Τμήμα Μηχανικών Η/Υ και Πληροφορικής

επιβλέπων: Μακρής Χρήστος, Επίκουρος Καθηγητής

Opinion Mining and Sentiment Analysis

– Επισκόπηση και πειραματική αξιολόγηση τεχνικών για opinion mining και sentiment analysis –

Παναγόπουλος Παναγιώτης

ΑΜ: 4622

ΠΕΡΙΕΧΟΜΕΝΑ

1 Εισαγωγή

Sentiment Analysis ή Opinion Mining είναι ο ερευνητικός τομέας που αναφέρεται στην χρήση τεχνικών natural language processing, text analysis και computational linguistics ώστε να αναγνωριστούν και να εξαχθούν υποκειμενικές πληροφορίες από υλικό που έχει παραχθεί από ανθρώπινη δραστηριότητα. Πιο συγκεκριμένα είναι ο ερευνητικός τομέας που αναλύει τις ανθρώπινες απόψεις, τα συναισθήματα, τις αξιολογήσεις και τις στάσεις, προς οντότητες όπως προϊόντα, υπηρεσίες, οργανισμούς, πρόσωπα, εκδηλώσεις, θεματικές ενότητες, καθώς και προς τις ιδιότητες αυτών.

Η εξαγωγή και ανάλυση απόψεων αποσκοπεί στο να προσδιορίσει τη στάση ενός ομιλητή ή συγγραφέα σε σχέση με κάποιο θέμα ή την συνολική πολικότητα ενός κειμένου. Η στάση του ομιλητή ή συγγραφέα μπορεί να είναι η κρίση ή η αξιολόγησή του, η υποκειμενική κατάσταση (δηλαδή η πολικότητα της συναισθηματικής κατάστασης του συγγραφέα την στιγμή που γράφει), ή η σκοπιμότητα της συναισθηματικά πολωμένης επικοινωνίας (δηλαδή το συναισθηματικό αντίκτυπο το οποίο ο συγγραφέας επιθυμεί να έχει στον αναγνώστη).

Οι απόψεις αποτελούν σημαντικό κομμάτι της ανθρώπινης δραστηριότητας επειδή επηρεάζουν την συμπεριφορά μας. Κάθε φορά που χρειάζεται να πάρουμε μία απόφαση, θέλουμε να γνωρίσουμε τις απόψεις των άλλων. Οι εταιρείες και οργανισμοί πάντα θέλουν να έχουν εικόνα των απόψεων των καταναλωτών ή της κοινής γνώμης για τα προϊόντα ή τις υπηρεσίες τους. Επίσης οι καταναλωτές θέλουν να γνωρίζουν τις απόψεις των άλλων χρηστών, πριν προχωρήσουν σε κάποια αγορά. Ακόμα σημαντικό ρόλο στην εξέλιξη της διαδικασίας μιας πολιτικής ψηφοφορίας, παίζει η γνώση των απόψεων της κοινή γνώμης σχετικά με τους πολιτικούς υποψηφίους και τις παρατάξεις.

Η ραγδαία αύξηση των τεχνολογιών του διαδικτύου και των κοινωνικών δικτύων, έχει δημιουργήσει ένα τεράστιο όγκο από review κείμενα για προϊόντα και υπηρεσίες, καθώς και απόψεις για γεγονότα και άτομα. Το πλήθος αυτό των απόψεων απαιτεί την κατανάλωση σημαντικού χρόνου και ενέργειας, στην προσπέλαση και επεξεργασία, για την κατανόηση της συνολικής εικόνας του εύρους των απόψεων. Είναι φανερό πως το πλήθος των απόψεων αποτελεί μία πρόκληση για τον καταναλωτή, καθώς επίσης και για τα συστήματα ανάλυσης απόψεων, και αναγνωρίζεται πως η ανάπτυξη

υπολογιστικών τεχνικών, οι οποίες βοηθούν τον χρήστη στην εξόρυξη και αποδοτική χρήση όλων των απόψεων, είναι σημαντική και ενδιαφέρουσα ερευνητική πρόκληση.

Όταν αναφερόμαστε στον όρο opinion mining, συνήθως εννοούμε μία διαδικασία text mining (εξόρυξης κειμένου). Text mining ή text data mining, είναι ο ερευνητικός τομέας που αναφέρεται στην διαδικασία άντλησης πληροφοριών υψηλής ποιότητας από κείμενικές πηγές. Για τις διαδικασίες του text mining συνήθως χρησιμοποιούνται τεχνικές ανάλυσης κειμένου όπως τεχνικές ανάκτησης πληροφορίας (information retrieval), λεξικής ανάλυσης (lexical analysis), αναγνώρισης προτύπων (pattern recognition), και εξόρυξης δεδομένων (data mining), καθώς επίσης και τεχνικές επεξεργασίας φυσικής γλώσσας (natural language processing, NLP).

Σε αυτό το κείμενο πραγματοποιείται μία αναλυτική αναφορά στο αντικείμενο και τις συνήθης τεχνικές της εξόρυξης και ανάλυσης απόψεων, και παρουσιάζεται το πρόβλημα της αξιολόγησης οντοτήτων χρησιμοποιώντας τις απόψεις που υπάρχουν σε review κείμενα χρηστών, για το οποίο προτείνονται νέες τεχνικές προσέγγισης, και πραγματοποιούνται πειράματα για την αξιολόγηση της απόδοσής τους. Πιο συγκεκριμένα, στη δεύτερη ενότητα παρατίθεται αναλυτική παρουσίαση του αντικειμένου της εξόρυξης και ανάλυσης απόψεων, των εφαρμογών που απαιτούν την ύπαρξη συστημάτων ανάλυσης απόψεων, καθώς και των παραγόντων που φράσσουν την απόδοση αυτών των συστημάτων. Στην τρίτη ενότητα παρουσιάζονται τα τρία βασικά επίπεδα ανάλυσης απόψεων, κειμένου, πρότασης και aspect, καθώς επίσης και τεχνικές που συνήθως χρησιμοποιούνται για την πραγματοποίηση της διαδικασίας της εξόρυξης και ανάλυσης απόψεων σε κάθε ένα από αυτά. Στην τέταρτη ενότητα παρουσιάζεται αναλυτικά το πρόβλημα της αξιολόγησης οντοτήτων χρησιμοποιώντας τις απόψεις που υπάρχουν σε review κείμενα χρηστών, και αναφέρονται γνωστές τεχνικές προσέγγισής της λύσης του. Για αυτό το πρόβλημα προτείνονται νέες τεχνικές για την βελτίωση της απόδοσης της διαδικασίας της αξιολόγησης των οντοτήτων, και αναφέρονται τα πειράματα και οι μετρήσεις που πραγματοποιήθηκαν, όπως επίσης και οι παρατηρήσεις που εξήχθησαν.

2 Opinion Mining and Sentiment Analysis

2.1 Εισαγωγή στην εξόρυξη και ανάλυση απόψεων

Όπως αναφέρθηκε παραπάνω στην εισαγωγή, ο τομέας του opinion mining and sentiment analysis είναι ερευνητικός τομέας που αναλύει τις ανθρώπινες απόψεις, τα συναισθήματα, τις αξιολογήσεις και τις στάσεις, προς οντότητες όπως προϊόντα, υπηρεσίες, οργανισμούς, πρόσωπα, εκδηλώσεις, θεματικές ενότητες, καθώς και προς τις ιδιότητες αυτών. Αυτός ο τομέας αντιπροσωπεύει ένα μεγάλο χώρο προβλημάτων, και για αυτό το λόγο υπάρχουν πολλά ονόματα και ελαφρώς διαφορετικές διαδικασίες. Για παράδειγμα, sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining. Ωστόσο, τώρα, όλες αυτές οι διαδικασίες είναι κάτω από την ομπρέλα του Opinion Mining and Sentiment Analysis. Αν και στον κόσμο των εταιρειών, είναι συχνότερος ο όρος sentiment analysis, στην ακαδημαϊκή κοινότητα και οι δύο όροι, opinion mining και sentiment analysis, χρησιμοποιούνται το ίδιο συχνά. Για συντακτικούς κυρίως λόγους, σε αυτό το κείμενο υϊοθετούμε, σχεδόν αυθαίρετα, τον όρο “εξόρυξη και ανάλυση απόψεων” ως την ελληνική μετάφραση του όρου “Opinion mining and Sentiment Analysis”. Ο λόγος που επιλέξαμε αυτήν την μετάφραση αντί της “Εξόρυξη απόψεων και Ανάλυση συναισθήματος” είναι πως η ανάλυση συναισθήματος πραγματοποιείται αποκλειστικά με την ανάλυση και επεξεργασία των απόψεων.

Ο επιστημονικός τομέας του opinion mining και sentiment analysis σχετίζεται με τον τομέα του natural language processing (NLP), διότι ο τομέας του NLP ασχολείται με την επεξεργασία της ανθρώπινης γλώσσας και την εξαγωγής νοήματος από αυτή. Αν και η γλωσσολογία, καθώς και η επεξεργασία φυσικής γλώσσας (NLP), έχουν μία μακρά ερευνητική ιστορία, έχει πραγματοποιηθεί λίγη έρευνα σχετικά με τις τεχνικές εξαγωγής και ανάλυσης απόψεων από κειμενικές πηγές, πριν από το 2000. Από τότε, ο τομέας του Opinion Mining έχει γίνει μία πολύ ενεργή ερευνητική περιοχή. Υπάρχουν αρκετοί λόγοι για αυτό. Πρώτον, έχει ένα ευρύ φάσμα εφαρμογών, σχεδόν σε κάθε θεματικό τομέα. Η βιομηχανία γύρω από το Opinion Mining άνθισε λόγω των εμπορικών εφαρμογών. Αυτό παρέχει ένα ισχυρό κίνητρο για την έρευνα. Δεύτερον, παρέχει πολλά ενδιαφέροντα ερευνητικά προβλήματα, τα οποία δεν έχουν μελετηθεί παλιότερα. Τρίτον, για πρώτη φορά στην ανθρώπινη ιστορία, έχουμε ένα τεράστιο όγκο από δεδομένα απόψεων στα κοινωνικά δίκτυα στο διαδίκτυο. Χωρίς αυτά τα δεδομένα, με-

γάλο μέρος της έρευνα δεν θα ήταν πραγματοποιήσιμο. Δεν είναι τυχαίο το γεγονός πως η έναρξη και η ταχεία ανάπτυξη της εξόρυξης και ανάλυσης απόψεων συμπίπτουν με αυτές των κοινωνικών δικτύων. Επίσης παρέχει ένα χώρο για τους ερευνητές του τομέα της επεξεργασίας φυσικής γλώσσας όπου μπορούν να μελετήσουν συγκεκριμένα συντακτικά φαινόμενα τα οποία ύστερα μπορούν να χρησιμοποιήσουν για την επίλυση γενικότερων συντακτικών προβλημάτων. Οπότε μπορούμε να πούμε πως η εξόρυξη και ανάλυση απόψεων έχει να προσφέρει δεδομένα, παρατηρήσεις και εργαλεία τα οποία θα βοηθήσουν στην καλύτερη προσέγγιση της λύσης του προβλήματος της άντλησης νοήματος από είσοδο σε ανθρώπινη ή φυσική γλώσσα.

Είναι γεγονός πως αυτός ο τομέας είναι στο επίκεντρο της ερευνητικής δραστηριότητας που έχει να κάνει με τα κοινωνικά δίκτυα. Ως εκ τούτου, η έρευνα πάνω στις απόψεις χρηστών δεν έχει μόνο σημαντικό αντίκτυπο για τον τομέα της επεξεργασίας της φυσικής γλώσσας, αλλά μπορεί επίσης να έχει μια βαθιά επίδραση στις επιστήμες διαχείρισης (management sciences), τις πολιτικές επιστήμες, στα οικονομικά και στις κοινωνικές επιστήμες, καθώς όλες αυτές επηρεάζονται από τις απόψεις των ανθρώπων.

2.2 Εφαρμογές της εξόρυξης και ανάλυσης απόψεων

Οι απόψεις αποτελούν σημαντικό κομμάτι της ανθρώπινης δραστηριότητας επειδή επηρεάζουν την συμπεριφορά μας. Κάθε φορά που χρειάζεται να πάρουμε μία απόφαση, θέλουμε να γνωρίσουμε τις απόψεις των άλλων. Οι εταιρείες και οργανισμοί θέλουν να έχουν εικόνα των απόψεων των καταναλωτών ή της κοινής γνώμης για τα προϊόντα ή τις υπηρεσίες τους. Επίσης οι καταναλωτές θέλουν να γνωρίζουν τις απόψεις των άλλων χρηστών, πριν προχωρήσουν σε κάποια αγορά. Ακόμα σημαντικό ρόλο στην εξέλιξη της διαδικασίας μιας πολιτικής ψηφοφορίας, παίζει η γνώση των απόψεων της κοινή γνώμης σχετικά με τους πολιτικούς υποψηφίους και τις παρατάξεις.

Πριν από περίπου μία δεκαετία, όταν κάποιος χρειαζόταν απόψεις σχετικά με κάποιο θέμα, τις αναζητούσε στον κοινωνικό κύκλο των φίλων ή της οικογένειάς του. Όταν μία επιχείρηση χρειαζόταν τις απόψεις των καταναλωτών, διεξήγαγε έρευνες κυρίως με την μορφή των δημοσκοπήσεων. Η απόκτηση της εικόνας των απόψεων της κοινή γνώμης και των καταναλωτών, αποτελούσε πάντα μία τεράστια επιχείρηση για το marketing, τις δημόσιες σχέσεις και τις εκστρατείες πολιτικών κινήσεων. Με την

εκρηκτική ανάπτυξη των μέσων κοινωνικής δικτύωσης στον διαδίκτυο (reviews, forums συζητήσεων, blogs, twitter, facebook), τα άτομα και οι επιχειρήσεις χρησιμοποιούν όλο και περισσότερο περιεχόμενο από αυτά τα μέσα για την λήψη αποφάσεων. Στις μέρες μας, ο καταναλωτής για την πραγματοποίηση μιας αγοράς, δεν είναι περιορισμένος πλέον στο στενό κύκλο των φίλων και της οικογένειας για την αναζήτηση απόψεων διότι υπάρχουν πολλά reviews και συζητήσεις σε forums στο διαδίκτυο σχετικά με το προϊόν της αγοράς. Για έναν οργανισμό, ίσως δεν είναι πλέον απαραίτητο να διεξάγει δημοσκοπήσεις προκειμένου να συγκεντρώσει τις απόψεις της κοινής γνώμης διότι υπάρχει μία αφθονία υποκειμενικών πληροφοριών, καταχωρημένη από τους ίδιους τους χρήστες, δημόσια διαθέσιμη online.

Ωστόσο, το να εντοπίσεις και να παρακολουθήσεις ιστοσελίδες με απόψεις στο διαδίκτυο, και να αξιολογήσεις τις πληροφορίες που περιέχουν, παραμένει μία δύσκολη διαδικασία εξαιτίας του πλήθους των ιστοσελίδων. Κάθε ιστότοπος συνήθως περιέχει τεράστιο όγκο απόψεων σε μορφή κειμένου το οποίο δεν είναι πάντα εύκολο να εξαχθεί. Ο μέσος άνθρωπος αντιμετωπίζει πρόβλημα στο να εντοπίσει αξιόλογες ιστοσελίδες με υποκειμενικό περιεχόμενο, καθώς και στο να εξάγει ή να συνοψίσει τις απόψεις που παρουσιάζονται σε αυτές. Οπότε μπορούμε εύκολα να δούμε πως παρουσιάζεται η ανάγκη για την δημιουργία συστημάτων αυτόματης εξαγωγής και ανάλυσης απόψεων.

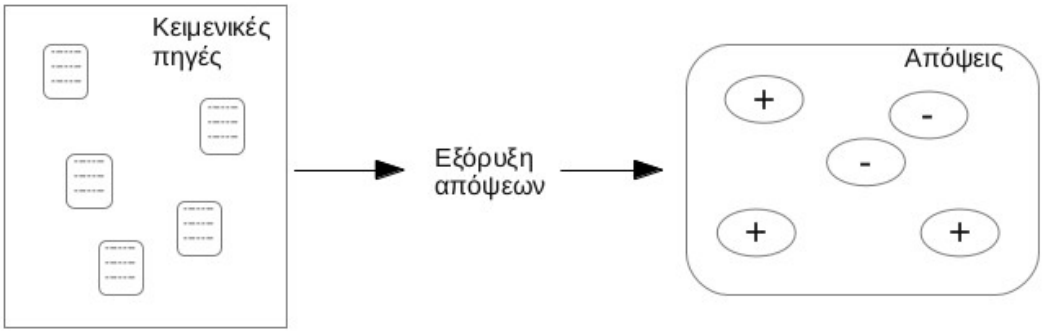
Τα τελευταία χρόνια έχουμε δει πως η ανάρτηση υποκειμενικού περιεχομένου στα κοινωνικά μέσα από τους χρήστες, έχει οδηγήσει στην αναμόρφωση των μοντέλων των επιχειρήσεων και έχει επηρεάσει τα κοινωνικά και πολιτικά μας συστήματα. Επίσης έχουμε δει αυτό το φαινόμενο να συνεισφέρει στην κινητοποίηση των μαζών για πολιτική και κοινωνική αλλαγή. Λόγω των παραπάνω, οι εφαρμογές της εξόρυξης και ανάλυσης απόψεων έχουν εξαπλωθεί σε κάθε πιθανό θεματικό τομέα, από τα προϊόντα, τις υπηρεσίες και την οικονομία στις κοινωνικές εκδηλώσεις και τις πολιτικές εκλογές.

Οι πρακτικές εφαρμογές και τα βιομηχανικά συμφέροντα έχουν δημιουργήσει ισχυρά κίνητρα για την έρευνα στον τομέα του opinion mining και sentiment analysis. Εκτός όμως από τις πρακτικές εφαρμογές έχουν επίσης δημοσιευθεί αρκετές έρευνες προσανατολισμένες σε επιπλέον πιθανές εφαρμογές. Για παράδειγμα, στην έρευνα Liu et al. (2007) παρουσιάζεται ένα sentiment μοντέλο το οποίο προβλέπει την απόδοση των πωλήσεων. Στην έρευνα McGlohon et al. (2010), χρησιμοποιούνται reviews για

την παραγωγή κατάταξης προϊόντων και προμηθευτών. Στην έρευνα O'Connor et al. (2010), μελετήθηκε η συσχέτιση του Twitter sentiment με τις δημοσκοπήσεις απόψεων της κοινής γνώμης. Επίσης το Twitter sentiment χρησιμοποιήθηκε για να προβλεφθούν αποτελέσματα εκλογών. Στην έρευνα Asur and Huberman (2010), Joshi et al. (2010) and Sadikov et al. (2009), δεδομένα από το Twitter, reviews ταινιών, και κείμενα από blogs χρησιμοποιήθηκαν για να προβλεφθεί η εισπρακτική επιτυχία ταινιών. Στην έρευνα Miller et al. (2011), ερευνήθηκε η ροή υποκειμενικών απόψεων (sentiment flow) στα κοινωνικά δίκτυα. Στην έρευνα Zhang and Skiena (2010), χρησιμοποιήθηκε η sentiment πληροφορία από blogs και ειδησεογραφικές πηγές για να μελετηθούν στρατηγικές marketing. Στην έρευνα Groh and Hauffa (2011), χρησιμοποιήθηκε sentiment analysis για να χαρακτηριστούν κοινωνικές σχέσεις.

2.3 Το προβλήμα της εξόρυξης και ανάλυσης απόψεων

Η εξαγωγή και επεξεργασία απόψεων έχει ως στόχο την αναγνώριση και να εξαγωγή υποκειμενικών πληροφοριών από κειμενικές πηγές.



Αυτές οι υποκειμενικές πληροφορίες ονομάζονται απόψεις. Πριν προχωρήσουμε στον ορισμό της διαδικασίας της εξόρυξης και ανάλυσης απόψεων, είναι απαραίτητο πρώτα να ορίσουμε τι είναι άποψη. Ακολουθεί ο ορισμός της γενικής άποψης ως το κύριο μοντέλο άποψης, και η αναφορά των διάφορων κατηγοριών απόψεων.

2.3.1 Ορισμός της άποψης

Μία άποψη αποτελείται από ένα στόχο και ένα συναίσθημα. Για παράδειγμα μία πρόταση που περιέχει μία άποψη *"Ο καφές είναι καλός"*. Εδώ ο στόχος είναι ο *"κα-*

φές" και το συναίσθημα είναι το *"είναι καλός"*. Σε αυτή την πρόταση εκφράζεται μία άποψη θετικής πολικότητας. Έστω *(g,s)* ο ορισμός της άποψης, όπου *g* είναι ο στόχος της άποψης και *s* είναι το συναίσθημα που εκφράζεται για το στόχο *g*. Ο στόχος *g* μπορεί να είναι μία οντότητα, ή μέρος ή ιδιότητα μιας οντότητας. Το *s* είναι ένα θετικό, αρνητικό ή ουδέτερο συναίσθημα, ή αριθμητική βαθμολογία η οποία εκφράζει την ένταση του συναισθήματος (1-10 ratings ή 0-5 stars).

Ο στόχος της άποψης μπορεί να είναι μία οντότητα, ή μέρος ή ιδιότητα μιας οντότητας. Για παράδειγμα *"η γεύση του καφέ είναι καλή"*. Εδώ ο στόχος της άποψης είναι μία ιδιότητα της οντότητας καφές, *"η γεύση του καφέ"*. Μία οντότητα ορίζεται ως *e(T,W)*, όπου *T* είναι η ιεραρχία με τα υπομέρη (sub-parts) που αποτελούν την οντότητα, και *W* είναι το σύνολο των ιδιοτήτων της οντότητας. Η αναγνώριση των μερών και των ιδιοτήτων μιας οντότητας είναι δύσκολο NLP πρόβλημα και συνήθως χρησιμοποιείται μια ιεραρχία 2 επιπέδων: entity (επίπεδο 1) και aspects (επίπεδο 2) , όπου aspects είναι το σύνολο των υπό-μερών και ιδιοτήτων της οντότητας

Ο αναλυτικός ορισμός της άποψης είναι ο εξής (Liu, 2006, 2011):

Μια άποψη είναι μία πεντάδα (e_i, a_{ij}, s_{ijkl}, h_k, t_l), όπου e_i είναι μία οντότητα, a_{ij} είναι ένα χαρακτηριστικό (aspect/feature) της οντότητας e_i, s_{ijkl} είναι το συναίσθημα το οποίο εκφράζεται για το aspect a_{ij} της οντότητας e_i, h_k είναι αυτός που εκφράζει την άποψη (opinion holder), και t_l είναι η χρονική στιγμή την οποία εκφράζεται η άποψη από τον h_k.

Το συναίσθημα s_{ijkl} είναι θετικό, αρνητικό ή ουδέτερο, ή εκφράζεται με διαφορετικά επίπεδα έντασης, για παράδειγμα 1–5 stars όπως χρησιμοποιείται συνήθως σε ιστοσελίδες με reviews στο διαδίκτυο. Σε αυτό τον ορισμό τα e_i και a_{ij} μαζί αντιπροσωπεύουν τον στόχο της άποψης. Όταν μία άποψη έχει ως στόχο την οντότητα ως ολότητα το a_{ij} μπορεί να πάρει την τιμή *"general"*.

Ο ορισμός αυτός καλύπτει τις περισσότερες αλλά όχι όλες τις πιθανές οπτικές της σημασιολογικής έννοιας μιας άποψης, η οποία μπορεί να είναι οσοδήποτε σύνθετη. Για παράδειγμα, δεν καλύπτει την εξής περίπτωση *"Το σκόπευτρο και ο φακός είναι πολύ κοντά"*, η οποία εκφράζει άποψη για την απόσταση μεταξύ των δύο μερών. Επίσης αυτό το μοντέλο της άποψης είναι ένας από τους τύπους απόψεων που χρησιμο-

ποιούνται αρκετά συχνά και ονομάζεται γενική άποψη. Παρά τους περιορισμούς που εισάγει στην ανάλυση, μπορούμε ακόμα να το χρησιμοποιούμε μιας και αποδεικνύεται επαρκές για τις περισσότερες εφαρμογές. Άλλωστε ένα πολύ σύνθετο μοντέλο μπορεί να κάνει το πρόβλημα εξαιρετικά δύσκολο ώστε να λυθεί.

2.3.1.1 Κατηγορίες απόψεων

Παραπάνω δόθηκε ο ορισμός της γενικής άποψης. Η άλλη μεγάλη κατηγορία απόψεων είναι η συγκριτικές απόψεις (comparative opinion), (Jindal and Liu, 2006b).
Οι γενικές απόψεις κατηγοριοποιούνται σε δύο υποκατηγορίες, τις άμεσες (direct) και τις έμμεσες (indirect) (Liu, 2006, 2011). Οι συγκριτικές απόψεις μπορούν να κατηγοριοποιηθούν σε δύο κύριες υποκατηγορίες, αυτές που εκφράζουν διαβαθμισμένη σύγκριση (gradable comparison) και αυτές που δεν χρησιμοποιούν διαβάθμιση για να πραγματοποιήσουν την σύγκριση (non-gradable comparison). Όλες οι απόψεις μπορούν να κατηγοριοποιηθούν επίσης σε δύο υποκατηγορίες, ανάλογα με το πως αυτές εκφράζονται στο κείμενο, τις σαφής (explicit) και τις υπονοούμενες (implicit).

Πιο συγκεκριμένα, γενική άποψη είναι η απλή άποψη όπως ορίστηκε πιο πάνω, όπου υπάρχει ένας στόχος και εκφράζεται ένα συναίσθημα για αυτό τον στόχο. Συχνά μπορεί να αναφέρεται και η οντότητα που εκφράζει την άποψη (opinion holder) αλλά και ο χρόνος που διατυπώνεται η άποψη. Μία γενική άποψη έχει δύο υποκατηγορίες, τις άμεσες και τις έμμεσες. Άμεση είναι μία γενική άποψη η οποία εκφράζεται με άμεσο τρόπο για μία οντότητα ή ένα χαρακτηριστικό μιας οντότητας (aspect). Για παράδειγμα *"Η γεύση του καφέ είναι καλή"*. Έμμεση είναι μία γενική άποψη η εκφράζεται για μία οντότητα ή ένα χαρακτηριστικό μιας οντότητας με βάση τα αποτελέσματα της οντότητας στόχου σε άλλες οντότητες, δηλαδή με έμμεσο τρόπο. Για παράδειγμα *"Μετά απο την κατανάλωση μεγάλης ποσότητας καφέ, ο ύπνος δεν είναι ευχάριστος"*. Εδώ περιγράφεται ένα μη επιθυμητό αποτέλεσμα του *"καφέ"* στον *"ύπνο"*, το οποίο έμμεσα δίνει μία αρνητική άποψη για τον *"καφέ"*. Σε αυτή την περίπτωση η οντότητα στόχος είναι ο *"καφές"* και το χαρακτηριστικό είναι το αποτελέσματά της στον *"ύπνο"*. Όπως γίνεται φανερό, οι έμμεσες απόψεις είναι συχνά πιο δύσκολο να τις διαχειριστούμε, λόγω της περιπλοκής δομής τους, και για αυτό το λόγο οι περισσότερες έρευνες επικεντρώνονται στις άμεσες απόψεις.

Η άλλη μεγάλη κατηγορία απόψεων που χρησιμοποιούνται είναι οι συγκριτικές απόψεις. Μία συγκριτική άποψη εκφράζει μία σχέση ομοιοτήτων ή διαφορών μεταξύ δύο ή περισσότερων οντοτήτων, και/ή μία προτίμηση του opinion holder με βάση

κάποια κοινά χαρακτηριστικά των οντοτήτων (Jindal and Liu, 2006a, 2006b). Για παράδειγμα, *"Η coca-cola έχει καλύτερη γεύση από την pepsi"* ή *"Η pepsi έχει την καλύτερη γεύση"*. Μία συγκριτική άποψη συχνά εκφράζεται χρησιμοποιώντας την συγκριτική ή υπερθετική μορφή ενός επιθέτου ή ενός επιρρήματος, αλλά όχι πάντα. Αυτές οι απόψεις κατηγοριοποιούνται σε δύο κύριες υποκατηγορίες, αυτές που εκφράζουν διαβαθμισμένη σύγκριση (gradable comparison) και αυτές που δεν χρησιμοποιούν διαβάθμιση για να πραγματοποιήσουν την σύγκριση (non-gradable comparison). Οι απόψεις διαβαθμισμένης σύγκρισης είναι αυτές που εκφράζουν μία σχέση ταξινόμισης των οντοτήτων που συγκρίνονται. Για παράδειγμα *"Προτιμώ τον espresso από τον frappe"*. Οι απόψεις μη-διαβαθμισμένης σύγκρισης διατυπώνουν μία σχέση δύο ή περισσότερων οντοτήτων, αλλά δεν τις διατάσσουν (δεν δείχνουν προτίμηση σε κάποια από τις οντότητες). Για παράδειγμα, *"Ο espresso έχει διαφορετική γεύση από το frappe"*. Οι απόψεις μη-διαβαθμισμένης σύγκρισης, αν και τις περισσότερες φορές δεν περιέχουν χρήσιμη πληροφορία, μιας και δεν εκφράζουν κάποια προτίμηση, μπορεί ακόμα κάποιες φορές να διατυπώνουν άποψη, αλλά είναι αρκετά δύσκολο να την διαγνώσουμε.

2.3.2 Ορισμός της διαδικασίας της εξόρυξης και ανάλυσης απόψεων

Έχοντας δώσει τον ορισμό της άποψης και τους συνήθεις τρόπους με τους οποίους αυτές εμφανίζονται μπορούμε να προχωρήσουμε στον ορισμό της διαδικασίας της εξόρυξης και ανάλυσης απόψεων.

Δοθέντος ενός κειμένου d (ή μιας συλλογής κειμένων D), ανακάλυψε όλες τις πλειάδες απόψεων (e_i, a_{ij}, s_{ijkl}, h_k, t_l) που υπάρχουν μέσα στο d. Η διαδικασία περιέχει 6 διεργασίες (tasks):

- *Task1: εξαγωγή όλων των οντοτήτων (entities) από το D και κατηγοριοποίηση των συνώνυμων σε ομάδες (entity clusters). Σε κάθε cluster δίνεται το αναγνωριστικό e_i*
- *Task2: εξαγωγή όλων των χαρακτηριστικών (aspects) από το D και κατηγοριοποίηση των συνώνυμων σε ομάδες (aspects clusters). Σε κάθε cluster δίνεται το αναγνωριστικό a_{ij} του e_i.*
- *Task3: εξαγωγή των opinion holders και κατηγοριοποίηση σε ομάδες με αναγνωριστικό h_k.*
- *Task4: εξαγωγή του χρόνου και κανονικοποίηση σε ένα πρότυπο.*

- *Task5: προσδιορισμός της πολικότητας της άποψης (θετική/αρνητική/ουδέτερη), ή βαθμολόγηση (σε ποσοστό επί τις % ή 1-5 stars).*
- *Task6: παραγωγή όλων των πλειάδων απόψεων (e_i, a_{ij}, s_{ijkl}, h_k, t_l).*

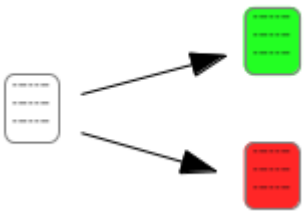
Ο παραπάνω ορισμός περιγράφει την πλήρη διαδικασία εξόρυξης και ανάλυσης απόψεων. Πρέπει να τονιστεί πως δεν απαιτείται από όλες τις εφαρμογές η πλήρης διαδικασία. Συνήθως ζητούνται πλειάδες του τύπου (e_i, a_{ij}, s_{ijkl}), οπότε οι διεργασίες 3 και 4 παραλείπονται.

Η κατηγοριοποίηση των συνωνύμων σε ομάδες που πραγματοποιείται στα παραπάνω tasks είναι απαραίτητη διότι, όπως συμβαίνει σε όλα τα κείμενα φυσικής γλώσσας, χρησιμοποιούνται διαφορετικές λέξεις (συνώνυμα) ή φράσεις για να γίνει αναφορά στις ίδιες οντότητες ή τα ίδια χαρακτηριστικά. Επίσης η κατηγοριοποίηση μπορεί να υπαγορεύεται από την ίδια την εφαρμογή ως μηχανισμός αφαίρεσης.

2.3.3 Διαφορετικά επίπεδα ανάλυσης

Η εξόρυξη και ανάλυση απόψεων έχει ερευνηθεί κυρίως σε τρία επίπεδα. Το επίπεδο κειμένου (document level), το επίπεδο πρότασης (sentence level), και το επίπεδο οντότητας ή χαρακτηριστικού (entity or aspect level).

Στο επίπεδο κειμένου στόχος είναι η κατηγοριοποίηση ενός ολόκληρου εγγράφου σε κείμενο θετικής πολικότητας ή αρνητικής, ανάλογα με το αν εκφράζει, ως ολότητα, θετικό ή αρνητικό συναίσθημα (document sentiment classification).



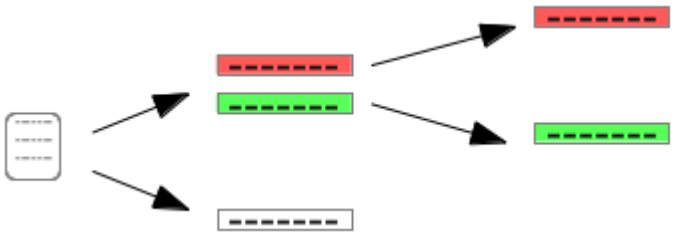
Σε αυτό το επίπεδο υποθέτουμε πως το έγγραφο αναφέρεται σε μία και μόνο οντότητα. Έτσι η διαδικασία είναι ευκολότερη διότι δεν χρειάζεται να πραγματοποιηθεί εξαγωγή οντοτήτων, και είναι ασφαλές να κάνουμε την υπόθεση πως όποια aspects υπάρχουν στο κείμενο, αναφέρονται στην μοναδική οντότητα του κειμένου. Σε αυτήν την περίπτωση ο ορισμός της διαδικασίας της εξόρυξης και ανάλυσης απόψεων διαμορφώνεται ως εξής:

Δοθέντος ενός κειμένου απόψεων d, να προσδιοριστεί η συνολική sentiment πολικότητα s που εκφράζει ο συγγραφέας (opinion holder) για την οντότητα, ως σύνολο (GENERAL). Σε αυτό το επίπεδο ανάλυσης η ζητούμενη πεντάδα είναι η (_, GENERAL, s, _, _), δηλαδή δεν ενδιαφερόμαστε για την εξαγωγή της οντότητας e, του opinion holder h, και του χρόνου t που διατυπώθηκε η άποψη.

Στο επίπεδο ανάλυσης κειμένου το πρόβλημα της εξαγωγής και ανάλυσης απόψεων παρουσιάζει δύο μορφές, ανάλογα με τη είδος του πεδίου τιμών της επιθυμητής εξόδου. Αν το πεδίο τιμών αποτελείται από διακριτές (κατηγορηματικές) τιμές, όπως θετικό ή αρνητικό, τότε είναι ένα πρόβλημα κατηγοριοποίησης. Αν το πεδίο τιμών αποτελείται από αριθμητικές τιμές ή βαθμολογίες κατάταξης σε ένα προκαθορισμένο εύρος, όπως για παράδειγμα 1-5 stars, τότε είναι ένα regression πρόβλημα. Στην στατιστική, regression analysis, είναι μία στατιστική διαδικασία για την εκτίμηση των σχέσων μεταξύ μεταβλητών, και περιλαμβάνει τεχνικές για την μοντελοποίηση και την ανάλυση των μεταβλητών και των σχέσεών τους. Στον τομέα του machine learning χρησιμοποιείται συνήθως για να προσαρμοστεί μία συνάρτηση σε ένα σύνολο δεδομένων. Η απλούστερη μορφή regression, linear regression, χρησιμοποιεί το μοντέλο μιας ευθείας γραμμής ($y = mx + b$) και καθορίζει τις κατάλληλες τιμές για τα m και b για να προβλέψει την τιμή y σε μία είσοδο x. Σε αυτό το επίπεδο ανάλυσης, όπως αναφέρθηκε πιο πάνω, γίνεται η υπόθεση πως σε κάθε κείμενο εκφράζονται απόψεις για μία συγκεκριμένη οντότητα από ένα συγκεκριμένο συγγραφέα. Αυτή η υπόθεση μπορεί να εφαρμοστεί στις περιπτώσεις όπου τα κείμενα τα οποία φέρουν τις απόψεις είναι reviews προϊόντων ή υπηρεσιών, διότι κάθε review συνήθως εστιάζει σε μία συγκεκριμένη οντότητα και ο συγγραφέας είναι ένα συγκεκριμένο άτομο. Ωστόσο, αυτή η υπόθεση συνήθως δεν μπορεί να τεθεί σε περιπτώσεις όπου τα κείμενα προέρχονται από blogs ή συζητήσεις διότι σε τέτοιου είδους κείμενα ο συγγραφέας μπορεί να αναφέρεται σε περισσότερες από μία οντότητες, όπως για παράδειγμα στην αρκετά συχνή περίπτωση της σύγκρισης οντοτήτων του ίδιου είδους.

Στο επίπεδο πρότασης στόχος είναι η κατηγοριοποίηση των προτάσεων σε θετικής ή αρνητικής πολικότητας, ανάλογα με το αν εκφράζει θετικό ή αρνητικό συναίσθημα (sentence sentiment classification). Πρέπει να προηγηθεί κατηγοριοποίηση της πρότα-

σης σε ουδέτερη ή πρόταση με πολικότητα, ανάλογα με το αν εκφράζει ή όχι κάποιο συναίσθημα (subjectivity classification).

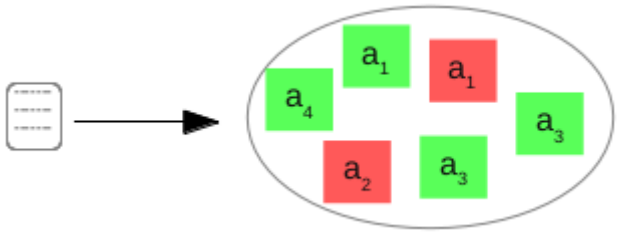


Σε αυτήν την περίπτωση ο ορισμός της διαδικασίας της εξόρυξης και ανάλυσης απόψεων μετατρέπεται ως εξής:

Δοθέντος μίας πρότασης x, το ζητούμενο είναι να προσδιοριστεί αν η πρόταση εκφράζει θετική, αρνητική, ή ουδέτερη άποψη.

Αν και με βάση την παρατήρηση ότι οι προτάσεις είναι κείμενα μικρού μήκους, δεν φαίνεται να υπάρχει σημαντική διαφορά μεταξύ της ανάλυσης σε επίπεδο κειμένου και πρότασης, η υπόθεση που κάνουν συχνά οι ερευνητές, και είναι αυτή που διαφοροποιεί τα δύο επίπεδα, είναι πως μία πρόταση συχνά φέρει μία άποψη. Αυτή η υπόθεση απλοποιεί την ανάλυση, αλλά όπως είναι φανερό, δεν ισχύει σε πολλές περιπτώσεις.

Στο επίπεδο ανάλυσης aspect ή χαρακτηριστικού, στόχος είναι ο προσδιορισμός των χαρακτηριστικών (aspects) της οντότητας, δηλαδή των υπομερών και των ιδιοτήτων της, καθώς και της πολικότητας της άποψης που εκφράζεται για αυτά τα χαρακτηριστικά. Μία πρόταση συνήθως περιέχει περισσότερα του ενός aspects. Για παράδειγμα: "*Αν και ο καφές δεν είναι ο καλύτερος, μου αρέσει αυτή η καφετέρια*". Σε αυτό το επίπεδο γίνεται ανάλυση κατευθείαν στην ίδια την άποψη. Για την εξαγωγή αυτών των λεπτομερειών πρέπει να πραγματοποιήσουμε ανάλυση απόψεων σε επίπεδο aspect. Αυτό σημαίνει πως πρέπει να χρησιμοποιήσουμε τον πλήρη ορισμό της εξόρυξης και ανάλυσης απόψεων με τις έξι διεργασίες, όπως αυτός αναφέρεται πιο πάνω σε αυτή την ενότητα.



Στο aspect επίπεδο ανάλυσης θεωρούμε πως το σύνολο των στόχων των απόψεων αποτελείται από τις οντότητες και τα χαρακτηριστικά τους (υπομέρη και ιδιότητες). Σε αυτό το επίπεδο η διαδικασία του opinion mining γίνεται πιο δύσκολη, διότι η αναγνώριση των οντοτήτων και των συναισθημάτων που εκφράζονται για αυτές, απαιτεί βαθύτερη NLP ανάλυση.

2.4 Παράγοντες που καθιστούν δύσκολη την εξόρυξη και ανάλυση απόψεων

Ο τομέας του opinion mining είναι ως ένας σχετικά νέος ερευνητικός τομέας, και η διαδικασία της εξόρυξης και ανάλυσης απόψεων θεωρείται ως μία δύσκολη διαδικασία για την οποία δεν υπάρχει κάποια συγκεκριμένη προσέγγιση που να προσφέρει ικανοποιητικά αποτελέσματα. Οι δυσκολίες απορρέουν από το γεγονός οτί απαιτείται από τους αλγορίθμους να είναι σε θέση να κατανοούν πλήρως την σημασιολογική πληροφορία των κειμένων, καθώς επίσης και από την φύση των απόψεων, οι οποίες συχνά εκφράζονται με σύνθετο και πολύπλοκο τρόπο. Φυσικά, μέχρι στιγμής, δεν βρισκόμαστε στην ευχάριστη θέση να απαιτούμε από τις μηχανές, και τους αλγορίθμους, να κατανοούν πλήρως την σημασιολογική πληροφορία των κειμένων. Όμως, ακόμη και αν βρισκόμασταν σε αυτήν την θέση, τα αποτελέσματα δεν θα ήταν 100% ικανοποιητικά, διότι και οι ίδιο οι άνθρωποι αντιμετωπίζουν προβλήματα. Έχει δειχθεί πως οι ανθρώπινοι βαθμολογητές συμφωνούν μόνο με ποσοστό 80%.

