stemmer*. ACM SIGIR Forum, Volume 24, No. 3,pp. 56-61. 1990.

Pang Bo, Lillian Lee and Shivakumar Vaithyanathan. *Thumbs up?: sentiment classification using machine learning techniques*. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), pp. 79–86. 2002.

Pang Bo and Lillian Lee. *A sentimental education: Sentiment analysis using subjectivity*. Proceedings of ACL, pp. 271-278, 2004.

Pang Bo and Lee Lillian. Seeing stars: *Exploiting class relationships for sentiment categorization with respect to rating scales*. In Proceedings of Meeting of the Association for Computational Linguistics (AC L-2005), 2005.

Pavlopoulos J. and Androutsopoulos I., *Multi-granular aspect aggregation in aspect-based sentiment analysis*. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp 78–87. 2014.

Popescu Ana-Maria and Oren Etzioni. *Extracting product features and opinions from reviews*. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2005). 2005. doi:10.3115/1220575.1220618

Porter Martin F., *An algorithm for suffix stripping*, Program, Vol. 14 Issue: 3, pp.130-137, 1980. doi: 10.1108/eb046814.

Potts, C. *Sentiment Symposium Tutorial: Lexicons (2011)*. Stanford Linguistics. Available: http://sentiment.christopherpotts.net/lexicons.html

Qiu Guang, Bing Liu, Jiajun Bu and Chun Chen. *Opinion Word Expansion and Target Extraction through Double Propagation*. Computational Linguistics, Vol. 37, No. 1: 9.27, 2011. doi: 10.1162/coli_a_00034

Rabiner Lawrence R. *A tutorial on hidden Markov models and selected applications in speech recognition*. Proceedings of the IEEE. 77(2): pp. 257–286, 1989. doi:10.1109/5.18626

Raschka Sebastian. *Python Machine Learning*, Packt Publishing, 2015.

Ricci, F., Rokach, L., Shapira, B. *Recommender Systems: Introduction and Challenges*, in: Ricci, F., Rokach, L., Shapira, B. (Eds.), Recommender Systems Handbook. Springer US, Boston, MA, pp. 1–34. 2015.

Riloff Ellen and Janyce Wiebe. *Learning extraction patterns for subjective expressions*. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2003). 2003.

Riloff Ellen, Siddharth Patwardhan and Janyce Wiebe. *Feature subsumption for opinion analysis*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP- 2006). 2006. doi:10.3115/1119355.1119369

Scaffidi Christopher, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng and Chun Jin. *Red opal: Product-feature scoring from reviews*. In Proceedings of EC'07, pp 182–191. 2007.

Scherer K. R. and Wallbott H. G., *Evidence for universality and cultural variation of differential emotion response patterning*. Journal of personality and social psychology, vol. 66, p. 310, 1994.

Schlosberg, H. *Three dimensions of emotion*. Psychological Review, 61(2), 81-88. 1954. http://dx.doi.org/10.1037/h0054570.

Sebastiani F., *Machine learning in automated text categorization*, ACM Computin Surveys (CSUR), ACM Press, 2002.

Snyder B. and Barzilay R. *Multiple aspect ranking using the good grief algorithm*, Proc. NAACL HLT, pp. 300–307, 2007.

Socher R., Pennington J., Huang E. H., Ng A. Y., and Manning C. D. *Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions*, no. i, 2008.

Somasundaran Swapna and Wiebe Janyce. *Recognizing stances in online debates*. In Proceedings of the 47th Annual Meeting of the AC L and the 4th IJCNLP of the AFNLP (AC L-IJCNLP-2009). 2009.

Stone Philip J., Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie with Associates. *General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, Cambridge, Massachusetts, Pages 375-376. 1966.

Stoyanov Veselin and Claire Cardie. *Topic identification for fine-grained opinion analysis.* In Proceedings of COLING'08. 2008.

Strapparava, C., and Valitutti, A. *WordNet-Affect: an affective extension of WordNet*. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pp. 1083-1086. 2004.

Strapparava C. and Mihalcea R. *SemEval-2007 Task 14: Affective Text*, in Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007), Prague, Czech Republic, 2007.

Su, Qi, Xinying Xu, Honglei Guo, Zhili Guo, Xian Wu, Xiaoxun Zhang, Bin Swen, and Zhong Su. *Hidden sentiment association in Chinese web opinion mining*. In Proceedings of International Conference on World Wide Web (WWW-2008). 2008. doi:10.1145/1367497.1367627

Taboada Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll and Manfred Stede. *Lexiconbased methods for sentiment analysis*. Computational Linguistics. 37(2): pp. 267–307. 2011. doi: 10.1162/COLI_a_00049

Takamura Hiroya, Takashi Inui and Manabu Okumura. *Extracting semantic orientations of words using spin model*. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (AC L-2005). 2005. doi:10.3115/1219840.1219857

Thet T., Na J. C., and Khoo C., *Sentiment classification of movie reviews using multiple perspectives*, Digital Libraries: Universal and Ubiquitous Access to Information, pp. 184–193, 2008.

Titov Ivan and Ryan McDonald. *Modeling online reviews with multi-grain topic models*. In Proceedings of International Conference on World Wide Web (WWW-2008). 2008. doi:10.1145/1367497.1367513

Toh Zhiqiang and Wang Wenting. *DLIREC: Aspect Term Extraction and Term Polarity Classification System*. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 235–240, 2014.