Αρχικά μπορούμε να αναζητήσουμε τις δυσκολίες της εξόρυξης και ανάλυσης απόψεων στις διεργασίες που απαιτούνται με βάση τον ορισμό της διαδικασίας, όπως αυτός ορίστηκε πιο πάνω σε αυτή την ενότητα. Η διαδικασία για την εξόρυξη των απόψεων και τον προσδιορισμό της πολικότητας αυτών, απαιτεί τις εξής διεργασίες:

- εξαγωγή όλων των οντοτήτων
- εξαγωγή όλων των χαρακτηριστικών

- εξαγωγή των opinion holders
- εξαγωγή της χρονικής στιγμής την οποία διατυπώθηκε η άποψη
- προσδιορισμός της πολικότητας της άποψης (θετική/αρνητική/ουδέτερη)

Στις παραπάνω διεργασίες απαιτείται η εξαγωγή συγκεκριμένων, πληροφοριών από κείμενο. Η εξόρυξη και ανάλυση απόψεων είναι μία διαδικασία text data mining (text mining ή text analysis). Στόχος του text mining είναι να μετατρέπει το κείμενο σε δεδομένα τα οποία είναι κατάλληλα για την πραγματοποίηση ανάλυσης από αλγορίθμους, και περιέχει, επίσης, διαδικασίες όπως είναι η κατηγοριοποίηση κειμένου (text categorization), η συσταδοποίηση κειμένου (text clustering), η σύνοψη κειμένου (document summarization), και άλλες. Ο στόχος του opinion mining είναι να αναγνωριστούν και να εξαχθούν υποκειμενικές πληροφορίες από κειμενικές πηγές. Ωστόσο, το opinion mining απαιτεί "λεπτότερη" εξαγωγή και ανάλυση από ότι οι άλλες text mining διαδικασίες. Για τις διαδικασίες του text mining συνήθως χρησιμοποιούνται τεχνικές ανάλυσης κειμένου όπως τεχνικές ανάκτησης πληροφορίας (information retrieval), λεξικής ανάλυσης (lexical analysis), αναγνώρισης προτύπων (pattern recognition), και εξόρυξης δεδομένων (data mining), καθώς επίσης και τεχνικές natural language processing (NLP). Οπότε μπορούμε να πούμε πως αυτά είναι τα διαθέσιμα εργαλεία που μπορούμε να χρησιμοποιήσουμε και για την εξόρυξη και ανάλυση απόψεων.

Στις περισσότερες έρευνες σχετικά με την εξόρυξη και ανάλυση απόψεων χρησιμοποιούνται κυρίως τεχνικές ανάλυσης κειμένου, ενώ λίγες είναι αυτές στις οποίες χρησιμοποιούνται τεχνικές NLP. Ο επιστημονικός τομέας του natural language processing, γενικά, ασχολείται με την αλληλεπίδραση μεταξύ ανθρώπου και υπολογιστή, και έχει ως στόχο να επιτρέψει στους υπολογιστές να αντλούν εννοιολογικές και σημασιολογικές πληροφορίες από τον άνθρωπο ή από είσοδο σε ανθρώπινη φυσική γλώσσα. Όπως είναι φανερό, ο τομέας του NLP μπορεί να δώσει τα απαραίτητα εργαλεία για την λύση του προβλήματος της εξόρυξης και ανάλυσης απόψεων, διότι ασχολείται με την επεξεργασία της ανθρώπινης γλώσσας και την εξαγωγής νοήματος από αυτή. Το πρόβλημα είναι πως πολλά από αυτά τα προβλήματα, μέχρι αυτή την στιγμή δεν έχουν αντιμετωπιστεί ικανοποιητικά. Οπότε οι κύριες τεχνικές για εξόρυξη και ανάλυση απόψεων βασίζονται σε τεχνικές ανάλυσης κειμένου, όπως τεχνικές ανάκτησης πληροφορίας (information retrieval), λεξικής ανάλυσης (lexical analysis), ανα-

γνώρισης προτύπων (pattern recognition), εξόρυξης δεδομένων (data mining), και άλλες υπολογιστικές τεχνικές, οι οποίες όμως δεν μπορούν παρά μόνο να παράγουν μία προσέγγιση της πλήρης λύσης του προβλήματος.

Οι εξαγωγή των οντοτήτων, των aspects, των opinion holders, και του χρόνου στον οποίο εκφράζονται οι απόψεις, μπορούμε να πούμε πως είναι κλασικά προβλήματα Named-entity recognition (NER). NER είναι μία διεργασία της εξαγωγής πληροφορίας, η οποία αναζητά και κατηγοριοποιεί ατομικά στοιχεία του κειμένου σε ονόματα ανθρώπων, οργανισμών, τοποθεσιών, εκφράσεων χρόνου, ποσοτήτων, και άλλων. Υπάρχουν δύο κύριες προσεγγίσεις για μία NER διαδικασία, ruled-based πορσεγγίσεις και στατιστικές προσεγγίσεις. Στις πιο πρόσφατες έρευνες χρησιμοποιούνται προσεγγίσεις οι οποίες βασίζονται σε κανόνες, αλλά γενικά δεν υπάρχει ικανοποιητική λύση για την αποδοτική πραγματοποίηση μιας NER διεργασίας. Επίσης, όσον αφορά τις διεργασίες της εξαγωγής των οντοτήτων και των aspects, βασικό πρόβλημα είναι η χρήση συνώνυμων λέξεων, από τους χρήστες, οι οποίοι χρησιμοποιούν διαφορετικά ονόματα για να αναφερθούν στις ίδιες οντότητες και aspects. Αυτό απαιτεί από τους αλγορίθμους να διαθέτουν μηχανισμούς αποσαφήνισης.

Για την διεργασία της αναγνώρισης της υποκειμενικότητας και της κατηγοριοποίησης της πολικότητας των απόψεων, οι οποίες εκφράζονται για οντότητες και aspects (ιδιότητες και υπομέρη οντοτήτων), σε απόψεις θετικής και αρνητικής πολικότητας, σημαντικά σημεία αποτελούν οι sentiment λέξεις. Η ύπαρξη sentiment λέξεων σε μία πρόταση, συνήθως υποδουλώνει την ύπαρξη κάποιας άποψης. Όμως δεν συμβαίνει πάντα αυτό. Ορισμένες απόψεις εκφράζονται χωρίς την χρήση sentiment λέξεων, ή μπορεί μία πρόταση να περιέχει sentiment λέξεις αλλά να μην εκφράζει άποψη. Πιο εξειδικευμένες τεχνικές για την αναγνώριση και κατηγοριοποίηση των απόψεων βασίζονται σε συντακτικά πρότυπα τα οποία χρησιμοποιούνται συχνά για την απόδοση των απόψεων στον ανθρώπινο λόγο. Επίσης υπάρχουν αρκετά είδη απόψεων όπως γενικές, συγκριτικές, άμεσες, έμμεσες. Το πλήθος των διαφορετικών μορφών με τους οποίους μπορεί να εκφραστεί μία άποψη καθιστά την αναγνώριση της υποκειμενικότητας (subjectivity detection) μία δύσκολη διαδικασία. Επίσης, η ύπαρξη sentiment λέξεων θετικής ή αρνητικής πολικότητας, σε μια πρόταση η οποία εκφράζει άποψη, δεν συνεπάγεται απαραίτητα πως η πολικότητας της άποψης θα είναι θετική ή αρνητική αντίστοιχα. Ο προσδιορισμός της πολικότητας των απόψεων, πρέπει να πραγματοποιηθεί με προσοχή, διότι πρέπει να ληφθούν υπόψιν σχήματα άρνησης, σαρκασμού, σχήματα

σύγκρισης και άλλα. Γενικά η ποικιλομορφία της ανθρώπινης έκφρασης αποτελεί έναν από τους βασικότερους παράγοντες οι οποίοι καθιστούν δύσκολη μία διαδικασία ανάλυσης απόψεων.

Ακόμα, ένα σημαντικό σημείο, το οποίο πρέπει να ληφθεί υπόψιν και καθιστά την διαδικασία της εξόρυξης και ανάλυσης απόψεων μία δύσκολη διαδικασία, είναι πως οι πολικότητες των απόψεων σχετικά με συγκεκριμένα aspects εξαρτώνται από το domain στο οποίο εργαζόμαστε. Σε ένα domain κάποιο aspect μπορεί να είναι επιθυμητό ενώ σε κάποιο άλλο όχι. Οπότε είναι απαραίτητο οι sentiment λέξεις και φράσεις, να αντιμετωπίζονται ως domain-dependent και ως context-dependent.

Στην διαδικασία της sentiment κατηγοριοποίησης ενός κειμένου απόψεων, δηλαδή στον καθορισμό της συνολικής sentiment πολικότητας του κειμένου, παίζει σημαντικό ρόλο η σωστή μοντελοποίηση της δομής του κειμένου. Ενώ το θέμα του κειμένου μπορεί να εξαχθεί ως η θεματική ενότητα στην οποία επικεντρώνεται η πλειοψηφία του περιεχομένου του κειμένου ανεξάρτητα από την διάταξη του περιεχομένου, όταν ζητείται να εξαχθεί η συνολική sentiment πολικότητα του κειμένου, η σειρά με την οποία παρουσιάζονται οι απόψεις μπορεί να οδηγήσει σε διαφορετική πολικότητα από ότι δείχνει να έχει αρχικά.

Σε μία διαδικασία sentiment ανάλυσης, πρέπει να προσπελαστούν οι απόψεις πολλών χρηστών διότι, λόγω της υποκειμενικής φύσης των απόψεων, αναλύοντας μόνο τις απόψεις ενός χρήστη ή ενός μικρού συνόλου χρηστών, δεν αναγνωρίζεται η συνολική εικόνα της κατανομής των απόψεων. Εξαιτίας αυτού είναι αναγκαίο να πραγματοποιηθεί κάποιο είδος σύνοψης των πληροφοριών των απόψεων. Επίσης είναι απαραίτητος ο προσδιορισμός της ποιότητας των απόψεων. Για παράδειγμα, όταν ο αριθμός των reviews σε ένα ιστότοπο είναι μεγάλος, είναι σημαντικό για τον αναγνώστη να εντοπίζει έυκολα και γρήγορα χρήσιμα reviews. Επίσης σε μία διαδικασία sentiment ανάλυσης σε review κείμενα, πληροφορίες οι οποίες προέρχονται από ποιοτικά reviews μπορούν να επισημαίνονται ως μεγαλύτερης βαρύτητας. Η βαθμολογία της ποιότητας ενός review κειμένου γίνεται με τον προσδιορισμό της ποιότητας, της εξυπηρετικότητας και της χρησιμότητας των περιεχομένων του κειμένου. Ακόμα πρέπει να ληφθεί υπόψιν και ο θόρυβος ο οποίος υπάρχει στα κείμενα τα οποία περιέχουν απόψεις, όπως για παράδειγμα τα κείμενα των blogs (blog noise). Τέτοια κείμενα συνήθως δεν είναι σωστά δομημένα και περιέχουν γραμματικά λάθη και αντιφα-

τικές πληροφορίες. Το γεγονός αυτό καθιστά την εφαρμογή τεχνικών, οι οποίες βασίζονται στο parsing και στο ταίριασμα προτύπων, μία δύσκολη διαδικασία.

Επίσης παράγοντες που καθιστούν δύσκολη την διαδικασία της εξόρυξης και ανάλυσης απόψεων, και σχετίζονται με την ποιότητα των απόψεων, είναι η διαφήμιση, και τα φαινόμενα του opinion spamming και sock puppeting. Λόγω της ραγδαίας αύξησης της χρήσης του διαδικτύου και των κοινωνικών δικτύων, ο αριθμός των διαθέσιμων απόψεων συνεχώς μεγαλώνει. Τα άτομα και οι οργανισμοί στηρίζουν συχνά τις αποφάσεις τους σε αυτές τις απόψεις. Τα άτομα αναζητούν απόψεις για την πραγματοποίηση κάποιας αγοράς ή για την επιλογή του σχήματος που πρέπει να στηρίξουν στις εκλογές. Οι οργανισμοί αναζητούν απόψεις χρηστών για τα προϊόντα τους, ώστε να βελτιώσουν το σχεδιασμό τους και την πολιτική προώθησής τους. Πολλές θετικές απόψεις για κάποια οντότητα συχνά σημαίνει έσοδα και φήμη, κάτι το οποίο δίνει ισχυρό κίνητρο στους ανθρώπους να ξεγελάσουν το σύστημα ανεβάζοντας ψεύτικες απόψεις ή reviews, για να προωθήσουν ή να δυσφημήσουν προϊόντα, υπηρεσίες, οργανισμούς, πρόσωπα, ακόμα και ιδέες, χωρίς να αποκαλύπτουν τις πραγματικές τους προθέσεις, ή τα άτομα ή τους οργανισμούς για τους οποίους κρυφά εργάζονται. Τέτοια άτομα ονομάζονται opinion spammers και οι ενέργειές τους ονομάζονται opinion spamming. Το opinion spamming για κοινωνικά και πολιτικά θέματα μπορεί να θεωρηθεί αρκετά επικίνδυνο, διότι μπορεί να στρεβλώσει απόψεις και να κινητοποιήσει τις μάζες σε παράνομες και ανήθικες κατευθύνσεις. Μπορούμε να πούμε πως όσο περισσότερο χρησιμοποιούνται οι απόψεις στα κοινωνικά δίκτυα, τόσο περισσότερο το opinion spamming θα γίνεται ανεξέλεγκτο και εξειδικευμένο.

Κλείνοντας αυτή την ενότητα, μπορούμε να πούμε πως αν και έχει πραγματοποιηθεί αρκετή έρευνα από την επιστημονική κοινότητα στον τομέα της εξόρυξης και επεξεργασίας, και επίσης έχουν δημιουργηθεί κάποια συστήματα, αυτή την στιγμή βρισκόμαστε αρκετά μακριά από την πλήρη λύση του προβλήματος. Κάθε υπο-πρόβλημα παραμένει μία πρόκληση. Όπως είχε αναφέρει κάποιος, περιγράφοντας τις δυσκολίες και την κατάσταση που επικρατεί αυτή την στιγμή στον τομέα του opinion mining, “*our sentiment analysis is as bad as everyone else's*”.

3 Τεχνικές Opinion Mining και Sentiment Analysis

Η εξόρυξη και ανάλυση απόψεων έχει ερευνηθεί κυρίως σε τρία επίπεδα. Το επίπεδο κειμένου (document level), το επίπεδο πρότασης (sentence level), και το επίπεδο οντότητας ή χαρακτηριστικού (entity or aspect level). Σε αυτήν την ενότητα παρουσιάζονται αναλυτικά τα τρία αυτά βασικά επίπεδα ανάλυσης απόψεων, καθώς επίσης και τεχνικές που συνήθως χρησιμοποιούνται για την πραγματοποίηση της διαδικασίας της εξόρυξης και ανάλυσης απόψεων σε κάθε ένα από αυτά. Επίσης αναλύονται σχετικές διεργασίες, όπως η δημιουργία ενός sentiment λεξικού, η ανίχνευση του opinion spamming, η αναζήτηση και ανάκτηση απόψεων, και η σύνοψη απόψεων, και παραθέτονται τεχνικές για την προσέγγισή τους.

3.1 Ανάλυση σε επίπεδο κειμένου

Στο επίπεδο κειμένου στόχος είναι η κατηγοριοποίηση ενός ολόκληρου εγγράφου σε κείμενο θετικής πολικότητας ή αρνητικής, ανάλογα με το αν εκφράζει θετικό ή αρνητικό συναίσθημα (document sentiment classification). Σε αυτό το επίπεδο υποθέτουμε πως το έγγραφο αναφέρεται σε μία και μόνο οντότητα. Έτσι η διαδικασία είναι ευκολότερη διότι δεν χρειάζεται να πραγματοποιηθεί εξαγωγή οντοτήτων, και είναι ασφαλής η υπόθεση πως όποια aspects υπάρχουν στο κείμενο, αναφέρονται στην μοναδική οντότητα του κειμένου. Ακόμα υποθέτουμε πως το κείμενο περιέχει απόψεις ενός μόνο συγγραφέα, opinion holder, οπότε επίσης δεν είναι απαραίτητη η αναγνώριση και εξαγωγή των opinion holders. Σε αυτήν την περίπτωση ο ορισμός της διαδικασίας της εξόρυξης και ανάλυσης απόψεων μετατρέπεται ως εξής:

Δοθέντος ενός κειμένου απόψεων d, να προσδιοριστεί η συνολική sentiment πολικότητα s που εκφράζει ο συγγραφέας (opinion holder) για την οντότητα, ως σύνολο (GENERAL). Σε αυτό το επίπεδο ανάλυσης η ζητούμενη πεντάδα είναι η (_, GENERAL, s, _, _), δηλαδή δεν ενδιαφερόμαστε για την εξαγωγή της οντότητας e, του opinion holder h, και του χρόνου που διατυπώθηκε η άποψη.

Στην πραγματικότητα, αν ένα opinion κείμενο αξιολογεί παραπάνω από μία οντότητες, τότε οι sentiment πολικότητες των απόψεων για κάθε μία από αυτές θα είναι διαφορε-

τικές. Για παράδειγμα, ο συγγραφέας μπορεί να εκφράζει θετικές απόψεις για κάποιες οντότητες. και αρνητικές για κάποιες άλλες, με στόχο να τις συγκρίνει μεταξύ τους. Σε αυτήν την περίπτωση δεν έχει νόημα να ανατεθεί μία συγκεκριμένη πολικότητα σε ολόκληρο το κείμενο, όπως και όταν στο κείμενο υπάρχουν παραπάνω του ενός opinion holders. Οι παραπάνω υποθέσεις ισχύουν για κείμενα, όπως τα review κέιμενα προϊόντων ή υπηρεσιών, διότι κάθε κείμενο αναφέρεται σε μία μόνο οντότητα, και επίσης υπάρχουν απόψεις ενός μόνο opinion holder.

Η sentiment κατηγοριοποίηση στο επίπεδο κειμένου, παρέχει μία συνολική άποψη για μία οντότητα, ένα θέμα, ή ένα γεγονός. Η sentiment ανάλυση στο επίπεδο κειμένου έχει μελετηθεί από ένα μεγάλο αριθμό ερευνητών. Ωστόσο, η ανάλυση σε αυτό το επίπεδο, παρουσιάζει αδυναμίες και ελλείψεις σε πολλές εφαρμογές του πραγματικού κόσμου. Για παράδειγμα, ο χρήστης συνήθως χρειάζεται επιπλέον πληροφορίες, εκτός της συνολικής πολικότητας του κειμένου, όπως ποια χαρακτηριστικά της οντότητας συγκεντρώνουν θετικές απόψεις και ποια αρνητικές. Επίσης, η sentiment κατηγοριοποίηση σε επίπεδο κειμένου δεν μπορεί να εφαρμοστεί για κείμενα τα οποία δεν είναι reviews όπως blog κείμενα και άρθρα ειδήσεων, διότι σε τέτοιου είδους κείμενα αναφέρονται συνήθως περισσότερες από μία οντότητες. Ακόμα αξίζει να αναφερθεί πως σε εφαρμογές του πραγματικού κόσμου, τα κείμενα στα οποία χρειάζεται περισσότερο να εφαρμοστεί sentiment ανάλυση είναι τα κείμενα των blogs και τα κείμενα συζητήσεων, διότι τα review κείμενα συνήθως συνοδεύονται ήδη με βαθμολογίες σε ένα μικρό σύνολο από σημαντικά aspects, όπως η μορφή 1-5 stars.

Στο επίπεδο ανάλυσης κειμένου το πρόβλημα της εξαγωγής και επεξεργασίας απόψεων παρουσιάζει δύο μορφές, ανάλογα με τη είδος του πεδίου τιμών της επιθυμητής εξόδου. Αν το πεδίο τιμών αποτελείται από διακριτές (κατηγορηματικές) τιμές, όπως θετικό ή αρνητικό, τότε είναι ένα πρόβλημα κατηγοριοποίησης. Αν το πεδίο τιμών αποτελείται από αριθμητικές τιμές ή βαθμολογίες κατάταξης σε ένα προκαθορισμένο εύρος, όπως για παράδειγμα 1-5 stars, τότε είναι ένα regression πρόβλημα. Στην στατιστική, regression analysis, είναι μία στατιστική διαδικασία για την εκτίμηση των σχέσεων μεταξύ μεταβλητών, και περιλαμβάνει τεχνικές για την μοντελοποίηση και την ανάλυση των μεταβλητών και των σχέσεών τους. Στον τομέα του machine learning χρησιμοποιείται συνήθως για να προσαρμοστεί μία συνάρτηση σε ένα σύνολο δεδομένων. Η απλούστερη μορφή regression, linear regression, χρησιμοποιεί το μο-

ντέλο μιας ευθείας γραμμής (y = mx + b) και καθορίζει τις κατάλληλες τιμές για τα m και b για να προβλέψει την τιμή y σε μία είσοδο x.

3.1.1 Εποπτευόμενες τεχνικές sentiment κατηγοριοποίησης

Για την λύση του προβλήματος της εξόρυξης και ανάλυσης απόψεων όταν αυτό έχει την μορφή ενός προβλήματος κατηγοριοποίησης, μπορεί να χρησιμοποιηθεί οποιαδήποτε εποπτευόμενη μέθοδος, όπως η κατηγοριοποίηση Naïve Bayes και η τεχνική Support Vector Machines (SVM) (Appendix II.). Αυτή η προσέγγιση χρησιμοποιήθηκε για πρώτη φορά στην έρευνα Pang et al. (2002). Όσον αφορά την απόδοση αυτών των δύο τεχνικών κατηγοριοποίησης, η τεχνική SVM τείνει να είναι καλύτερη από τη Naïve Bayes. Επίσης η χρήση unigrams ή bag of words (BOW) ως features (χαρακτηριστικά) στην διαδικασία της κατηγοριοποίησης, καθώς και η χρήση feature presence αντί για feature frequency βελτιώνει την απόδοση των δύο παραπάνω τεχνικών.

Όπως συμβαίνει σε όλες τις εφαρμογές της εποπτευόμενης μηχανικής μάθησης, το κλείδι για την κατηγοριοποίηση υποκειμενικής άποψης είναι η σύσταση ενός συνόλου αποτελεσματικών features (feature engineering). Δημοφιλή παραδείγματα features είναι τα ακόλουθα:

- λέξεις και η συχνότητά τους. Αυτά τα χαρακτηριστικά είναι μεμονωμένες λέξεις (unigrams) ή n-grams, με τις συχνότητες εμφάνισής τους. Είναι τα πιο συχνά χρησιμοποιούμενα στην παραδοσιακή κατηγοριοποίηση κειμένου με βάση το θέμα περιεχομένου (topic-based text classification). Επίσης μπορεί να χρησιμοποιείται, εκτός της συχνότητας, και η θέση των όρων στο κείμενο ή η μετρική βάρους TF-IDF από τον τομέα του Information Retrieval.
- part of speech (POS tags). Το μέρος του λόγου της λέξης. Λέξεις που αποτελούν διαφορετικά μέρη του λόγου πρέπει να διαχειρίζονται διαφορετικά. Για παράδειγμα, έχει δειχθεί πως τα τα επίθετα αποτελούν σημαντικούς δείκτες υποκειμενικού περιεχομένου. Επίσης ως features μπορούν να χρησιμοποιηθούν και n-grams από POS tags.
- Sentiment λέξεις και φράσεις. Οι υποκειμενικές λέξεις σε μία γλώσσα χρησιμοποιούνται για να εκφράσουν θετική ή αρνητική πολικότητα. Για παράδειγμα, λέξεις όπως “καλό”, “υπέροχο”, “καταπληκτικό” είναι sentiment λέξεις που εκφράζουν θετική πολικότητα, και λέξεις όπως “κακό”, “άσχημο”, είναι

sentiment λέξεις που εκφράζουν αρνητική πολικότητα. Οι περισσότερες sentiment λέξεις είναι επίθετα και επιρρήματα, όμως μπορεί να είναι και ουσιαστικά και ρήματα. Εκτός από λέξεις μπορούν να χρησιμοποιηθούν επίσης και συγκεκριμένες φράσεις οι οποίες διαφέρουν ανάλογα με την κοινωνική ομάδα και την χρονολογική περίοδο.

- Κανόνες έκφρασης απόψεων και συντακτική εξάρτηση λέξεων. Συχνά υπάρχουν συγκεκριμένοι κανόνες για την απόδοση απόψεων. Τέτοιοι κανόνες μπορούν να εκφραστούν ως ακολουθίες pos tags σε συγκεκριμένη σειρά.

Ακολουθούν διάφορες προσεγγίσεις και παρατηρήσεις από έρευνες στην sentiment κατηγοριοποίηση σε επίπεδο κειμένου.

Στην έρευνα Martineau and Finin (2009), παρουσιάστηκε η Delta TFIDF ως μία νέα τεχνική απόδοσης βάρους στις λέξεις των κειμένων, η οποία αποδίδει καλύτερα την σημαντικότητα των όρων ανάμεσα στα κείμενα, για την διαδικασία της sentiment κατηγοριοποίησης των κειμένων. Για την δημιουργία ενός συνόλου από features για την διαδικασία της εποπτευόμενης μηχανικής μάθησης χρησιμοποιούνται συχνά unigrams (a bag of words) και κάθε λέξη σχετίζεται με κάποια τιμή. Αυτή η τιμή είναι συνήθως η συχνότητα του όρου στο κείμενο σε συνδυασμό με την ανεστραμμένη συχνότητα όρου, η οποία αποδίδει το πόσο σπάνιος είναι ο όρος για όλη την συλλογή κειμένων, (TF-IDF). Όμως σε μία συλλογή κειμένων τα οποία περιέχουν sentiment πληροφορία, οι sentiment λέξεις τείνουν να χρησιμοποιούνται σε όλα σχεδόν τα κείμενα, οπότε τους αποδίδεται μικρές IDF βαθμολογίες. Επίσης τέτοιες λέξεις τείνουν να έχουν μικρή συχνότητα εμφάνισης (TF) σε ένα συγκεκριμένο κείμενο, διότι ο συγγραφέας συνήθως χρησιμοποιεί συνώνυμα για να αποδώσει καλή σύνταξη στο κείμενο ή για να μην κουράσει τους αναγνώστες. Οπότε οι sentiment λεξεις καταλήγουν να λαμβάνουν μικρές τιμές TF-IDF. Σε αυτή την έρευνα προτείνεται να αποδίδονται τιμές στους όρους του συνόλου χαρακτηριστικών ανάλογα με την προδιάθεση που παρουσιάζουν οι όροι στην συλλογή κειμένων. Η Delta TFIDF διαχωρίζει το σύνολο των χαρακτηριστικών σε θετικά και αρνητικά, και προωθεί την σημαντικότητα των όρων οι οποίοι είναι άνισα κατανεμημένοι ανάμεσα στις δύο κλάσεις, ενώ δίνει λιγότερο βάρος σε όρους οι οποίοι είναι ομοιόμορφα κατανεμημένοι. Με αυτή την μετρική η τιμή ενός ομοιόμορφα κατανεμημένου όρου στις δύο κλάσεις πολικότητας είναι μηδέν. Όσο πιο άνιση η κατανομή, τόσο πιο σημαντικό πρέπει να είναι ένα χαρακτηρι-

στικό. Χαρακτηριστικά τα οποία εμφανίζονται περισσότερο σε κείμενα αρνητικής από ότι σε θετικής πολικότητας έχουν θετική βαθμολογία, ενώ χαρακτηριστικά που εμφανίζονται περισσότερο σε κείμενα θετικής από ότι σε κείμενα αρνητικής έχουν αρνητική βαθμολογία. Σε μετρήσεις που έγιναν η Delta TFIDF χρησιμοποιήθηκε στην sentiment κατηγοριοποίηση κειμένων με Support Vector Machines και έδειξε πως αυξάνει την απόδοση της διαδικασίας.

Στην έρευνα Gamon (2004), πραγματοποιήθηκε κατηγοριοποίηση σε δεδομένα ανάδρασης πελατών (customer feedback), τα οποία είναι μικρότερα σε μήκος και περιέχουν περισσότερο θόρυβο από τα κανονικά reviews κείμενα. Δοκιμάστηκαν διάφορα μεγάλου μήκους features για την κατηγοριοποίηση με γραμμικά Support Vector Machines (SVMs). Ένα από τα αποτελέσματα της έρευνας ήταν πως η χρήση παρουσίας (precense) γλωσσολογικών features βασισμένα στην συγκεκριμένη δομή και τη σημασιολογική εξάρτηση βελτιώνει την απόδοση των κατηγοριοποιητών.

Στην έρευνα Bespalov et al. (2011), τονίζεται το γεγονός πως η χρήση unigrams (bag of words) διαχωρίζει τις λέξεις και με αυτό τον τρόπο καταστρέφει sentiment πληροφορία η οποία έχει αποτυπωθεί σε φράσεις. Για την αντιμετώπιση αυτού του φαινομένου, χρησιμοποιήθηκε η ιδέα να αντιμετωπιστούν τα κείμενα ως σύνολα από μικρές ακολουθίες λέξεων και η χρήση n-grams μεγάλης τάξης ($n \geq 3$), bag-of-ngram (BON). Στις έρευνες Pang et al. (2008) και Cui et al. (2006) έχει δειχθεί πως η χρήση 2-grams στην αναπαράσταση του "σάκου" των χαρακτηριστικών βελτιώνει την απόδοση της sentiment κατηγοριοποίησης. Ωστόσο, προσθέτοντας μεγαλύτερη τάξης n-grams, $n \geq 3$, στο μοντέλο BON μπορεί να έχει απαγορευτικό κόστος, διότι αυξάνει η διάσταση του BON διανύσματος εκθετικά καθώς αυξάνει το n. Για την αντιμετώπιση των προβλημάτων που παρουσιάζονται με την αύξηση της τάξης στην αναπαράσταση των n-grams, μπορούν να χρησιμοποιηθούν δύο μέθοδοι. Η πρώτη έχει να κάνει με τον καθορισμό συγκεκριμένων μοντέλων ακολουθιών σύμφωνα με κάποιους κανόνες, για παράδειγμα σημασιολογικούς, και την επιλογή ακολουθιών λέξεων με βάση αυτούς. Η δεύτερη αφορά την μείωση της διάστασης στην πραγματική τάξη προβλήματος, για παράδειγμα χρήση της τεχνικής της λανθάνουσας σημασιολογικής δεικτοδότησης (LSI) (Appendix IV.). Ωστόσο σε αυτή την έρευνα παρουσιάζεται μία νέα τεχνική, η "latent n-grams", για την αντιμετώπιση αυτού του προβλήματος.

Στην έρευνα Pang and Lee (2004), προτάθηκε να επισημαίνονται οι προτάσεις ενός κειμένου αρχικά σε υποκειμενικές και αντικειμενικές, και αφού απορριφθούν οι τε-

λευταίες, να εφαρμόζεται κάποιος κατηγοριοποιητής για την sentiment κατηγοριοποίηση review κειμένων. Δοκιμάστηκαν Naïve Bayes και Support Vector Machines κατηγοριοποιητές. Αυτή η διαδικασία μπορεί αποτρέψει τον sentiment κατηγοριοποιητή από το να λάβει υπόψιν του μη σχετικό ή παραπλανητικό κείμενο. Τα αποτελέσματα έδειξαν πως η χρήση τεχνικών subjectivity detection μπορεί να συμπιέσει τα κείμενα σε μικρότερα, αγνοώντας τις αντικειμενικές προτάσεις, και η απόδοση των κατηγοριοποιητών να είναι ίδια με την περίπτωση της ανάλυσης όλου του κειμένου των reviews. Επίσης σε αυτή την έρευνα εξετάστηκαν μέθοδοι για την αποδοτική εξαγωγή υποκειμενικού περιεχομένου. Προτάθηκε το μοντέλο cut-based subjectivity detector, το οποίο λαμβάνει ως είσοδο το σύνολο προτάσεων ενός κειμένου και αποφαίνεται για την υποκειμενική κατάσταση όλων των προτάσεων ταυτόχρονα, χρησιμοποιώντας πληροφορίες σχέσεων κατά αντικείμενο και κατά ζεύγη. Η διαδικασία δημιουργίας των cut-based subjectivity detectors είναι η ακόλουθη. Αρχικά, δημιουργείται ένας SVM υποκειμενικός κατηγοριοποιητής και εκπαιδεύεται με αντικειμενικές και υποκειμενικές προτάσεις από review κείμενα. Ύστερα τα reviews διασπώνται στις προτάσεις τους οι οποίες εισάγονται ως κόμβοι σε ένα γράφο. Στον γράφο επίσης εισάγονται δύο επιπλέον κόμβοι, ένας θετικός και ένα αρνητικός. Τοποθετούνται βάρη στις ακμές μεταξύ των κόμβων-προτάσεων και των δύο επιπλέον κόμβων (θετικού και αρνητικού) χρησιμοποιώντας την απόσταση της sentiment βαθμολογίας των προτάσεων από τα όρια βαθμολογίας του κατηγοριοποιητή (η πιθανότητες η πρόταση-κόμβος να είναι θετική και αρνητική αντίστοιχα). Τέλος τοποθετούνται βάρη στις ακμές μεταξύ ζευγών προτάσεων-κόμβων με βάση την απόσταση της sentiment βαθμολογίας των προτάσεων και ένα κατώφλι (αποδίδει το πόσο σημαντικό είναι να ανήκουν και οι δύο προτάσεις-κόμβοι στην ίδια κλάση). Αφού σχηματιστεί ο γράφος και αποδοθούν βάρη, χρησιμοποιούνται τα minimum cuts στον γράφο για να αφαιρεθεί το αντικειμενικό περιεχόμενο από τα review κείμενα. Τέλος εκπαιδεύεται ένας SVM κατηγοριοποιητής στο συμπιεσμένο κείμενο των reviews χρησιμοποιώντας ως feature, unigrams (bag of words).

Στην έρευνα Li et al. (2009) τονίζεται πως το υποκειμενικό κειμενικό περιεχόμενο συχνά αποτυπώνεται με γλωσσικούς μηχανισμούς και εκφράσεις εξαρτώμενες από το πεδίο συζήτησης (domain-specific phrases). Αυτό καθιστά την διαδικασία μάθησης αξιόπιστων μοντέλων δύσκολη και επιρρεπή σε λάθη. Ακόμα η εξέλιξη και οι αλλαγές στις κοινότητες χρηστών που παρατηρούνται με την πάροδο του χρόνο, επιφέρουν και

αλλαγές στις εκφράσεις που χρησιμοποιούνται για την απόδοση απόψεων, καθιστώντας τα μοντέλα αναποτελεσματικά. Προτάθηκε μέθοδος δημιουργίας μοντέλου το οποίο “μαθαίνει” χρησιμοποιώντας sentiment λέξεις ανεξάρτητες από το domain (domain-independent) σε συνεργασία με domain-dependent κείμενα. Το μοντέλο αυτό βασίζεται στην παραγοντοποίηση του μητρώου όρων-κειμένων σε τρία μητρώα (non-negative tri-factorization) (Appendix VI.), η οποία μπορεί να πραγματοποιηθεί με απλούς κανόνες ανανέωσης (dual-supervision in matrix factorization models).

Στην έρευνα Becker and Aharonson (2010) τονίζεται πως στην διαδικασία της εξόρυξης πολικότητας από κειμενικές πηγές, πρέπει να δοθεί επίσης προσοχή στην γλωσσολογία καθώς υπάρχουν ιδιότητες που έχουν παρατηρηθεί, και μπορούν να χρησιμοποιηθούν αποκαλύψουν τα σημεία, μέσα σε ένα κείμενο, στα οποία εκδηλώνεται συνήθως η πολικότητα. Χρησιμοποιώντας μία παρατήρηση από τον τομέα της ψυχολογίας γνωστή ως “recency effect”, προτείνεται πως αρκεί να εξετάζονται μόνο οι τελευταίες προτάσεις αντί για όλο το μήκος του κειμένου για την κατηγοριοποίηση του με βάση την πολικότητά. Ωστόσο, οι πειραματικές μετρήσεις έδειξαν πως υπάρχει μικρή διαφορά στην απόδοση της διαδικασίας της βαθμολογίας review κειμένων όταν λαμβάνονται υπόψιν μόνο οι τελευταίες προτάσεις και όταν χρησιμοποιείται όλο το κείμενο.

3.1.2 Μη-εποπτευόμενες τεχνικές sentiment κατηγοριοποίησης

Για την λύση του προβλήματος της εξόρυξης και ανάλυσης απόψεων όταν αυτό έχει την μορφή ενός προβλήματος κατηγοριοποίησης, δεν χρησιμοποιείται κάποιο μοντέλο το οποίο έχει δημιουργηθεί με την διαδικασία της μάθησης όπως στην εποπτευόμενη, αλλά απλούστερες μέθοδοι οι οποίες εντοπίζουν στο κείμενο τις υποκειμενικές πληροφορίες και αποδίδουν σε αυτό, με κάποια τεχνική, μία τελική βαθμολογία, κατηγοριοποιώντας το ως θετικό ή αρνητικό. Οι περισσότερες τεχνικές για τον εντοπισμό υποκειμενικού περιεχομένου, σε μη-εποπτευόμενες διαδικασίες, βασίζονται στην χρήση συντακτικών patterns (syntactic patterns based) και sentiment λεξικού (lexicon based).

Ένα παράδειγμα μη εποπτευόμενης μεθόδου για την πραγματοποίηση της sentiment κατηγοριοποίησης εγγράφων η οποία βασίζεται σε συντακτικά πρότυπα είναι αυτή που παρουσιάζεται στην έρευνα Turney (2002). Αυτή η μέθοδος πραγματοποιεί sentiment ανάλυση βασιζόμενη σε συντακτικά πρότυπα τα οποία χρησιμοποιούνται

συνήθως για να εκφραστεί άποψη. Τα συντακτικά πρότυπα αναγνωρίζονται μέσα σε μία πρόταση με βάση τα pos tags των όρων, τα οποία εμφανίζονται σε συγκεκριμένη σειρά. Ενδεικτικά, συντακτικά πρότυπα για την εξαγωγή φράσεων μήκους δύο λέξεων (two-word phrases) είναι τα ακόλουθα:

First Word	**Second Word**	**Third Word(not extracted)**
JJ	NN/NNS	anything
RB/RBR/RBS	JJ	not(NN/NNS)
JJ	JJ	not(NN/NNS)
NN/NNS	JJ	not(NN/NNS)
RB/RBR/RBS	VB/VBD/VBN/VBG	anything

Ο αλγόριθμος είναι ο εξής:

Βήμα 1: επεξεργαζόμαστε τα review κείμενα πρόταση προς πρόταση και αποδίδουμε στις λέξεις POS tags. Σε κάθε μία πρόταση προσπαθούμε να αναγνωρίσουμε κάποιο από τα πρότυπα και να εξάγουμε τις 2 πρώτες λέξεις της φράσης στην οποία ταίριαξε το πρότυπο.

Βήμα 2: για κάθε μία από τις εξαγόμενες εκφράσεις υπολογίζουμε την πολικότητα της υποκειμενικότητας (sentiment orientation, SO) χρησιμοποιώντας την μετρική point wise mutual inforamtion (PMI):

$$PMI(term1, term2) = \log_2 \frac{Pr(term1 \wedge term2)}{Pr(term1) * Pr(term2)}$$

Η μετρική PMI μετρά την στατιστική εξάρτηση μεταξύ των δύο όρων. Το γινόμενο $Pr(term1 \wedge term2)$ είναι η πιθανότητα συνεμφάνισης των όρων term1 και term2, και το γινόμενο $Pr(term1) * Pr(term2)$ είναι η πιθανότητα συνεμφάνισης των δύο όρων αν είναι στατιστικά ανεξάρτητοι. Το sentiment orientation (SO) της φράσης υπολογίζεται με βάση την σχέση της με την θετική λέξη αναφοράς “excellent” και την αρ-

νητική λέξη αναφοράς “poor”. Οι πιθανότητες υπολογίζονται εκτελώντας ερωτήματα σε μία μηχανή αναζήτησης και συλλέγοντας τον αριθμό των hits (αριθμός των σχετικών κειμένων).

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

Οπότε:

$$SO(phrase) = \log_2 \frac{hits(phrase\ NEAR\ 'excellent')\ hits('poor')}{hits(phrase\ NEAR\ 'poor')\ hits('excellent')}$$

Βήμα 3: τέλος υπολογίζεται ο μέσος όρος των SO όλων των φράσεων, ως η βαθμολογία της συνολικής πολικότητας της υποκειμενικότητας του κειμένου.

Επίσης υπάρχουν μη εποπτευόμενες μέθοδοι για την πραγματοποίηση της sentiment κατηγοριοποίησης εγγράφων οι οποίες κάνουν χρήση sentiment λεξικών (lexicon-based). Τα sentiment λεξικά αποτελούνται από συνήθης sentiment λέξεις και φράσεις με τις σχετικές πολικότητές τους και το βαθμό της πόλωσής τους. Για την sentiment κατηγοριοποίηση ενός κειμένου, εντοπίζονται οι sentiment λέξεις ή φράσεις του λεξικού στο κείμενο και χρησιμοποιούνται οι βαθμολογίες τους για την παραγωγή με κάποιο τρόπο μιας βαθμολογίας, θετικής ή αρνητικής, για το κείμενο (Taboada et al., 2011).

3.1.3 Πρόβλεψη sentiment βαθμολογιών

Για την λύση του προβλήματος της εξόρυξης και ανάλυσης απόψεων όταν αυτό έχει την μορφή ενός προβλήματος regression, μπορεί να χρησιμοποιηθεί κάποια τεχνική regression, αν και υπάρχουν έρευνες οι οποίες ακολουθούν και άλλες προσεγγίσεις. Στην στατιστική, regression analysis, είναι μία στατιστική διαδικασία για την εκτίμηση των σχέσων μεταξύ μεταβλητών, και περιλαμβάνει τεχνικές για την μοντελοποίηση και την ανάλυση των μεταβλητών και των σχέσεών τους. Στον τομέα του machine learning χρησιμοποιείται συνήθως για να προσαρμοστεί μία συνάρτηση σε ένα σύνο-

λο δεδομένων. Η απλούστερη μορφή regression, linear regression, χρησιμοποιεί το μοντέλο μιας ευθείας γραμμής ($y = mx + b$) και καθορίζει τις κατάλληλες τιμές για τα m και b για να προβλέψει την τιμή y σε μία είσοδο x. Αυτή η προσέγγιση απαιτεί την δημιουργία και εκπαίδευση ενός μοντέλου με βάση κάποιο σύνολο δεδομένων εκπαίδευσης. Ύστερα το μοντέλο μπορεί να χρησιμοποιηθεί για την πρόβλεψη σε νέα σύνολα δεδομένων.

Ένα απλό παράδειγμα πάνω σε αυτή την προσέγγιση είναι το ακόλουθο, όπου χρησιμοποιείται το PRank μοντέλο Crammer et al. (2001) για την πρόβλεψη μιας τελικής βαθμολογίας για ένα σύνολο από review κείμενα. Σε αυτήν την διαδικασία, κάθε review κείμενο, αναπαριστάται ως ένα feature vector $x \in Rn$. Ο χώρος χαρακτηριστικών μπορεί να είναι unigrams, n-grams, POS-tag patterns, ή οποιοσδήποτε άλλος χώρος. Το μοντέλο προβλέπει ένα $y \in \{1, ..., k\}$ για κάθε $x \in Rn$, η οποία είναι η βαθμολογία του κειμένου (k πιθανές βαθμολογίες). Πιο συγκεκριμένα, το μοντέλο αποθηκεύει ένα διάνυσμα βάρους $w \in Rn$ και ένα διάνυσμα από όρια $b0 = -\infty \leq b1 \leq ... \leq bk-1 \leq bk = \infty$ τα οποία διαιρούν τη γραμμή των πραγματικών αριθμών σε k όρια, ένα για κάθε πιθανή βαθμολογία. Για την πρόβλεψη της βαθμολογίας του review κειμένου, αρχικά το μοντέλο δίνει ένα score για κάθε είσοδο χρησιμοποιώντας το διάνυσμα βάρους: $score(x) = w * x$, όπου το σύμβολο "*" συμβολίζει το εσωτερικό γινόμενο μεταξύ των δύο διανυσμάτων. Ύστερα, τοποθετεί το score(x) στον άξονα των πραγματικών αριθμών και επιστρέφει το κατάλληλο rank σύμφωνα με τα όρια bk. Συγκεκριμένα, επιστρέφει ένα rank r τέτοιο ώστε $br-1 \leq score(x) < br$. Το μοντέλο αυτό εκπαιδεύεται χρησιμοποιώντας τον αλγόριθμο Prank (Perceptron Ranking algorithm) (Appendix I.), ο οποίος αντιδρά σε λανθασμένες προβλέψεις κατά την διάρκεια της εκπαίδευσης, ανανεώνοντας τα διανύσματα βάρους (w) και ορίων (b).

Στην έρευνα Snyder and Barzilay (2007), εξετάζεται το πρόβλημα της πρόβλεψης των βαθμολογιών για κάθε aspect ξεχωριστά. Μία απλή προσέγγιση θα ήταν να χρησιμοποιηθούν m ανεξάρτητα regression μοντέλα, όπως το PRank που παρουσιάζεται πιο πάνω, ένα για κάθε aspect. Όμως αυτή η προσέγγιση δεν λαμβάνει υπόψιν της, τις εξαρτήσεις μεταξύ των βαθμολογιών που δίνουν οι χρήστες στα διάφορα χαρακτηριστικά. Η γνώση αυτών των εξαρτήσεων είναι σημαντική για την απόδοση της πρόβλεψης. Προτείνεται ένα μοντέλο το οποίο ελαχιστοποιεί την δυσαρέσκεια (grief) των επιμέρους μερών με μια κοινή πρόβλεψη (Good Grief decoding procedure). Το μοντέλο Good Grief αποτελείται από ένα μοντέλο βαθμολογίας για κάθε ένα από τα

aspects, και ένα μοντέλο συμφωνίας το οποίο προβλέπει πότε οι βαθμολογίες των χαρακτηριστικών είναι ίδιες ή όχι. Η διαδικασία μοντελοποιεί τις εξαρτήσεις μεταξύ διάφορων βαθμολογιών, μέσω μιας σχέσης συμφωνίας (aggrement relation). Αυτή η σχέση συμφωνίας καταγράφει πότε ο χρήστης προτιμά το ίδιο όλα τα aspects ή αποδίδει σε αυτά διαφορετικές βαθμολογίες.

Στην έρευνα Long et al. (2010) χρησιμοποιήθηκε ένας Baysian network κατηγοριοποιητής για την πρόβλεψη των βαθμολογιών των χαρακτηριστικών (aspects) σε review κείμενα. Για καλύτερη ακρίβεια, αντί να γίνει πρόβλεψη για όλο το σύνολο των reviews, προτείνεται να προβλεφθούν μόνο οι βαθμολογίες των χαρακτηριστικών σε ένα υποσύνολο των reviews τα οποία αξιολογούνται συνολικά. Οι υπολογισμοί για αυτά τα reviews θα είναι πιο ακριβείς. Η επιλογή των reviews έγινε χρησιμοποιώντας μία μετρική πληροφορίας βασισμένη στην Kolmogorov complexity.

Στην έρευνα Pang and Lee (2005) πραγματοποιήθηκαν πειράματα με SVM regression, SVM multiclass classification με εφαρμογή της στρατηγικής one-vs.-all (OVA), και μία meta-learning μέθοδο με την ονομασία metric labeling. Τα πειράματα έδειξαν πως η κατηγοριοποίηση η οποία ήταν βασισμένη στο OVA έχει χειρότερη απόδοση από τις άλλες δύο τεχνικές, κάτι το οποίο σύμφωνα με τους συγγραφείς της έρευνας, εξηγείται από το γεγονός πως δεν είναι καλή τακτική να αντιμετωπίζονται οι rating βαθμολογίες ως κατηγορικές τιμές.

Στην έρευνα Qu et al. (2010) προτείνεται η αναπαράσταση των κειμένων ως bag-of-opinions, αντί για την αναπαράσταση bag-of-words, με βάση το γεγονός πως οι απόψεις εμφανίζονται συχνά ως n-grams patterns. Προτείνεται κάθε άποψη να αναπαρίσταται ως μία τριάδα, η οποία να αποτελείται από ένα sentiment όρο, ένα modifier όρο, και ένα negator όρο. Για παράδειγμα “not very good”, όπου “good” είναι ο sentiment όρος, “very” είναι ο modifier, και “not” είναι ο negator. Τονίζεται πως η χρήση των modifier και negator όρων για την κατηγοριοποίηση σε δύο κλάσεις (θετική και αρνητική) μπορεί να μην είναι και τόσο σημαντική, αλλά σε μία διαδικασία πρόβλεψης βαθμολογιών, αυτοί οι όροι πρέπει να λαμβάνονται υπόψιν. Δημιουργείται μία regression μέθοδος για την διαδικασία της μάθησης των sentiment βαθμολογιών των απόψεων από μία ήδη βαθμολογημένη domain-independent (ανεξάρτητη θεματικής ενότητας-domain) συλλογή κειμένων. Η βασική ιδέα πίσω από την διαδικασία της μάθησης είναι η χρήση ενός λεξικού απόψεων (opinion lexicon), και των βαθμολογιών των review κειμένων. Για την μεταφορά του regression μοντέλου που δημιουργή-

θηκε σε ένα νέο domain (θεματική ενότητα), ο αλγόριθμος εξάγει ένα σύνολο στατιστικών από τις βαθμολογίες των απόψεων, τα οποία χρησιμοποιεί ως επιπλέον χαρακτηριστικά στην διαδικασία της πρόβλεψης.

Στην έρευνα Liu and Seneff (2009) προτείνεται μία μέθοδος η οποία εξάγει φράσεις της μορφής (επίρρημα-επίθετο-ουσιαστικό). Οι sentiment βαθμολογίες αποδίδονται με μία ευρετική μέθοδο η οποία υπολογίζει την συνεισφορά των επιρρημάτων, των επιθέτων και των αρνήσεων, με βάση τις βαθμολογίες των review κειμένων στα οποία αυτές εμφανίζονται.

3.1.4 Cross-domain sentiment κατηγοριοποίηση

Η κατηγοριοποίηση υποκειμενικού περιεχομένου είναι αρκετά ευαίσθητη στην θεματική ενότητα που αυτή πραγματοποιείται. Όταν χρησιμοποιείται εποπτευόμενη μάθηση, ο κατηγοριοποιητής που έχει εκπαιδευτεί με δεδομένα από μία συγκεκριμένη θεματική ενότητα, αδυνατεί να αποδώσει το ίδιο καλά και σε άλλα domains. Αυτό οφείλεται στο γεγονός πως οι λέξεις και γενικά η “γλώσσα”, που χρησιμοποιούνται για να εκφραστούν απόψεις σε διαφορετικές θεματικές ενότητες, είναι διαφορετικές. Επίσης, κάποιες φορές η ίδια λέξη μπορεί σε ένα domain να υποδουλώνει θετική πολικότητα ενώ σε ένα άλλο domain να έχει αρνητική σημασία. Οπότε δημιουργείται η ανάγκη για τον ορισμό cross-domain σχημάτων τα οποία να προσαρμόζονται στις θεματικές αλλαγές.

Προηγούμενες έρευνες πάνω σε αυτό το πρόβλημα βασίζονται κυρίως στην δημιουργία καρηγοριοποιητών χρησιμοποιώντας δεδομένα από δύο domains (target και source), ή στην δημιουργία domain independent κατηγοριοποιητών. Στην πρώτη περίπτωση εκπαιδεύονται κατηγοριοποιητές με σύνολα κειμένων και από τα δύο domains (target και source), ή λειτουργούν με features τα οποία έχουν επιλεγεί και από τα δύο domains με κάποια τεχνική, ή χρησιμοποιούνται semi-supervised τεχνικές. Στην περίπτωση των domain independent κατηγοριοποιητών χρησιμοποιούνται για την μάθηση σύνολα κειμένων από πολλά domains, ή επιλέγεται ένας domain independent χώρο χαρακτηριστικών με χαρακτηριστικά από πολλά domains.
Πιο συγκεκριμένα:

Στην έρευνα Aue and Gamon (2005), προτείνεται να μεταφέρονται οι sentiment κατηγοριοποιητές σε νέα domains. Έγιναν πειράματα με τέσσερις στρατηγικές:

1. εκπαίδευση του κατηγοριοποιητή σε ένα σύνολο από reviews τα οποία προέρχονται από πολλά διαφορετικά domains.
2. εκπαίδευση του κατηγοριοποιητή σε ένα σύνολο από reviews από διαφορετικά domains, αλλά με περιορισμό του συνόλου των χαρακτηριστικών σε αυτά που παρατηρούνται στο νέο domain.
3. χρήση κατηγοριοποιητών από διαφορετικά domains.
4. συνδυασμός μικρών συνόλων labeled δεδομένων με μεγάλα σύνολα unlabeled δεδομένων, όπως ακριβώς σε ένα σχήμα semi-supervised learning.

Για τα τρία πρώτα σχήματα χρησιμοποιήθηκαν Support Vector Machines, ενώ για το το τέταρτο semi-supervised μάθηση EM (Nigam et al., 2000). Τα πειράματά τους έδειξαν πως η τέταρτη στρατηγική είχε την καλύτερη απόδοση.

Στην έρευνα Yang et al. (2006), προτείνεται μία απλή στρατηγική για την μετάδοση της μάθησης στην sentiment κατηγοριοποίηση, η οποία βασίζεται στην επιλογή συγκεκριμένων χαρακτηριστικών. Σε αυτή την μέθοδο αρχικά χρησιμοποιούνται δύο labeled σύνολα δεδομένων εκπαίδευσης από δύο διαφορετικά domains, για την επιλογή χαρακτηριστικών τα οποία είναι σημαντικά και για τα δύο domains. Αυτά τα χαρακτηριστικά θεωρούνται πως είναι ανεξάρτητα θεματικής ενότητας (domain independent features). Ο κατηγοριοποιητής, ο οποίος δημιουργείται με βάση αυτά τα χαρακτηριστικά, θεωρείται domain independent και μπορεί να εφαρμοστεί σε οποιοδήποτε domain.

Στην έρευνα Tan et al. (2007), προτείνεται μία απλή μέθοδος, στην οποία αρχικά εκπαιδεύεται ένας κατηγοριοποιητής χρησιμοποιώντας labeled δεδομένα από το source domain, και μετά χρησιμοποιείται για την σήμανση μερικών δεδομένων στο domain στόχο. Με βάση αυτά τα επιλεγμένα δεδομένα εκπαιδεύεται ένας νέος κατηγοριοποιητής ο οποίος μπορεί να χρησιμοποιηθεί στο νέο domain.

Στην έρευνα Blitzer et al. (2007), χρησιμοποιείται μία μέθοδος η οποία ονομάζεται structural correspondence learning (SCL) για προσαρμογή στο domain. Δοθέντος ενός συνόλου με labeled reviews από το source domain και ενός συνόλου unlabeled reviews από τα source και target domains, το SCL αρχικά επιλέγει ένα σύνολο από m χαρακτηριστικά τα οποία εμφανίζονται πιο συχνά και στα δύο domains. Αυτά τα χαρακτηριστικά ονομάζονται χαρακτηριστικά οδηγοί “pivot features” και αναπαριστούν τον κοινό χώρο χαρακτηριστικών των δύο domains. Ύστερα το SCL υπολογίζει τις συσχετίσεις κάθε pivot χαρακτηριστικού με κάθε non-pivot χαρακτηριστικού και στα δύο

domains. Αυτό παράγει ένα μητρώο συσχετίσεων W όπου η γραμμή i είναι το διάνυσμα των τιμών των συσχετίσεων των non-pivot χαρακτηριστικών με το i χαρακτηριστικό οδηγό. Μετά από αυτό, εφαρμόζεται διάσπαση ιδιαζουσών τιμών για να υπολογιστεί μία προσέγγιση μικρότερης τάξης (k) του μητρώου W (q). Το τελικό σύνολο χαρακτηριστικών παράγεται από το αρχικό σύνολο χαρακτηριστικών συνδυασμένο με το q (k feature set). Ο κατηγοριοποιητής δημιουργείται με βάση αυτό το σύνολο των τελικών k χαρακτηριστικών, και μπορεί να χρησιμοποιηθεί και στα δύο domains.

Στην έρευνα He et al. (2011) χρησιμοποιήθηκε topic modeling για την αναγνώριση των opinion topics στα δύο domains (source και target). Τα topics τα οποία αναφέρονταν και στα δύο domains χρησιμοποιήθηκαν για να εμπλουτιστεί το αρχικό σύνολο των χαρακτηριστικών της sentiment κατηγοριοποίησης.

Στην έρευνα Gao and Li (2011), χρησιμοποιήθηκε topic modeling για να βρεθεί ένας σημασιολογικός χώρος με βάση τις αντιστοιχίες όρων και τις συνυπάρξεις όρων στις δύο θεματικές ενότητες. Ύστερα αυτός ο χώρος χρησιμοποιήθηκε για να εκπαιδευτεί ένας κατηγοριοποιητής για το target domain.

Στην έρευνα Bollegala et al. (2011) προτείνεται μία μέθοδος για την αυτόματη δημιουργία ενός sentiment “θησαυρού” (thesaurus) χρησιμοποιώντας labeled και unlabeled δεδομένα από πολλά domains, με στόχο την ανακάλυψη συσχετίσεων μεταξύ των λέξεων οι οποίες χρησιμοποιούνται για να εκφράσουν συναίσθημα της ίδιας πολικότητας σε διαφορετικά domains. Ο θησαυρός αυτός χρησιμοποιείται για να την μάθηση ενός δυαδικού sentiment κατηγοριοποιητή.

Στην έρευνα Wu et al. (2009), προτείνεται μία graph-based μέθοδος, η οποία χρησιμοποιεί την ιδέα του label propagation σε ένα γράφο ομοιότητας για την πραγματοποίηση της μεταφοράς μεταξύ των domains. Στο γράφο, κάθε έγγραφο είναι ένας κόμβος και κάθε σύνδεσμος μεταξύ δύο κόμβων είναι ένα βάρος το οποίο έχει υπολογιστεί χρησιμοποιώντας της ομοιότητα συνημιτόνου των δύο εγγράφων. Επίσης, κάθε έγγραφο στο παλιό domain έχει μία ετικέτα με βαθμολογία +1 (θετική) ή -1 (αρνητική) και κάθε έγγραφο στο νέο domain έχει μία ετικέτα με βαθμολογία η οποία έχει αποδοθεί από ένα sentiment κατηγοριοποιητή ο οποίος έχει εκπαιδευτεί στο παλιό domain. Ο αλγόριθμος επαναληπτικά ανανεώνει τις βαθμολογίες στις ετικέτες για κάθε κόμβο-έγγραφο του νέου domain i βρίσκοντας τους k κοντινότερους γείτονες από το παλιό domain και τους k κοντινότερους γείτονες από το νέο domain. Ένας γραμμικός συνδυασμός των βαθμολογιών των ετικετών και των βαρών των συνδέσμων χρησιμο-

ποιούνται για να αποδοθεί μια νέα βαθμολογία στον κόμβο i. Η επαναληπτική διαδικασία σταματά όταν όταν οι βαθμολογίες των ετικετών converge. Η sentiment πολικότητες των εγγράφων του νέου domain καθορίζονται από τις βαθμολογίες των ετικετών τους.

Στην έρευνα Xia and Zong (2011) τονίζεται πως μεταξύ διαφορετικών domains υπάρχουν κάποια POS tag χαρακτηριστικά τα οποία είναι συχνά domains-dependent, ενώ υπάρχουν κάποια άλλα τα οποία είναι domain-independent ή domain-free. Με βάση αυτή την παρατήρηση, προτείνεται ένα POS-based μοντέλο για την ενσωμάτωση χαρακτηριστικών με διαφορετικά είδη POS tags ώστε να βελτιωθεί η απόδοση της κατηγοριοποίησης.

3.1.5 Ποιότητα των review κειμένων

Ο προσδιορισμός της ποιότητας ενός review κειμένου είναι σημαντικός για τα συστήματα που διαχειρίζονται reviews χρηστών σχετικά με προϊόντα ή υπηρεσίες, αλλά και για τους αλγορίθμους εξόρυξης πληροφορίας και ειδικά για τις διαδικασίες εξόρυξης και ανάλυσης απόψεων. Όταν ο αριθμός των reviews σε ένα ιστότοπο είναι μεγάλος, είναι σημαντικό για τον αναγνώστη να εντοπίζει εύκολα και γρήγορα χρήσιμα reviews. Επίσης σε μία διαδικασία sentiment ανάλυσης σε review κείμενα, πληροφορίες οι οποίες προέρχονται από ποιοτικά reviews μπορούν να επισημαίνονται ως μεγαλύτερης βαρύτητας.

Η βαθμολογία της ποιότητας ενός review κειμένου γίνεται με τον προσδιορισμό της ποιότητας, της εξυπηρετικότητας και της χρησιμότητας των περιεχομένων του κειμένου. Η κατάταξη των reviews με βάση την ποιότητά τους, προέρχεται από το γεγονός πως είναι επιθυμητό να παρέχονται στον χρήστη τα πιο χρήσιμα reviews, όπως ακριβώς του παρέχονται και οι περισσότερο σχετικές πληροφορίες σε ένα ερώτημα ανάκτησης πληροφορίας. Σε ένα online σύστημα το οποίο διαθέτει μεγάλο αριθμό από reviews, οι διαχειριστές εφαρμόζουν διάφορες τεχνικές για να επιτύχουν το παραπάνω. Η πιο συνηθισμένη τεχνική είναι να διαμορφώνουν μία βαθμολογία εξυπηρετικότητας ή ποιότητας για κάθε review, ζητώντας από τους χρήστες να παρέχουν feedback. Για παράδειγμα στον ιστότοπο amazon, ο αναγνώστης μπορεί να παρέχει feedback δηλώνοντας αν ήταν ή όχι το review που μόλις διάβασε χρήσιμο (“Was the review helpful to you?”). Τα αποτελέσματα από αυτή την διαδικασία συγκεντρώνονται και κάθε review επισημαίνεται με την βαθμολογία του (“9 of 10 people found the following review helpful.”). Αν και σε όλα σχεδόν τα συστήματα αυτής της κατηγορί-

ας χρησιμοποιείται αυτή η τεχνική, η αναζήτηση μίας αυτόματης διαδικασίας έχει ακόμα νόημα, διότι με την τεχνική του feedback για να προσδιοριστεί η ποιότητα ενός review πρέπει να συγκεντρωθεί ένας καλός αριθμός από βαθμολογίες χρηστών, και συχνά παρατηρείται το γεγονός νέα reviews να μην έχουν την βαθμολογία που πρέπει.

Ο προσδιορισμός της ποιότητας ενός review κειμένου προσεγγίζεται συνήθως ως ένα regression πρόβλημα. Το μοντέλο το οποίο σχηματίζεται, καταχωρεί μία βαθμολογία με βάση την ποιότητα σε κάθε review κείμενο, η οποία μπορεί να χρησιμοποιηθεί για την κατάταξη των review κειμένων. Τα δεδομένα εκπαίδευσης απαρτίζονται από τα το feedback των χρηστών για τα review κείμενα, τα οποία δίνονται με την μορφή που περιγράφεται πιο πάνω ή παρόμοια. Τα σύνολα χαρακτηριστικών που χρησιμοποιούνται για την δημιουργία των μοντέλων μπορούν να είναι διάφορων ειδών. Για παράδειγμα στην έρευνα Kim et al. (2006) χρησιμοποιήθηκαν Support Vector Machines (SVM) και regression για την λύση του προβλήματος του προσδιορισμού της ποιότητας review κειμένων. Τα σύνολα των χαρακτηριστικών αποτελούνταν από:

- χαρακτηριστικά δομής: μέγεθος του review κειμένου, αριθμός των προτάσεων, ποσοστό των προτάσεων ερώτησης και των επιφωνημάτων, αριθμός των HTML bold tags <b> και αλλαγής γραμμής
- λεξικά χαρακτηριστικά: unigrams και bigrams με TF_IDF βάρη
- συντακτικά χαρακτηριστικά: ποσοστό των λέξεων οι οποίες είναι ουσιαστικά και ρήματα, ποσοστό των λέξεων οι οποίες είναι ρήματα πρώτου προσώπου, ποσοστό των λέξεων οι οποίες είναι επίθετα ή επιρρήματα
- σημασιολογικά χαρακτηριστικά: aspects των οντοτήτων, sentiment λέξεις
- meta-data χαρακτηριστικά: βαθμολογίες του review κειμένου

Επίσης στην έρευνα Zhang and Varadarajan (2006), το πρόβλημα του προσδιορισμού της ποιότητας review κειμένων προσεγγίζεται πάλι ως πρόβλημα regression, και τα χαρακτηριστικά που χρησιμοποιήθηκαν είναι παρόμοια με αυτά που αναφέρονται παραπάνω.

Στην έρευνα Ghose and Ipeirotis (2007, 2010), χρησιμοποιήθηκαν επιπλέον σύνολα χαρακτηριστικών, τα οποία έχουν να κάνουν με πληροφορίες για τον συγγραφέα του review κειμένου. Αυτά τα χαρακτηριστικά είναι το ιστορικό της εξυπηρετικότητας των

review κειμένων που έχει γράψει ο συγγραφέας και ένα σύνολο από χαρακτηριστικά αναγνωσιμότητας, για παράδειγμα ορθογραφικά λάθη.

Στην έρευνα Lu et al. (2010), το πρόβλημα του προσδιορισμού της ποιότητας ενός review κειμένου προσεγγίζεται με διαφορετικό τρόπο. Ερευνάται πως το κοινωνικό πλαίσιο των reviewers μπορεί να βοηθήσει στην βελτίωση της ακρίβειας ενός text-based συστήματος πρόβλεψης της ποιότητας review κειμένων. Τονίζεται πως το κοινωνικό πλαίσιο των reviewers μπορεί να αποκαλύψει πληροφορίες σχετικά με την ποιότητα των reviewers, η οποία αποτυπώνεται και διαμορφώνει τα reviews που γράφουν. Πιο συγκεκριμένα, η προσέγγισή τους βασίζεται στις ακόλουθες υποθέσεις:

- υπόθεση σχετικά με τη συνοχή του συγγραφέα: τα review κείμενα από τον ίδιο συγγραφέα έχουν όλα την ίδια ποιότητα
- υπόθεση σχετικά με τη συνοχή της εμπιστοσύνης: ένας σύνδεσμος από τον reviewer r1 στον reviewer r2 είναι μία άμεση ή έμμεση δήλωση εμπιστοσύνης. Ο reviewer r1 εμπιστεύεται τον reviewer r2 μόνο αν η ποιότητα του reviewer r2 είναι το τουλάχιστον τόσο καλή όσο του reviewer r1
- υπόθεση σχετικά με τη συνοχή της ομαδικής παραπομπής: οι άνθρωποι είναι συνεπής στον τρόπο που εμπιστεύονται άλλους ανθρώπους. Οπότε αν δύο reviewers, r1 και r2, είναι έμπιστοι σύμφωνα με ένα τρίτο reviewer r3, τότε και η ποιότητά τους θα είναι ίδια
- υπόθεση σχετικά με τη συνοχή των συνδέσμων: αν δύο άνθρωποι είναι συνδεδεμένοι στον γράφο ενός κοινωνικού δικτύου (ο r1 εμπιστεύεται τον r2, ή ο r2 εμπιστεύεται τον r1, ή και τα δύο), τότε η ποιότητά τους θα είναι ίδια

Οι παραπάνω υποθέσεις εισήχθησαν σε ένα text-based γραμμικό regression μοντέλο για την πραγματοποίηση της πρόβλεψης της ποιότητας review κειμένων.

Σημαντικό για ένα σύστημα το οποίο παράγει μία κατάταξη των review κειμένων με βάση την ποιότητά τους, είναι να μην διαταράσσει την φυσική κατανομή των θετικών και αρνητικών απόψεων. Δεν πρέπει να εμφανίζονται ψηλά στην κατάταξη μόνο κείμενα τα οποία έχουν θετικές ή αρνητικές απόψεις, επειδή η ποιότητάς τους είναι καλύτερη, διότι με αυτό τον τρόπο υπάρχει κίνδυνος να δοθεί στον χρήστη μία παραπλανητική και λανθασμένη εικόνα την κατανομής των απόψεων. Το σύστημα πρέπει να λαμβάνει υπόψιν του και να εισάγει, με κάποιο τρόπο, αυτή την κατανομή στην

διαδικασία της δημιουργίας της κατάταξης των review κειμένων. Επιπλέον πρέπει να ληθεί υπόψιν και το φαινόμενο του spamming ψεύτικων reviews, καθώς και του spamming ψεύτικών review feedbacks. Το φαινόμενο του spamming είναι συχνό σε τέτοια συστήματα, και είναι δύσκολο να αναγνωριστεί. Αυτό το φαινόμενο και οι τεχνικές αντιμετώπισής του αναλύονται, πιο κάτω, στην ενότητα opinion spam detection.

3.2 Ανάλυση σε επίπεδο πρότασης

Η ανάλυση και επεξεργασία απόψεων σε επίπεδο πρότασης είναι μία πιο “λεπτή” διαδικασία από αυτήν της ανάλυσης σε επίπεδο κειμένου, διότι κινείται πιο κοντά στους στόχους των απόψεων (opinion targets), και στα συναισθήματα τα οποία εκφράζονται για αυτούς. Αν και με βάση την παρατήρηση ότι οι προτάσεις είναι κείμενα μικρού μήκους, δεν φαίνεται να υπάρχει σημαντική διαφορά μεταξύ της ανάλυσης σε επίπεδο κειμένου και πρότασης, η υπόθεση που κάνουν συχνά οι ερευνητές, και είναι αυτή που διαφοροποιεί τα δύο επίπεδα, είναι πως μία πρόταση συχνά φέρει μία άποψη. Αυτή η υπόθεση απλοποιεί την ανάλυση, αλλά όπως είναι φανερό, δεν ισχύει σε πολλές περιπτώσεις. Η ανάλυση απόψεων σε επίπεδο πρότασης πολλές φορές χρησιμοποιείται ως ένα ενδιάμεσο βήμα για την ανάλυση σε επίπεδο κειμένου, και σε μία τέτοια διαδικασία η υπόθεση της μιας άποψης ανά πρόταση γίνεται συχνά, απλοποιώντας την διαδικασία.

Σε αυτό το επίπεδο ο ορισμός της διαδικασίας της εξόρυξης και ανάλυσης απόψεων μετατρέπεται ως εξής:

Δοθέντος μίας πρότασης x, το ζητούμενο είναι να αποφασιστεί αν η πρόταση εκφράζει θετική, αρνητική, ή ουδέτερη άποψη.

Στο επίπεδο πρότασης στόχος είναι η κατηγοριοποίηση των προτάσεων σε θετικής ή αρνητικής πολικότητας, ανάλογα με το αν εκφράζει θετικό ή αρνητικό συναίσθημα (sentence sentiment classification). Πρέπει να προηγηθεί κατηγοριοποίηση της πρότασης σε ουδέτερη ή πρόταση με πολικότητα, ανάλογα με το αν εκφράζει ή όχι κάποιο συναίσθημα (subjectivity classification). Η sentiment κατηγοριοποίηση μιας πρότασης μπορεί να αντιμετωπιστεί είτε ως ένα πρόβλημα κατηγοριοποίησης τριών κατηγοριών (θετική, αρνητική, ουδέτερη), είτε ως δύο ξεχωριστά προβλήματα κατηγοριοποίησης,

όπου το πρώτο είναι η κατηγοριοποίηση της πρότασης σε υποκειμενικού ή ουδέτερου περιεχομένου, και το δεύτερο είναι η κατηγοριοποίηση της υποκειμενικής πρότασης σε θετικής ή αρνητικής πολικότητας. Στην περίπτωση της προσέγγισης της ανάλυσης ως δύο διαδικασιών κατηγοριοποίησης, η πρώτη διαδικασία ονομάζεται κατηγοριοποίηση υποκειμενικότητας (subjectivity classification), ενώ η δεύτερη κατηγοριοποίηση πολικότητας της πρότασης (sentence sentiment classification).

Όπως αναφέρθηκε και πιο πάνω, η ανάλυση απόψεων σε επίπεδο πρότασης συχνά χρησιμοποιείται ως ένα ενδιάμεσο βήμα για την ανάλυση σε επίπεδο κειμένου, και αυτό διότι έχει αρκετές αδυναμίες ως αυτόνομη διαδικασία σε εφαρμογές του πραγματικού κόσμου. Στις περισσότερες εφαρμογές ο χρήστης επιθυμεί να γνωρίζει επιπλέον λεπτομέρειες, για παράδειγμα ποιες οντότητες ή χαρακτηριστικά οντοτήτων (aspects) είναι δημοφιλής ή όχι. Όπως και η ανάλυση σε επίπεδο κειμένου, έτσι και η ανάλυση σε επίπεδο πρότασης, γενικά, δεν μπορεί να δώσει άμεσα αυτές τις απαντήσεις. Αν όμως γνωρίζουμε το στόχο των απόψεων, όπως στην περίπτωση των reviews κειμένων όπου κάθε κείμενο αναφέρεται σε μία οντότητα, τότε μπορούμε εύκολα να αποδώσουμε την sentiment βαθμολογία του κειμένου ή της πρότασης σε αυτόν, και να απαντήσουμε στα παραπάνω ερωτήματα. Ωστόσο, γενικά αυτό είναι ανεπαρκές διότι:

- πολύπλοκες προτάσεις περιέχουν διαφορετικές απόψεις για διαφορετικούς στόχους
- ακόμα και αν μία πρόταση έχει γενικά θετική ή αρνητική πολικότητα, κάποια μέρη της ίσως εκφράζουν αντίθετες απόψεις. Για παράδειγμα, *"Αν και η ανεργία βρίσκεται σε υψηλά επίπεδα, η οικονομία είναι σε καλή κατάσταση"*
- η ανάλυση σε επίπεδο πρότασης δεν μπορεί να αντιμετωπίσει την περίπτωση των συγκριτικών προτάσεων (comparative sentences). Για παράδειγμα, "Η γεύση του καφέ είναι καλύτερη από αυτή του χυμού βύσσινο". Αν και η πρόταση ξεκάθαρα εκφράζει άποψη, δεν μπορούμε να την κατηγοριοποιήσουμε σε θετική, αρνητική ή ουδέτερη.

3.2.1 Κατηγοριοποίηση υποκειμενικότητας

Η κατηγοριοποίηση υποκειμενικότητας κατηγοριοποιεί τις προτάσεις σε δύο κλάσεις, υποκειμενικές και ουδέτερες. Μία αντικειμενική πρόταση διατυπώνει τεκμηριωμένες πληροφορίες, ενώ μία υποκειμενική πρόταση εκφράζει προσωπικές γνώμες, απόψεις, αξιολογήσεις, πεποιθήσεις, εικασίες, καταγγελίες, και στάσεις. Κάποιες από τις υπο-

κειμενικές προτάσεις φέρουν θετική πολικότητα, ενώ κάποιες άλλες αρνητική. Οι πρώτες έρευνες αντιμετώπιζαν την υποκειμενική κατηγοριοποίηση ως αυτόνομο πρόβλημα, ενώ σε πρόσφατες έρευνες, οι ερευνητές, την αντιμετωπίζουν ως το πρώτο βήμα της sentiment κατηγοριοποίησης, και την χρησιμοποιούν για να εντοπίσουν και να αφαιρέσουν τις προτάσεις που δεν φέρουν υποκειμενικό περιεχόμενο.

Οι περισσότερες υπάρχουσες προσεγγίσεις για την κατηγοριοποίηση υποκειμενικότητας, είναι βασισμένες στην εποπτευόμενη μάθηση. Κλειδί σε αυτή την διαδικασία είναι η επιλογή του αλγορίθμου μάθησης και του συνόλου χαρακτηριστικών. Παρακάτω παραθέτονται έρευνες με τις προσεγγίσεις για την λύση της κατηγοριοποίησης υποκειμενικότητας.

Στην έρευνα Wiebe et al. (1999) πραγματοποιείται κατηγοριοποίηση χρησιμοποιώντας Naïve Bayes κατηγοριοποιητή, και ένα δυαδικό σύνολο χαρακτηριστικών με βάση την παρουσία (presence) στην πρόταση συγκεκριμένων μερών του λόγου.

Στην έρευνα Wiebe (2000) προτάθηκε μία μη εποπτευόμενη μέθοδος για την υποκειμενική κατηγοριοποίηση προτάσεων, η οποία χρησιμοποιεί την παρουσία (presence) υποκειμενικών εκφράσεων στις προτάσεις για να καθορίσει την υποκειμενικότητα τους. Δεδομένου ότι δεν υπάρχει ένα πλήρες σύνολο τέτοιων εκφράσεων, παρέχει μερικά αρχικά seeds και χρησιμοποιεί μία κατανομή ομοιότητας για να εντοπίσει παρόμοιες λέξεις οι οποίες πιθανόν αποτελούν δείκτες υποκειμενικότητας. Ωστόσο οι λέξεις που εντοπίζονται με αυτή την τεχνική έχουν μικρή ακρίβεια και υψηλή ανάκληση.

Στην έρευνα Yu and Hatzivassiloglou (2003) πραγματοποιήθηκε υποκειμενική κατηγοριοποίηση προτάσεων χρησιμοποιώντας την ομοιότητα των προτάσεων και ένα Naïve Bayes κατηγοριοποιητή. Η μέθοδος μέτρησης της ομοιότητας των προτάσεων στηρίζεται πάνω στην υπόθεση πως οι προτάσεις που περιέχουν απόψεις είναι περισσότερο ίδιες με άλλες υποκειμενικές προτάσεις, από ότι είναι με προτάσεις οι οποίες διατυπώνουν τεκμηριωμένες πληροφορίες. Για την μέτρηση της ομοιότητας των προτάσεων χρησιμοποιούνται λέξεις, φράσεις και συνώνυμα από το σημασιολογικό δίκτυο WordNet. Για την Naïve Bayes κατηγοριοποίηση χρησιμοποιήθηκαν χαρακτηριστικά όπως unigrams, bigrams, trigrams, μέρη του λόγου, η παρουσία sentiment λέξεων, η αρθροιστική πολικότητα ακολουθιών με sentiment λέξεις (για παράδειγμα “++” για μία ακολουθία με δύο sentiment λέξεις θετικής πολικότητας), και μέρη του λόγου μαζί με την πολικότητα (για παράδειγμα “JJ+” για) θετικό επίθετο.

Στην έρευνα Riloff and Wiebe (2003), τονίζεται πως σε μία διαδικασία εποπτευόμενης μάθησης προτάσεων παρουσιάζεται το “bottleneck” του μη-αυτόματου σχολιασμού (επισήμανσης) ενός μεγάλου αριθμού από δεδομένα εκπαίδευσης, και προτείνεται μία μέθοδος για την αυτόματη επισήμανση των δεδομένων εκπαίδευσης. Ο αλγόριθμος αποτελείται από μία επαναληπτική διαδικασία. Αρχικά χρησιμοποιεί δύο κατηγοριοποιητές υψηλής ακρίβειας για την αυτόματη αναγνώριση μερικών υποκειμενικών και αντικειμενικών προτάσεων, HP-Subj και HP-Obj αντίστοιχα. Αυτοί οι κατηγοριοποιητές χρησιμοποιούν λίστες με λεξικά αντικείμενα (unigrams ή n-grams) τα οποία είναι καλά στοιχεία υποκειμενικότητας. Ο κατηγοριοποιητής HP-Subj κατηγοριοποιεί μία πρόταση ως υποκειμενική αν αυτή περιέχει δύο ή περισσότερες ισχυρές υποκειμενικές αποδείξεις. Ο κατηγοριοποιητής HP-Obj κατηγοριοποιεί μία πρόταση ως αντικειμενική αν δεν υπάρχουν αποδείξεις υποκειμενικότητας. Οι παραπάνω κατηγοριοποιητές δίνουν ηψηλή ακρίβεια αλλά χαμηλή ανάκληση. Οι εξαγώμενες προτάσεις προστίθενται στο στο σύνολο των δεδομένων εκπαίδευσης για την εκμάθηση προτύπων από αυτές. Τα εξαγόμενα πρότυπα διαμορφώνουν τους κατηγοριοποιητές, οι οποίοι χρησιμοποιούνται για την αυτόματη εξαγωγή επιπλέον υποκειμενικών και αντικειμενικών προτάσεων. Οι νέες προτάσεις προστίθενται στο σύνολο εκπαίδευσης και ο αλγόριθμος ξεκινά μία νέα επανάληψη.

Στην έρευνα Barbosa and Feng (2010) πραγματοποιείται υποκειμενική κατηγοριοποίηση tweets με βάση απλά features σε συνδυασμό με κάποια twitter specific στοιχεία όπως retweets, hashtags, links, uppercase words ,emoticon, exclamation, και question marks. Το ίδιο σύνολο χαρακτηριστικών χρησιμοποιήθηκε επίσης και για την κατηγοριοποίηση των tweets σε θετική και αρνητικής πολικότητας.

3.2.2 Sentiment κατηγοριοποίηση πρότασης

Αν μία πρόταση έχει κατηγοριοποιηθεί ως πρόταση που περιέχει υποκειμενική πληροφορία, το επόμενο βήμα είναι να κατηγοριοποιηθεί ως πρόταση που περιέχει θετική ή αρνητική άποψη. Πρέπει να τονιστεί πως κάποιοι αλγόριθμοι δεν χρησιμοποιούν το πρώτο βήμα της κατηγοριοποίησης της πρότασης σε υποκειμενικού περιεχομένου ή ουδέτερη πρόταση. Για αυτήν την διαδικασία της sentiment κατηογριοποίησης προτάσεων μπορεί να χρησιμοποιηθεί, όπως και στην document-level sentiment κατηγοριοποίηση, εποπτευόμενη μάθηση, όπως επίσης και lexicon-based μέθοδοι.

Στην έρευνα Yu and Hatzivassiloglou (2003) για την sentiment κατηγοριοποίηση των προτάσεων χρησιμοποιήθηκε μία παρόμοια τεχνική με αυτή στην έρευνα Turney

(2002), η οποία αναφέρεται αναλυτικά παραπάνω στην ενότητα μη-εποπτευόμενες τεχνικές sentiment κατηγοριοποίησης. Οι διαφορές είναι στον υπολογισμό του sentiment orientation. Αντί να χρησιμοποιηθεί μία θετική και αρνητική λέξη, χρησιμοποιείται ένα μεγάλο σύνολο από επίθετα. Επιπλέον αντί για την μετρική PMI, χρησιμοποιείται μία τροποποιημένη έκδοση του log-likelihood ratio ώστε να για τον προσδιορισμό της πολικότητας κάθε επιθέτου, επιρρήματος, ουσιαστικού και ρήματος. Επιλέχθηκαν δύο κατώφλια, χρησιμοποιώντας το σύνολο δεδομένων εκπαίδευσης, για τον καθορισμό της πολικότητας της πρότασης (θετική, αρνητική, ουδέτερη).

Στην έρευνα Hu and Liu (2004) προτάθηκε ένας lexicon-based αλγόριθμος για την πραγματοποίηση sentiment κατηγοριοποίησης σε επίπεδο aspect, ο οποίος μπορεί να χρησιμοποιηθεί επίσης για τον προσδιορισμό της sentiment πολικότητας προτάσεων. Ο αλγόριθμος είναι βασισμένος σε ένα sentiment λεξικό το οποίο παράγεται χρησιμοποιώντας μία bootstrapping στρατηγική με κάποιες λέξεις ως θετικές και αρνητικές λέξεις ως seeds, καθώς επίσης και τα συνώνυμα και αντώνυμα αυτών από το σημασιολογικό δίκτυο WordNet. Η sentiment πολικότητα μιας πρότασης υπολογίζεται αθροίζοντας τις βαθμολογίες βαθμολογίες των λέξεων της πρότασης. Μία θετική λέξη λαμβάνει την sentiment βαθμολογία +1, ενώ μία αρνητική λέξη λαμβάνει την sentiment βαθμολογία -1. Στην διαδικασία λαμβάνονται υπόψιν επίσης και οι όροι άρνησης.

Στην έρευνα Kim and Hovy (2004), χρησιμοποιήθηκε μία παρόμοια διαδικασία με την παραπάνω. Η μέθοδος για την παραγωγή του sentiment λεξικού είναι η ίδια. Ωστόσο, ο προσδιορισμός της sentiment πολικότητας μιας πρότασης, πραγματοποιείται πολλαπλασιάζοντας της βαθμολογίες των λέξεων της πρότασης. Όπως και στην παραπάνω διαδικασία, μία θετική λέξη λαμβάνει την sentiment βαθμολογία +1, ένω μία αρνητική την sentiment βαθμολογία -1. Επίσης έγιναν πειράματα και με άλλες μεθόδους την συνάρθροιση των sentiment βαθμολογίων μιας πρότασης (sentiment aggrefation), αλλά όλες έδειξαν πως δίνουν χαμηλότερη απόδοση.

Στην έρευνα Gamon et al. (2005), χρησιμοποιείται ένας semi-supervised αλγόριθμος μάθησης για την μάθηση από ένα μικρό σύνολο labeled προτάσεων και ένα μεγάλο σύνολο unlabeled προτάσεων. Ο αλγόριθμος μάθησης είναι βασισμένος στον Expectation Maximization (EM) χρησιμοποιώντας ως κύριο κατηγοριοποιητή τον Naive Bayes. Αυτή το σχήμα πραγματοποιεί κατηγοριοποίηση σε τρεις κλάσεις, θετική, αρνητική, και "'αλλο" (δεν περιέχει άποψη ή περιέχει μεικτές απόψεις).

Στην έρευνα McDonald et al. (2007), προτείνεται ένα ιεραρχικο ακολουθιακό μοντέλο μάθησης, παρόμοιο με το conditional random fields (CRF), ώστε να γίνει από κοινού μάθηση και εξαγωγή υποκειμενικής πληροφορίας και από τα δύο επίπεδα, πρότασης και κειμένου. Στα δεδομένα εκπαίδευσης, κάθε πρόταση είναι επισημασμένη (labeled) με μία πολικότητα, όπως και κάθε κείμενο. Αυτή η έρευνα δείχνει πως η από κοινού μάθηση βελτιώνει την ακρίβεια και στα δύο επίπεδα κατηγοριοποίησης.

Στην έρευνα Davidov et al. (2010), μελετήθηκε η κατηγοριοποίηση σε tweets. Κάθε tweet είναι μπορεί να θεωρηθεί ως μία πρόταση. Η προσέγγιση που ακολουθήθηκε ήταν η εποπτευόμενη μάθηση. Εκτός από κλασικά χαρακτηριστικά, χρησιμοποιήθηκαν επιπλέον και hashtags, smileys, σημεία στίξης, και τα πρότυπα συχνοτήτων τους. Αυτά τα χαρακτηριστικά έδεξαν πως είναι αρκετά αποδοτικά.