Turney P.D. *Mining the Web for synonyms*: *PMI-IR versus LSA on TOEFL*. Proceedings of the Twelfth European Conference on Machine Learning. pp. 491-502. Berlin: Springer-Verlag. 2001.

Turney Peter D., *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews*. In Proceedings of Annual Meeting of the Association for Computational Linguistics (AC L-2002). 2002.

Turney Peter D. and Micharel L. Littman. *Measuring praise and criticism: Inference of semantic orientation from association.* ACM Transactions on Information Systems, 2003. doi:10.1145/944012.944013

Wang Bo and Liu Min. *Deep Learning for Aspect Based Sentiment Analysis*, Stanford University report. 2015.

Wei Wei and Jon Atle Gulla. *Sentiment learning on product reviews via sentiment ontology tree*. In Proceedings of Annual Meeting of the Association for Computational Linguistics (AC L-2010). 2010.

Wiebe Janyce, Theresa Wilson, Rebecca F. Bruce, Matthew Bell and Melanie Martin. *Learning subjective language*. Computational Linguistics. 30(3): pp. 277–308. 2004. doi: 10.1162/089120 1041850885

Wilson Theresa, Wiebe Janyce and Paul Hoffmann. *Recognizing contextual polarity in phrase-level sentiment analysis*. In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005). 2005. doi:10.3115/1220575.1220619

Wilson Theresa, Wiebe Janyce and Hwa Rebecca. *Recognizing strong and weak opinion clauses*. Computational Intelligenc. 22(2): pp. 73–99. 2006. doi:10.1111/j.1467-8640.2006.00275.x

Wu Yuanbin, Qi Zhang, Xuanjing Huang and Lide Wu. *Phrase dependency parsing for opinion mining*. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2009). 2009. doi:10.3115/1699648.1699700

Yu Hong and Hatzivassiloglou Vasileios. *Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences*. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2003), 2003.

Yu, J., Zha, Z., Wang, M., Wang, K., Chua, T.: *Domain-Assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews*. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, EMNLP. 2011.

Zhai Zhongwu, Bing Liu, Hua Xu and Peifa Jia. Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints. In Proceedings of International Conference on Computational Linguistics (CO LING-2010). 2010. doi: 10.1007/978-3-642-19460-3_11

Zhai Zhongwu, Bing Liu, Hua Xu and Peifa Jia. *Constrained LDA for Grouping Product Features in Opinion Mining*. In Proceedings of PAKDD-2011. 2011. doi:10.1109/MIS.2011.38

Zhang Lei and Liu Bing. *Identifying Noun Product Features that Imply Opinions*. ACL-2011, 2011.

Zhang Lei and Liu Bing. *Identifying noun product features that imply opinions*. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (short paper) (AC L-2011). 2011b.

Zhao Wayne Xin, Jing Jiang, Hongfei Yan and Xiaoming Li. *Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid*. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010). 2010.

Zhu Jingbo, Wang Huizhen, Tsou Benjamin K. and Zhu Muhua. *Multi-aspect opinion polling from textual reviews*. In Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2009). 2009. doi:10.1145/1645953.1646233

Zhuang Li, Feng Jing , Xiao-Yan Zhu, *Movie review mining and summarization*, Proceedings of the 15th ACM international conference on Information and knowledge management, November 06-11, Arlington, Virginia, USA , 2006. Doi:10.1145/1183614.1183625

5.2 Ηλεκτρονικές Πηγές

[1] https://en.wikipedia.org/wiki/Pattern_recognition

[2] https://en.wikipedia.org/wiki/Artificial_intelligence

[3] https://en.wikipedia.org/wiki/Cognitive_science

[4] https://www.datasciencecentral.com/profiles/blogs/types-of-machine-learning-algorithms-in-one-picture

[5] https://en.wikipedia.org/wiki/Web_crawler

[6] https://en.wikipedia.org/wiki/Instant_messaging

[7] https://en.wikipedia.org/wiki/Recommender_system

[8] http://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications

[9] https://en.wikipedia.org/wiki/Application_programming_interface

[10] https://www.languageconnect.net/blog/market-research/the-language-of-web-content-infographic

[11] https://en.wikipedia.org/wiki/Overfitting

[12] https://nlp.stanford.edu/projects/coref.shtml

[13] https://en.wikipedia.org/wiki/Multinomial_distribution

[14] https://en.wikipedia.org/wiki/Bernoulli_distribution

[15] https://en.wikipedia.org/wiki/Sigmoid_function

[16] https://eclass.upatras.gr/modules/document/file.php/CEID1094/Open Courses Edition (Διαφάνειες Διαλέξεων 14-15)/6_Μεγιστοποίηση της Εντροπίας.pdf

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

[18] https://en.wikipedia.org/wiki/Hamming_distance

[19] https://en.wikipedia.org/wiki/Covariance_matrix

[20] https://en.wikipedia.org/wiki/Bootstrap_aggregating

[21] http://en.proft.me/2017/01/22/classification-using-k-nearest-neighbors-r/

[22] https://www.mathworks.com/matlabcentral/fileexchange/52579-k-means-clustering?requestedDomain=www.mathworks.com

[23] https://en.wikipedia.org/wiki/Thesaurus

[24] https://en.wikipedia.org/wiki/Plate_notation

[25] https://en.wikipedia.org/wiki/Dimension_(vector_space)

[26] https://www.researchgate.net/Concepts-of-must-link-and-cannot-link-constriants_fig1_310767815