3.2.3 Διαχείριση ειδικών μορφών προτάσεων

Οι περισσότερες από τις έρευνες πάνω στην sentiment κατηγοριοποίηση στο επίπεδο πρότασης επικεντρώνονται στην λύση του γενικού προβλήματος, χωρίς να λαμβάνουν υπόψιν τους το γεγονός πως υπάρχουν διαφορετικά είδη προτάσεων τα οποία επιβάλουν ειδική μεταχείριση. Όπως τονίζεται και πιο πάνω σε αυτή την ενότητα, οι περισσότερες έρευνες κάνουν την υπόθεση πως μία πρόταση φέρει μία μόνο άποψη, και αυτό διότι η ανάλυση γίνεται ευκολότερη. Στην έρευνα Narayanan et al. (2009) τονίζεται πως είναι απίθανο να υπάρχει μία τεχνική η οποία να μπορεί να διαχειριστεί όλα τα είδη προτάσεων, διότι διαφορετικά είδη προτάσεων εκφράζουν απόψεις με διαφορετικούς τρόπους. Προτείνεται μία divide-and-conquer προσέγγιση, όπου οι έρευνες να επικεντρώνονται σε συγκεκριμένα είδη προτάσεων. Η συγκεκριμένη έρευνα επικεντρώνεται στις υπό συνθήκη (conditional) προτάσεις, οι οποίες παρουσιάζουν κάποια ξεχωριστά χαρακτηριστικά και δυσκολεύουν τα συστήματα στο να προσδιορίσουν την πολικότητά τους. Οι υπό συνθήκη προτάσεις είναι προτάσεις οι οποίες περιγράφουν επιπτώσεις ή υποθετικές καταστάσεις και τις συνέπειές τους. Αυτές οι προτάσεις συνήθως περιέχουν δύο μέρη, τη συνθήκη και το το αποτέλεσμα ή συνέπεια, τα οποία είναι εξαρτημένα μεταξύ τους. Η σχέση μεταξύ των δύο αυτών μερών, έχει καθοριστικό ρόλο στον αν η πρόταση, ως σύνολο, εκφράζει θετική ή αρνητική άποψη. Η υπόθεση που γίνεται στις περισσότερες έρευνες για την sentiment κατηγοριοποίηση μιας πρότασης, πως μία πρόταση περιέχει μόνο μία ή καμία άποψη, οδηγεί στην απλή προσέγγιση πως οι sentiment λέξεις της πρότασης (αν υπάρχουν), θα αποκαλύπτουν την πολικότητα αυτής της άποψης ή την ουδετερότητα της αντίστοιχα. Αυτή η προσέγγιση,

γενικά, δεν μπορεί να εφαρμοστεί σε μία υπό συνθήκη πρόταση όπως *"Αν η x βιβλιοθήκη δεν σε καλύπτει, προσπάθησε να δουλέψεις με την y."* Σε αυτή την πρόταση το πρώτο μέρος δεν εκφράζει κάποια άποψη, αλλά το δεύτερο εκφράζει θετική άποψη. Η παραπάνω πρόταση δεν περιέχει καμία sentiment λέξη, αλλά περιέχει άποψη. Για την διαχείριση των υπο συνθήκη προτάσεων συνήθως χρησιμοποιούνται εποπτευόμενες τεχνικές οι οποίες χρησιμοποιούν σύνολα γλωσσολογικών χαρακτηριστικών (linguistic features), όπως η θέση συγκεκριμένων λέξεων στις προτάσεις, POS tags, πρότυπα, και συντακτικοί σύνδεσμοι.

Ένα άλλο είδος δύσκολων προτάσεων, ως προς την διαχείριση, είναι οι προτάσεις ερωτήσεων, όπως για παράδειγμα *"Μπορεί κάποιος να μου προτείνει ένα μοντέλο smartphone με καλή οθόνη;"*, η οποία δεν περιέχει κάποια άποψη, αν και περιέχει μία sentiment λέξη η οποία στοχεύει ένα χαρακτηριστικό μιας οντότητας (καλή οθόνη smartphone).

Επίσης υπάρχουν και οι προτάσεις οι οποίες περιέχουν σαρκασμό. Αυτό το είδος προτάσεων είναι αρκετά συχνό στις συζητήσεις στο διαδίκτυο, όπως και στα tweets, αλλά σπάνιο σε review κείμενα. Οι προτάσεις σαρκασμού είναι μία ειδική μορφή του λόγου, με την οποία ο ομιλητής ή συγγραφέας λέει ή γράφει το αντίθετο από αυτό που πραγματικά εννοεί. Ο σαρκασμός έχει μελετηθεί στην γλωσσολογία, την ψυχολογία και την γνωστική επιστήμη. Όσον αφορά την διαδικασία της sentiment ανάλυσης, όταν εντοπίζεται ένα τέτοιο είδος πρότασης, η πολικότητά της πρέπει να αντιστρέφεται. Αν και αυτό το είδος προτάσεων, όπως είναι φανερό, είναι δύσκολο να εντοπιστεί και να διαχειριστεί από έναν αλγόριθμο, υπάρχουν έρευνες οι οποίες προτείνουν τρόπους για την διαχείρισή του.

Στην έρευνα Tsur et al. (2010), χρησιμοποιείται μία semi-supervised προσέγγιση για την αναγνώριση του σαρκασμού. Χρησιμοποιεί ένα μικρό σύνολο labeled προτάσεων (seeds), αλλά δεν χρησιμοποιεί σύνολο unlabeled προτάσεων. Αντί για αυτό, το σύνολο των seeds επεκτείνεται αυτόματα χρησιμοποιώντας αναζήτηση στο διαδίκτυο. Η υπόθεση πάνω στην οποία στηρίζεται η επέκταση του seed συνόλου, είναι πως οι προτάσεις σαρκασμού συνήθως συνυπάρχουν στο κείμενο μαζί με άλλες προτάσεις σαρκασμού. Οπότε πραγματοποιείται μία αυτόματη αναζήτηση στο διαδίκτυο για κάθε πρόταση του seed συνόλου ως ερωτήματος αναζήτησης. Το σύστημα συλλέγει τα snippets (σύντομο κείμενο περιγραφής των αποτελεσμάτων αναζήτησης) για κάθε ερώτημα και προσθέτει τις προτάσεις του στο σύνολο εκπαίδευσης. Αυτό το σύ-

νολο εκπαίδευσης χρησιμοποιείται για την μάθηση του κατηγοριοποιητή. Τα χαρακτηριστικά που χρησιμοποιούνται βασίζονται σε πρότυπα και σημεία στίξης. Ένα πρότυπο είναι απλά μία διατεταγμένη ακολουθία λέξεων, οι οποίες έχουν μεγάλη συχνότητα εμφάνισης. Τα χαρακτηριστικά τα οποία βασίζονται σε σημεία στίξης περιέχουν των αριθμό των “!”, “?” και “"” στην πρόταση, καθώς επίσης τον αριθμό των λέξεων οι οποίες αρχίζουν με κεφαλαίο. Για την κατηγοριοποίηση χρησιμοποιήθηκε μία μέθοδος βασισμένη στους k κοντινότερους γείτονες (k-NN-based).

Στην έρευνα González-Ibáñez et al. (2011), μελετήθηκε το πρόβλημα της αναγνώρισης των σαρκαστικών προτάσεων σε δεδομένα από το Twitter. Προτάθηκε μία τεχνική εποπτευόμενης μάθησης η οποία χρησιμοποιεί Support Vector Machines (SVM) και παλινδρόμηση (logistic regression). Ως χαρακτηριστικά χρησιμοποιήθηκαν unigrams και κάποιες dictionary-based πληροφορίες, όπως κατηγορίες λέξεων, επιφωνήματα, σημεία στίξης, καθώς επίσης και emoticons και ToUser (το οποίο δείχνει πως ένα tweet είναι απάντηση σε κάποιο tweet που έθεσε ο χρήστης <@user>). Τα πειράματα έδειξαν πως η κατηγοριοποίηση των προτάσεων στις τρεις κλάσεις (σαρκαστικές, θετικές, αρνητικές), μπορεί να επιτύχει ακρίβεια μόνο 57%.

3.3 Ανάλυση σε επίπεδο χαρακτηριστικού

Η κατηγοριοποίηση των απόψεων σε επίπεδο κειμένου ή σε επίπεδο πρότασης είναι συχνά αναποτελεσματική για πολλές εφαρμογές, επειδή δεν αναγνωρίζονται οι συγκεκριμένοι στόχοι των απόψεων, και δεν αποδίδεται άμεσα η πολικότητα των απόψεων σε αυτούς τους στόχους. Ακόμα και αν υποθέσουμε πως ένα κείμενο αναφέρεται σε μία συγκεκριμένη οντότητα, όπως τα review κείμενα, το γεγονός πως ένα review κείμενο το οποίο εκφράζει στο σύνολό τους θετική άποψη για μία οντότητα, δεν σημαίνει απαραίτητα πως ο συγγραφέας αυτού του review έχει θετική άποψη για όλα τα aspects (χαρακτηριστικά) αυτής της οντότητας. Το ίδιο μπορούμε να πούμε και για ένα αρνητικό review κείμενο. Για μία εφαρμογή η οποία απαιτεί πιο λεπτή ανάλυση, πρέπει να ανακαλύψουμε αυτά τα aspects και να αποδώσουμε σε αυτά τις πολικότητες των απόψεων. Τα aspects μπορεί να είναι τα μέρη από τα οποία αποτελείται μία οντότητα καθώς και οι ιδιότητες που έχει, και συνήθως αποτελούν τους στόχους των απόψεων. Για την εξαγωγή αυτών των λεπτομερειών πρέπει να πραγματοποιήσουμε ανάλυση απόψεων σε επίπεδο aspect. Αυτό σημαίνει πως πρέπει να χρησιμοποιήσουμε τον πλήρη ορισμό της εξόρυξης και ανάλυσης απόψεων, όπως αυτός αναφέρθηκε στην ενότη-

τα 2 πιο πάνω. Στο aspect επίπεδο ανάλυσης θεωρούμε πως το σύνολο των στόχων των απόψεων αποτελείται από τις οντότητες και τα χαρακτηριστικά τους (υπομέρη και ιδιότητες) ή aspects. Πρέπει να τονιστεί πως αν και η ανάλυση σε αυτό το επίπεδο είναι πιο πλήρης, διότι ακολουθεί τον πλήρη ορισμό του προβλήματος της εξόρυξης και ανάλυσης απόψεων, είναι πιο δύσκολη να πραγματοποιηθεί επειδή ερχόμαστε αντιμέτωποι με τα δύσκολα προβλήματα του natural language processing (NLP).

Πιο συγκεκριμένα, ο ορισμός της εξόρυξης και ανάλυσης απόψεων στο aspect επίπεδο ανάλυσης είναι ο ακόλουθος:

Δοθέντος ένος κειμένου d (ή μιας συλλογής κειμένων D), ανακάλυψε όλες τις πλειάδες απόψεων (e_i, a_{ij}, s_{ijkl}, h_k, t_l) που υπάρχουν μέσα στο d. Η διαδικασία περιέχει 6 διεργασίες (tasks):

- *Task1: εξαγωγή όλων των οντοτήτων (entities) από το D και κατηγοριοποίηση των συνόνυμων σε ομάδες (entity clusters). Σε κάθε cluster δίνεται το αναγνωριστικό e_i.*
- *Task2: εξαγωγή όλων των χαρακτηριστικών (aspects) από το D και κατηγοριοποίηση των συνόνυμων σε ομάδες (aspects clusters). Σε κάθε cluster δίνεται το αναγνωριστικό a_{ij} του e_i.*
- *Task3: εξαγωγή των opinion holders και κατηγοριοποίηση σε ομάδες με αναγνωριστικό h_k.*
- *Task4: εξαγωγή του χρόνου και κανονικοποίηση σε ένα πρότυπο.*
- *Task5: προσδιορισμός της πολικότητας της άποψης (θετική/αρνητική/ουδέτερη). ή βαθμολόγηση (σε ποσοστό επί τις % ή 1-5 stars).*
- *Task6: παραγωγή όλων των πλειάδων απόψεων (e_i, a_{ij}, s_{ijkl}, h_k, t_l).*

Στην συνέχεια παρουσιάζεται αναλυτικότερα κάθε μία από τις έξι διεργασίες που αποτελούν την ανάλυση απόψεων στο aspect επίπεδο, καθώς και κάποιες τεχνικές προσέγγισής τους.

3.3.1 Εξαγωγή των οντοτήτων, των opinion holders, και του χρόνου

Η πρώτη, ή τρίτη και η τέταρτη διεργασία της ανάλυσης απόψεων στο aspect επίπεδο, αναφέρονται στην εξαγωγή των οντοτήτων, των opinion holders (αυτών που εκφράζουν την άποψη), και του χρόνου στον οποίο εκφράζεται η άποψη, αντίστοιχα. Αυτές οι διεργασίες είναι κλασικά προβλήματα named entity recognition (NER). Τέτοια προβλήματα έχουν μελετηθεί στους τομείς της ανάκτησης πληροφορίας, του text

mining, του machine learning, και natural language processing, και κατατάσονται στην κατηγορία των προβλημάτων εξαγωγής πληροφορίας. Υπάρχουν δύο κύριες μέθοδοι εξαγωγής πληροφορίας, αυτές οι οποίες βασίζονται σε κανόνες (rule-based), και στατιστικές μέθοδοι. Παλιότερα συστήματα εξαγωγής πληροφορίας βασίζονταν κυρίως σε κανόνες. Οι στατιστικές μέθοδοι βασίζονται κυρίως σε Hidden Markov Models (HMM) (Rabiner, 1989; Jin and Ho, 2009) (Appendix III.) και Conditional Random Fields (CRF) (Lafferty et al., 2001), οι οποίες είναι εποπτευόμενες τεχνικές.

Από αυτές τις τρεις διεργασίες στην πλειοψηφία των συστημάτων sentiment ανάλυσης πραγματοποιείται μόνο η πρώτη, εξαγωγή των οντοτήτων, διότι αυτή αποκαλύπτει τους γενικούς στόχους των απόψεων. Η εξαγωγή του χρόνου και των opinions holders στις περισσότερες εφαρμογές παραλείπεται, είτε επειδή δεν χρειάζονται στην διαδικασία, είτε επειδή είναι ήδη διαθέσιμοι. Για παράδειγμα, σε εφαρμογές οι οποίες χρησιμοποιούν τα κοινωνικά δίκτυα δεν υπάρχει ανάγκη για την εξαγωγή των opinion holders και του χρόνου, επειδή είναι ήδη διαθέσιμοι. Οι opinion holders είναι οι συγγραφείς των reviews, των blogs, ή των μηνυμάτων σε μία συζήτηση, και διαθέτουν συνήθως ένα αναγνωριστικό το οποίο όμως δεν συνδέεται πάντα με το πραγματικό τους όνομα. Επίσης ο χρόνος που διατυπώθηκε το κείμενο συνήθως είναι διαθέσιμος από το ίδιο το σύστημα και μπορέι να εξαχθεί άμεσα.

Όσον αφορά την εξαγωγή οντοτήτων υπάρχουν επίσης περιπτώσεις στις οποίες μπορεί να αποφευχθεί. Για παράδειγμα, σε ένα σύστημα αναζήτησης απόψεων για οντότητες, ο χρήστης εισάγει το όνομα ή το είδος της οντότητας για την οποία ενδιαφέρεται, οπότε το σύστημα μπορεί να επεξεργάζεται μόνο τα κείμενα ή τις προτάσεις οι οποίες αναφέρουν αυτή την οντότητα. Θα πρέπει όμως να τονιστεί πως οι χρήστες συνήθως χρησιμοποιούν διαφορετικά ονόματα για να αναφερθούν στην ίδια οντότητα. Οπότε είναι επιθυμητό από το σύστημα να έχει μία λίστα με τις διαφορετικές ονομασίες μιας οντότητας.

3.3.2 Εξαγωγή και ομαδοποίηση των aspects

Αυτή η διεργασία της ανάλυσης απόψεων στο aspect επίπεδο εξάγει τα υπομέρη ή τις ιδιότητες των οντοτήτων (aspects) από το κείμενο, τα οποία, aspects, όπως αναφέρεται και πιο πάνω αποτελούν τους συγκεκριμένους στόχους των απόψεων. Για τις απόψεις οι οποίες έχουν ως στόχο τη οντότητα ως σύνολο χρησιμοποιούμε το χαρακτηριστικό GENERAL. Για παράδειγμα στην πρόταση *“Η ποιότητα της οθόνης αυτού του κινητού είναι καλή”*, ο στόχος της άποψης είναι *“η ποιότητα της οθόνης”*. Όμως οι λέξεις *“αυ-*

τού του κινητού" δεν σημαίνει πως ο στόχος της άποψης στην πρόταση είναι το χαρακτηριστικό GENERAL. Αντίθετα, στην πρόταση *"Αυτό το κινητό είναι καλό"*, ο στόχος της άποψης είναι το κινητό ως σύνολο, οπότε η άποψη αναφέρεται στο γενικό χαρακτηριστικό GENERAL.

Η διεργασία της εξαγωγής των aspects μιας οντότητας από κείμενο είναι μία διαδικασία εξαγωγής πληροφορίας, όπως και η εξαγωγή των οντοτήτων, των opinion holders, και του χρόνου της άποψης, όμως υπάρχουν ορισμένα χαρακτηριστικά τα οποία καθιστούν την συγκεκριμένη διεργασία ευκολότερη. Το κύριο χαρακτηριστικό είναι πως μία άποψη έχει πάντα ένα στόχο. Ο στόχος αυτός είναι συνήθως το aspect το οποίο πρέπει να εξαχθεί από την πρόταση. Οπότε αναγνωρίζοντας τις sentiment λέξεις ή τις εκφράσεις οι οποίες χρησιμοποιούνται για να εκφραστεί η άποψη, μπορούμε εύκολα να εντοπίσουμε και τον στόχο της άποψης, το aspect, ο οποίος τοποθετείται, μέσα στην πρόταση, συνήθως κοντά σε αυτές τις λέξεις. Πρέπει όμως να τονιστεί πως υπάρχουν προτάσεις οι οποίες εκφράζουν μία άποψη για ένα aspect μιας οντότητας, έμμεσα, χωρίς να το αναφέρουν. Για παράδειγμα στην πρόταση *"Αυτό το κινητό είναι ακριβό"* ο στόχος της άποψης είναι η τιμή του κινητού, αλλά δεν αναφέρεται άμεσα πουθενά στην πρόταση η λέξη *"τιμή"*.

Στις περιπτώσεις τις οποίες οι απόψεις είναι άμεσες, υπάρχουν τέσσερις κύριες προσεγγίσης για την διεργασία της εξαγωγής των aspects:

- Εξαγωγή χρησιμοποιώντας την συχνότητα των ουσιαστικών και των ονοματικών εκφράσεων
- Εξαγωγή χρησιμοποιώντας την σχέση μεταξύ άποψης και του στόχου της άποψης
- Εξαγωγή χρησιμοποιώντας εποπτευόμενη μάθηση
- Εξαγωγή χρησιμοποιώντας μοντελοποίηση θέματος (topic modeling)

3.3.2.1 Εξαγωγή χρησιμοποιώντας την συχνότητα των ουσιαστικών και των ονοματικών εκφράσεων

Αυτή η προσέγγιση εντοπίζει ως aspects τα ουσιαστικά και τις ονοματικές εκφράσεις οι οποίες έχουν μεγάλη συχνότητα εμφάνισης. Η υπόθεση πάνω στην οποία στηρίζεται αυτή η προσέγγιση είναι πως όταν οι άνθρωποι σχολιάζουν τα aspects μιας οντότητας το λεξιλόγιο συχνά συγκλίνει. Οπότε ουσιαστικά τα οποία εμφανίζονται συχνά είναι σημαντικά aspects, ενώ ουσιαστικά με μικρή συχνότητα εμφάνισης είναι λιγότεροση-

μαντικά aspects ή δεν είναι aspects. Αν και αυτή η προσέγγιση είναι πολύ απλή, έχει δειχθεί πως είναι αρκετά αποδοτική. Για την αναγνώριση των ουσιαστικών ή των ονοματικών φράσεων μπορεί να χρησιμοποιηθεί ένας POS tagger (part-of-speech tagger). Για κάθε ένα ουσιαστικό ή ονοματική φράση, μετριούνται οι συχνότητες εμφάνισης, και ως aspects εξάγονται τα στοιχεία των οποίων η συχνότητα εμφάνισης ξεπερνά ένα κατώφλι.

Στην έρευνα Popescu and Etzioni (2005) προτείνονται κάποιες αλλαγές για την βελτίωση της ακρίβειας της παραπάνω διεργασίας. Ο αλγόριθμος ο οποίος προτείνεται προσπαθεί να αφαιρέσει τα στοιχεία τα οποία μπορεί να μην είναι aspects οντοτήτων. Πιο συγκεκριμένα, γίνεται αξιολόγηση κάθε στοιχείου στην λίστα των πιθανών aspects, υπολογίζοντας την pointwise mutual information (PMI) βαθμολογία μεταξύ του κάθε στοιχείου και ενός συνόλου εκφράσεων, οι οποίες χρησιμοποιούνται για την αναφορά σε aspects κάποιας οντότητας της συγκεκριμένης κλάσης όταν γίνεται αναζήτηση στο διαδίκτυο. Για παράδειγμα αν η κλάση της οντότητας είναι αυτή των κινητών συσκευών tablet, τότε οι εκφράσεις οι οποίες μπορούν να χρησιμοποιηθούν είναι “of tablet”, “tablet has”, “tablet comes with”, κτλ. Η μετρική PMI υπολογίζεται ως εξής:

$$PMI\left(a,d\right)=\frac{hits\left(a\wedge d\right)}{hits\left(a\right)hits\left(d\right)}$$

όπου *a* είναι κάποιο υποψήφιο aspect το οποίο έχει αναγνωριστεί με βάση την συχνότητα εμφάνισης, και *d* είναι κάποια από τις εκφράσεις διευκρίνησης. Για τον υπολογισμό των *hits()*, μοναδικού όρου *hits(i)* και συνεμφάνισης *hits(i^j)*, χρησιμοποιήθηκε αναζήτηση στο διαδίκτυο. Πιο συγκεκριμένα, *hits(k)* είναι ο αριθμός των αποτελεσμάτων που επιστρέφει η μηχανή αναζήτησης όταν ο όρος ή οι όροι αναζήτησης είναι ο *k*. Αν η PMI βαθμολογία του υποψηφίου aspect *a* και της φράσης *d*, *PMI(a,d)*, είναι μικρή τότε το *a* πιθανόν δεν είναι aspect επειδή τα *a* και *d* δεν εμφανίζονται συχνά μαζί. Ο αλγόριθμος, επίσης, διακρίνει τα εξαγόμενα aspects τα οποία είναι υπομέρη της οντότητας από αυτά που αποτελούν ιδιότητες, χρησιμοποιώντας το σημασιολογικό δίκτυο WordNet.

Στην έρευνα Blair-Goldensohn et al. (2008) χρησιμοποιείται επίσης η προσέγγιση της εξαγωγής ουσιαστικών και ονοματικών φράσεων με μεγάλη συχνότητα εμφάνισης

ως aspects. Η διαφορά είναι πως εξάγονται ουσιαστικά και ονοματικές φράσεις οι οποίες εμφανίζονται με μεγάλη συχνότητα σε προτάσεις που φέρουν άποψη. Τα υποψήφια aspects τα οποία δεν αναφέρονται πολλές φορές σε sentiment προτάσεις, απορρίπτονται.

Στην έρευνα Ku et al. (2006), ακολουθείται επίσης η προσέγγιση εξαγωγής aspects με βάση την συχνότητα εμφάνισης, αλλά χρησιμοποιείτια επιπλέον και η μετρική TF-IDF για την βαθμολογία των όρων.

Ακόμα, στην έρευνα Scaffidi et al. (2007), συγκρίνεται η συχνότητα εμφάνισης των εξαγόμενων ουσιαστικών και φράσεων, με τα ποσοστά εμφάνισής τους σε μία συλλογή κειμένων, για την πραγματοποίηση της αναγνώρισης των aspects.

Στην έρευνα Long et al. (2010) εξάγονται aspects με βάση την συχνότητα εμφάνισης και την απόσταση πληροφορίας. Η προτεινόμενη μέθοδος αρχικά εντοπίζει ένα μικρό σύνολο από aspects χρησιμοποιώντας την συχνότητα. Ύστερα, χρησιμοποιεί την απόσταση πληροφορίας, Cilibrasi and Vitanyi (2007), για κάθε ένα από τα στοιχεία του μικρού συνόλου για να εντοπίσει περισσότερα aspects, τα οποία θα είναι σχετικά με τα αρχικά.

3.3.2.2 Εξαγωγή χρησιμοποιώντας την σχέση μεταξύ άποψης και στόχου

Κάθε άποψη έχει ένα στόχο. Οι απόψεις εκφράζονται με sentiment λέξεις και εκφράσεις, οι οποίες εντοπίζονται εύκολα μέσα σε μία πρόταση. Τα aspects είναι συνήθως οι στόχοι των απόψεων και με μεγάλη πιθανότητα βρίσκονται, μέσα στην πρόταση, κοντά στις sentiment λεξεις και εκφράσεις. Οπότε μπορούμε να χρησιμοποιήσουμε την σχέση μεταξύ άποψης και στόχου για να εντοπίσουμε aspects σε κείμενο. Στην έρευνα Hu and Liu (2004), χρησιμοποιείται αυτή η προσέγγιση για την εξαγωγή aspect με μικρή συχνότητα εμφάνισης. Για παράδειγμα στην πρόταση *"The software is amazing"*, αν γνωρίζουμε πως το "amazing" είναι sentiment λέξη τότε η λέξη "software" εξάγεται ως aspect (σε αυτή την περίπτωση ο στόχος της άποψης είναι το χαρακτηριστικό GENERAL, η οντότητα στο σύνολό της). Αυτή η προσέγγιση είναι επίσης αρκετά χρήσιμη για τον εντοπισμό σημαντικών θεματικών ενοτήτων (topics) σε opinion κείμενα, επειδή ένα aspect ή topic δεν μπορεί να είναι σημαντικό αν κανένας δεν εκφράζει άποψη για αυτό.

Στην έρευνα Zhuang et al. (2006), χρησιμοποιείται ένας dependency parser (parser εξαρτήσεων), ο οποίος εκμεταλλεύεται αυτήν την σχέση εξάρτησης μεταξύ άποψης και στόχου για την εξαγωγή aspects. Στην έρευνα Somasundaran and Wiebe (2009)

χρησιμοποιήθηκε παρόμοια προσέγγιση, όπως επίσης και στην Kobayashi et al. (2006).

Στην έρευνα Wu et al. (2009), χρησιμοποιείται ένας dependency parser για εκφράσεις, αντί για λέξεις, οι οποίες μπορεί να είναι πιθανά aspects. Ένας dependency parser αναγνωρίζει εξαρτήσεις μεταξύ λέξεων μόνο, ενώ ένας dependency parser εκφράσεων αναγνωρίζει εξαρτήσεις μεταξύ εκφράσεων.

Στην έρευνα Qiu et al. (2011), η ιδέα της εξάρτησης άποψης και στόχου γενικεύτηκε περισσότερο, και οδήγησε στην δημιουργία double-propagation μεθόδου για την ταυτόχρονη εξαγωγή sentiment λέξεων και aspects.

3.3.2.3 Εξαγωγή χρησιμοποιώντας εποπτευόμενη μάθηση

Η εξαγωγή των aspects μπορεί να προσεγγιστεί επίσης ως μία ειδική περίπτωση εξαγωγής πληροφορίας. Έχουν προταθεί αρκετοί αλγόριθμοι εποπτευόμενης μάθησης για εξαγωγή πληροφορίας στο παρελθόν. Οι κυρίαρχες μέθοδοι βασίζονται στην ακολουθιακή μάθηση (sequential learning ή sequential labeling). Δημοφιλής τεχνικές sequential learning είναι τα Hidden Markov Models (HMM) (Rabiner, 1989) και Conditional Random Fields (CRF) (Lafferty et al., 2001).

Στην έρευνα Jin and Ho (2009), χρησιμοποιήθηκε ένα lexicalized HMM μοντέλο ώστε να εξαχθούν aspects και opinion εκφράσεις. Στην έρευνα Jakob and Gurevych (2010) χρησιμοποιήθηκαν CRF, τα οποία εκπαιδεύτηκαν σε review προτάσεις από διαφορετικά domains, για την domain independent εξαγωγή. Επίσης χρησιμοποιήθηκε ένα σύνολο domain independent features, για παράδειγμα λέξεις, POS tags, συντακτική εξάρτηση, απόσταση λέξεων, και opinion προτάσεις.

Στην έρευνα Li et al (2010) συνδυάστηκαν δύο CRF παραλλαγές, το μοντέλο Skip-CRF και το Tree-CRF, για την εξαγωγή aspects, καθώς και απόψεων. Αντίθετα με τα απλά μοντέλα CRF, τα οποία μπορούν να χρησιμοποιήσουν μόνο ακολουθίες λέξεων στην μάθηση, τα Skip-CRF και Tree-CRF μοντέλα χρησιμοποιούν χαρακτηριστικά δομής.

Στην έρευνα Yu et al. (2011), χρησιμοποιήθηκε μία μερικώς εποπτευόμενη μέθοδος μάθησης η οποία ονομάζεται one-class SVM (Manevitz and Yousef, 2002), για την εξαγωγή των aspects. Χρησιμοποιώντας one-class SVM, χρειάζεται η σήμανση μόνο μερικών παραδειγμάτων (οχι παραδείγματα μη aspect).

Στην έρευνα Ghani et al. (2006), χρησιμοποιήθηκε εποπτευόμενη, καθώς και semi-supervised μάθηση, για την εξαγωγή των χαρακτηριστικών.

3.3.2.4 Εξαγωγή χρησιμοποιώντας θεματικά μοντέλα

Τα τελευταία χρόνια, στατιστικά θεματικά μοντέλα (topic models) έχουν αναδειχθεί ως η κύρια μέθοδος για την ανακάλυψη θεμάτων σε μεγάλες συλλογές κειμένων. Topic modeling είναι μία μη-εποπτευόμενη τεχνική μάθησης η οποία στηρίζεται στην υπόθεση πως κάθε κείμενο αποτελείται από ένα μείγμα από θεματικές ενότητες, και πως κάθε θεματική ενότητα είναι μία πιθανοτική κατανομή σε λέξεις. Ένα topic μοντέλο είναι ένα μοντέλο παραγωγής εγγράφων, το οποίο καθορίζει μία πιθανοτική διαδικασία με την οποία μπορούν να αναπαραχθούν κείμενα. Η έξοδος ενός είναι ένα σύνολο από συστάδες λέξεων. Κάθε συστάδα διαμορφώνει μία θεματική ενότητα και είναι μία πιθανοτική κατανομή λέξεων σε μία συλλογή κειμένων.

Υπάρχουν δύο κύρια μοντέλα, το pLSA (Probabilistic Latent Semantic Analysis) (Hofmann, 1999) (Appendix V.), και το LDA (Latent Dirichlet allocation) (Blei et al., 2003). Τεχνικά, τα θεματικά μοντέλα είναι ένα είδος γραφικών μοντέλων βασισμένα στα Bayesian networks. Αν και κυρίως χρησιμοποιούνται για την μοντελοποίηση και εξαγωγή θεματικών ενοτήτων από συλλογές κειμένων, μπορούν να επεκταθούν και να μοντελοποιούν επιπλέον τύπους πληροφοριών ταυτόχρονα. Για παράδειγμα, για την διαδικασία της sentiment ανάλυσης, μπορεί να σχεδιαστεί ένα μοντέλο για την μοντελοποίηση των sentiment λέξεων και των θεματικών ενοτήτων ταυτόχρονα, με βάση την παρατήρηση πως κάθε άποψη έχει ένα στόχο. Μπορούμε να πούμε, πως οι θεματικές ενότητες είναι τα aspects στον τομέα της sentiment ανάλυσης. Οπότε η θεματική μοντελοποίηση μπορεί να εφαρμοστεί για την εξαγωγή των aspects. Ωστόσο, υπάρχει επίσης μία διαφορά. Οι θεματικές ενότητες μπορούν να περιλαμβάνουν τα aspects και τις sentiment λέξεις. Για τη διαδικασία της sentiment ανάλυσης, αυτά πρέπει να διαχωριστούν. Αυτός ο διαχωρισμός μπορεί να επιτευχθεί επεκτείνοντας το βασικό μοντέλο ώστε να μοντελοποιεί από κοινού τα aspects και τις sentiments λέξεις. Επίσης τα θεματικά μοντέλα έχουν την δυνατότητα να εξάγουν aspects και να τα ομαδοποιούν σε ομάδες συνώνυμων.

Παρακάτω επισημαίνονται sentiment ανάλυσης έρευνες οι οποίες χρησιμοποιούν θεματικά μοντέλα για την εξαγωγή aspects.

Στην έρευνα Mei et al. (2007) προτείνεται ένα joint μοντέλο για sentiment ανάλυση. Πιο συγκεκριμένα, χτίζεται ένα aspect-sentiment mixture μοντέλο, το οποίο βασίζεται σε ένα aspect (topic) μοντέλο, ένα positive sentiment μοντέλο, και ένα negative sentiment μοντέλο. Αυτά τα μοντέλα είναι βασισμένα στο pLSA.

Στην έρευνα Titov and McDonald (2008) τονίζεται πως topic μοντέλα όπως το LDA (Blei et al., 2003) ίσως να μην είναι ιδανικά να αναγνωρίσουν aspects. Ο λόγος για αυτό είναι πως το LDA βασίζεται στις διαφορές της θεματικής κατανομής και τις συνεμφανίσεις λέξεων στα κείμενα, για να αναγνωρίσει θεματικές ενότητες και πιθανοτικές κατανομές λέξεων για κάθε θεματική ενότητα. Ωστόσο, κείμενα τα οποία περιέχουν απόψεις για ένα συγκεκριμένο είδος προϊόντος, όπως τα review κείμενα, είναι ομοιογενή, το οποίο σημαίνει πως κάθε κείμενο αναφέρεται στα ίδια aspects, κάτι που καθιστά τα global topic μοντέλα μη αποδοτικά, και τα περιορίζει μόνο στην ανακάλυψη οντοτήτων. Προτείνεται τα multigrain topic μοντέλα. Το global μοντέλο ανακαλύπτει οντότητες ενώ το local μοντέλο ανακαλύπτει aspects χρησιμοποιώντας λίγες προτάσεις (ή ένα κινούμενο παράθυρο προτάσεων) ως ένα κείμενο. Κάθε aspect το οποίο ανακαλύπτεται είναι ένα unigram language μοντέλο, για παράδειγμα μία πολυωνυμική κατανομή σε λέξεις. Διαφορετικές λέξεις που εκφράζουν την ίδια ή σχετικές όψεις, ομαδοποιούνται μαζί ως το ίδιο aspect. Ωστόσο, αυτή η τεχνική δεν καταφέρνει να ξεχωρίσει τα aspects από τις sentiment λέξεις.

Στην έρευνα Li et al. (2010) προτείνονται δύο joint μοντέλα, Sentiment-LDA και Dependency-sentiment-LDA, για την ανακάλυψη aspect με θετική και αρνητική πολικότητα. Αυτή η τεχνική όμως δεν έχει την δυνατότητα να ανακαλύπτει ανεξάρτητα aspect και να διαχωρίζει τα aspects από τις sentiment λέξεις.

Στην έρευνα Mukherjee and Liu (2012), προτείνεται ένα semi-supervised joint μοντέλο, το οποίο επιτρέπει στον χρήστη να παρέχει μερικά seed aspects, ώστε να καθοδηγήσει το μοντέλο στην παραγωγή aspect κατανομών οι οποίες είναι σύμφωνες με τις ανάγκες του χρήστη.

Αν και τα θεματικά μοντέλα είναι μία καλή επιλογή μεθόδου και μπορεί να επεκταθεί ώστε να μοντελοποιεί πολλούς τύπους πληροφοριών, έχουν κάποιες αδυναμίες οι οποίες περιορίζουν την χρήση τους σε πραγματικές εφαρμογές sentiment ανάλυσης. Ένα κύριο σημείο είναι πως ένα θεματικό μοντέλο χρειάζεται ένα μεγάλο σύνολο δεδομένων καθώς επίσης και ένα σημαντικό αριθμό επεμβάσεων/βελτιώσεων ώστε να αποδίδει αξιόλογα αποτελέσματα. Επιπλέον, ενώ δεν είναι δύσκολο για ένα θεματικό μοντέλο για εντοπίσει πολύ γενικά και συχνά θέματα ή aspects από μία μεγάλη συλλογή κειμένων, είναι δύσκολο να εντοπίσει τοπικά συχνά και μη συχνά γενικά aspects. Τα τοπικά και συχνά aspects είναι, στις περισσότερες περιπτώσεις, τα περισσότερα χρήσιμα διότι είναι σχετικότερα με τις συγκεκριμένες οντότητες για τις οποίες ενδια-

φέρεται ο χρήστης. Οπότε μπορούμε να πούμε πως τα θεματικά μοντέλα δεν είναι αρκετά συγκεκριμένα για τις περισσότερες πρακτικές εφαρμογές τις sentiment ανάλυσης.

3.3.2.5 Ομαδοποίηση των aspects

Αφού πραγματοποιηθεί η εξαγωγή των aspects, πρέπει να ακολουθήσει η ομαδοποίησή τους σε κατηγορίες συνωνύμων. Κάθε κατηγορία αντιπροσωπεύει ένα συγκεκριμένο aspect. Σε κάθε γραπτό κείμενο, οι άνθρωποι συχνά χρησιμοποιούν διαφορετικές λέξεις και φράσεις για να περιγράψουν το ίδιο χαρακτηριστικό. Για παράδειγμα, *"call quality"* και *"voice quality"* αναφέρονται στο ίδιο aspect για τα κινητά τηλέφωνα. Οπότε είναι φανερό πως η ομαδοποίηση αυτών των φράσεων είναι σημαντική για την ακρίβεια και την απόδοση της διαδικασίας της ανάλυσης απόψεων στο aspect επίπεδο.

Για αυτήν την διεργασία μπορεί να χρησιμοποιηθεί το σημασιολογικό δίκτυο WordNet, καθώς επίσης και άλλα λεξικά ή "θησαυροί", αλλά αυτό δεν είναι επαρκές διότι πολλά συνώνυμα είναι domain dependent. Για παράδειγμα, *"movie"* και *"picture"* είναι συνώνυμα σε review κείμενα ταινιών, αλλά δεν είναι συνώνυμα στο domain των ψηφιακών συσκευών όπου η λέξη *"picture"* είναι περισσότερο συνώνυμη με την λέξη *"photo"*, και η λέξη *"movie"* με την λέξη *"video"*. Επίσης πολλές aspect εκφράσεις αποτελούνται από πολλαπλές λέξεις, οι οποίες δεν είναι εύκολο να διαχειριστούν με λεξικά.

3.3.3 Sentiment κατηγοριοποίηση στο aspect επίπεδο

Η sentiment κατηγοριοποίηση των aspects προσδιορίζει την πολικότητα των απόψεων οι οποίες εκφράζονται για aspects, θετική, αρνητική ή ουδέτερη. Υπάρχουν δύο κύριες προσεγγίσεις για αυτή την διαδικασία, η προσέγγιση της εποπτευόμενης μάθησης, και αυτή που βασίζεται σε κάποιο λεξικό.

Για την προσέγγιση την εποπτευόμενη μάθησης, η μάθηση μπορεί να πραγματοποιηθεί με τις ίδιες μεθόδους οι οποίες χρησιμοποιούνται για την sentiment κατηγοριοποίηση σε επίπεδο πρότασης ή παραγράφου. Το βασικό ζήτημα είναι πως θα προσδιοριστεί η βαθμολογία κάθε sentiment έκφρασης. Η κύρια μέθοδος είναι να πραγματοποιηθεί parsing για να προσδιοριστεί η sentiment πληροφορία και οι άλλες σχετικές πληροφορίες. Για παράδειγμα, στην έρευνα Jiang et al. (2011), χρησιμοποιήθηκε ένας parser εξαρτήσεων για την δημιουργία ενός συνόλου από aspect dependent features για την κατηγοριοποίηση. Μία παρόμοια προσέγγιση χρησιμοποιήθηκε στην έρευνα

Boiy and Moens (2009), στην οποία δίνεται βάρος σε κάθε feature με βάση την θέση του σε σχέση με το στόχο, το aspect, σε ένα parse tree. Για τις συγκριτικές εκφράσεις, η λέξη "than" ή παρόμοιες λέξεις μπορούν να χρησιμοποιηθούν, ώστε να τμηματοποιηθούν αυτές οι προτάσεις. Η εποπτευόμενη μάθηση εξαρτάται από τα δεδομένα εκπαίδευσης, οπότε ένα μοντέλο το οποίο έχει εκπαιδευτεί από labeled δεδομένα από ένα domain, συχνά αποδίδει μη ικανοποιητικά σε διαφορετικά domains.

Η προσέγγιση της διαδικασίας του προσδιορισμού της πολικότητας των απόψεων οι οποίες εκφράζονται για aspects, η οποία βασίζεται σε λεξικά μπορεί να αντιμετωπίσει μερικά από τα προβλήματα, τα οποία παρουσιάζονται, σε μία προσέγγιση εποπτευόμενης μάθησης. Για παράδειγμα, οι τεχνικές οι οποίες βασίζονται σε λεξικά έχουν δείξει πως αποδίδουν αρκετά καλά σε ένα μεγάλο αριθμό από domains, αντίθετα με τις τεχνικές που βασίζονται στην εποπτευόμενη μάθηση. Οι τεχνικές προσδιορισμού της πολικότητας των απόψεων, οι οποίες εκφράζονται για aspects, και βασίζονται σε λεξικά, είναι κυρίως μη εποπτευόμενες τεχνικές. Χρησιμοποιούν συνήθως ένα sentiment λεξικό (το οποίο περιέχει μία λίστα με λέξεις και εκφράσεις), σύνθετες εκφράσεις, κανόνες απόψεων, ή το parse tree της πρότασης, για να προσδιορίσουν την sentiment πολικότητα των απόψεων για κάθε aspect σε μία πρόταση. Επίσης λαμβάνουν υπόψιν όρους, οι οποίοι είναι sentiment shifters (λέξεις οι οποίες αντιστρέφουν την πολικότητα της άποψης), και άλλα στοιχεία τα οποία επηρεάζουν την πολικότητα των απόψεων.

Παρακάτω παρουσιάζεται μία απλή lexicon-based μέθοδος. Αυτή η μέθοδος είναι από την έρευνα Ding et al. (2008) και αποτελείται από τέσσερα βήματα. Η μέθοδος υποθέτει πως οι οντότητες και τα aspects είναι γνωστά:

1. Εντοπισμός των sentiment λέξεων και εκφράσεων. Για κάθε πρόταση η οποία περιέχει ένα ή περισσότερα aspects, σε αυτό το βήμα, εντοπίζονται οι sentiment λέξεις και φράσεις. Στις sentiment λέξεις θετικής πολικότητας αποδίδεται η sentiment βαθμολογία +1, ενώ σε αυτές με αρνητική πολικότητα η sentiment βαθμολογία -1. Για παράδειγμα, στην πρόταση *"The voice quality of this phone is not good, but the battery life is long."*, υπάρχουν δύο aspects, τα "voice quality" και "battery life". Η sentiment λέξη είναι το *"good"*, η οποία αναφέρεται στο aspect "voice quality", είναι θετικής πολικότητας και λαμβάνει την βαθμολογία +1. Για το δεύτερο aspect δεν μπορούμε ακόμα να διακρίνουμε κάποια πολικότητα.

2. Εφαρμογή των sentiment shifters: Sentiment shifters είναι λέξεις και φράσεις οι οποίες μπορούν να αντιστρέψουν τις sentiment πολικότητες. Υπάρχουν πολλοί τύποι τέτοιων shifters. Λέξεις άρνησης (negation words) όπως οι [not, never, none, nobody, nowhere, neither, and cannot] είναι οι πιο συχνά χρησιμοποιούμενοι. Για το παράδειγμα της πρότασης *"The voice quality of this phone is not good, but the battery life is long."*, στο οποίο η sentiment λέξη η οποία εντοπίστηκε στο παραπάνω βήμα ήταν η *"good"* και η οποία έλαβε την sentiment βαθμολογία +1, σε αυτό το βήμα, επειδή υπάρχει η λέξη *"not"* αντιστρέφεται η πολικότητα της sentiment βαθμολογίας από +1 σε -1. Πρέπει να τονιστεί πως η εμφάνιση ενός όρου στην πρόταση ο οποίος ανήκει στο σύνολο των sentiment shifters, δεν συνεπάγεται απαραίτητα την αντιστροφή της sentiment πολικότητας, για παράδειγμα *"not only . . . but also."*. Τέτοιες περιπτώσεις πρέπει να αντιμετωπίζονται με προσοχή, και πρέπει να ορίζονται ειδικοί κανόνες, όπως για παράδειγμα πρότυπα (pattern-based κανόνες), οι οποίοι να καθορίζουν την αντιστροφή της πολικότητας.

3. Διαχείριση των but-περιπτώσεων: Υπάρχουν επίσης λέξεις και εφράσεις οι οποίες δείχνουν αντίθεση, και οι οποίες χρειάζονται ειδική διαχείριση, διότι μπορεί να αντιστρέφουν επίσης την πολικότητα. Από αυτές, η πιο συχνά χρησιμοποιούμενη σε κείμενα είναι η λέξη *"but"*. Μία πρόταση η οποία περιέχει μία λέξη αντίθεσης μπορεί να διαχειριστεί με τον εξής κανόνα: οι sentiment πολικότητες πριν την λέξη αντίθεσης και μετά την λέξη αντίθεσης, είναι αντίθετες μεταξύ τους αν η πολικότητα κάποιας από τις δύο πλευρές δεν μπορεί να καθοριστεί. Το 'αν' στον παραπάνω κανόνα υπάρχει, διότι οι λέξεις αντίθεσης δεν χρησιμοποιούνται για να εκφράσουν πάντα αντίθεση. Για παράδειγμα *"Car-x is great, but Car-y is better."*. Όσον αφορά το παράδειγμα *"The voice quality of this phone is not good, but the battery life is long."*, μετά από αυτό το βήμα οι πολικότητες έχουν διαμορφωθεί ως εξής *"The voice quality of this phone is not good[-1], but the battery life is long[+1]"* εξαιτίας του "but". Παρατηρούμε εδώ πως αν και δεν υπάρχει sentiment λέξη στο δευτερο μέρος της πρότασης, έχει αποδοθεί θετική sentiment βαθμολογία, η εκφράζεται για το aspect "battery life". Εκτός από το λέξεις όπως το "but", υπάρχουν και φράσεις όπως οι "with the exception of," "except that," και "except for", οι οποίες μπορούν να αντιμετωπίσουν με παρόμοιο τρόπο.

4. Συγκέντρωση των opinion βαθμολογιών: Σε αυτό το βήμα εφαρμόζεται μία συνάρτηση συγκέντρωσης για να υπολογιστεί η τελική sentiment βαθμολογία κάθε aspect σε μία πρόταση. Έστω s η πρόταση, η οποία περιέχει ένα σύνολο από aspects $\{a_1, \ldots, a_m\}$ και ένα σύνολο από sentiment λέξεις και φράσεις $\{sw_1, \ldots, sw_n\}$ με τις sentiment βαθμολογίες τους όπως αυτές υπολογίστηκαν στα παραπάνω βήματα. Η sentiment πολικότητα ενός aspect a_i στην πρόταση s υπολογίζεται ως εξής:

$$score(a_i, s) = \sum_{sw_j \in s} \frac{sw_j.so}{dist(sw_j, a_i)}$$

όπου sw_j είναι μία sentiment λέξη ή φράση στην πρόταση s, $dist(sw_j, a_i)$ είναι η απόσταση μεταξύ του aspect a_i και της sentiment λέξης sw_j στην πρόταση s, και $sw_j.so$ είναι η sentiment πολικότητα της sw_j. Σε αυτόν τον υπολογισμό δίνεται μικρότερο βάρος σε sentiment λέξεις οι οποίες βρίσκονται "μακριά" από το aspect a_i. Εάν η τελική βαθμολογία ενός aspect είναι θετική, τότε η άποψη για το aspect a_i στην πρόταση s έχει θετική πολικότητα, ενώ αντίστοιχα εάν είναι αρνητική η άποψη στην πρόταση s για το a_i έχει αρνητική πολικότητα. Εάν είναι μηδέν, δεν υπάρχει άποψη.

Αυτός ο απλός αλγόριθμος έχει δείξει πως αποδίδει αρκετά καλά σε πολλές περιπτώσεις. Είναι ικανός να διαχειριστεί χωρίς πρόβλημα προτάσεις όπως *"Google is doing very well in this bad economy"*. Πρέπει να τονιστεί πως υπάρχουν αρκετές μέθοδοι συγκέντρωσης των απόψεων. Για παράδειγμα, στην έρευνα Hu and Liu (2004) απλά αθροίζονται οι sentiment βαθμολογίες όλων των sentiment λέξεων σε μία πρόταση, ενώ στην έρευνα Kim and Hovy (2004) πολλαπλασιάζονταν οι sentiment βαθμολογίες των sentiment λέξεων. Για να δίνει αυτή η μέθοδος πιο αποτελεσματική, μπορούμε να προσδιορίσουμε το σκοπό κάθε sentiment λέξης αντί της απόστασης από τα aspects. Σε αυτή την περίπτωση χρειάζεται να εντοπιστούν οι εξαρτήσεις μεταξύ sentiment λέξεων και aspect, όπως παρουσιάστηκε πιο πάνω στις εποπτευόμενες τεχνικές. Για παράδειγμα, στην έρευνα, Blair-Goldensohn et al. (2008) συνδυάζεται μία lexicon-based μέθοδος με εποπτευόμενη μάθηση.

3.4 Sentiment Λεξικό

Σε πολλές τεχνικές sentiment ανάλυσης χρησιμοποιούνται λίστες με λέξεις ή φράσεις οι οποίες συνήθως εκφράζουν θετική ή αρνητική πολικότητα στο λόγο. Αυτές οι λίστες ονομάζονται sentiment λεξικά ή opinion λεξικά. Οι λέξεις θετικής πολικότητας εκφράζουν επιθυμητές καταστάσεις ή ιδιότητες, ενώ οι αρνητικές μη επιθυμητές καταστάσεις ή ιδιότητες. Επίσης υπάρχουν sentiment εκφράσεις. Ένα sentiment λεξικό μπορεί να περιέχει μόνο λέξεις, ή μόνο εκφράσεις, ή λέξεις και εφράσεις μαζί, από μία ή και από τις δύο πολικότητες. Όταν σε μία τεχνική χρησιμοποιείται ένα sentiment λεξικό θα πρέπει να λαμβάνεται υπόψιν, όπως έχει τονιστεί και στα προηγούμενα κεφάλαια, πως οι sentiment λέξεις και φράσεις είναι πολλές φορές domain-dependent και context-dependent, καθώς και το γεγονός πως η ύπαρξη μιας sentiment λέξης στο κείμενο δεν συνεπάγεται απαραίτητα την ύπαρξη κάποιας άποψης. Ένα domain μπορεί να έχει πολλά contexts, και μία sentiment λέξη μπορεί να έχει διαφορετικές πολικότητες στα διάφορε contexts του ίδιου domain.

Πολλοί ερευνητές έχουν δημιουργήσει τέτοια λεξικά, μερικά από τα οποία είναι δημόσια διαθέσιμα:

- General Inquirer lexicon (Stone, 1968):
  (http://www.wjh.harvard.edu/~inquirer/spread sheet_guide.htm)
- Sentiment lexicon (Hu and Liu, 2004):
  (http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html)
- MPQA subjectivity lexicon (Wilson et al., 2005):
  (http://www.cs.pitt.edu/mpqa/subj_ lexicon.html)
- SentiWordNet (Esuli and Sebastiani, 2006):
  (http://sentiwordnet.isti.cnr.it/)
- Emotion lexicon (Mohammad and Turney, 2010):
  (http://www.purl.org/net/emolex)

Για την δημιουργία ενός sentiment λεξικού έχουν προταθεί πολλές τεχνικές. Υπάρχουν τρεις κύριες προσεγγίσεις, manual προσέγγιση (μη αυτόματη), dictionary-based προσέγγιση, and corpus-based προσέγγιση. Η manual προσέγγιση απαιτεί ανθρώπινη, χειροκίνητη, εργασία και όπως είναι φανερό είναι χρονοβόρα, αλλά μπορούμε να πούμε πως με αυτή την προσέγγιση εξασφαλίζουμε την απόδοση της ανθρώπινης ακρίβειας.

Συνήθως όμως χρησιμοποιείται σε συνδυασμό με αυτόματες προσεγγίσεις, ως το βήμα του τελικού ελέγχου των αποτελεσμάτων, διότι οι αυτόματες προσεγγίσεις έχουν συχνά λάθη.

Οι dictionary-based προσεγγίσεις χρησιμοποιούν ένα dictionary, όπως για παράδειγμα το σημασιολογικό δίκτυο WrodNet, για την δημιουργία του sentiment λεξικού. Τα dictionaries αυτής της μορφής διαθέτουν λίστες με τα συνώνυμα και αντώνυμα των λέξεων. Οπότε μία απλή τεχνική για την δημιουργία ενός sentiment λεξικού, είναι η επιλογή ενός μικρού συνόλου λέξεων από κάποιο dictionary (seeds) και η χρήση των λιστών με τα συνώνυμα και τα αντώνυμα αυτών των λέξεων για των εμπλουτισμό του λεξικού (οι νέες λέξεις προστίθενται στο λεξικό). Αυτή η διαδικασία γίνεται επαναληπτικά και σταματά όταν σε κάποια επανάληψη δεν έχουν προστεθεί νέες λέξεις στο λεξικό. Συνήθως στο τέλος της διαδικασίας πραγματοποιείται έλεγχος της λίστας των λέξεων του λεξικού για τον εντοπισμό λαθών. Το πλεονέκτημα αυτής της προσέγγισης είναι ο εύκολος και γρήγορος εντοπισμός μεγάλου αριθμού sentiment λέξεων. Πρέπει να τονιστεί πως οι sentiment λέξεις οι οποίες συγκεντρώνονται με αυτό τον τρόπο είναι domain-independent και context-independent. Είναι δύσκολο να χρησιμοποιηθεί μία dictionary-based προσέγγιση δημιουργίας ενός sentiment λεξικού το οποίο να περιέχει domain-dependent και context-dependent sentiment λέξεις.

Οι corpus-based προσεγγίσεις έχουν κυρίως δύο εφαρμογές. Η πρώτη είναι, με βάση μίας λίστα με seed sentiment λέξεις οι οποίες είναι domain-independent να βρεθούν sentiment λέξεις και η πολικότητά τους από μία domain-specific συλλογή κειμένων (corpus). Η δεύτερη είναι η προσαρμογή ενός domain-dependent sentiment λεξικού σε ένα συγκεκριμένο domain χρησιμοποιώντας μία συλλογή κειμένων από αυτό το domain (corpus). Πρέπει να τονιστεί πως αν και το το νέο λεξικό είναι domain-dependent, μία λέξη μπορεί να έχει θετική πολικότητα σε ένα context του domain και αρνητική σε ένα άλλο context του ίδιου domain. Οι corpus-based προσεγγίσεις μπορούν να χρησιμοποιηθούν επίσης για την δημιουργία domain-independent λεξικών (γενικού σκοπού), αν στη συλλογή των κειμένων (corpus) περιέχονται κείμενα από πολλά domains (αν και για αυτό το σκοπό είναι πιο αποδοτικές οι dictionary-based προσεγγίσεις).

3.5 Ανίχνευση του opinion spam

Λόγω της ραγδαίας αύξησης της χρήσης του διαδικτύου και των κοινωνικών δικτύων, ο αριθμός των διαθέσιμων απόψεων συνεχώς μεγαλώνει. Τα άτομα και οι οργανισμοί στηρίζουν συχνά τις αποφάσεις τους σε αυτές τις απόψεις. Τα άτομα αναζητούν απόψεις για την πραγματοποίηση κάποιας αγοράς ή για την επιλογή του σχήματος που πρέπει να στηρίξουν στις εκλογές. Οι οργανισμοί αναζητούν απόψεις χρηστών για τα προϊόντα τους, ώστε να βελτιώσουν το σχεδιασμό τους και την πολιτική προώθησής τους. Πολλές θετικές απόψεις για κάποια οντότητα συχνά σημαίνει έσοδα και φήμη, κάτι το οποίο δίνει ισχυρό κίνητρο στους ανθρώπους να ξεγελάσουν το σύστημα ανεβάζοντας πλαστές απόψεις ή reviews, για να προωθήσουν ή να δυσφημήσουν προϊόντα, υπηρεσίες, οργανισμούς, πρόσωπα, ακόμα και ιδέες, χωρίς να αποκαλύπτουν τις πραγματικές τους προθέσεις, ή τα άτομα ή τους οργανισμούς για τους οποίους κρυφά εργάζονται. Τέτοια άτομα ονομάζονται opinion spammers και οι ενέργειές τους ονομάζονται opinion spamming. Το opinion spamming για κοινωνικά και πολιτικά θέματα μπορεί να θεωρηθεί αρκετά επικίνδυνο, διότι μπορεί να στρεβλώσει απόψεις και να κινητοποιήσει τις μάζες σε παράνομες και ανήθικες κατευθύνσεις. Μπορούμε να πούμε πως όσο περισσότερο χρησιμοποιούνται οι απόψεις στα κοινωνικά δίκτυα, τόσο περισσότερο το opinion spamming θα γίνεται ανεξέλεγκτο και εξειδικευμένο. Αυτό το φαινόμενο πρέπει να αντιμετωπιστεί για να εξασφαλιστεί πως τα κοινωνικά μέσα θα συνεχίσουν να είναι μία έμπιστη πηγή απόψεων. Η αντιμετώπιση του φαινομένου του opinion spamming μπορεί να ξεκινήσει από την ανίχνευσή του και αυτό παρουσιάζει την πρόκληση για την δημιουργία συστημάτων ανίχνευσης του opinion spamming.

Η ανίχνευση του spam έχει μελετηθεί σε πολλά επιστημονικά πεδία. Το web spam και το email spam είναι δύο περισσότερο μελετημένα είδη spam. Η ανίχνευσης του opinion spam, αντίθετα με τις άλλες μορφές spam, είναι πολύ δύσκολη. Είναι δύσκολο, έως αδύνατο, να αναγνωριστούν ψεύτικες απόψεις απλά διαβάζοντας το κείμενο. Ο άνθρωπος, κάποιες φορές, βρίσκεται στην θέση να αναγνωρίζει το opinion spam, όταν προσπελαύνει κάποιο κείμενο, διότι διαθέτει πληροφορίες εκτός του κειμένου. Για παράδειγμα, κάποιος μπορεί να γράψει μία ειλικρινή κριτική για ένα καλό εστιατόριο, αλλά να την τοποθετήσει ως ψεύτικο review σε ένα όχι και τόσο αξιόλογο εστιατόριο, για να το προωθήσει. Είναι αδύνατο να αναγνωριστεί αυτό το ψεύτικο review χωρίς να ληφθούν υπόψιν πληροφορίες εκτός του review κειμένου, επειδή το ίδιο review κείμενο μπορεί να είναι ταυτόχρονα ειλικρινές αλλά και ψεύτικο.

Το opinion spamming εμφανίζεται όχι μόνο σε reviews, αλλά και σε άλλες μορφές στα κοινωνικά μέσα, όπως τα blogs, οι συζητήσεις στα forums, τα σχόλια, καθώς και στο Twitter. Ωστόσο, λίγη έρευνα έχει πραγματοποιηθεί για την ανίχνευση του opinion spamming στα τελευταία. Η περισσότερη ερευνητική δραστηριότητα, όσον αφορά το φαινόμενο του opinion spam, έχει γίνει σε review κείμενα. Στην έρευνα Jindal and Liu (2008) έχουν αναγνωριστεί τρία είδη opinion spam σε reviews:

- ψεύτικα reviews: Αυτά είναι μη αξιόπιστα reviews τα οποία έχουν γραφτεί με βάση τις αυθεντικές εμπειρίες του χρήστη από την χρήση κάποιων προϊόντων ή υπηρεσιών, αλλά με κρυφά κίνητρα. Συνήθως περιέχουν αβάσιμες θετικές απόψεις σχετικά με κάποιες οντότητες, με σκοπό την προώθηση αυτών, ή αβάσιμες αρνητικές απόψεις με σκοπό την προσβολή της φήμης τους. Τα πλαστά review κείμενα μπορούν να θεωρηθούν ως μία ειδική μορφή εξαπάτησης. Οι ερευνητές έχουν εντοπίσει πολλές ενδείξεις εξαπάτησης σε τέτοια κείμενα. Για παράδειγμα, οι έρευνες έχουν δείξει πως όταν οι άνθρωποι λένε ψέματα, συνηθίζουν να απομακρύνουν το εαυτό τους, χρησιμοποιώντας δεύτερο και τρίτο πρόσωπο στις εκφράσεις τους αντί για πρώτο. Επίσης χρησιμοποιούν συχνότερα λέξεις οι οποίες σχετίζονται με βεβαιότητα, για να αποκρύψουν το πλαστό και να τονίσουν το αληθές.

- reviews σχετικά με μάρκες ή κατασκευαστές μόνο: Αυτά τα review κείμενα δεν περιέχουν σχόλια σχετικά με τα προϊόντα ή υπηρεσίες, τα οποία υποτίθεται πως πρέπει να σχολιάζουν, αλλά επικεντρώνονται στο να σχολιάζουν τις μάρκες ή τους κατασκευαστές αυτών των προϊόντων ή των υπηρεσιών. Αν και πολλές φορές οι απόψεις μπορεί να είναι ειλικρινείς, αυτά τα reviews θεωρούνται spam, διότι απομακρύνονται από τον πραγματικό στόχο των απόψεων, καθιστώντας το κείμενο προκατειλημμένο.

- non-reviews: Αυτά τα κείμενα δεν είναι reviews. Συχνά αποτελούνται από κείμενα διαφημίσεων ή άσχετα κείμενα χωρίς απόψεις. Αυτό το

είδος spam κειμένων μπορεί εύκολα να αναγνωριστεί από τους χρήστες, αλλά και από τους αλγορίθμους αναγνώρισης.

Τα παλστά review κείμενα μπορεί να έχουν γραφτεί από πολλά διαφορετικά είδη ανθρώπων, για παράδειγμα φίλους, υπαλλήλους εταιρειών, ανταγωνιστές, ή και επιχειρήσεις οι οποίες παρέχουν υπηρεσίες παροχής ψεύτικων reviews. Στις άλλες μορφές κοινωνικών μέσων, δημόσιες ή ιδιωτικές εταιρείες ή πολιτικοί οργανισμοί προσλαμβάνουν άτομα τα οποία πραγματοποιούν opinion spamming, για να επηρεάσουν κρυφά συζητήσεις στα κοινωνικά μέσα και να διαδώσουν ψέματα και να παραπληροφορήσουν. Γενικά ένας spammer μπορεί να εργάζεται ατομικά ή να ανήκει σε μία ομάδα από spammers.

Ο στόχος της ανίχνευσης του opinion spam είναι αναγνωριστούν οι ψεύτικες απόψεις, οι spammers ή οι ομάδες των spammers. Οι τρεις αυτές διεργασίες είναι εξαρτημένες μεταξύ τους, διότι η επιτυχής αναγνώριση μίας οντότητας από το ένα είδος μπορεί να οδηγήσει στην αναγνώριση οντοτήτων στα άλλα δύο είδη. Για την αναγνώριση οντοτήτων από τα τρία παραπάνω είδη σε μία διαδικασία ανίχνευσης opinion spam σε review κείμενα, μπορούν να χρησιμοποιηθούν τα παρακάτω είδη δεδομένων:

- περιεχόμενο του review: Από το κείμενο κάθε review μπορούν να εξαχθούν γλωσσολογικά χαρακτηριστικά (features) όπως λέξεις και POS n-grams και άλλα συντακτικά και γλωσσολογικά στοιχεία, τα οποία να υποδεικνύουν εξαπάτηση. Ωστόσο, τα γλωσσολογικά features μπορεί να μην είναι αρκετά, διότι ένα παραπλανητικό review κείμενο μπορεί να δημιουργηθεί με τον ίδιο τρόπο με ένα αξιόπιστο. Για παράδειγμα, κάποιος μπορεί να γράψει ένα ψεύτικο θετικό review κείμενο για μη αξιόλογο εστιατόριο, με βάση τις πραγματικές εμπειρίες του από ένα καλό εστιατόριο.

- Metadata του review: Δεδομένα όπως βαθμολογίες με την μορφή stars περιέχουν συχνά μεταδεδομένα όπως το id του χρήστη που έδωσε την βαθμολογία, το χρόνο της βαθμολόγησης, την IP και την MAC διεύθυνση του υπολογιστή του χρήστη, την τοποθεσία (geolocation) του χρήστη, καθώς και την ακολουθία των clicks του χρήστη στην review ιστοσελίδα. Από αυτά τα δεδομένα μπορούν να εξαχθούν πρότυπα ασυνήθιστης συμπεριφοράς των reviewers και

ασυνήθιστης συμπεριφοράς βαθμολογιών. Για παράδειγμα, μπορεί να εντοπιστούν συμπεριφορές όπου ο χρήστης δίνει θετικές βαθμολογίες σε προϊόντα μίας συγκεκριμένης εταιρείας, ενώ πάντα βαθμολογεί αρνητικά προϊόντα της ανταγωνίστριας εταιρείας. Επίσης ασυνήθιστη συμπεριφορά είναι αν πολλαπλοί χρήστες βαθμολογούν ένα προϊόν από τον ίδιο υπολογιστή.

- πληροφορίες σχετικά με το προϊόν : Πληροφορίες σχετικά με την οντότητα η οποία βαθμολογείται όπως είδος του προϊόντος και ο αριθμός των πωλήσεων. Για παράδειγμα, αν ένα προϊόν έχει μεγάλο αριθμό πωλήσεων, αλλά έχει πολλές θετικές βαθμολογίες, είναι πολύ πιθανόν να υπάρχουν πολλές ψεύτικες βαθμολογίες.

Η ανίχνευση του opinion spamming μπορεί να προσεγγιστεί ως ένα πρόβλημα κατηγοριοποίησης δύο κλάσεων, ψεύτικα και αξιόπιστα. Με βάση αυτήν την προσέγγιση μπορούν να εφαρμοστούν τεχνικές εποπτευόμενης μάθησης. Το σύνολο των χαρακτηριστικών το οποίο μπορεί να χρησιμοποιηθεί είναι το εξής:

- features σχετικά με το review κείμενο: Αυτά τα χαρακτηριστικά μπορούν να αποτελούνται από λέξεις, n-grams, η συχνότητα αναφορά στο κείμενο ονομάτων εταιρειών, το ποσοστό των λέξεων που εκφράζουν άποψη, το μέγεθος των review κειμένων, ο αριθμός των helpful feedbacks.

- features σχετικά με τον reviewer: Αυτά τα χαρακτηριστικά μπορούν να αποτελούνται από το μέσο όρο των βαθμολογιών που έχει δώσει ο συγκεκριμένος χρήστης, ο λόγος του αριθμού των reviews τα οποία έχει γράψει ο συγκεκριμένος reviewer ως πρώτα reviews σε κάποια προϊόντα προς τον συνολικό αριθμό των reviews που έχει γράψει ο συγκεκριμένος χρήστης, καθώς επίσης και ο αριθμός των περιπτώσεων στις οποίες ο συγκεκριμένος reviewer είναι ο μόνος που έχει δώσει review σε κάποιο προϊόν.

- features σχετικά με το προϊόν: Αυτά τα χαρακτηριστικά μπορούν να αποτελούνται από τον αριθμό των πωλήσεων του συγκεκριμένου προϊόντος, το μέσο όρο και την τυπική απόκλιση των βαθμολογιών των reviews του προϊόντος.

Εκτός από τις τεχνικές εποπτευόμενης μάθησης, μπορούν να εφαρμοστούν και μη εποπτευόμενες τεχνικές για την ανίχνευση του opinion spamming. Στην έρευνα Lim et al. (2010), προτείνεται μία τεχνική η οποία προσπαθεί να αναγνωρίσει ασυνήθιστες συμπεριφορές των reviewers, χρησιμοποιώντας κάποια μοντέλα ασυνήθιστης συμπεριφοράς με βάση διαφορετικά review πρότυπα τα οποία υποδηλώνουν opinion spamming κάποιου τύπου. Κάθε μοντέλο αποδίδει μία βαθμολογία spamming συμπεριφοράς σε ένα reviewer, μετρώντας το βαθμό στον οποίο ο reviewer πραγματοποιεί spamming συμπεριφορές κάποιου τύπου. Όλες οι βαθμολογίες συνδυάζονται για να παράξουν την τελική spamming βαθμολογία του reviewer. Αυτή η μέθοδος επικεντρώνεται στο να εντοπίζει spammers και όχι ψεύτικα review κείμενα. Τα μοντέλα opinion spamming συμπεριφοράς είναι τα εξής:

- στοχοποίηση προϊόντος: καθορίζει το μοντέλο συμπεριφοράς του spammer ο οποίος παραποιεί τη βαθμολογία προϊόντων, ανεβάζοντας πλαστά reviews την κατάλληλη χρονική στιγμή.
- στοχοποίηση ομάδων: καθορίζει το μοντέλο συμπεριφοράς του spammer ο οποίος παραποιεί τις βαθμολογίες ενός συνόλου προϊόντων, τα οποία έχουν κάποια κοινά χαρακτηριστικά, σε σύντομο χρονικό διάστημα.
- γενική απόκλιση βαθμολογίας: καθορίζει το μοντέλο συμπεριφοράς του spammer ο οποίος προσπαθεί να προωθήσει κάποιο προϊόν. Όταν συμβαίνει αυτό, οι βαθμολογία του προϊόντος αποκλίνει σημαντικά από αυτή των άλλων reviewers.
- πρόωρης βαθμολόγηση: καθορίζει το μοντέλο συμπεριφοράς του spammer ο οποίος ανεβάζει ένα ψεύτικο review μόλις το προϊόν εμφανιστεί.

Επίσης, στην έρευνα Wu et al. (2010), προτείνεται μία μη εποπτευόμενη μέθοδος για την ανίχνευση πλαστών review κειμένων με βάση ένα κριτήριο παραμόρφωσης (distortion criterion). Η ιδέα είναι πως τα ψεύτικα review κείμενα θα παραμορφώνουν την συνολική κατάταξη δημοτικότητας μίας συλλογής οντοτήτων. Οπότε, διαγράφοντας ένα σύνολο τυχαία επιλεγμένων reviews δεν πρέπει να επηρεάζεται η κατάταξη των ονοτήτων, ενώ διαγράφοντας παραπλανητικά review κείμενα η κατάταξη των οντοτήτων θα διαφοροποιείται. Η παραμόρφωση μπορεί να μετρηθεί συγκρίνοντας τις κατατάξεις των οντοτήτων, πριν και μετά την διαγραφή, χρησιμοποιώντας rank correlation.

3.6 Αναζήτηση και ανάκτηση απόψεων

Η αναζήτηση στο διαδίκτυο έχει αποδειχθεί πως είναι μία πολύτιμη υπηρεσία. Παρόμοια μπορούμε να πούμε πως μία υπηρεσία αναζήτησης απόψεων θα είναι το ίδιο χρήσιμη. Η υπηρεσία θα μπορούσε να προσφέρει αναζήτηση απόψεων σχετικά με μία συγκεκριμένη οντότητα ή κάποιο χαρακτηριστικό μιας οντότητας, ή αναζήτηση απόψεων συγκεκριμένων ατόμων ή οργανισμών σχετικά με κάποια οντότητα ή κάποιο χαρακτηριστικό μιας οντότητας.

Παρόμοια με την αναζήτηση στο διαδίκτυο, η αναζήτηση απόψεων χρειάζεται να πραγματοποιεί δύο διαδικασίες. Πρώτον, αναζήτηση κειμένων ή προτάσεων οι οποίες είναι σχετικές με το ερώτημα που τίθεται, όπως αυτή πραγματοποιείται στην απλή αναζήτηση στο διαδίκτυο αυτή την στιγμή. Δεύτερον, εξαγωγή του υποκειμενικού περιεχομένου από το κείμενο, και σήμανση αυτού ως θετικής πολικότητας ή αρνητικής. Αυτή η διαδικασία αφορά τον τομέα της εξαγωγής και επεξεργασίας απόψεων. Στην κλασικές μηχανές αναζήτησης, οι ιστοσελίδες βαθμολογούνται με βάση την αυθεντία και την σχετικότητα με το θέμα. Η βασική αρχή πάνω στην οποία βασίζονται οι μηχανές αναζήτησης, είναι πως οι ιστοσελίδες οι οποίες βρίσκονται ψηλά στην κατάταξη περιέχουν επαρκείς πληροφορίες για να καλύψουν την ανάγκη πληροφόρησης του χρήστη, οπότε αρκεί να ανακτηθούν μόνο αυτές, ή μόνο η πρώτη σ στην κατάταξη. Αυτό απορρέει από το ότι ένα γεγονός ισοδυναμεί με οποιοδήποτε αριθμό ίδιων γεγονότων. Οπότε αν το πρώτο αποτέλεσμα στην κατάταξη περιέχει την πληροφορία που αναζητά ο χρήστης, δεν χρειάζεται η προσπέλαση των υπόλοιπων αποτελεσμάτων. Για ένα σύστημα αναζήτησης απόψεων όμως αυτό δεν είναι αρκετό, διότι τα πρώτα αποτελέσματα περιέχουν μόνο ένα μέρος των απόψεων, έστω για μία οντότητα, ή μόνο

θετικές/αρνητικές απόψεις, ή απόψεις μόνο από μιας συγκεκριμένη κοινότητα. Ιδανικά, θα πρέπει να προσπελαστούν όλα τα αποτελέσματα για να δημιουργηθεί η πραγματική εικόνα της κατανομής των απόψεων (θετικές-αρνητικές). Μία απλή λύση θα ήταν να δημιουργούνται δύο λίστες αποτελεσμάτων, μία για θετικές και μία για αρνητικές απόψεις, αλλά και πάλι το πρόβλημα της προσπέλασης όλων των αποτελεσμάτων παραμένει.

Η τρέχουσα έρευνα στο τομέα της ανάκτησης απόψεων αντιμετωπίζει την διαδικασία ως μία εργασία δύο σταδίων. Στο πρώτο στάδιο, τα κείμενα βαθμολογούνται με βάση την σχετικότητά τους με το θέμα του ερωτήματος. Στο δεύτερο στάδιο, τα υποψήφια σχετικά κείμενα επαναξιολογούνται με βάση τις βαθμολογίες των απόψεων που φέρουν. Οι βαθμολογίες των απόψεων μπορούν να αποδοθούν χρησιμοποιώντας τεχνικές οι οποίες βασίζονται στην μηχανική μάθηση, όπως Support Vector Machines, ή τεχνικές οι οποίες χρησιμοποιούν κάποιο sentiment λεξικό, ή κάποια άλλη τεχνική εξαγωγής υποκειμενικού περιεχομένου. Πιο εξειδικευμένα μοντέλα πραγματοποιούν και τις δύο διαδικασίες ταυτόχρονα.

Ένα παράδειγμα opinion search συστήματος παρουσιάζεται στην έρευνα (Zhang and Yu, 2007). Αυτό το σύστημα αποτελείται από δύο μέρη. Το πρώτο αναλαμβάνει την ανάκτηση σχετικών κειμένων για το ερώτημα που τίθεται. Το δεύτερο μέρος αναλαμβάνει την εξαγωγή και κατηγοριοποίηση κειμένων σε υποκειμενικά πολωμένα και ουδέτερα. Τα κείμενα που περιέχουν υποκειμενικό περιεχόμενα, κατηγοριοποιούνται σε θετικά πολωμένα, αρνητικά πολωμένα, ή μεικτά (περιέχουν και θετικές και αρνητικές απόψεις). Πιο συγκεκριμένα τα δύο μέρη του συστήματος:

Retrieval component: Αυτό είναι το πρώτο κομμάτι του συστήματος, το οποίο πραγματοποιεί την διαδικασία της κλασικής ανάκτησης πληροφορίας (IR). Για την δημιουργία του ερωτήματος χρησιμοποιούνται λέξεις κλειδιά, ονόματα οντοτήτων, καθώς και λέξεις και φράσεις οι οποίες καθορίζουν το θεματική ενότητα της αναζήτησης. Κατά την επεξεργασία του ερωτήματος, αρχικά αναγνωρίζονται και αποσαφηνίζονται οι θεματικές ενότητες του ερωτήματος του χρήστη, και το ερώτημα εμπλουτίζεται με συνώνυμα. Ύστερα, τίθεται το ερώτημα και ανακτούνται τα έγγραφα. Στα έγγραφα που έχουν ανακτηθεί, πραγματοποιείται αναγνώριση των θεματικών ενοτήτων, και εκτελείται pseudo-feedback για την αυτόματη εξαγωγή σχετικών λέξεων από τα top-ranked κείμενα στην κατάταξη και τον εμπλουτισμό του ερωτήματος. Τέλος, υπολογί-

ζεται μία βαθμολογία σχετικότητας για κάθε κείμενο με βάση το εμπλουτισμένο ερώτημα, χρησιμοποιώντας τις λεξεις κλειδιά και τις θεματικές ενότητες.

Opinion classification component: Αυτό είναι το δεύτερο κομμάτι του συστήματος, το οποίο, πρώτον, πραγματοποιεί κατηγοριοποίηση των ανακτηθέντων κειμένων σε υποκειμενικά πολωμένα και ουδέτερα, και δεύτερον, πραγματοποιεί κατηγοριοποίηση των υποκειμενικά πολωμένων κειμένων σε θετικά, αρνητικά, ή μεικτά πολωμένα. Και για τις δύο διαδικασίες, το σύστημα χρησιμοποιεί εποπτευόμενη μάθηση. Για την πρώτη διαδικασία προτείνεται ένας SVM κατηγοριοποιητής. Κάθε κείμενο αποδομείται σε προτάσεις τις οποίες κατηογοριοποιεί ο SVM κατηγοριοποιητής σε υποκειμενικά πολωμένες και ουδέτερες. Στις προτάσεις οι οποίες φέρουν απόψεις ο κατηγοριοποιητής αποδίδει επίσης μία βαθμολογία. Ένα κείμενο είναι υποκειμενικά πολωμένο, άν περιέχει τουλάχιστον μία υποκειμενικά πολωμένη πρόταση. Για να αποφασιστεί η πολικότητα ενός εγγράφου, χρησιμοποιείται ένας δεύτερος κατηγοριοποιητής. Τα κείμενα τελικά βαθμολογούνται με βάση την συνολική βαθμολογία των προτάσεων οι οποίες φέρουν απόψεις, καθώς και την βαθμολογία σχετικότητας του κειμένου με το ερώτημα.

3.7 Σύνοψη απόψεων

Σε μία διαδικασία sentiment ανάλυσης, πρέπει να προσπελαστούν οι απόψεις πολλών χρηστών διότι, λόγω της υποκειμενικής φύσης των απόψεων, αναλύοντας μόνο τις απόψεις ενός χρήστη ή ενός μικρού συνόλου χρηστών, δεν αναγνωρίζεται η συνολική εικόνα της κατανομής των απόψεων. Εξαιτίας αυτού είναι αναγκαίο να πραγματοποιηθεί κάποιο είδος σύνοψης των πληροφοριών των απόψεων. Η σύνοψη των απόψεων μπορεί να έχει πολλές μορφές, για παράδειγμα σύνοψη συγκεκριμένης δομής ή σύνοψη με τη μορφή σύντομου κειμένου. Η σύνοψη απόψεων πρέπει να περιέχει τις απόψεις για τις οντότητες και για τα aspects αυτών, και πρέπει να αποδίδει την ποσοτική κατανομή των απόψεων. Η ποσοτική κατανομή των απόψεων είναι σημαντική, διότι, για παράδειγμα, ένα ποσοστό που δείχνει πως το 20% των χρηστών έχει θετική άποψη για κάποιο προϊόν, είναι διαφορετικό από ένα ποσοστό που δείχνει πως το 80% των χρηστών έχει θετική άποψη.

Το πιο διαδεδομένο μοντέλο σύνοψης απόψεων είναι αυτό της aspect-based opinion σύνοψης ή feature-based opinion σύνοψης (Hu and Liu, 2004; Liu et al., 2005). Η aspect-based σύνοψη απόψεων έχει δύο χαρακτηριστικά, Πρώτον, συλλαμ-

βάνει τους στόχους των απόψεων, τις οντότητες και τα aspects αυτών, καθώς και τις πολικότητες των απόψεων. Δεύτερον, είναι ποσοτική, το οποίο σημαίνει πως αποδίδει τα ποσοστά των θετικών και αρνητικών απόψεων των χρηστών σχετικά με τις οντότητες και τα aspects αυτών. Με αυτό το είδος της σύνοψης είναι εύκολο να εντοπιστεί η συνολική κατανομή των απόψεων των χρηστών για τις οντότητες, καθώς επίσης και η κατανομή των απόψεων για τα aspects των οντοτήτων. Οι περισσότερες έρευνες στο opinion mining χρησιμοποιούν αυτό το μοντέλο σύνοψης ή παρόμοιο. Αυτό το μοντέλο εφαρμόζεται επίσης ευρέως σε συστήματα, τα οποία αποτελούν σημαντικές εφαρμογές sentiment ανάλυσης, του πραγματικού κόσμου, όπως τα συστήματα αναζήτησης προϊόντων της Microsoft Bing και Google Product Search.

Κάποιοι ερευνητές επιλέγουν την παραδοσιακή μορφή σύνοψης κειμένου, όπου παράγεται ένα σύντομο κείμενο σύνοψης. Ωστόσο, η σύνοψη απόψεων, είναι αρκετά διαφορετική από την σύνοψη κειμένου ή συλλογής κειμένων. Η σύνοψη απόψεων πρέπει να επικεντρώνεται στις οντότητες και τα aspects αυτών, καθώς και στο sentiment πολικότητες των απόψεων για αυτές. Η παραδοσιακή σύνοψη κειμένου παράγει ένα σύντομο κείμενο από ένα ή περισσότερα μεγάλα κείμενα, εξάγοντας από αυτά τις σημαντικές εκφράσεις. Αυτή η μορφή σύνοψης απόψεων δεν περιέχει ποσοτικές τιμές για την κατανομή των θετικών και αρνητικών απόψεων.

4 Βελτιώνοντας την opinion-based κατάταξη των οντοτήτων

Σε αυτή την ενότητα παρουσιάζεται η έρευνα που πραγματοποίησε ο γράφων μαζί με τον επίκουρο καθηγητή του τμήματος Μηχανικών Η/Υ και Πληροφορικής, του πανεπιστήμιο Πατρών, Χρήστο Μακρή, (Christos Makris, Panagiotis Panagopoulos, 2014). Σε αυτή την έρευνα παρουσιάζουμε αναλυτικά το πρόβλημα της αξιολόγησης οντοτήτων χρησιμοποιώντας τις απόψεις που υπάρχουν σε review κείμενα χρηστών, και αναφέρουμε γνωστές τεχνικές προσέγγισής της λύσης του. Ακόμα προτείνουμε νέες τεχνικές για την βελτίωση της απόδοσης της διαδικασίας της αξιολόγησης των οντοτήτων, και αναφέρουμε τα πειράματα και τις μετρήσεις που πραγματοποιήθηκαν, όπως επίσης και τις παρατηρήσεις που εξήχθησαν.

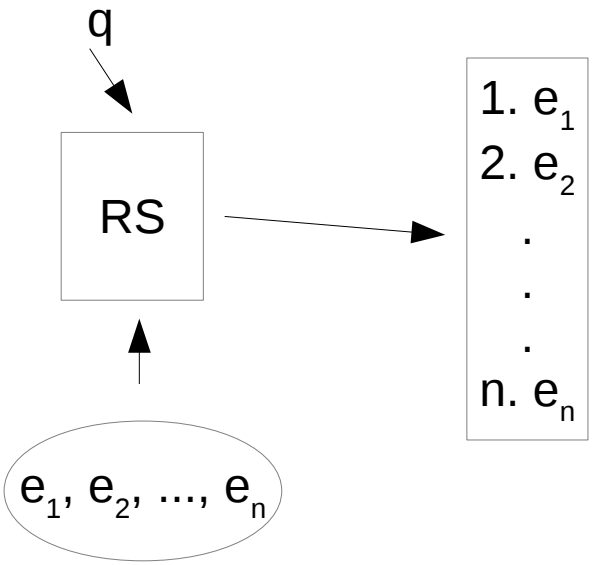
Όπως έχουμε αναφέρει και σε προηγούμενες ενότητες σε αυτό το κείμενο, στις μέρες μας παρατηρείται μεγάλη ανάπτυξη του όγκου των απόψεων και των reviews στο διαδίκτυο, τα οποία περιλαμβάνουν reviews για προϊόντα και υπηρεσίες, καθώς και απόψεις για γεγονότα και άτομα. Για προϊόντα ειδικότερα, υπάρχουν χιλιάδες reviews χρηστών, τα οποία οι καταναλωτές συνήθως συμβουλεύονται πριν προχωρήσουν σε κάποια αγορά. Στην ερευνά μας ακολουθούμε την ιδέα της μετατροπής του προβλήματος της αξιολόγησης οντοτήτων σε πρόβλημα ταιριάσματος προτιμήσεων. Αυτή η ιδέα μας επιτρέπει να προσεγγίσουμε την λύση του προβλήματος, χρησιμοποιώντας οποιοδήποτε κλασικό μοντέλο ανάκτησης πληροφορίας. Χτίζοντας πάνω σε αυτό το framework, εξετάζουμε τεχνικές οι οποίες χρησιμοποιούν sentiment και clustering πληροφορίας, και προτείνουμε το naïve consumer μοντέλο. Πραγματοποιούμε μετρήσεις σε δύο σύνολα πειραμάτων, οι οποίες δείχνουν πως οι τεχνικές που προτείνουμε δίνουν αξιόλογα αποτελέσματα.

4.1 Αξιολόγηση οντοτήτων χρησιμοποιώντας τις απόψεις χρηστών

Η ραγδαία αύξηση των τεχνολογιών του διαδικτύου και των κοινωνικών δικτύων, έχει δημιουργήσει ένα τεράστιο όγκο από review κείμενα για προϊόντα και υπηρεσίες, καθώς και απόψεις για γεγονότα και άτομα. Οι απόψεις αποτελούν σημαντικό μέρος της ανθρώπινης δραστηριότητας επειδή επηρεάζουν την συμπεριφορά μας στην λήψη αποφάσεων. Οι καταναλωτές ενημερώνονται από τα reviews άλλων χρηστών, πριν προχωρήσουν στην αγορά κάποιου προϊόντος ή υπηρεσίας. Οι επιχειρήσεις επίσης

θέλουν να είναι σε θέση να γνωρίζουν τις απόψεις των καταναλωτών για τα προϊόντα τους, ή τις υπηρεσίες τους, και να προσαρμόζουν τις στρατηγικές προώθησης και την περεταίρω ανάπτυξή τους.

Ο καταναλωτής, ωστόσο, για να διαμορφώσει μία συνολική εικόνα για ένα σύνολο οντοτήτων μίας συγκεκριμένης κατηγορίας, πρέπει να προσπελάσει και να επεξεργαστεί πολλά reviews. Από αυτά τα reviews πρέπει να εξάγει όσο το δυνατόν περισσότερες απόψεις, έτσι ώστε να διαμορφώσει μία τελική εικόνα για κάθε μία από τις οντότητες της συγκεκριμένης κατηγορίας, και να αναγνωρίσει αυτά τα οποία ξεχωρίζουν. Είναι ξεκάθαρο πως το πλήθος των απόψεων αποτελεί μία πρόκληση για τον καταναλωτή, καθώς επίσης και για τα συστήματα αξιολόγησης και κατάταξης οντοτήτων. Για παράδειγμα, έστω ένα σύνολο από συσκευές smartphones, ως οντότητες οι οποίες πρέπει να αξιολογηθούν. Για κάθε μία συσκευή υπάρχει ένα σύνολο από review κείμενα τα οποία είναι γραμμένα από χρήστες με βάση τις εμπειρίες τους από τα συγκεκριμένα προϊόντα. Ένα υποψήφιος αγοραστής θέλει να γνωρίζει αυτές τις απόψεις πριν προχωρήσει στην πραγματοποίηση της αγοράς κάποιου smartphone. Ο καταναλωτής συνήθως αναζητούν συγκεκριμένα χαρακτηριστικά σε ένα προϊόν. Σε αυτό το παράδειγμα, θα μπορούσε να ενδιαφέρεται για συσκευές οι οποίες έχουν android λειτουργικό σύστημα, multi-touch οθόνη, ή αξιόλογη διάρκεια μπαταρίας. Συνήθως αυτός ο καταναλωτής ανατρέχει σε review κείμενα άλλων χρηστών αναζητώντας συσκευές, οι απόψεις για τις οποίες σε συγκεκριμένα χαρακτηριστικά είναι θετικές. Όμως όπως είναι φανερό αυτή η διαδικασία είναι κουραστική και απαιτεί την προσοχή και το χρόνο του χρήστη.



Οπότε, αναγνωρίζουμε πως η ανάπτυξη υπολογιστικών τεχνικών, οι οποίες βοηθούν τον χρήστη στην εξόρυξη και αποδοτική χρήση όλων των απόψεων, είναι σημαντική και ενδιαφέρον ερευνητική πρόκληση. Στην έρευνα (Ganesan and ChengXiang, 2012) παρουσιάζεται ένα η δομή για ένα opinion-based σύστημα αξιολόγησης οντοτήτων. Η ιδεά, η οποία προτείνεται σε αυτή την έρευνα, είναι πως κάθε οντότητα αντιπροσωπεύεται από το κείμενο όλων των review κειμένων τα οποία αναφέρονται σε αυτήν, και πως οι χρήστες, ενός τέτοιου συστήματος, εκφράζουν τις προτιμήσεις τους σε διάφορα χαρακτηριστικά (aspects) κατά την διαδικασία της αξιολόγησης. Οπότε μπορούμε να περιμένουμε πως το ερώτημα ενός χρήστη θα αποτελείται από προτιμήσεις σε πολλαπλά aspects. Για το παράδειγμα με τις συσκευές smartphones το οποίο αναφέρεται παραπάνω, το ερώτημα του χρήστη θα μπορούσε να είναι "*android, multi-touch, good battery life*". Με αυτό το ερώτημα, ο χρήστης εκφράζει την προτίμησή του σε τρία διαφορετικά χαρακτηριστικά της οντότητας smartphone, το λειτουργικό σύστημα, την οθόνη και την μπαταρία. Προσεγγίζοντας το πρόβλημα της αξιολόγησης οντοτήτων ως πρόβλημα ταιριάσματος προτιμήσεων, μπορούμε να χρησιμοποιήσουμε, ώστε να το λύσουμε, οποιοδήποτε standard μοντέλο ανάκτησης πληροφορίας. Δοθέντος ενός ερωτήματος χρήστη, το οποίο αποτελείται από λέξεις και εφράσεις οι οποίες αποτυπώνουν τα επιθυμητά χαρακτηριστικά τα οποία πρέπει να έχει μία οντότητα, μπορούμε να αξιολογήσουμε τις οντότητες με βάση το πόσο καλά οι απόψεις για αυτές ταιριάζουν τις προτιμήσεις του χρήστη.

Πάνω σε αυτή την ιδέα, αναπτύσσουμε σχήματα τα οποία χρησιμοποιούν clustering και sentiment πληροφορία, για τις απόψεις σε review κείμενα. Επίσης προτείνουμε το naïve consumer μοντέλο, ως ένα setup το οποίο χρησιμοποιεί πληροφορίες από το διαδίκτυο για να συγκεντρώσει την γνώση της κοινότητας για το ποια χαρακτηριστικά (aspects) είναι περισσότερο σημαντικά για να χρησιμοποιηθούν στην αξιολόγηση οντοτήτων μίας συγκεκριμένης κατηγορίας.

Το setup το οποίο παρουσιάζεται στην έρευνα (Ganesan and ChengXiang, 2012), είναι μία προσέγγιση ανάκτησης πληροφορίας. Χρησιμοποιεί την σημαντικότητα των aspect λέξεων κλειδιών, και ορισμένων λέξεων θετικής πολικότητας, για τα review κείμενα έτσι ώστε να επιβραβεύσει, στην κατάταξη, οντότητες για τις οποίες υπάρχουν πολλές θετικές απόψεις. Στις δικά μας σχήματα, χρησιμοποιούμε δύο γνωστές μη

εποπτευόμενες τεχνικές sentiment analysis, και επιχειρούμε να χρησιμοποιήσουμε την sentiment πληροφορία η οποία υπάρχει στις απόψεις οι οποίες εκφράζονται στα review κείμενα, για τα συγκεκριμένα χαρακτηριστικά. Αυτές οι τεχνικές λαμβάνουν υπόψιν τις θετικές αλλά και τις αρνητικές απόψεις, οι οποίες εκφράζονται για μία οντότητα, στην διαδικασία τις κατάταξης των οντοτήτων. Επίσης ερευνούμε την συμπεριφορά της διαδικασίας της κατάταξης οντοτήτων, ακολουθώντας την προσέγγιση της ανάκτησης πληροφορίας. Επιδιώκουμε να χρησιμοποιήσουμε τις sentiment βαθμολογίες των χαρακτηριστικών έτσι ώστε να αναγνωρίσουμε παρόμοιες συμπεριφορές αξιολόγησης ανάμεσα στους χρήστες, και να χρησιμοποιήσουμε αυτή την πληροφορία για την δημιουργία μιας καλύτερης κατάταξης οντοτήτων. Τέλος, τονίζουμε πως δεν είναι όλα τα aspects το ίδιο σημαντικά όταν χρησιμοποιούνται στην κατάταξη οντοτήτων. Για αυτό το λόγο παρουσιάζουμε το Naïve Consumer Model, ως ένα μη εποπτευόμενο σχήμα το οποίο χρησιμοποιεί πληροφορίες από το διαδίκτυο για να διαμορφώσει ένα βάρος σημαντικότητας για κάθε ένα από τα χαρακτηριστικά τα οποία χρησιμοποιούνται για την κατάταξη των οντοτήτων.

4.2 Σχετική Εργασία και προτάσεις βελτίωσης

Σε αυτή την έρευνα καταπιανόμαστε με το πρόβλημα της δημιουργίας μιας λίστας-κατάταξης οντοτήτων χρησιμοποιώντας review κείμενα χρηστών. Για να προσεγγίσουμε την λύση αυτού του προβλήματος, κινούμαστε στην κατεύθυνση της aspect-oriented, ή feature-based, εξόρυξης απόψεων, όπως προτείνεται στην (Ganesan and ChengXiang, 2012). Σε αυτή την προσέγγιση, κάθε οντότητα αναπαρίσταται με το συνολικό κείμενο από όλα τα διαθέσιμα reviews που αναφέρονται σε αυτή την οντότητα, και οι χρήστες εκφράζουν τα ερωτήματά τους ως προτιμήσεις σε πολλαπλά aspects. Οι οντότητες αξιολογούνται με βάση πόσο καλά οι απόψεις, που εκφράζονται στα reviews, ταιριάζουν τις προτιμήσεις του χρήστη.

Όσον αφορά τα reviews, μεγάλο μέρος της έρευνας έχει γίνει στην κατηγοριοποίησή τους σε θετικά και αρνητικά με βάση την συνολική sentiment πληροφορία που περιέχουν (document level sentiment classification). Έχουν προταθεί πολλές εποπτευόμενες τεχνικές (Pang, Lee and Vaithyanathan, 2002), (Gamon, 2005), (Pang and Lee, 2004), (Dave, Lawrence and Pennock, 2003), μη-εποπτευόμενες (Turney and Littman, 2002), (Nasukawa and Yi, 2003), καθώς επίσης και υβριδικές τεχνικές (Pang and Lee, 2005), (Prabowo and Thelwall, 2009).

Σχετική ερευνητική περιοχή είναι και αυτή της ανάκτησης απόψεων (Liu, 2012). Ο στόχος της ανάκτησης απόψεων είναι η αναγνώριση κειμένων τα οποία περιέχουν απόψεις. Ένα σύστημα ανάκτησης απόψεων δημιουργείται συνήθως στο επίπεδο πάνω από τα κλασικά μοντέλα ανάκτησης, όπου σχετικά αρχικά κείμενα ανακτώνται και ύστερα εφαρμόζεται κάποια τεχνική ανάλυσης απόψεων για την εξαγωγή μόνο των κειμένων τα οποία περιέχουν απόψεις. Στην δική μας προσέγγιση, υποθέτουμε πως έχουμε ήδη διαθέσιμα τα κείμενα τα οποία περιέχουν απόψεις.

Ακόμα μία σχετική ερευνητική περιοχή είναι αυτή του Expert Finding. Σε αυτή την ερευνητική περιοχή, στόχος είναι η ανάκτηση μιας λίστας κατάταξης ατόμων, τα οποία είναι ειδικοί (experts) σε ένα συγκεκριμένο θέμα (Fang and Zhai, 2007), (Baeza-Yates and Ribeiro-Neto, 2011), (Wang, Lu and Zhai, 2010). Εμείς προσπαθούμε να ανακτήσουμε μία λίστα κατάταξης οντοτήτων, αλλά αντί να αξιολογούμε τις οντότητες με βάση το πόσο καλά ταιριάζουν σε ένα θέμα, χρησιμοποιούμε τις απόψεις για τις οντότητες, και παρατηρούμε το πόσο καλά ταιριάζουν τις προτιμήσεις του χρήστη.

Επίσης, υπάρχει αρκετή ερευνητική δραστηριότητα στην κατεύθυνση της πρόβλεψης aspect based βαθμολογιών (Wang et al.,2010), (Lu, Zhai, and Sundaresan, 2009), (Snyder and Barzilay, 2007). Αυτή η κατεύθυνση είναι επίσης σχετική με την δική μας, διότι εφαρμόζοντας aspect based ανάλυση, μπορούμε να προβέψουμε τις βαθμολογίες για τα διάφορα aspect από τα review κείμενα. Οπότε μπορούμε να κατατάξουμε τις οντότητες με βάση τις βαθμολογίες τους στα aspects, τα οποία ενδιαφέρουν, κάθε φορά, τον χρήστη.

Τέλος σε ένα από τα σχήματα που προτείνουμε για την βελτίωση της απόδοσης της κατάταξης οντοτήτων με βάση review χρηστών, εξετάζουμε το naïve consumer μοντέλο, ως ένα μη εποπτευόμενο σχήμα το οποίο χρησιμοποιεί πληροφορίες από το διαδίκτυο για να διαμορφώσει ένα βάρος σημαντικότητας για κάθε ένα από τα χαρακτηριστικά τα οποία χρησιμοποιούνται για την κατάταξη των οντοτήτων. Επιλέγουμε να χρησιμοποιήσουμε μία φόρμουλα η οποία έχει κάποια ομοιότητα με αυτές οι οποίες χρησιμοποιούνται στην item response theory (ITL) (Hambleton, Swaminathan, and Rogers, 1991) και στο Rasch μοντέλο (Rasch), (Rasch, 1960/1980). Η item response θεωρία χρησιμοποιείται για τον σχεδιασμό, την ανάλυση και την βαθμολογία εξετάσεων, ερωτηματολογίων και παρόμοιων οργάνων μέτρησης δεξιοτήτων, στάσεων και άλλων μεταβλητών. Η μαθηματική θεωρία στην οποία στηρίζεται το Rasch μοντέλο είναι μία ειδική περίπτωση της item response theory. Υπάρχουν προσεγγίσεις στο text

mining οι οποίες χρησιμοποιούν το Rasch μοντέλο και την item response theory, όπως οι (Tikves et al., 2012), (He, 2013), (Drehmer, Belohlav and Coye , 2000). Για παράδειγμα στην έρευνα (Tikves et al., 2012), οι συγγραφείς χρησιμοποίησαν την κλίμακα Guttman (Guttman 1950) και το Rasch μοντέλο για να κατατάξουν θρησκευτικές οργανώσεις με βάση την ποιότητα του βαθμού του ριζοσπαστισμού τους στην ρητορική τους. Στην μοντελοποίησή τους τα υποκείμενα αντιστοιχούν σε ομάδες θρησκευτικών οργανώσεων και τα αντικείμενα αντιστοιχούν σε επιλεγμένα σύνολα λέξεων κλειδιών. Ωστόσο, η προσέγγισή μας διαφοροποιείται στις επιλεγμένες μετρικές και στην εφαρμογή της μεθοδολογίας, και δεν χρησιμοποιούμε καμία από τις ικανότητες μοντελοποίησης των παραπάνω θεωριών.

Τέλος, προσεγγίσεις οι οποίες λαμβάνουν υπόψιν τους το βάρος των χαρακτηριστικών, αναφέρονται στο (Liu, 2012), και συγκεκριμένα στην έρευνα (Yu, Jianxing, Zha, Wang, Chua, 2011), ωστόσο η τεχνική μας είναι απλούστερη και εντάσσεται στο setup το οποίο παρουσιάζεται στην έρευνα (Ganesan and ChengXiang, 2012).

4.2.1 Προτάσεις σχημάτων για την βελτίωση της απόδοσης

Στην έρευνα (Ganesan and ChengXiang, 2012), παρουσιάζεται ένα setup για την κατάταξη οντοτήτων, όπου οι οντότητες αξιολογούνται με βάση το πόσο καλά οι απόψεις, οι οποίες εκφράζονται στα review κείμενα ταιριάζουν τις προτιμήσεις του χρήστη. Σε αυτή την έρευνα μελετήθηκε η χρήση διάφορων state-of-the-art μοντέλων ανάκτησης πληροφορίας, για αυτήν την διαδικασία. Πιο συγκεκριμένα, δοκιμάστηκαν οι φόρμουλες ανάκτησης πληροφορίας, BM25 (Baeza-Yates and Ribeiro-Neto, 2011), (Robertson and Zaragoza, 2009), Dirichlet (Zhai and Lafferty, 2001), και η PL2 (Amati, and van Rijsbergen, 2002). Επίσης προτάθηκε η χρήση κάποιων νέων επεκτάσεων για αυτά τα μοντέλα, όπως το query aspect modeling (QAM) και το opinion expansion. Με αυτές τις επεκτάσεις δίνεται η δυνατότητα στα κλασικά μοντέλα ανάκτησης πληροφορίας να εντοπίζουν υποκειμενική πληροφορία, όπως για παράδειγμα απόψεις, η οποία υπάρχει στα review κείμενα. Η opinion expansion επέκταση εισάγει στο ερώτημα του χρήστη όρους όπως “πολύ, καταπληκτικό, θαυμάσιο” και λέξεις με θετική έννοια, τοποθετώντας στην κορυφή τις κατάταξης οντοτήτων, οντότητες με πολλές θετικές απόψεις χρηστών για αυτές. Αυτή η επέκταση ευνοεί κείμενα, και τις αντίστοιχες οντότητες, με θετικές απόψεις στα χαρακτηριστικά, το οποίο

και είναι ο στόχος της διαδικασίας της αξιολόγησης. Ωστόσο αυτή η προσέγγιση *δεν επιβάλλει κυρώσεις για τις αρνητικές απόψεις*.

Επιχειρούμε να βελτιώσουμε ακόμα περισσότερο αυτή την προσέγγιση και αναπτύσσουμε σχήματα τα οποία λαμβάνουν υπόψιν sentiment (υποενότητα 4.4) και clustering πληροφορία (υποενότητα 4.5) σχετικά με τις απόψεις στα review κείμενα. Επίσης προτείνουμε το naïve consumer μοντέλο στην υποενότητα 4.6.

Πιο συγκεκριμένα, στην υποενότητα 4.4, αντί της χρήσης της σημαντικότητας των aspect keywords του ερωτήματος για ένα review κείμενο (r_i), δοκιμάζουμε να χρησιμοποιήσουμε την sentiment πληροφορία των απόψεων, οι οποίες εκφράζονται στο r_i, για τα συγκεκριμένα aspects. Εξάγοντας την sentiment πληροφορία των απόψεων, διαμορφώνουμε μία βαθμολογία για κάθε aspect, τις οποίες ύστερα χρησιμοποιούμε για να αποδώσουμε μία βαθμολογία στο κείμενο r_i, λαμβάνοντας υπόψιν τις aspect προτιμήσεις του χρήστη. Στην υποενότητα 4.5 επιχειρούμε να αναγνωρίσουμε παρόμοιες συμπεριφορές βαθμολόγησης ανάμεσα στους χρήστες, και να χρησιμοποιήσουμε αυτήν την πληροφορία για την βελτίωση της κατάταξης. Σε αυτή την προσέγγιση, χρησιμοποιούμε το σχήμα το οποίο παρουσιάζεται στην έρευνα (Kurland, 2006) για να διαμορφώσουμε ομάδες (clusters), έτσι ώστε κάθε cluster να αποτελείται από review κείμενα με παρόμοιες βαθμολογίες στα aspects. Στην υποενότητα 4.6, επιχειρούμε να προσομοιώνουμε την συμπεριφορά ενός καταναλωτή ο οποίος προσπαθεί να κατατάξει οντότητες μίας συγκεκριμένης κατηγορίας, αλλά δεν γνωρίζει ποια από τα aspect, των οντοτήτων αυτής της κατηγορίας, είναι περισσότερο σημαντικά για να χρησιμοποιηθούν ως κριτήρια στην αξιολόγηση. Προτείνουμε το naïve consumer μοντέλο ως ένα μη εποπτευόμενο σχήμα, το οποίο χρησιμοποιεί πληροφορίες από το διαδίκτυο ώστε να αποδώσει ένα βάρος σημαντικότητας σε κάθε ένα από τα aspects τα οποία χρησιμοποιούνται στην αξιολόγηση των οντοτήτων. Τέλος στην υποενότητα 4.7, παρουσιάζουμε τα πειράματα που πραγματοποιήσαμε, για την αξιολόγηση των τεχνικών μας, καθώς επίσης και αναλυτική παρουσίαση των μετρήσεων και των αποτελεσμάτων.

4.3 Το πρόβλημα της αξιολόγησης οντοτήτων ως πρόβλημα ανάκτησης πληροφορίας

Έστω ένα σύστημα αξιολόγησης οντοτήτων, RS, και μία συλλογή οντοτήτων E= {e1, e2, ..., en} του ίδιου είδους. Υποθέτουμε πως κάθε μία από τις οντότητες στο σύνολο

E συνοδεύεται από ένα κείμενο με όλα τα διαθέσιμα reviews για αυτήν. Έστω R = {r1, r2, ..., rn} το σύνολο αυτών των κειμένων. Υπάρχει μία “1-1” σχέση μεταξύ των στοιχείων του συνόλου E και αυτών του συνόλου R. Δοθέντος ενός συνόλου απαιτήσεων q, το σύστημα RS παράγει μία λίστα κατάταξης οντοτήτων E.

Η ιδέα για την διάταξη των οντοτήτων, είναι να αναπαρασταθεί κάθε οντότητα με το κείμενο όλων των διαθέσιμων review κειμένων τα οποία αναφέρονται για αυτήν. Δοθέντος ενός ερωτήματος του χρήστη, το οποίο εκφράζει τα επιθυμητά aspects, τα οποία μία οντότητα πρέπει να διαθέτει, μπορούμε να αξιολογήσουμε τις οντότητες με βάση το πόσο καλά τα review κείμενα (ri) ταιριάζουν τις προτιμήσεις του χρήστη. Οπότε το πρόβλημα της αξιολόγησης οντοτήτων μετατρέπεται σε πρόβλημα ανάκτησης πληροφορίας. Έτσι μπορούμε να χρησιμοποιήσουμε οποιοδήποτε γνωστό μοντέλο ανάκτησης πληροφορίας, όπως το BM25 (Baeza-Yates and Ribeiro- Neto, 2011). Αυτό το setup παρουσιάζεται στην έρευνα (Ganesan and ChengXiang, 2012). Πιο συγκεκριμένα, δοκιμάστηκαν οι φόρμουλες ανάκτησης πληροφορίας, BM25 (Baeza-Yates and Ribeiro-Neto, 2011), (Robertson and Zaragoza, 2009), Dirichlet (Zhai and Lafferty, 2001), και η PL2 (Amati, and van Rijsbergen, 2002). Επίσης προτάθηκε η χρήση κάποιων νέων επεκτάσεων για αυτά τα μοντέλα, όπως το query aspect modeling (QAM) και το opinion expansion, και πραγματοποιήθηκε μία σειρά πειραμάτων τα οποία έδειξαν πως αυτές οι προσεγγίσεις δίνουν καλύτερα αποτελέσματα. Η QAM επέκταση, χρησιμοποιεί κάθε aspect του ερωτήματος για να βαθμολογήσει τις οντότητες και ύστερα συναθροίζει τα αποτελέσματα των βαθμολογιών για τα διάφορα aspects στο ερώτημα χρησιμοποιώντας μία συνάρτηση συνάθροισης όπως ο μέσος όρος των βαθμολογιών. Η opinion expansion επέκταση, επεκτένει το ερώτημα με σχετικές opinion λέξεις από ένα online θησαυρό. Τα αποτελέσματα των πειραμάτων έδειξαν πως και τα τρία μοντέλα ανάκτησης πληροφορίας αποδίδουν καλύτερα με τις παραπάνω επεκτάσεις, ένω το μοντέλο BM25 είναι περισσότερο συνεπή και λειτουργεί καλύτερα με αυτές τις επεκτάσεις.

Ο στόχος των δικών μας πειραμάτων είναι να μελετήσουμε αυτή η ενδιαφέρουσα προσέγγιση και να συγκρίνουμε την απόδοσή της με γνωστές τεχνικές sentiment analysis, καθώς και να ερευνήσουμε την συμπεριφορά της όσο αφορά το clustering και το naïve consumer μοντέλο το οποίο παρουσιάζουμε στην υποενότητα 4.6. Για τα πειράματά μας επιλέγουμε να χρησιμοποιήσουμε το μοντέλο BM25 ως την συνάρτηση ανάκτησης πληροφορίας. Επίσης, αναπτύσσουμε τις επεκτάσεις query aspect

modeling (QAM) και opinion expansion, όπως αυτές περιγράφονται στο (Ganesan and ChengXiang, 2012). Η BM25 φόρμουλα που χρησιμοποιήσαμε για τις μετρήσεις μας είναι η ακόλουθη:

$$BM25(q,d)=\sum_{t\in q}(\log\frac{N}{df_t})\frac{(k_1+1)tf_{td}}{k_1((1-b)+b*(\frac{L_d}{L_{ave}}))+tf_{td}}$$

όπου tf_{td} είναι η συχνότητα του όρου t στο κείμενο d, L_d είναι το μήκος του κειμένου d, L_{ave} είναι ο μέσος όρος του μήκους των κειμένων στην συλλογή, df_t είναι ο αριθμός των εγγράφων τα οποία περιέχουν τον όρο t, και b και k_1 είναι μεταβλητές οι οποίες λαμβάνουν τις τιμές b=0.75 και k_1=1.2.

4.4 Opinion-based Aspect Βαθμολογίες

Για να δημιουργήσουμε ένα μοντέλο το οποίο λαμβάνει υπόψιν του την sentiment πληροφορία των απόψεων οι οποίες εκφράζονται στα review κείμενα, χρησιμοποιούμε δύο μη εποπτευόμενες τεχνικές sentiment ανάλυσης. Με βάση το γεγονός πως κάθε review κείμενο r_i περιέχει απόψεις χρηστών για μία συγκεκριμένη οντότητα, εφαρμόζουμε aspect-based τεχνικές sentiment ανάλυσης για να εξάγουμε την sentiment πληροφορία η οποία εκφράζεται για τα aspects στις προτάσεις, και aspect-based summarization (ή feature-based summarization) για να υπολογίσουμε την συνολική βαθμολογία των aspects σε κάθε κείμενο.

4.4.1 Sentiment ανάλυση βασισμένη σε sentiment λεξικό

Έστω A = {a1, a2, ..., an} το σύνολο των aspects του ερωτήματος. Πραγματοποιούμε sentiment ανάλυση στο aspect επίπεδο, εξάγοντας από τα review κείμενα r_i την πολικότητα $s(a_j)$, η οποία εκφράζεται για κάθε ένα από τα aspect keywords στου ερωτήματος.

Η συνολική βαθμολογία την οποία λαμβάνει το review κείμενο r_i είναι το άθροισμα των sentiment βαθμολογιών των aspects, στο συγκεκριμένο κείμενο, κανονικοποιημένη με τον πλήθος των aspects στο ερώτημα.

Για τον υπολογισμό της sentiment βαθμολογίας $s(a_j)$, εντοπίζουμε στο κείμενο r_i τις προτάσεις στις οποίες αναφέρεται κάθε ένα από τα aspects a_j του ερωτήματος, και αποδίδουμε σε αυτές μία sentiment βαθμολογία. Η βαθμολογία $s(a_j)$ ειναι το άθροισμα των sentiment βαθμολογιών των προτάσεων στις οποίες αναφέρεται το aspect a_j. Για τον υπολογισμό της sentiment βαθμολογίας μίας πρότασης, χρησιμοποιούμε μία διαδικασία pos tagging (Tsuruoka), η οποία αποδίδει σε κάθε όρο ένα pos tag, και επεξεργαζόμαστε μόνο τους όρους στους οποίους έχει αποδοθεί ένα από τα ακόλουθα pos tags (λίστα με τα pos tags υπάρχει στο Appendix):

{ RB / RBR / RBS / VBG / JJ / JJR / JJS }

ως στοιχεία τα οποία συνήθως φέρουν sentiment πληροφορία. Για κάθε έναν από αυτούς τους όρους, εντοπίζουμε την sentiment βαθμολογία τους σε ένα sentiment λεξικό (Liu, Sentiment Lexicon), 1 εάν ο όρος είναι επισημασμένος ως θετικής πολικότητας όρος, και -1 εάν είναι επισημασμένος ως αρνητικής πολικότητας όρος. Τέλος, αθροίζουμε τις βαθμολογίες όλων των όρων. Αν το άθροισμα είναι θετικό, τότε η sentiment βαθμολογία της πρότασης είναι 1, ενώ αν είναι αρνητικό η sentiment βαθμολογία της πρότασης είναι -1.

Επίσης διαχειριζόμαστε την άρνηση, και χρησιμοποιούμε sentiment shifters. Αν σε μία πρόταση υπάρχει κάποιος από τους ακόλουθους όρους:

{not, don't, none, nobody, nowhere, neither, cannot}

αντιστρέφουμε την πολικότητα της τελικής sentiment βαθμολογίας της πρότασης.

4.4.1.1 Επέκταση των ερωτημάτων

Σε αυτή την διαδικασία προσπελάσουμε κάθε κείμενο r_i , πρόταση προς πρόταση, και επεξεργαζόμαστε μόνο τις προτάσεις οι οποίες περιέχουν τουλάχιστον ένα από aspect keyword του ερωτήματος, Αλλά συχνά οι χρήστες χρησιμοποιούν διαφορετικές λέξεις ή εκφράσεις για να εκφράσουν τις απόψεις τους για ένα aspect μιας οντότητας. Για να διαχειριστούμε αυτό το φαινόμενο, πραγματοποιούμε επέκταση του αρχικού ερωτήματος (query expansion), το οποίο εμπλουτίζουμε με συνώνυμα των aspects $\sigma(a_j)$ από το σημασιολογικό δίκτυο WordNet (Fellbaum, 1998). Για παράδειγμα, έστω ένα ερώτημα q το οποίο αποτελείται από τα aspect keywords $\{a_1, a_2, a_3\}$. Για κάθε μία από τις λέξεις

κλειδιά του ερωτήματος q, προσπαθούμε να εντοπίσουμε συνώνυμα σ(a_j) χρησιμοποιώντας το σημασιολογικό δίκτυο WordNet, τα οποία εισάγουμε στο αρχικό ερώτημα.

$$a_1 \rightarrow (\sigma_{1a1}, \sigma_{2\,a1})$$
$$a_2 \rightarrow (\sigma_{1a2}, \sigma_{2a2}, \sigma_{3\,a2})$$
$$a_3 \rightarrow (\sigma_{1a3})$$

Το τελικό ερώτημα που διαμορφώνεται σε αυτή την περίπτωση είναι:

$$q = (a_1 , \sigma_{1a1} , \sigma_{2a1} , a_2 , \sigma_{1a2} , \sigma_{2a2} , \sigma_{3\,a2} , a_3 , \sigma_{1\,a3})$$

Σε αυτήν την περίπτωση, η sentiment βαθμολογία του aspect a_i, $s_{exp}(a_i)$, είναι το άθροισμα της sentiment βαθμολογίας του όρου a_i και των sentiment βαθμολογιών των επιπλέον συνώνυμων που εισήχθησαν σj_{ai}, $s(\sigma j_{ai})$.

$$S_{\exp}(a_i) = s(a_i) + \sum_{j=1,2,\ldots,n} s(\sigma j_{ai})$$

4.4.2 Sentiment ανάλυση βασισμένη σε συντακτικά πρότυπα

Σε αυτό το σχήμα χρησιμοποιούμε ως βάση τον αλγόριθμο ο οποίος παρουσιάστηκε στην έρευνα (Turney, 2002), για να υπολογίσουμε την sentiment βαθμολογία κάθε πρότασης. Εδώ, αντί να χρησιμοποιήσουμε τις sentiment βαθμολογίες των λέξεων μιας πρόταση, χρησιμοποιούμε την sentiment πολικότητα (SO) συντακτικών προτύπων μίας πρότασης, τα οποία συνήθως χρησιμοποιούνται για να εκφραστεί μία άποψη.

Τα συντακτικά πρότυπα εντοπίζονται μέσα στην πρόταση χρησιμοποιώντας pos tags όρων, τα οποία εμφανίζονται σε συγκεκριμένη σειρά. Τα ακόλουθα είναι συντακτικά πρότυπα τα οποία χρησιμοποιούνται για την εξαγωγή φράσεων δύο λέξεων:

1st Word	**2nd Word**	**3nd Word**
JJ	NN/NNS	anything
RB/RBR/RBS	JJ	not(NN/NNS)
JJ	JJ	not(NN/NNS)
NN/NNS	JJ	not(NN/NNS)
RB/RBR/RBS	VB/VBD/VBN/VBG	anything

Για να υπολογίσουμε την sentiment πολικότητα των εξαγόμενων φράσεων, χρησιμοποιούμε την μετρική point wise mutual information (PMI):

$$PMI\left(term1, term2\right)=\log_2\frac{Pr\left(term1\wedge term2\right)}{Pr\left(term1\right)*Pr\left(term2\right)}$$

Η μετρική PMI μετρά την στατιστική ανεξαρτησία μεταξύ δύο όρων. $Pr\left(term1\wedge term2\right)$ είναι η πιθανότητα συνεμφάνισης των όρων term1 και term2, και το γινόμενο $Pr\left(term1\right)*Pr\left(term2\right)$ είναι η πιθανότητα συνεμφάνισης των δύο όρων, term1 και term2, όταν αυτοί είναι στατιστικά ανεξάρτητοι. Η sentiment πολικότητα μίας φράσεις υπολογίζεται με βάση την σχέση της, με τη θετικής πολικότητας λέξη αναφοράς "excellent", και την αρνητικής πολικότητας λέξη αναφοράς "poor".

$$SO\left(phrase\right)=PMI\left(phrase,'excellent'\right)-PMI\left(phrase,'poor'\right)$$

Οι πιθανότητες υπολογίζονται πραγματοποιώντας ερωτήματα σε μία μηχανή αναζήτησης, και συλλέγοντας τον αριθμό των απαντήσεων οι οποίες επιστρέφονται (ο αριθμός των σχετικών με το ερώτημα κειμένων):

$$SO\left(phrase\right)=\log_2\frac{hits\left(phrase\ NEAR\ 'excellent'\right)hits\left('poor'\right)}{hits\left(phrase\ NEAR\ 'poor'\right)hits\left('excellent'\right)}$$

Στην περίπτωσή μας χρησιμοποιούμε όλα τα κείμενα τα οποία έχουμε συλλέξει από το διαδίκτυο εκτελώντας ερωτήματα στην μηχανή αναζήτησης Google, και υπολογίζουμε αυτές τις πιθναότητες, ως των αριθμό των κειμένων τα οποία περιέχουν κάθε μία από τις φράσεις.

Επιπλέον αντί να χρησιμοποιήσουμε μόνο τις λέξεις αναφορά “excellent” και “poor”, δουλεύουμε με δύο σύνολα λέξεων αναφοράς. Το σύνολο '+' = {excellent, good} ως σύνολο λέξεων αναφοράς θετικής πολικότητας, και το σύνολο '-' = {horrible , bad} ως σύνολο λέξεων αναφοράς αρνητικής πολικότητας. Επιπλέον εμπλουτίζουμε αυτά τα σύνολα με συνώνυμα από το σημασιολογικό δίκτυο WordNet.

Οπότε ο υπολογισμός της sentiment πολικότητας μίας φράσης διαμορφώνεται ως εξής:

$$SO(phrase) = \log_2 \frac{hits(phrase\ NEAR\ '+')\ hits('-')}{hits(phrase\ NEAR\ '-')\ hits('+')}$$

όπου αθροίζονται τα hits( ) για όλα τα στοιχεία ενός συνόλου. Για παράδειγμα:

$$hits('+') = hits('excellent') + hits('good') + \sum_{j=1,2,\ldots,n} hits(\sigma_j)$$

όπου με σ_j αναπαρίστανται τα συνώνυμα τα οποία προστίθενται στο σύνολο λέξεων αναφοράς από το WordNet κατά την διάρκεια της διαδικασίας.

4.5 Εξομάλυνση της κατάταξης με Opinion-based Clusters

Σε αυτό το σχήμα προσπαθούμε να χρησιμοποιήσουμε clustering πληροφορία σχετικά με τα reviews ώστε να βελτιώσουμε την κατάταξη των οντοτήτων. Χρησιμοποιούμε τον αλγόριθμο ClustFuse του Kurland (Kurland, 2006), ο οποίος χρησιμοποιεί δύο συνιστώσες για να αποδώσει μία βαθμολογία σε ένα κείμενο d, την πιθανότητα σχετικότητας του κειμένου με το ερώτημα, και την υπόθεση πως οι συστάδες μπορούν να χρησιμοποιηθούν ως proxies για τα κείμενα, με βάση την οποία ανταμείβονται κείμενα τα οποία ανήκουν “strongly” σε μία συστάδα η οποία είναι περισσότερο σχετική με το ερώτημα.

Ο αλγόριθμος ClustFuse χρησιμοποιεί cluster πληροφορία για να βελτιώσει την κατάταξη των κειμένων. Περιληπτικά, ο αλγόριθμος για να δημιουργήσει την κατάταξη των κειμένων σε ένα ερώτημα q, παράγει, με βάση το q, ένα σύνολο παρόμοιων ερωτημάτων με το q, έστω το Q = {q_1, q_2, …, q_n}, και για κάθε ένα από αυτά λαμβάνει μία κατάταξη κειμένων L_i. Ύστερα, προσπαθεί να αναγνωρίσει συστάδες σε όλο το σύνολο των κειμένων των κατατάξεων Li. Τέλος, παράγει μία τελική κατάταξη κειμένων χρησιμοποιώντας των ακόλουθη εξίσωση:

$$p(q|d) = (1-\lambda)\, p(d|q) + \lambda \sum_{c \in Cl(CL)} p(c|q)\, p(d|c)$$

Στην περίπτωσή μας ως σύνολο CL θεωρούμε όλα τα review κείμενα r_i, Για την δημιουργία του συνόλου των ερωτημάτων $Q = \{q_1, q_2, \ldots, q_n\}$ για κάθε ερώτημα q, χρησιμοποιούμε συνδυασμούς συνωνύμων των όρων από το σημασιολογικό δίκτυο WordNet. Επίσης χρησιμοποιούμε το μοντέλο διανυσματικού χώρου (Vector Space) για την αναπαράσταση των review κειμένων r_i, την ομοιότητα συνημιτόνου ως μετρική απόστασης των κειμένων, τον αλγόριθμο k-means για την συστασοποίηση, την FcombSUM (d, q) ως μέθοδο fusion (Kurland, 2006), και την μετρική BM25 για την διάταξη των κειμένων r_i και την παραγωγή των λιστών κατάταξης L_i. Ακόμα, ως χαρακτηριστικά των κειμένων r_i, χρησιμοποιούμε τις sentiment βαθμολογίες των aspects, όπως αυτές παράγονται από την διαδικασία η οποία περιγράφεται πιο πάνω στην ενότητα 4. Οπότε, με βάση τα παραπάνω, μπορούμε να πούμε πως κάθε συστάδα θα αποτελείται από review κείμενα με παρόμοιες βαθμολογίες στα aspects. Λεπτομερή περιγραφή του σχήματος του Kurland και ερμηνεία των εξισώσεων που χρησιμοποιούνται υπάρχει στο Appendix.

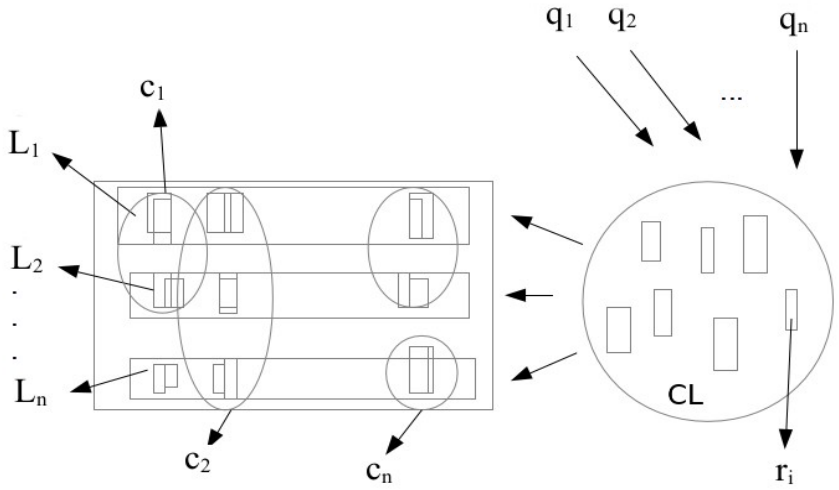


Figure 1
Η διαδικάσια παραγωγής των κατατάξεων Li και των συστάδων ci στον αλγόριθμο ClustFuse

4.6 Naive Consumer μοντέλο

Στα προηγούμενα σχήματα χρησιμοποιούμε ένα σύνολο από aspects λέξεις κλειδιά (χαρακτηριστικά) ως ερωτήματα, και αξιολογούμε τα review κείμενα με βάση την

σχετικότητά τους με αυτά τα ερωτήματα. Αλλά με αυτό τον τρόπο θεωρούμε πως όλα τα aspect είναι το ίδιο σημαντικά για να την κατάταξη των οντοτήτων, κάτι το οποίο δεν ισχύει πάντα. Για παράδειγμα, στο domain των αυτοκινήτων, μπορούμε να πούμε πως το aspect "κατανάλωση καυσίμου" είναι πιο σημαντικό από το aspect "δερματινα καθίσματα". Μπορεί όμως και να μην είναι. Πιστεύουμε πως η απάντηση σε κάτι τέτοιο μπορεί να δοθεί μόνο από την κοινότητα των χρηστών. Οπότε, ανακτούμε αυτή την πληροφορία από το διαδίκτυο. Έστω D_{inf} το σύνολο αυτών των κειμένων.

Με αυό το μοντέλο επιδιώκουμε να προσομοιώσουμε την συμπεριφορά ενός καταναλωτή ο οποίος προσπαθεί να κατατάξει οντότητες μιας συγκεκριμένης κατηγορίας, αλλά δεν γνωρίζει ποια από τα aspects, αυτών των οντοτήτων, είναι σημαντικά για να χρησιμοποιηθούν ως κριτήρια στην διαδικασία της αξιολόγησης. Συνήθως, ένας τέτοιος χρήστης ανατρέχει στο διαδίκτυο, σε σχετικά άρθρα στην Wikipedia, σε blogs, σε φόρα, καθώς επίσης και σε ιστοτόπους οι οποίοι περιέχουν reviews άλλων χρηστών, για να αναγνωρίσει όλα αυτά τα χαρακτηριστικά.

Επιδιώκουμε να συλλέξουμε την γνώση του D_{inf}, για το ποια aspects είναι πιο σημαντικά. Δημιουργούμε ένα σύνολο ερωτημάτων $Q = \{q_1, q_2, \ldots, q_n\}$ τα οποία αποτελούνται από aspect λέξεις κλειδιά. Κάθε στοιχείο του συνόλου Q, δίνεται ως ερώτημα σε μία μηχανή αναζήτησης, και συλλέγονται τα πρώτα αποτελέσματα. Δεδομένου πως οι μηχανές αναζήτησης χρησιμοποιούν ένα γραμμικό συνδυασμό μετρικών όπως οι BM25 και ο PageRank, μπορούμε να υποθέσουμε πως όλα τα κείμενα (ιστοσελίδες) τα οποία συλλέγονται, είναι σχετικά με τις οντότητες του domain που εξετάζουμε, και πως επίσης αποτελούν σημαντικούς κόμβους στον γράφο του παγκοσμίου ιστού, οπότε σημαντικά και για την κοινότητα.

Η σημαντικότητα ενός όρου t για ένα κείμενο d, ως μέρος μιας συλλογής κειμένου, μπορούμε να πούμε πως υπολογίζεται από την BM25 βαθμολογία του όρου t στο d. Έχοντας το σύνολο όλων των κειμένων D_{inf}, προσπαθούμε να εξάγουμε πόσο σημαντικό είναι κάθε aspect keyword (a_i) για την οντότητα του domain που εξετάζουμε, υπολογίζοντας μία βαθμολογία σημαντικότητας. Για αυτό τον υπολογισμό μπορούμε να χρησιμοποιήσουμε διάφορες φόρμουλες, αλλά επιλέγουμε να χρησιμοποιήσουμε την ακόλουθη, η οποία περιέχει το ποσοστό συμμετοχής του aspect a_i στην βαθμολόγηση των reviews.

$$scoreA(a_i) = \frac{\sum_{d \in D_{inf}} BM25(a_i)_d}{\sum_{a' \in A} (\sum_{d \in D_{inf}} BM25(a')_d)} \quad (1)$$

Αυτή η φόρμουλα είναι παρόμοια με το μοντέλο Rasch. Στην ανάλυση δεδομένων με ένα μοντέλο Rasch, ο στόχος είναι να μετρηθεί το επίπεδο του κάθε εξεταζόμενου σε ένα λανθάνον γνώρισμα (για παράδειγμα, ικανότητα στα μαθηματικά, στάση απέναντι στην θανατική ποινή) το οποίο κρύβεται πίσω από τις βαθμολογίες του/της στα στοιχεία ενός test. Στην περίπτωσή μας το test είναι η κατάταξη των review κειμένων, οι εξεταζόμενοι είναι τα aspects, και τα στοιχεία του test είναι τα review κείμενα. Πάνω σε αυτήν την ιδέα αναπτύσσουμε τα δύο μοντέλα NCM1 και NCM2.

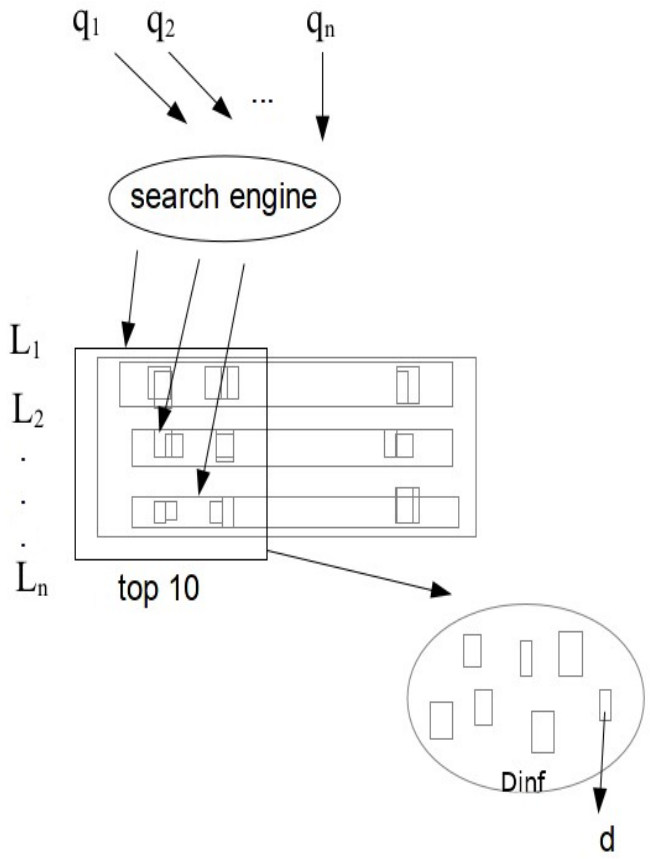


Figure 2
Η διαδικάσια της συλλογής των κειμένων από το διαδίκτυο

4.6.1 NCM 1

Έχοντας το ποσοστό scoreA(t) για κάθε aspect t, το οποίο εκφράζει πόσο σημαντικό είναι το συγκεκριμένο aspect όταν χρησιμοποιείται για την αξιολόγηση οντοτήτων μίας συγκεκριμένης κλάσης, μπορούμε να το συνδυάσουμε με τον όρο που εκφράζει πόσο σημαντικό είναι το aspect, ως όρος κειμένου, για ένα συγκεκριμένο review κείμενο r_i ως μέρος μιας συλλογής κειμένων (R), έτσι ώστε να κατατάξουμε τα review κείμενα σε ερωτήματα τα οποία αποτελούνται από aspect λέξεις κλειδιά, ως εξής:

$$p(q, ri) = \sum_{t \in q} score(t)_{ri} * scoreA(t) \quad (2)$$

όπου $score(t)_{ri}$ είναι η BM25 βαθμολογία της λέξης/aspect t για το review κείμενο r_i και το $scoreA(t)$ προέρχεται από την (1).

4.6.2 NCM 2

Στο NCM2 εργαζόμαστε όπως και στο NCM1, εφαρμόζοντας επιπλέον το σχήμα του Kurland (Kurland, 2006), για να εκμεταλλευτούμε οποιαδήποτε cluster δομή, μπορεί να υπάρχει στα review κείμενα.

Εφαρμόζουμε την εξίσωση (2) για να κατατάξουμε τα review κείμενα στα ερωτήματα Q = {q1, q2, ..., qn} και να δημιουργήσουμε τις Li λίστες (figure 2), όπου scoreA(t) είναι η βαθμολογία σημαντικότητας των aspects στην χρήση τους στην διαδικασία της αξιολόγησης των οντοτήτων, όπως αυτή υπολογίστηκε από το σύνολο των κειμένων που συλλέξαμε από το διαδίκτυο. Χρησιμοποιούμε τον αλγόριθμο ClustFuse, όπως αναφέρεται πιο πάνω, στην ενότητα 5.

4.7 Πειράματα

Σε αυτή την έρευνα καταπιανόμαστε με το πρόβλημα της κατάταξης οντοτήτων χρησιμοποιώντας τα review κείμενα χρηστών. Ακολουθούμε την προσέγγιση της μετατροπής του προβλήματος σε ένα πρόβλημα ταιριάσματος προτιμήσεων, όπως αυτή παρουσιάζεται στην έρευνα (Ganesan and ChengXiang, 2012). Ο στόχος των πειραμάτων μας είναι να μελετήσουμε αυτήν την ενδιαφέρουσα προσέγγιση, να συγκρίνου-

με την απόδοσή της με γνωστές τεχνικές sentiment analysis, καθώς και να εξετάσουμε την συμπεριφορά της όσο αφορά το clustering και το naïve consumer μοντέλο.

Όπως αναφέρεται στην υποενότητα 4.3, για την λύση του προβλήματος ακολουθώντας αυτήν την προσέγγιση, μπορούμε να χρησιμοποιήσουμε μερικά από τα γνωστά μοντέλα ανάκτησης πληροφορίας. Επιλέγουμε να χρησιμοποιήσουμε το BM25 μοντέλο. Στην έρευνα (Ganesan and ChengXiang, 2012) επίσης προτείνονται κάποιες νέες επεκτάσεις για τα κλασικά μοντέλα ανάκτησης πληροφορίας, όπως το query aspect modeling (QAM) και το opinion expansion. Τα αποτελέσματα των πειραμάτων στην έρευνα (Ganesan and ChengXiang, 2012), έδειξαν πως το BM25 μοντέλο είναι περισσότερο σταθερό και αποδίδει αρκετά καλά με αυτές τις επεκτάσεις. Για τα δικά μας πειράματα επιλέγουμε να υλοποιήσουμε τις δύο αυτές επεκτάσεις, το AvgScoreQAM (query aspect modeling με την μέθοδο συνάθροισης "Average Score") και το opinion expansion. Πρώτα εξετάζουμε και συγκρίνουμε την απόδοση του BM25 μοντέλου με, και χωρίς, τις επεκτάσεις. Θα πρέπει να αναφερθεί πως δεν χρησιμοποιήσαμε κάποιο μηχανισμό pseudo feedback.

Προσεγγίζοντας το πρόβλημα της κατάταξης οντοτήτων ως ένα πρόβλημα ταιριάσματος προτιμήσεων σε συγκεκριμένα χαρακτηριστικά χρησιμοποιώντας ένα μοντέλο ανάκτησης πληροφορίας, όπως παρουσιάζεται στην έρευνα (Ganesan and ChengXiang, 2012), ευνοούμε review κείμενα, και τις αντίστοιχες οντότητες, με θετικές απόψεις στα aspects, αλλά δεν θέτουμε ποινές για τις αρνητικές απόψεις. Αυτό γίνεται εξαιτίας της opinion expansion επέκτασης, όπου και εισάγονται όροι όπως "πολύ, πάρα πολύ, κτλ." καθώς και λέξεις με θετική σημασία όπως "φανταστικό, καλό, κτλ.", κάτι το οποίο τοποθετεί ψηλά στην κατάταξη των οντοτήτων, οντότητες με πολλές θετικές απόψεις. Οπότε σε αυτήν την έρευνα προσπαθούμε να εξετάσουμε γνωστές τεχνικές sentiment analysis, οι οποίες λαμβάνουν υπόψιν τις θετικές καθώς και τις αρνητικές, απόψεις, και να συγκρίνουμε τις αποδόσεις τους με αυτήν της προσέγγισης της ανάκτησης πληροφορίας. Υλοποιούμε δύο τεχνικές sentiment analysis, έτσι ώστε να εξάγουμε την sentiment πληροφορία από τις απόψεις των χρηστών, όπως περιγράφουμε παραπάνω στην υποενότητα 4.4. Η πρώτη τεχνική είναι βασισμένη σε λεξικό (lexicon based), ένω η δεύτερη σε συντακτικά πρότυπα (syntactic patterns based).

Προκειμένου να εκμεταλλευτούμε οποιαδήποτε cluster πληροφορία, η οποία μπορεί να υπάρχει ή να διαμορφώνεται στο σύνολο των review κειμένων, χρησιμοποιούμε

το setup του Kurland με τον αλγόριθμο ClustFuse. Επίσης εκμεταλλευόμαστε την αναπαράσταση των review κειμένων, χρησιμοποιώντας ως features τις sentiment βαθμολογίες των aspects. Έτσι κατασκευάζουμε opinion-based συστάδες, και παρατηρούμε τα αποτελέσματα στην απόδοση της διαδικασίας.

Επίσης συγκρίνουμε τα δύο naïve consumer μοντέλα, NCM1 και NCM2, με τα οποία προσπαθούμε να συλλέξουμε την γνώση του διαδίκτυο (της κοινότητας) ώστε να αποδώσουμε περισσότερο βάρος στα σημαντικότερα aspects στην διαδικασία της αξιολόγησης.

Χρησιμοποιούμε δύο επιπλέον τεχνικές για την κατάταξη των οντοτήτων, ως σημεία αναφοράς απόδοσης. Την μετρική BM25 ως πρότυπο μοντέλο ανάκτησης πληροφορία, και το m-aspect prank μοντέλο ως εποπτευόμενη τεχνική η οποία αποδίδει αρκετά καλά στην πρόβλεψη aspect βαθμολογιών.

Πραγματοποιούμε δύο σύνολα πειραμάτων, για να εξετάσουμε την απόδοση των σχημάτων μας, χρησιμοποιώντας δύο διαφορετικά σύνολα δεδομένων αντίστοιχα. Τα σύνολα δεδομένων αποτελούνται από σύνολα οντοτήτων οι οποίες συνοδεύονται με review κείμενα χρηστών, τα οποία προέρχονται από online ιστοσελίδες. Οι ερωτήσεις αποτελούνται από aspect λέξεις κλειδία. Για κάθε μία από τις ερωτήσεις παράγουμε την ιδανική κατάταξη οντοτήτων με βάση τις βαθμολογίες των aspects οι οποίες δίνονται από τους χρήστες μαζί με τα review κείμενα. Υπολογίζεται ο μέσος όρος των βαθμολογιών που δίνουν οι χρήστες για ένα συγκεκριμένο aspect ως το Average Aspect Rating (AAR). Για ερωτήσεις οι οποίες αποτελούνται από περισσότερα του ενός aspects, υπολογίζεται ο μέρος όρος των AAR aspect βαθμολογιών, ως το Multi-Aspect AAR (MAAR). Πιο συγκεκριμένα, έστω ένα ερώτημα $Q=\{Q_1, Q_2, ..., Q_k\}$ με k aspects ως λέξεις κλειδιά, και μία οντότητα e, τότε $r_i(e)$ είναι το AAR της οντότητας e για το i-οστό aspect. Οπότε το MAAR, της οντότητας e, υπολογίζεται ως εξής:

$$MAAR(e,Q)=\frac{1}{k}\sum_{i=1}^{k} r_i(e)$$

Για το πρώτο σύνολο πειραμάτων, χρησιμοποιούμε review κείμενα από το domain των αυτοκινήτων, τα οποία προέρχονται από την ιστοσελίδα Edmunds.com, και για το δεύτερο σύνολο πειραμάτων χρησιμοποιούμε review κείμενα από το domain των ξενοδοχείων, τα οποία προέρχονται από την ιστοσελίδα we8there.com. Στο πρώτο σύνο-

λο πειραμάτων πραγματοποιούμε 300 ερωτήματα, και στο δεύτερο 30, και αξιολογούμε την απόδοση των σχημάτων μας να αποδώσουν την σωστή κατάταξη οντοτήτων, υπολογίζοντας την nDCG στα πρώτα 10 αποτελέσματα.

Στο δεύτερο σύνολο πειραμάτων συγκρίνουμε την απόδοση των σχημάτων μας με ένα multiple aspect online ranking μοντέλο το οποίο παρουσιάζεται στην έρευνα (Snyder and Barzilay, 2007), και είναι βασισμένο στον αλγόριθμο PRank ο οποίος παρουσιάστηκε από τους by Crammer and Singer in (Crammer and Singer, 2001). Αυτή η εποπτευόμενη τεχνική έχει δείξει πως αποδίδει αρκετά καλά στο να προβλέπει τις βαθμολογίες σε συγκεκριμένα aspects μιας οντότητας χρησιμοποιώντας τα review κείμενα των χρηστών για αυτήν.

Για να δημιουργήσουμε ένα m-aspect ranking μοντέλο χρησιμοποιούμε m ανεξάρτητα PRank μοντέλα, ένα για κάθε ένα από τα aspects. Κάθε ένα από τα m μοντέλα, εκπαιδεύεται για να προβλέπει σωστά ένα μόνο από τα m aspects. Έχοντας αναπαραστήσει τα review κείμενα r_i ως ένα feature διάνυσμα $x \in Rn$, αυτό το μοντέλο προβλέπει μία τιμή $y \in \{1, ..., k\}$ για κάθε $x \in Rn$. Το μοντέλο αποθηκεύει ένα διάνυσμα βάρους $w \in Rn$ και ένα διάνυσμα ορίων $b_0 = -\infty \leq b_1 \leq ... \leq b_{k-1} \leq b_k = \infty$ το οποίο διαιρεί την γραμμή των πραγματικών αριθμών σε k στάθμες, μία για κάθε πιθανή βαθμολογία. Το μοντέλο αποδίδει μία βαθμολογία για κάθε είσοδο χρησιμοποιώντας το διάνυσμα βάρους: score(x) = w * x. Τέλος, προβάλει την βαθμολογία score(x) στον άξονα των πραγματικών αριθμών, και επιστρέφει την κατάλληλη βαθμολογία σύμφωνα με τα όρια b_k. Πιο συγκεκριμένα, το μοντέλο επιστρέφει μία βαθμολογία r τέτοια ώστε $b_{r-1} \leq score(x) < b_r$. Το μοντέλο εκπαιδεύεται χρησιμοποιώντας τον αλγόριθμο PRank (Perceptron Ranking algorithm), ο οποίος αντιδρά σε λανθασμένες προβλέψεις κατά την διάρκεια της εκπαίδευσης, ενημερώνοντας τα διανύσματα βάρους (w) και ορίων (b).

4.7.1 Datasets

Το πρώτο σύνολο δεδομένων πειραμάτων είναι το OpinRank Dataset, το οποίο παρουσιάζεται στην έρευνα (Ganesan and ChengXiang, 2012) και αποτελείται από οντότητες οι οποίες συνοδεύονται review κείμενα χρηστών από δύο διαφορετικά domains (αυτοκίνητα και ξενοδοχεία). Τα review κείμενα προέρχονται από τις ιστοσελίδες Edmunds.com και Tripadvisor.com αντίστοιχα. Στο πρώτο σύνολο πειραμάτων χρησιμοποιούμε review κείμενα από το domain των αυτοκινήτων τα οποία αποτελούνται από μοντέλα αυτοκινήτων και τα αντίστοιχα review κείμενα, για τις χρονιές 2007 (227

μοντέλα), 2008 (228 μοντέλα), και 2009 (143 μοντέλα). Τα κείμενα των reviews αποτελούνται κατά μέσο όρο από 3000 λέξεις.

Το δεύτερο σύνολο δεδομένων πειραμάτων είναι μία συλλογή review κειμένων για εστιατόρια από την ιστοσελίδα www.we8there.com. Κάθε review κείμενο συνοδεύεται από ένα σύνολο από 5 βαθμολογίες, κάθε μία στο διάστημα τιμών 1-5, μία για κάθε ένα από τα ακόλουθα πέντε aspects:

{food. ambience. service, value, experience}

Αυτές οι βαθμολογίες έχουν δοθεί από τους καταναλωτές που έχουν γράψει τα reviews. Στο δεύτερο σύνολο πειραμάτων χρησιμοποιούμε 420 review κείμενα, με μέσο όρο 115 λέξεις, όπως αυτά δημοσιεύονται στο σύνδεσμο:

http://people.csail.mit.edu/bsnyder/naacl07/data/

Το παραπάνω σύνολο έχει υποστεί προεπεξεργασία ώστε να απομονωθεί το κείμενο των reviews από τις βαθμολογίες των aspects, και να υπολογιστούν οι MAAR βαθμολογίες. Για να αξιολογήσουμε τα σχήματά μας, παράγουμε ένα σύνολο 31 ερωτήσεων από το σύνολο των 5 διαθέσιμων aspects (food, ambience, service, value, experience), και για κάθε μία από αυτές παράγουμε την ιδανική κατάταξη οντοτήτων με βάση τις βαθμολογίες των aspects που έχουν δόσει οι χρήστες.

4.7.2 Αποτελέσματα, παρατηρήσεις και συμπεράσματα

Αρχικά συγκρίνουμε την απόδοση την BM25 μοντέλου με, και χωρίς, τις AvgScoreQAM and opinion expansion επεκτάσεις, οι οποίες παρουσιάζονται στην έρευνα (Ganesan and ChengXiang, 2012). Παρατηρούμε πως και στα δύο σύνολα πειραμάτων που πραγματοποιήσαμε, η χρήση του BM25 μοντέλου με τις επεκτάσεις δίνει καλύτερα αποτελέσματα. Αυτή είναι μία από τις κύριες παρατηρήσεις στην έρευνα (Ganesan and ChengXiang, 2012), και εδώ μπορούμε να πούμε πως επαληθεύεται. Στα πειράματά μας, ωστόσο, δεν είδαμε την αναμενόμενη αύξηση στην απόδοση στο πρώτο σύνολο πειραμάτων. Πρέπει να τονιστεί πως στα πειράματά μας δεν χρησιμοποιήσαμε κάποιο μηχανισμό pseudo feedback, όπως στην έρευνα (Ganesan and ChengXiang, 2012).

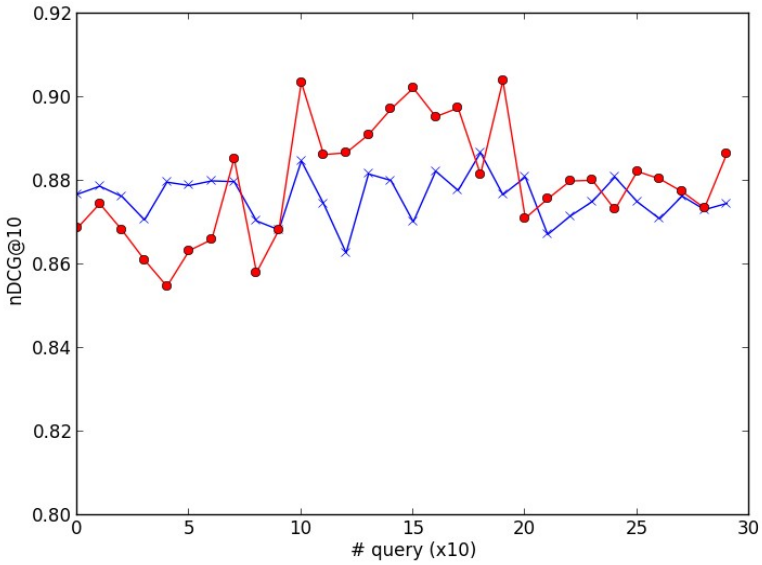


Figure 3
Μετρήσεις της nDCG@10 από το πρώτο σύνολο πειραμάτων. Συγκρίνεται η απόδοση του BM25std+AvgScoreQAM+opinExp (κόκκινο) με αυτήν τουBM25std μοντέλου (μπλε).

Τα πειραματικά αποτελέσματα δείχνουν πως η sentiment πληροφορίας από τα review κείμενα για την αξιολόγηση των οντοτήτων, μπορεί να χρησιμοποιηθεί το ίδιο καλά με τις κλασικές τεχνικές ανάκτησης πληροφορίας, όπως η μετρική BM25. Και τα δύο σχήματα sentiment analysis πραγματοποιούν ανάλυση σε επίπεδο πρότασης, κάθε ένα με διαφορετικό τρόπο. Στο πρώτο σύνολο πειραμάτων, τα σχήματά αποδίδουν σχεδόν το ίδιο καλά, ένω στο δεύτερο σύνολο πειραμάτων, η τεχνική η οποία βασίζεται σε συντακτικά πρότυπα αποδίδει καλύτερα από αυτή που βασίζεται στο λεξικό. Πιστεύουμε πως η καλύτερη απόδοση της τεχνικής που βασίζεται στα συντακτικά πρότυπα στο δεύτερο σύνολο πειραμάτων, πιθανόν οφείλεται στο γεγονός πως στο δεύτερο dataset τα review κείμενα είναι μικρότερα σε μήκος (μέσος όρος λέξεων ανά review 115 λέξεις) και οι χρήστες εκφράζουν άμεσα και ξεκάθαρα τις απόψεις τους διαμορφώνοντας απλές εκφράσεις. Θα πρέπει να τονιστεί πως αν και επιλέξαμε να χρησιμοποιήσουμε μη εποπτευόμενες τεχνικές sentiment analysis, οι οποίες δεν πραγματοποιούν ανάλυση σε βάθος, πιστεύαμε πως θα μπορούσαν να ξεπεράσουν σε απόδοση την προσέγγιση της ανάκτησης πληροφορίας. Και αυτό διότι έχουν την ικανότητα να αναγνωρίζουν επιπλέον και τις αρνητικές, απόψεις, γνώση την οποία αγνοεί ένα μοντέλο όπως το BM25std+AvgScoreQAM+opinExp. Ωστόσο δεν καταφέραμε να

παρατηρήσουμε κάτι τέτοιο. Όμως ακόμα πιστεύουμε πως πιο εξειδικευμένες τεχνικές sentiment analysis, μπορούν να πετύχουν κάτι τέτοιο. Δεν πρέπει να ξεχνάμε πως οι τεχνικές sentiment analysis έχουν να αντιμετωπίσουν την ποικιλομορφία της ανθρώπινης έκφρασης. Οι άνθρωποι χρησιμοποιούν πολλούς τρόπους για να εκφράσουν τις απόψεις τους, και υπάρχουν πολλοί τύποι απόψεων. Από την άλλη, η απόδοση και η απλότητα της προσέγγισης της ανάκτησης πληροφορίας, την καθιστά μία ελκυστική επιλογή.

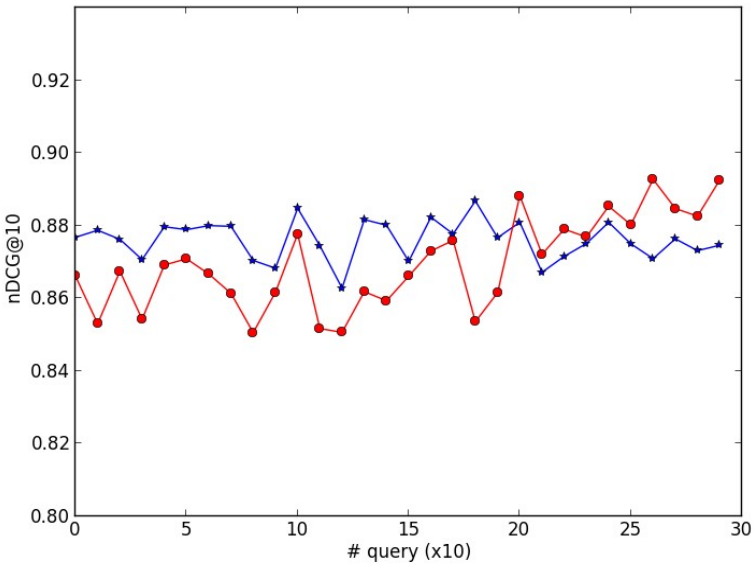


Figure 4
Μετρήσεις της nDCG@10 από το πρώτο σύνολο πειραμάτων. Συγκρίνεται η απόδοση του δεύτερου σχήματος sentiment analysis (syntactic patterns-based SA) (κόκκινο χρώμα) με αυτή του BM25std (μπλε χρώμα).

method	Average of nDCG@10
BM25std	0.87
BM25std+AvgScoreQAM+opinExp	0.88
lexicon-based SA	0.865
syntactic patterns-based SA	0.869
BM25std+kurland	0.88
BM25std+Kurland+opinion-based clusters	**0.89**

Table 1
Παρουσιάζεται ο μέσος όρος της nDCG@10 στο σύνολο των ερωτημάτων για όλα τα σχήματα στο πρώτο σύνολο πειραμάτων.

Επιπλέον παρατηρούμε την απόδοση των δύο clustering σχημάτων, του BM25std+Kurland και του BM25std+Kurland+opinion-based clusters, με τα οποία επιδιώκουμε να εκμεταλλευτούμε clustering πληροφορία από τα review κείμενα έτσι ώστε να βελτιώσουμε την κατάταξη των οντοτήτων. Στο πρώτο σύνολο πειραμάτων το σχήμα BM25std+Kurland αποδίδει καλά, ενώ στο δεύτερο εμφανίζει χαμηλή απόδοση. Η χαμηλή απόδοση του σχήματος BM25std+kurland οφείλεται πιθανόν στο γεγονός πως τα κείμενα στο δεύτερο dataset έχουν μικρό μήκος (μέσος όρος λέξεων ανά κείμενο 115 λέξεις). Οπότε η αναπαράστασή τους στο διανυσματικό χώρο των χαρακτηριστικών είναι παρόμοια, κάτι το οποίο εισάγει θόρυβο στην διαδικασία του clustering με τον αλγόριθμο k-means. Το σχήμα BM25std+kurland+opinion-based clusters, αποδίδει καλά και στα δύο σύνολα πειραμάτων. Επίσης είναι πάντα καλύτερο από την standard BM25 φόρμουλα και το σχήμα BM25std+kurland. Οπότε μπορούμε να πούμε πως το opinion-based clustering μπορεί να αναγνωρίσει παρόμοιες συμπεριφορές αξιολόγησης σε παρόμοια aspect ερωτήματα ανάμεσα στις οντότητες, και μπορεί να χρησιμοποιήσει αυτή την πληροφορία για να διαμορφώσει μία καλύτερη κατάταξη οντοτήτων. Αυτό επίσης δείχνει πως το opinion-based clustering είναι περισσότερο κατάλληλο για μία διαδικασία opinion-based κατάταξης οντοτήτων από το content clustering.

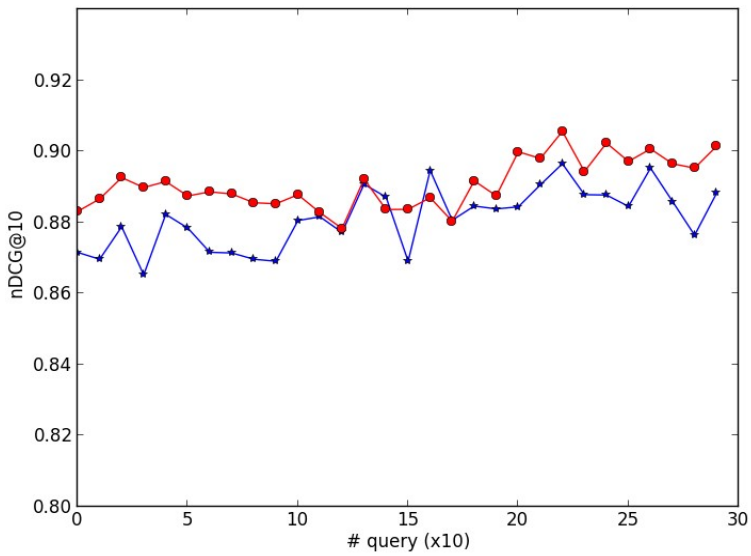


Figure 5
Μετρήσεις της nDCG@10 από το πρώτο σύνολο πειραμάτων.
Συγκρίνεται η απόδοση του σχήματος BM25+kurland+opinion-based clusters (κόκκινο) με αυτή του BM25+kurland (μπλε).

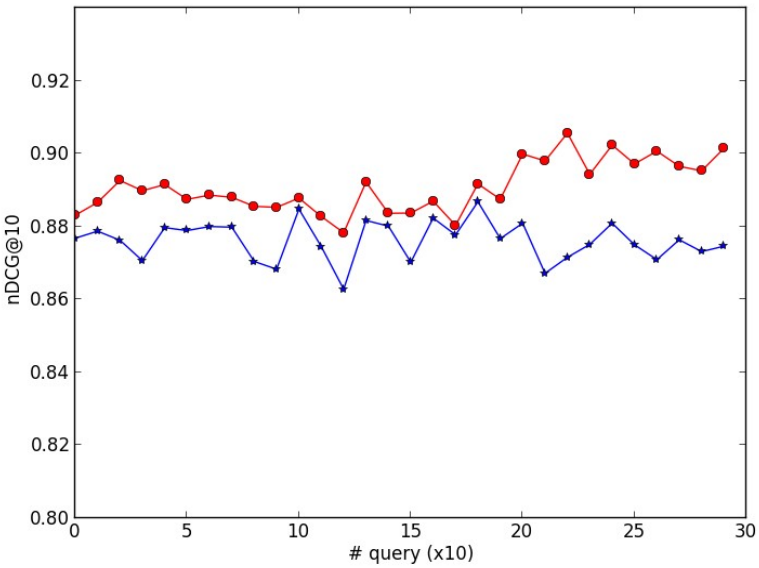


Figure 6
Μετρήσεις της nDCG@10 από το πρώτο σύνολο πειραμάτων.
Συγκρίνεται η απόδοση του σχήματος BM25+kurland+opinion-based clusters (κόκκινο) με αυτή του BM25std (μπλε).

method	Average of nDCG@10
BM25std	0.936
BM25std+AvgScoreQAM+opinExp	**0.955**
lexicon-based SA	0.91
syntactic patterns-based SA	0.94
BM25std+kurland	0.90
BM25std+Kurland+opinion-based clusters	**0.956**
NCM1	0.938
NCM2	**0.96**
m-aspect Prank	**0.95**

Table 2
Παρουσιάζεται ο μέσος όρος της nDCG@10 στο σύνολο των ερωτημάτων για όλα τα σχήματα στο δεύτερο σύνολο πειραμάτων.

Σύμφωνα με τις μετρήσεις και τα δύο naive consumer μοντέλα δείχνουν αποδίδουν καλύτερα στο πρώτο σύνολο πειραμάτων, και στο δεύτερο σύνολο πειραμάτων, το σχήμα το οποίο χρησιμοποιεί το setup του Kurland και κάνει χρήση clustering πληροφορίας, δείχνει να ξεπερνά σε απόδοση ακόμα και την εποπτευόμενη τεχνική κατηγοριοποίησης του m-aspect PRank μοντέλου. Οπότε βλέπουμε πως η γνώση του πόσο σημαντικό είναι ένα aspect, για να χρησιμοποιηθεί σε μία διαδικασία κατάταξης οντοτήτων με βάση τις απόψεις χρηστών πάνω σε συγκεκριμένα aspect, παίζει όντως σημαντικό ρόλο, όπως επίσης και η γνώση των ομάδων των aspects όπως αυτές καθορίζονται από την κοινότητα των χρηστών.

Προκειμένου να παρέχουμε πιο αναλυτικές υποδείξεις χρειαζόμαστε επιπλέον πειράματα με δεδομένα από διάφορους χώρους εφαρμογής, ωστόσο σε αυτή την έρευνα στόχος μας ήταν να δείξουμε ένα “proof of concept” της ορθότητας της προσέγγισής μας. Η προσέγγιση της ανάκτησης πληροφορίας με τις δύο επεκτάσεις, το query aspect modeling, και το opinion expansion, η οποία παρουσιάζεται στην έρευνα (Ganesan and ChengXiang, 2012), είναι μία ελκυστική επιλογή η οποία δουλεύει. Το NCM μοντέλο μπορεί να χρησιμοποιηθεί για την δημιουργία περισσότερων αξιόπιστων κατατάξεων οντοτήτων, λόγω της γνώσης που εξάγει από το διαδίκτυο. Το opinion- based clustering σχήμα μπορεί να χρησιμοποιηθεί για την δημιουργία περισσότερο ακριβών κατατάξεων οντοτήτων. Όσον αφορά τις τεχνικές senitment analysis,

οι οποίες είναι αυτές που πιθανόν να δώσουν την πλήρη λύση στο πρόβλημα της κατάταξης οντοτήτων με βάση τις απόψεις χρηστών σε διάφορα χαρακτηριστικά, μέχρι στιγμής, είναι εξαρτημένες από το επίπεδο της ανάλυσης και τα χαρακτηριστικά του κειμένου που περιέχει την υποκειμενική πληροφορία. Η τεχνική sentiment analysis η οποία βασίζεται στα συντακτικά πρότυπα, στο δεύτερο σύνολο των πειραμάτων μας, έχει καλύτερη απόδοση από αυτήν που βασίζεται στο λεξικό. Στο δεύτερο dataset τα review κείμενα είναι μικρά σε μήκος και οι χρήστες εκφράζουν άμεσα και ξεκάθαρα τις απόψεις τους, διαμορφώνοντας απλές εκφράσεις, ενώ στο πρώτο dataset τα review κείμενα έχουν μεγαλύτερο μήκος και η εξαγωγή των απόψεων γίνεται πολύπλοκη. Ωστόσο υπάρχουν datasets τα οποία αποτελούνται από κείμενα μικρού μήκους, όπως twitter datasets, στα οποία όμως η εξαγωγή των απόψεων είναι γενικά δύσκολη και απαιτεί τεχνικές οι οποίες πραγματοποιούν sentiment ανάλυση σε βάθος.

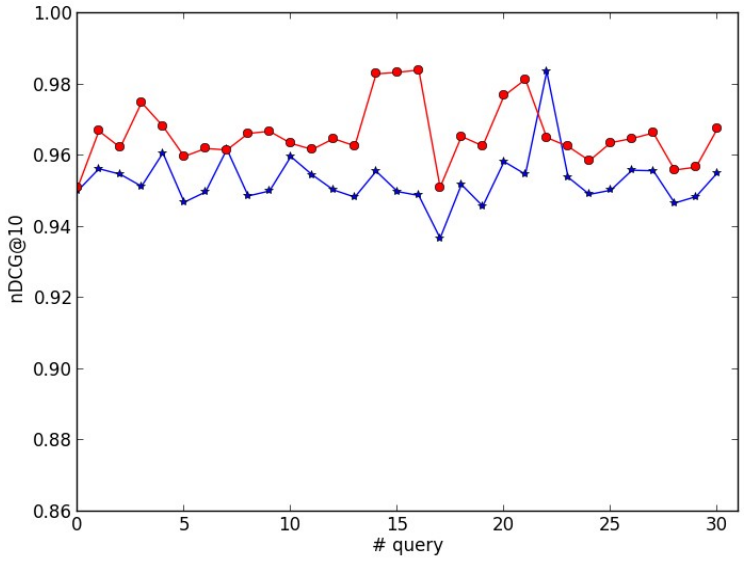


Figure 7
Μετρήσεις της nDCG@10 από το δεύτερο σύνολο πειραμάτων. Συγκρίνεται η απόδοση του NMC2 σχήματος (κόκκινο) με αυτή του PRank μοντέλου (μπλε).

5 ΕΠΙΛΟΓΟΣ

Σε αυτό το κείμενο πραγματοποιήθηκε μία αναλυτική αναφορά στο αντικείμενο και τις συνήθης τεχνικές της εξόρυξης και ανάλυσης απόψεων, και παρουσιάστηκε το πρόβλημα της αξιολόγησης οντοτήτων χρησιμοποιώντας τις απόψεις που υπάρχουν σε review κείμενα χρηστών, για το οποίο αναφέρθηκαν νέες τεχνικές προσέγγισης.

Ο τομέας του opinion mining and sentiment analysis είναι ο ερευνητικός τομέας που αναλύει τις ανθρώπινες απόψεις, τα συναισθήματα, τις αξιολογήσεις και τις στάσεις, προς οντότητες όπως προϊόντα, υπηρεσίες, οργανισμούς, πρόσωπα, εκδηλώσεις, θεματικές ενότητες, καθώς και προς τις ιδιότητες αυτών. Από το 2000 ο τομέας του Opinion Mining έχει γίνει μία αρκετά ενεργή ερευνητική περιοχή. Η ανάγκη όμως για την γνώση των απόψεων της κοινής γνώμης υπάρχει από πολύ παλιά. Η συγκέντρωση και η ανάλυση των απόψεων, της κοινής γνώμης ή συγκεκριμένων ομάδων ατόμων, δεν είναι κάτι νέο. Οι δημοσκοπήσεις και τα ερωτηματολόγια προϋπήρξαν των υπολογιστών. Ο κύριος λόγος που η εξόρυξη και ανάλυση απόψεων συγκεντρώνει το ερευνητικό ενδιαφέρον αρκετών ερευνητών τα τελευταία χρόνια είναι πως για πρώτη φορά στην ανθρώπινη ιστορία, έχουμε ένα τεράστιο όγκο από δεδομένα απόψεων στα κοινωνικά δίκτυα στο διαδίκτυο . Χωρίς αυτά τα δεδομένα, μεγάλο μέρος της έρευνας δεν θα ήταν πραγματοποιήσιμο.

Αν και η επιστημονική κοινότητα έχει ερευνήσει την εξόρυξη και ανάλυση απόψεων από πολλές διαφορετικές οπτικές γωνίες, και έχουν δημοσιευθεί αρκετές έρευνες κανένα από τα υποπροβλήματα δεν έχει λάβει ικανοποιητική λύση. Ο κύριος λόγος είναι πως η διαδικασία της εξόρυξης και ανάλυσης απόψεων είναι μία διαδικασία natural language processing. Οι διεργασίες της εξόρυξης και ανάλυσης απόψεων απαιτούν την ύπαρξη αλγορίθμων οι οποίοι να είναι ικανοί να εξάγουν νόημα και πληροφορία από την ανθρώπινη γλώσσα, καθώς και να κατανοούν την ανθρώπινη ψυχολογία και φύση. Ο τομέας του Natural Language Processing ασχολείται με την επεξεργασία της ανθρώπινης γλώσσας και την εξαγωγής νοήματος από αυτή, αλλά δεν έχει προσφέρει, μέχρι στιγμής, ισχυρά εργαλεία. Αυτό οδηγεί τους ερευνητές να επιλέγουν τεχνικές ανάλυσης κειμένου, όπως τεχνικές ανάκτησης πληροφορίας, λεξικής ανάλυσης, αναγνώρισης προτύπων, εξόρυξης δεδομένων, και άλλες υπολογιστικές και στατιστικές τεχνικές, για την πραγματοποίηση των διαδικασιών της εξόρυξης και ανάλυ-

σης απόψεων. Αυτές όμως οι τεχνικές δεν μπορούν παρά μόνο να παράγουν μία προσέγγιση της πλήρης λύσης του προβλήματος, διότι δεν είναι ικανές να διαχειριστούν την ποικιλομορφία της ανθρώπινης έκφρασης. Όμως οι χρήσιμες και ενδιαφέρουσες πρακτικές εφαρμογές, καθώς και οι προκλήσεις της εξόρυξης και ανάλυσης απόψεων, απαιτούν την συνέχιση της έρευνας, και φαίνεται πως θα διατηρήσουν αυτόν τον ερευνητικό τομέα ενεργό τα επόμενα χρόνια. Όσον αφορά τις πρακτικές εφαρμογές, μία τελείως αυτοματοποιημένη και αρκιβής λύση δεν φαίνεται να υπάρχει. Οι περισσότερες τεχνικές ακολουθούν semi-automated προσεγγίσεις. Ωστόσο με την βαθύτερη κατανόηση των διαφόρων υποπροβλημάτων, και των μεταξύ τους συσχετίσεων, μπορούμε να κινούμαστε όλο και περισσότερο προς την κατεύθυνση της δημιουργίας μιας πλήρους αυτοματοποιημένης διαδικασίας.

Όσον αφορα το πρόβλημα της αξιολόγησης οντοτήτων χρησιμοποιώντας τις απόψεις που υπάρχουν σε review κείμενα χρηστών, δεν μπορούμε να πούμε πως πλησιάσαμε στην πλήρη λύση του προβλήματος. Οι τεχνικές που εξετάσαμε και προτείναμε υπόσχονται μόνο προσεγγίσεις της λύσης του προβλήματος. Αρχικά επιλέξαμε να υιοθετήσουμε και να εξετάσουμε ένα απλό και αποδοτικό setup το οποίο μας επιτρέπει να μετατρέψουμε το πρόβλημα της αξιολόγησης οντοτήτων σε πρόβλημα ανάκτησης πληροφορίας, και έτσι να προσεγγίσουμε την λύση του εφαρμόζοντας τεχνικές ανάκτησης πληροφορίας. Συγκρίνοντας την απόδοση αυτού του σχήματος με την απόδοση δύο μη εποπτευόμενων τεχνικών sentiment analysis, ανακαλύψαμε πως δεν υστερεί σε απόδοση. Οι τεχνικές sentiment analysis που επιλέξαμε είναι απλές και δεν πραγματοποιούν ανάλυση σε βάθος. Πιστεύουμε πιο εξειδικευμένες τεχνικές sentiment analysis θα μπορούσαν να ξεπεράσουν σε απόδοση την προσέγγιση της ανάκτησης πληροφορίας. Λόγω της απλότητας και της απόδοσης του σχήματος που διαχειρίζεται του πρόβλημα της αξιολόγησης οντοτήτων ως πρόβλημα ανάκτησης πληροφορίας, αποφασίσαμε να μελετήσουμε περισσότερο αυτήν την προσέγγιση και να δοκιμάσουμε τρόπους βελτίωσης της απόδοσής της. Επιχειρήσαμε να εισάγουμε στο σχήμα επιπλέον πληροφορία χρησιμοποιώντας clustering και το naive consumer μοντέλο. Με το clustering προσπαθήσαμε να αναγνωρίσουμε παρόμοιες συμπεριφορές βαθμολόγησης ανάμεσα στους χρήστες. Προτείναμε, επίσης, το naïve consumer μοντέλο ως ένα μη εποπτευόμενο σχήμα, το οποίο χρησιμοποιεί πληροφορίες από το διαδίκτυο ώστε να αποδώσει ένα βάρος σημαντικότητας σε κάθε ένα από τα aspects τα οποία χρησιμοποιούνται στην αξιολόγηση των οντοτήτων.

Τα πειράματά μας έδειξαν πως η προσθήκη των παραπάνω πληροφοριών στην διαδικασία της αξιολόγησης οντοτήτων, βελτιώνει την απόδοση της διαδικασίας. Επίσης παρατηρήσαμε πως το opinion-based clustering είναι περισσότερο κατάλληλο για μία διαδικασία opinion-based κατάταξης οντοτήτων από το content clustering, και πως το η γνώση του πόσο σημαντικό είναι ένα aspect, παίζει όντως σημαντικό ρόλο, όπως επίσης και η γνώση των ομάδων των aspects όπως αυτές καθορίζονται από την κοινότητα των χρηστών. Πρέπει να τονίσουμε πως σε καμία περίπτωση δεν πρέπει να θεωρηθούν τα πειράματά μας πλήρη. Αν και πραγματοποιήσαμε δύο σύνολα πειραμάτων με δύο διαφορετικά σύνολα δεδομένων, ως προς τη θεματική ενότητα και τα χαρακτηριστικά τους, κατανοούμε πως τα σχήματα πρέπει να εξεταστούν και να αξιολογηθούν περισσότερο. Επίσης η προσέγγιση της αντιμετώπισης του προβλήματος της αξιολόγησης οντοτήτων ως πρόβλημα ανάκτησης πληροφορίας δείχνει πως δέχεται επιπλέον βελτιώσεις. Όσον αφορά τα clustering σχήματα αξίζει να δοκιμαστούν και άλλοι αλγόριθμοι συσταδοποίησης, καθώς και να γίνουν πειράματα με χώρους χαρακτηριστικών οι οποίοι να είναι ικανοί να συλλάβουν την υποκειμενική φύση των κειμένων. Όσον αφορά το naive consumer μοντέλο και την απόδοση βαρών στα aspects αξίζει να δοκιμαστούν και άλλες φόρμουλες υπολογισμού της σημαντικότητας. Επίσης αξίζει να ερευνηθούν περισσότερο οι οντότητες και ιδιαίτερα τα aspects αυτών, όσον αφορά τις συσχετίσεις μεταξύ τους, καθώς επίσης και τις εξαρτήσεις που ίσως εμφανίζονται στα σύνολα των aspect ερωτημάτων.

6 ΠΑΡΑΡΤΗΜΑ

I: Penceptron Ranking algorithm (PRank)

Το μοντέλο Penceptron Ranking ή PRank αλγόριθμος είναι ένα μοντέλο για την πρόβλεψη βαθμολογιών (Crammer και Singer, 2001). Έχοντας αναπαραστήσει κάθε ένα από τα review κείμενα r_i ως ένα feature διάνυσμα $x \in Rn$, αυτό το μοντέλο προβλέπει μία τιμή $y \in \{1, ..., k\}$ για κάθε $x \in Rn$. Το μοντέλο αποθηκεύει ένα διάνυσμα βάρους $w \in Rn$ και ένα διάνυσμα ορίων $b_0 = -\infty \leq b_1 \leq ... \leq b_{k-1} \leq b_k = \infty$ το οποίο διαιρεί την γραμμή των πραγματικών αριθμών σε k στάθμες, μία για κάθε πιθανή βαθμολογία. Το μοντέλο αποδίδει μία βαθμολογία για κάθε είσοδο χρησιμοποιώντας το διάνυσμα βάρους: score(x) = w * x. Τέλος, προβάλει την βαθμολογία score(x) στον άξονα των πραγματικών αριθμών, και επιστρέφει την κατάλληλη βαθμολογία σύμφωνα με τα όρια b_k. Πιο συγκεκριμένα, το μοντέλο επιστρέφει μία βαθμολογία r τέτοια ώστε $b_{r-1} \leq score(x) < b_r$. Το μοντέλο εκπαιδεύεται χρησιμοποιώντας τον αλγόριθμο PRank (Perceptron Ranking algorithm), ο οποίος αντιδρά σε λανθασμένες προβλέψεις κατά την διάρκεια της εκπαίδευσης, ενημερώνοντας τα διάνυσματα βάρους (w) και ορίων (b).

Στον τομέα του machine learning, ο penceptron (Frank Rosenblatt, 1958) είναι ένας αλγόριθμος για εποπτευόμενη κατηγοριοποίηση μιας εισόδου σε μία από τις δύο πιθανές κλάσεις. Είναι ένα είδος γραμμικού κατηγοριοποιητή. Ο αλγόριθμος πραγματοποιεί προβλέψεις με βάση μία γραμμική συνάρτηση πρόβλεψης η οποία συνδυάζει ένα σύνολο βαρών με το διάνυσμα εισόδου των χαρακτηριστικών. Ο αλγόριθμος μάθησης είναι ένας online αλγόριθμος και επεξεργάζεται σειριακά τα δεδομένα εκπαίδευσης. Ο penceptron κατηγοριοποιητής δέχεται ως είσοδο ένα διάνυσμα x και δίνει ως έξοδο μία τιμή σύμφωνα με την εξής συνάρτηση:

$$f(x) = \begin{cases} 1, w * x + b > 0 \\ 0 \end{cases}$$

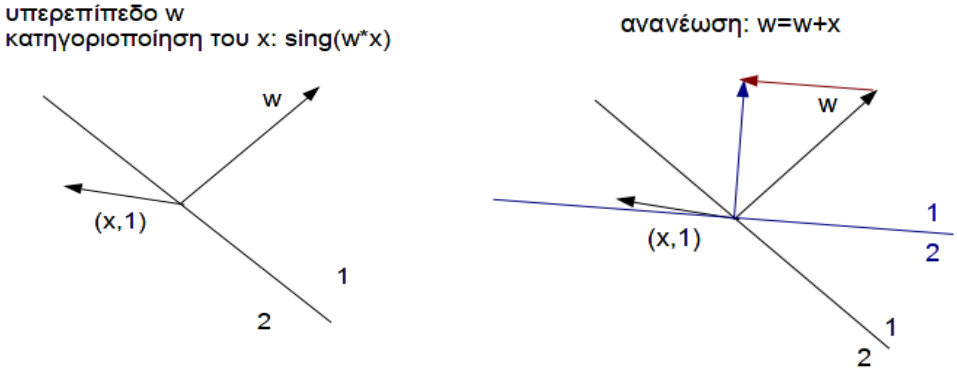


Figure 8
δυαδικός penceptron κατηγοριοποιητής: διανυσματική αναπαράσταση της εκπαίδευσης ενός δυαδικού κατηγοριοποιητή penceptron. Δεξιά φαίνεται η ανανέωση του διανύσματος βάρους w.

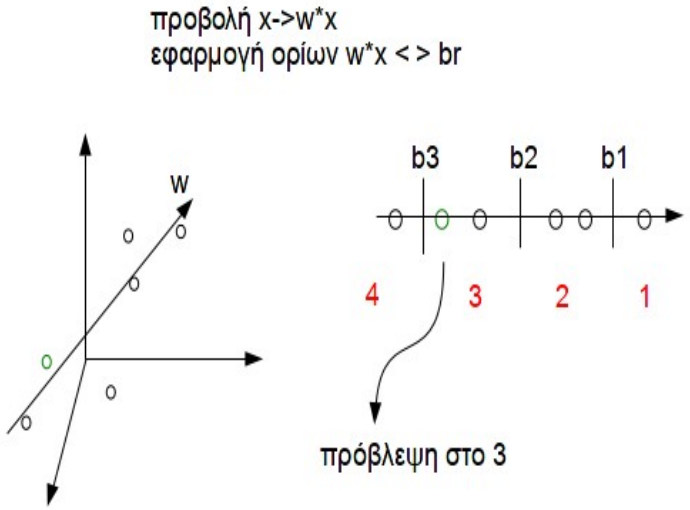


Figure 9
PRank μοντέλο: παρουσιάζεται η διαδικασία πρόβλεψης

II: Support Vector Machines (SVM)

Στον ερευνητικό τομέα του machine learning, τα support vector machines (SVM), είναι μοντέλα εποπτευόμενης μάθησης τα οποία χρησιμοποιούνται για κατηγοριοποίηση και για regression ανάλυση. Το βασικό SVM μοντέλο λαμβάνει ένα σύνολο δεδο-

μένων εισόδου και προβλέπει, για κάθε ένα στοιχείο αυτού του συνόλου, σε ποια από τις δύο πιθανές κλάσεις ανήκει.

Για την εκπαίδευση ενός SVM μοντέλου χρειάζεται ένα σύνολο από δεδομένα εκπαίδευσης, κάθε στοιχείο του οποίου πρέπει να είναι απισημασμένο (labeled) με την κλάση στην οποία ανήκει. Ο αλγόριθμος εκπαίδευσης δημιουργεί ένα μοντέλο το οποίο έχει την δυνατότητα να κατηγοριοποιεί νέα δεδομένα στις δύο πιθανές κατηγορίες. Ένα SVM μοντέλο είναι απλά η αναπαράσταση των δεδομένων εκπαίδευσης στο χώρο, με τέτοιο τρόπο ώστε τα στοιχεία των δύο κατηγοριών να χωρίζονται με όσο το δυνατόν μεγαλύτερο κενό μεταξύ τους. Τα δεδομένα προς κατηγοριοποίηση, τοποθετούνται στον ίδιο χώρο και ανάλογα σε ποια πλευρά του κενού βρίσκονται προβλέπεται σε ποια από τις δύο κλάσεις ανήκουν.

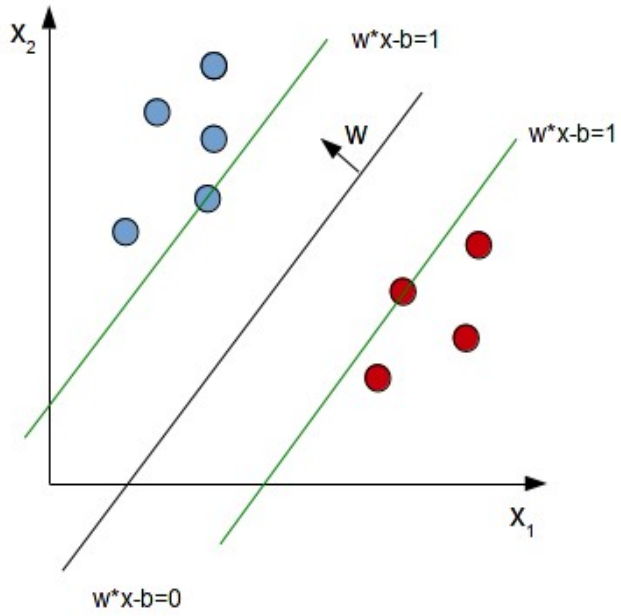


Figure 10
SVM: Maximum-margin υπερεπίπεδο και όρια για ένα SVM μοντέλο εκπαιδευμένο με δεδομένα δύο κλάσεων. Τα δείγματα πάνω στα όρια αποτελούν τα support vectors.

Συγκεκριμένα, ένα support vector machine δημιουργεί ένα υπερεπίπεδο ή σύνολο υπερεπιπέδων σε χώρο μεγαλύτερης ή άπειρης διάστασης, το οποίο μπορεί να χρησιμοποιηθεί για κατηγοριοποίηση ή regression. Τα δεδομένα εκπαίδευσης αναπαρίστα-

νται ως διανύσματα στο χώρο. Αν αυτός ο χώρος είναι διάστασης p τότε για να διαχωρίσουμε τα σημεία των δύο κλάσεων χρειαζόμαστε ένα υπερεπίπεδο διάστασης p-1 για την δημιουργία ενός γραμμικού κατηγοριοποιητή. Μπορούμε να επιλέξουμε πολλά τέτοια υπερεπίπεδα, αλλά ένας καλός διαχωρισμός των δεδομένων των δύο κλάσεων, επιτυγχάνεται από το υπερεπίπεδο αυτό το οποίο απέχει την μεγαλύτερη απόσταση από τα κοντινότερα δεδομένα εκπαίδευσης των δύο κλάσεων. Όσο μεγαλύτερο είναι το κενό μεταξύ των δεδομένων των δύο κλάσεων τόσο μικρότερο θα είναι το σφάλμα του κατηγοριοποιητή. Τα δείγματα πάνω στα όρια των κλάσεων ονομάζονται support vectors.

Εκτός της γραμμικής κατηγοριοποίησης, τα support vector machines μπορούν να χρησιμοποιηθούν και για μη-γραμμική κατηγοριοποίηση, χρησιμοποιώντας μία τεχνική η οποία ονομάζεται kernel trick, η οποία τοποθετεί, έμμεσα, τα δεδομένα σε χώρους μεγαλύτερη διάστασης. Εκτός της κατηγοριοποίησης δεδομένων σε δύο κατηγορίες, τα support vector machines μπορούν να χρησιμοποιηθούν και για την κατηγοριοποίηση σε περισσότερες κατηγορίες. Η κύρια προσέγγιση για να πραγματοποιηθεί αυτό είναι η μείωση του προβλήματος της πολλαπλής κατηγοριοποίησης σε πολλαπλά δυαδικά προβλήματα κατηγοριοποίησης.

III: Hidden Markov Model (HMM)

Το κρυφό Markov μοντέλο είναι μία επέκταση των Markov μοντέλων (MM), στο οποίο το σύστημα που μοντελοποιείται είναι μία διαδικασία Markov με μη παρατηρήσιμες (κρυφές-hidden) καταστάσεις. Μία διαδικασία Markov είναι μία στοχαστική διαδικασία η οποία ικανοποιεί την ιδιότητα Markov ή ιδιότητα της αμνησίας. Σε μία διαδικασία Markov η μελλοντική κατάσταση εξαρτάται μόνο από την παρούσα κατάσταση και η γνώση των προηγούμενων καταστάσεων δεν έχει κανένα αντίκτυπο στην μετάβαση. Το μέλλον και το παρελθόν είναι ανεξάρτητα. Ένα μοντέλο Markov είναι ένας κατευθυνόμενος γράφος, όπου οι κορυφές αναπαριστούν καταστάσεις και οι ακμές δείχνουν μεταβάσεις μεταξύ των καταστάσεων. Κάθε ακμή χαρακτηρίζεται από μία πιθανότητα μετάβασης.

Το hidden Markov μοντέλο διαφέρει από το απλό Markov μοντέλο στο ότι οι καταστάσεις σε ένα HMM δεν αντιστοιχούν σε παρατηρήσιμες καταστάσεις. Ένα HMM μοντελοποιεί μία διαδικασία η οποία παράγει ως έξοδο μία παρατηρήσιμη ακολουθία, αλλά η ακολουθία καταστάσεων που παρήγαγε αυτήν την παρατηρήσιμη ακολουθία

είναι δεν είναι γνωστή, είναι κρυφή. Δεν υπάρχει σχέση μεταξύ των καταστάσεων και των τιμών που παρατηρούνται. Επίσης μία ακολουθία παρατηρήσεων μπορεί να παραχθεί από πολλές ακολουθίες καταστάσεων. Όπως και το Markov μοντέλο το hidden Markov μοντέλο αποτελείται από ένα σύνολο καταστάσεων με πιθανότητες μετάβασης. Στο HMM υπάρχει επιπλέον σε κάθε κατάσταση μία κατανομή πιθανότητας παρατήρησης οι οποία καθορίζει την έξοδο της κατάστασης.

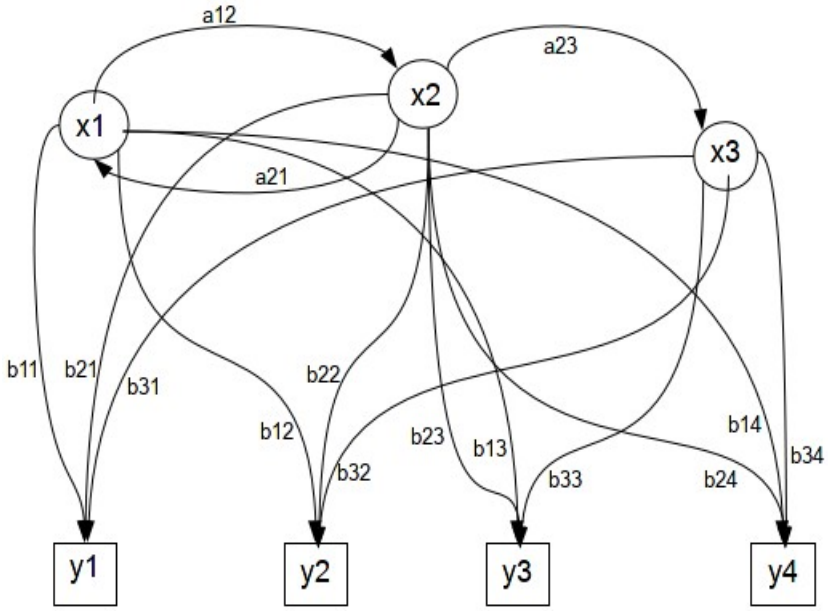


Figure 11
παράδειγμα HMM: x — καταστάσεις, y — πιθανές παρατηρήσεις, a — πιθανότητες μετάβασης καταστάσεων, b — πιθανότητες παρατήρησης/εξόδου.

Τα HMM, όπως και τα MM, μπορούν να αναγνωρίζουν αλλά και να προβλέπουν. Κατά την διάρκεια της μάθησης προσδιορίζεται ο αριθμός των καταστάσεων, οι πιθανότητες μετάβασης, και οι κρυμμένες πιθανότητες παρατήρησης.

Υπάρχουν τρία βασικά προβλήματα στα οποία χρησιμοποιούνται τα HMM:

- Εύρεση της πιθανότητας το μοντέλο HMM να παρήγαγε μία ακολουθία παρατήρησης. Δοθείσης μίας ακολουθίας από αντικείμενα που έχουν παρατηρηθεί και ενός HMM, ποια είναι η πιθανότητα το HMM να παρήγαγε την ακολου-

θία. Εάν η πιθανότητα είναι μικρή τότε το μοντέλο πιθανόν να μην παρήγαγε την ακολουθία.

- Εύρεση της ακολουθίας καταστάσεων που παρήγαγε μία ακολουθία παρατήρησης. Δοθείσης μίας ακολουθίας παρατηρήσεων και ένα HMM μοντέλο, ποια είναι η πιο πιθανή ακολουθία καταστάσεων που παρήγαγε την ακολουθία παρατηρήσεων.
- Βελτίωση των παραμέτρων του μοντέλου.

Τα δύο πρώτα προβλήματα σχετίζεται με τα προβλήματα αναγνώρισης και πρόβλεψης που υπάρχουν στην ανάλυση απόψεων, όπως η εξαγωγή των aspects ή των οντοτήτων.

IV: LSI

Το LSI είναι μία μέθοδος ευρετηρίασης και ανάκτησης, η οποία χρησιμοποιεί την τεχνική singular value decomposition (SVD) (διάσπαση ιδιαζουσών τιμών) για να εντοπίσει πρότυπα στις σχέσεις μεταξύ των όρων και των θεμάτων τα οποία περιέχονται σε μία συλλογή κειμένων. Το LSI είναι βασισμένο στην αρχή πως οι λέξεις που χρησιμοποιούνται στις ίδιες θεματικές ενότητες τείνουν να έχουν παρόμοιες έννοιες. Το χαρακτηριστικό κλείδι της μεθόδου LSI είναι η δυνατότητά της να εξάγει το εννοιολογικό περιεχόμενο του κειμένου, δημιουργώντας σχέσεις μεταξύ των όρων που εμφανίζονται σε παρόμοιες θεματικές ενότητες.

Το όνομα Latent Semantic Indexing, δηλώνει την δυνατότητα της μεθόδου να συσχετίζει σημασιολογικά σχετικούς όρους, των οποίων οι σχέσεις είναι κρυμμένες “λανθάνουσες” σε μία συλλογή κειμένων. Η LSI μέθοδος ονομάζεται επίσης και latent semantic analysis (LSA), διότι αποκαλύπτει την λανθάνουσα σημασιολογική δομή η οποία υπάρχει στην χρήση των λέξεων σε ένα κείμενο, και πως αυτή η δομή μπορεί να χρησιμοποιηθεί για να εξαχθούν οι έννοιες του κειμένου ως απόκριση σε ένα ερώτημα κάποιου χρήστη ή σε μία διαδικασία αναζήτησης εννοιών.

Τα ερωτήματα, ή οι αναζητήσεις εννοιών, σε μία συλλογή κειμένων, η οποία έχει επεξεργαστεί με την μέθοδο LSI, επιστρέφει αποτελέσματα τα οποία είναι θεματικά σχετικά με τις έννοιες που περιγράφονται στα ερωτήματα, ακόμα και αν τα αποτελέσματα δεν μοιράζονται μια συγκεκριμένη λέξη ή λέξεις με τα κριτήρια αναζήτησης. Για αυτό το λόγο το LSI μπορεί να χρησιμοποιηθεί για πολυγλωσσική ανάκτηση, και κυρίως για να αντιμετωπιστούν τα φαινόμενα της συνωνυμίας και της πολυσημίας τα

οποία παρουσιάζονται, κατά την αναπαράσταση των εγγράφων και των ερωτημάτων ως διανύσματα ενός διανυσματικού χώρου, ως επακόλουθο της υπόθεσης πως οι όροι είναι ανεξάρτητοι μεταξύ τους και πως πρέπει να αντιστοιχίζουμε κάθε έναν από αυτούς σε ξεχωριστή διάσταση.

Πιο συγκεκριμένα, έστω το μητρώο όρων-εγγραφών C διαστάσεων $m_x n$ (m όροι και n κείμενα). Το LSI πραγματοποιεί μία παραλλαγή της SVD, την ανηγμένη διάσπαση ιδιαζουσών τιμών (rank-reduced-μειωμένης τάξης Singular Value Decomposition) στο μητρώο C, ώστε να αναγνωριστούν τα πρότυπα στις σχέσεις μεταξύ των όρων και των θεματικών ενοτήτων που περιέχονται στα κείμενα. Η SVD παραγοντοποίηση υπολογίζει τους χώρους διανυσμάτων των όρων και των κειμένων παραγοντοποιώντας το μητρώο C, σε τρία άλλα μητρώα, ένα U $m_x r$ μητρώο όρου-θεματικής ενότητας, ένα S $r_x r$ μητρώο ιδιαζουσών τιμών, και ένα V $n_x r$ μητρώο θεματικής ενότητας-κειμένου, τα οποία ικανοποιούν της εξής σχέση:

$$C_r = U_r S_r V_r'$$

Ύστερα μειώνεται η διάσταση του μητρώου S_r των ιδιαζουσών τιμών από r σε k, συνήθως σε k=200-300, μειώνοντας έτσι και τα το μέγεθος των άλλων δύο μητρώων με τους χώρους όρων και κειμένων, σε $m_x k$ και $n_x k$ αντίστοιχα. Αυτό πραγματοποιείται αφαιρώντας τις r-k μικρότερες ιδιάζουσες τιμές. Αυτή η μείωση των διαστάσεων έχει ως αποτέλεσμα να παραμένουν μόνο οι περισσότερο σημαντικές σημασιολογικές πληροφορίες στα κείμενα, αφαιρώντας τον θόρυβο που υπάρχει στον αρχικό χώρο του μητρώου C. Με αυτή την μείωση, το αρχικό μητρώο C, προσεγγίζεται ως μία χαμηλόβαθμη προσέγγιση k τάξης ως εξής:

$$C_k = U_k S_k V_k'$$

Έχοντας τη παραπάνω διάσπαση ένα ερώτημα q μπορεί να αναπαρασταθεί ως ένα διάνυσμα στο χώρο LSI ως εξής, $\vec{q}_k = S_k^{-1} * U_k' * \vec{q}$, όπου το q είναι ένα διάνυσμα διάστασης $m_x 1$ (έχει υπολογιστεί ως στήλη του αρχικού πίνακα C). Τότε η ομοιότητα μεταξύ δύο κειμένων ή κειμένου και ερωτήματος είναι ο υπολογισμός της συνημιτονοειδής ομοιότητας των δύο διανυσμάτων:

$$score(d_i, d_j) = \frac{V(d_i) * V(d_j)}{|V(d_i)| * |V(d_j)|}$$

Η μέθοδος LSI μπορεί να θεωρηθεί ως ελαστική συσταδοποίηση, αν ερμηνεύσουμε κάθε διάσταση του περιορισμένου χώρου (k διαστάσεις) ως συστάδα και την τιμή που έχει κάθε έγγραφο για αυτήν την διάσταση ως την κλασματική συμμετοχή του στην αντίστοιχη συστάδα.

V: PLSI

Η πιθανοτική λανθάνουσα σημασιολογική ανάλυση PLSA (Probabilistic latent semantic analysis) ή PLSI (Probabilistic latent semantic indexing), όπως διαφορετικά ονομάζεται, είναι μία τεχνική για την ανάλυση δεδομένων συνεμφάνισης. Όπως και στην λανθάνουσα σημασιολογική ευρετηρίαση (LSI) και σε αυτή την τεχνική αντλείται μία μικρής διάστασης αναπαράσταση των παρατηρούμενων μεταβλητών όσον αφορά την σχέσεις τους με συγκεκριμένες κρυφές (λανθάνουσες) μεταβλητές. Ενώ η λανθάνουσα σημασιολογική ανάλυση χρησιμοποιεί την ανηγμένη διάσπαση ιδιαζουσών τιμών για να μικρύνει τις διαστάσεις του μητρώου συνεμφάνισης, η πιθανοτική λανθάνουσα σημασιολογική ανάλυση βασίζεται σε ένα μείγμα αποσύνθεσης σύμφωνα με τα λανθάνων κατηγορικά μοντέλα, τα οποία συσχετίζουν ένα σύνολο από παρατηρούμενες μεταβλητές σε ένα σύνολο λανθάνων μεταβλητών. Μία λαμθάνουσα μεταβλητή λαμβάνει διακριτές τιμές (κλάση) και χαρακτηρίζεται από πρότυπα υπο συνθήκη πιθανοτήτων τα οποία καθορίζουν την πιθανότητα οι παρατηρούμενες μεταβλητές να λαμβάνουν συγκεκριμένες τιμές.

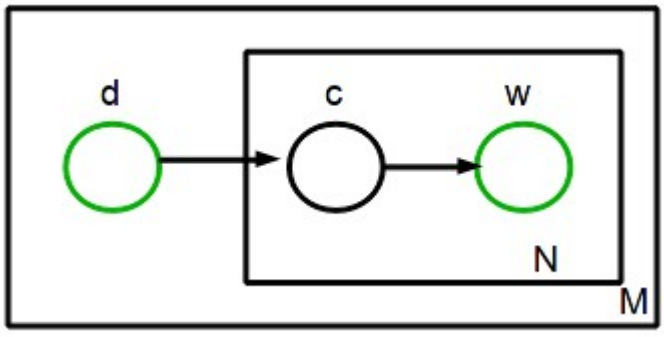


Figure 12
PLSA μοντέλο: παρουσιάζετια ένα PLSA μοντέλο, d είναι ένα κέιμενο ως μεταβλητή ευρετηρίασης, c είναι η θεματική ενότητα η οποία έχει παραχθεί σύμφωνα με την θεματική κατανομή του κειμένου P(c|d), w είναι η λέξη η οποία έχει παραχθεί από την κατανομή λέξεων της θεματικής ενότητας P(w|c). Τα d και w είναι παρατηρούμενες μεταβλητές, ενώ το c είναι λανθάνουσα

Για παράδειγμα, θεωρώντας παρατηρήσεις σε μορφή συνεμφανίσεων λέξεων και κειμένων (w, d), το PLSA μοντελοποιεί την πιθανότητα κάθε συνεμφάνισης, λέξη-κειμένου, ως ένα μείγμα από υπό συνθήκη ανεξάρτητες κατηγορικές κατανομές:

$$P(w,d)=\sum_c P(c)\,P(d|c)\,P(w|c)$$

Αυτός ο τύπος δηλώνει την πιθανότητα η λέξη w και το κείμενο d να έχουν παραχθεί από την λανθάνουσα κλάση c (θεματική ενότητα). Οι πιθανότητες P(d|c) και P(w|c) είναι οι υπό συνθήκη πιθανότητες.

VI: Non-Negative Matrix Tri-factorization

Σε αυτά τα μοντέλα ένα $m_x n$ μητρώο όρων-κειμένων X, προσεγγίζεται από τρεις παράγοντες οι οποίοι καθορίζουν ασθενή ιδιότητα μέλους όρων και κειμένων σε μία από τις k πιθανές κλάσεις:

$$X \sim FSG'$$

όπου F είναι ένα $m_x k$ μη αρνητικό μητρώο το οποίο αναπαριστά γνώση στο χώρο των λέξεων, για παράδειγμα η i-οστή σειρά του F αναπαριστά την εκ των υστέρων πιθανότητα να ανήκει μία λέξη σε μία από τις k πιθανές κλάσεις, G είναι ένα $n_x k$ μη-αρνητικό μητρώο το οποίο αναπαριστά την εκ των υστέρων πιθανότητα κείμενο i αν ανήκει σε μία από τις k πιθανές κλάσεις, και S είναι ένα $k_x k$ μη-αρνητικό μητρώο το οποίο παρέχει μία συνοπτική αναπαράσταση του X.

Αυτό το μοντέλο παραγοντοποίησης μητρώου είναι παρόμοιο με το PLSI, στο οποίο το μητρώο X αντιμετωπίζεται ως η ενοποιημένη κατανομή λέξεων και κειμένων. Στο PLSI, όταν X είναι το $m_x n$ μητρώο λέξεων-κειμένων, η διάσπαση που πραγματοποιείται είναι η X=WSD, όπου W είναι υπό συνθήκη πιθανότητα των κλάσε-

ων των λέξεων, D είναι η υπό συνθήκη πιθανότητα των κλάσεων των κειμένων, και S είναι η πιθανοτική κατανομή των κλάσεων. Το PLSI παρέχει μία ταυτόχρονη λύση για τις κατανομές των υπό συνθήκη πιθανότητων των κλάσεων των λέξεων και των κειμένων, ενώ τα μοντέλα non-negative matrix tri-factorization παρέχουν μία ταυτόχρονη λύση για την συσταδοποίηση των γραμμών και των στηλών του X.

VII: ClustFuse

Ο αλγόριθμος ClustFuse συγχωνεύει πολλές λίστες κατάταξης κειμένων, χρησιμοποιώντας συστάδες κειμένων οι οποίες δημιουργούνται στο σύνολο των κειμένων όλων των λιστών, για να παράξει μία τελική και πιο ακριβής κατάταξη. Ο αλγόριθμος στηρίζεται στην υπόθεση πως έγγραφα τα οποία συνδέονται στενά μετάξύ τους τείνουν να είναι σχετικά στα ίδια ερωτήματα (cluster hypothesis), και χρησιμοποιεί μία μέθοδο σύντηξης η οποία δίνει την δυνατότητα σε παρόμοια κείμενα, στο σύνολο των κειμένων όλων των λιστών, να παρέχουν μεταξύ τους υποστήριξη σχετικότητας.

Ο αλγόριθμος ClustFuse περιγράφεται ως εξής: Έστω CL το σύνολο των κειμένων και q_s ένα ερώτημα, το οποίο περιέχει τα aspect λέξεις κλειδιά. Από το ερώτημα q_s παράγεται ένα σύνολο ερωτημάτων $Q= \{q_1, q_2, ..., q_n\}$, τα οποία εκφράζουν περίπου την ίδια ανάγκη πληροφόρισης. Για παράδειγμα, το σύνολο των ερωτημάτων μπορεί να είναι αποτέλεσμα μιας διαδικασίας query expansion ή query disambiquation. Για κάθε ένα ερώτημα του συνόλου Q, παράγεται μία κατάταξη κειμένων με βάση των σχετικότητά τους με το ερώτημα. Έστω $L = \{L_1, L_2, ..., L_n\}$ το σύνολο των κατατάξεων των κειμένων που δημιουργήθηκε από n διαφορετικά ερωτήματα $Q= \{q_1, q_2, ..., q_n\}$. Στο σύνολο των κειμένων όλων των λιστών πραγματοποιείται συσταδοποίηση, με κάποιον αλγόριθμο συσταδοποίησης, για την ανακάλυψη συστάδων κειμένων. Έστω $Cl = \{c_1, c_2, ..., c_n\}$ ένα σύνολο από clusters στο σύνολο των κειμένων CL.

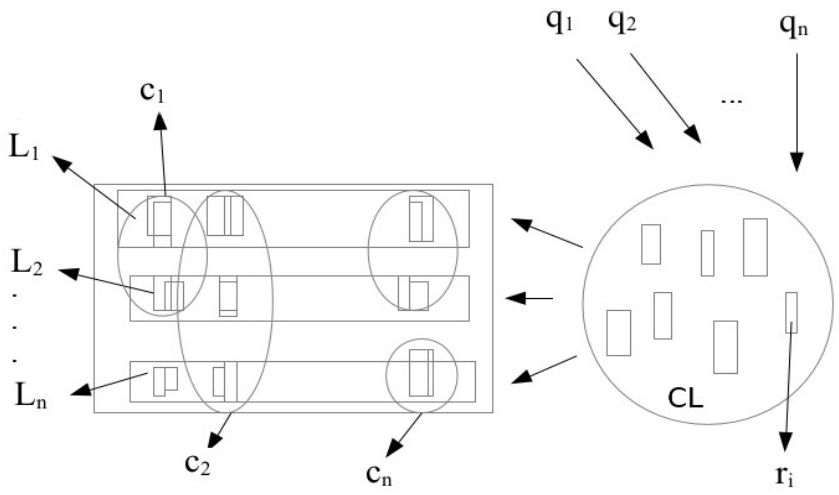


Figure 13
Η διαδικάσια παραγωγής των κατατάξεων Li και των συστάδων ci στον αλγόριθμο ClustFuse

Ο αλγόριθμος ClustFuse κάνει χρήση δύο component για να δώσει score σε ένα κέιμενο d στο ερώτημα q_s, της πιθανότητας σχετικότητας του κειμένου με το ερώτημα και της της θεώρησης πως clusters μπορούν να χρησιμοποιηθούν σαν proxies για τα κείμενα, η οποία ανταμείβει κείμενα τα οποία ανήκουν “strongly” σε ένα cluster το οποίο είναι πολύ σχετικό με το ερώτημα q_s.

$$p(q_s|d)=(1-\lambda)\,p(d|q_s)+\lambda\sum_{c\in Cl(CL)}p(c|q_s)\,p(d|c)$$

Το λ προτείνεται να είναι μία τιμή στο διάστημα (0.6 – 0.8),

και:

πιθανότητα σχετικότητας κειμένου με το ερώτημα: $p(d|q_s)=\dfrac{F(d;q_s)}{\sum_{d'\in CL}F(d';q_s)}$

πιθανότητα σχετικότητας cluster με το ερώτημα: $p(c|q)=\frac{\prod_{d\in c} F(d;q)}{\sum_{c'\in Cl(CL)} (\prod_{d'\in c'} F(d';q))}$

πιθανότητα σχετικότητας κειμένου με cluster: $p(d|c)=\frac{\sum_{d'\in c} similarity(d',d)}{\sum_{d_i\in CL} (\sum_{d'\in c} simlarity(d',d_i))}$

όπου:

F(d;q) είναι μία από τις παρακάτω μεθόδους σύντηξης:

$$F_{CombSUM}(d;q)=\sum_{L_i} S_{L_i}(d)$$

$$F_{CombMNZ}(d,q)=numberOf\,[L_i: d\in L_i]\,F_{CombSUM}(d;q)$$

$$F_{Borda}(d;q)=\sum_{L_i} numberOf\,[d'\in L_i: S_{L_i}(d')\leq S_{L_i}(d)]$$

(S_{L_i} είναι το score του κειμένου στην λίστα L_i)

και

similarity(d',d) είναι μία μετρική ομοιότητας δύο κειμένων

VIII: Penn TreeBank POS Tags

Penn Treebank POS Tags όπως αυτά παρουσιάζονται στον ιστότοπο:

http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Table 3
Penn Treebank POS tags: αλφαβητική λίστα των POS tags τα οποία χρησιμοποιούνται στο Penn Treebank project.

7 Αναφορές

Amati, G., and van Rijsbergen, C.J.,. *Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst., 20(4):357{389*, 2002.

Asur, Sitaram and Bernardo A. Huberman. *Predicting the future with social media. Arxiv preprint arXiv:1003.5699*, 2010.

Aue, Anthony and Michael Gamon. *Customizing sentiment classifiers to new domains: a case study. In Proceedings of Recent Advances in Natural Language Processing (RANLP-2005).* 2005

Barbosa, Luciano and Junlan Feng. *Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the International Conference on Computational Linguistics (COLING-2010).* 2010.

Becker, Israela and Vered Aharonson. *Last but definitely not least: on the role of the last sentence in automatic polarity-classification. In Proceedings of the ACL 2010 Conference Short Papers.* 2010.

Bespalov, Dmitriy, Bing Bai, Yanjun Qi, and Ali Shokoufandeh. *Sentiment classification based on supervised latent n-gram analysis. In Proceeding of the ACM conference on Information and knowledge management (CIKM-2011).* 2011

Blair-Goldensohn, Sasha, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis, and Jeff Reynar. *Building a sentiment summarizer for local service reviews. In Proceedings of WWW-2008 workshop on NLP in the Information Explosion Era.* 2008.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. *Latent dirichlet allocation. The Journal of Machine Learning Research*, 2003.

Blitzer, John, Mark Dredze, and Fernando Pereira. *Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2007).* 2007.

Boiy, Erik and Marie-Francine Moens. *A machine learning approach to sentiment analysis in multilingual Web texts. Information Retrieval*, 2009.

Bollegala, Danushka, David Weir, and John Carroll. *Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011).* 2011.

Christos Makris, Panagiotis Panagopoulos, *Improving Opinion-based Entity Ranking, WEBIST 2014, short paper, 223-230*, 2014.

Cilibrasi, Rudi L. and Paul M. B. Vitanyi. *The google similarity distance. IEEE Transactions on Knowledge and Data Engineering*, 2007.

Crammer K., Singer Y., *Pranking with ranking. NIPS* 2001, 641-647

Cui H., Mittal V., and M. Datar. *Comparative experiments on sentiment classification for online product reviews. In proceedings of the 21st national conference on Artificial intelligence - Volume 2, pages 1265–1270. AAAI Press*, 2006.

Davidov, Dmitry, Oren Tsur, and Ari Rappoport. *Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of Coling-2010.* 2010.

Drehmer D, Belohlav J, Coye R, An exploration of employee participation using a scaling approach. Group Org Manage 25(4):397, 2000.

Ding, Xiaowen, Bing Liu, 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.

Esuli, Andrea and Fabrizio Sebastiani. *Determining term subjectivity and term orientation for opinion mining. In Proceedings of Conf. of the European Chapter of the Association for Computational Linguistics (EACL-2006).* 2006.

Gamon, Michael. S*entiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In Proceedings of International Conference on Computational Linguistics (COLING-2004).* 2004.

Gamon, Michael, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. Pulse: *Mining customer opinions from free text. Advances in Intelligent Data Analysis VI*, 2005.

Gao, Sheng and Haizhou Li. *A cross-domain adaptation method for sentiment classification using probabilistic latent analysis. In Proceeding of the ACM Conference on Information and knowledge management (CIKM-2011).* 2011.

Ghani, Rayid, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano. *Text mining for product attribute extraction. ACM SIGKDD Explorations Newsletter,* 2006.

Ghose, Anindya and Panagiotis G. Ipeirotis. *Designing novel review ranking systems: predicting the usefulness and impact of reviews. In Proceedings of the International Conference on Electronic Commerce.* 2007.

González-Ibáñez, Roberto, Smaranda Muresan, and Nina Wacholder. *Identifying sarcasm in Twitter: a closer look. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers (ACL-2011)*. 2011.

Groh, Georg and Jan Hauffa. *Characterizing Social Relations Via NLP-based Sentiment Analysis. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM- 2011).* 2011.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. Fundamentals of Item Response Theory. Newbury Park, CA: Sage Press, 1991.

He, Q., Text Mining and IRT for Psychiatric and Psychological Assessment. Ph.D. thesis, University of Twente, Enschede, the Netherlands, 2013.

He, Yulan, Chenghua Lin, and Harith Alani. *Automatically extracting polarity-bearing topics for cross- domain sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011). 2011.*

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

Hu, Minqing and Bing Liu. *Mining and summarizing customer reviews. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004).* 2004.

Jakob, Niklas and Iryna Gurevych. *Extracting Opinion Targets in a Single-and Cross-Domain Setting with Conditional Random Fields. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010).* 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 (ACL-2011).* 2011.

Jin, Wei and Hung Hay Ho. *A novel lexicalized HMM-based learning framework for web opinion mining. In Proceedings of International Conference on Machine Learning (ICML-2009).* 2009.

Jindal, Nitin and Bing Liu. *Identifying comparative sentences in text documents. In Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR-2006).* 2006a.

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

Jindal, Nitin and Bing Liu. *Opinion spam and analysis. In Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008).* 2008.

Joshi, Mahesh, Dipanjan Das, Kevin Gimpel, and Noah A. Smith. *Movie reviews and revenues: An experiment in text regression. In Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference (NAACL 2010).* 2010.

Kim, Soo-Min and Eduard Hovy. *Determining the sentiment of opinions. In Proceedings of Interntional Conference on Computational Linguistics (COLING-2004).* 2004.

Kim, Soo-Min, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. *Automatically assessing review helpfulness. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2006).* 2006.

Kim, Soo-Min and Eduard Hovy. *Determining the sentiment of opinions. In Proceedings of Interntional Conference on Computational Linguistics (COLING-2004).* 2004.

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

Lu, Yue, Huizhong Duan, Hongning Wang, and ChengXiang Zhai. *Exploiting Structured Ontology to Organize Scattered Online Opinions. In Proceedings of Interntional Conference on Computational Linguistics (COLING-2010).* 2010.

Li, Fangtao, Minlie Huang, and Xiaoyan Zhu. *Sentiment analysis with global topics and local dependency. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2010).* 2010.

Li, Shasha, Chin-Yew Lin, Young-In Song, and Zhoujun Li. *Comparable entity mining from comparative questions. In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2010).* 2010.

Li, Tao, Yi Zhang, and Vikas Sindhwani. A *non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2009).* 2009.

Lim, Ee-Peng, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady W. Lauw. *Detecting Product Review Spammers using Rating Behaviors. In Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2010).* 2010.

Liu, Bing. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2006 and 2011: Springer.*

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, Jing jing, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. *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.

Liu, Jingjing and Stephanie Seneff. *Review sentiment scoring via a parse-and-paraphrase paradigm. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009).* 2009

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

Manevitz, Larry M. and Malik Yousef. *One-class SVMs for document classification. The Journal of Machine Learning Research*, 2002.

Martineau, Justin and Tim Finin. *Delta tfidf: An improved feature space for sentiment analysis. In Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM- 2009).* 2009.

McDonald, Ryan, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. *Structured models for fine-to-coarse sentiment analysis. In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2007).* 2007.

McGlohon, Mary, Natalie Glance, and Zach Reiter. *Star quality: Aggregating reviews to rank products and merchants. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010)*. 2010.

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.

Miller, Mahalia, Conal Sathi, Daniel Wiesenthal, Jure Leskovec, and Christopher Potts. *Sentiment Flow Through Hyperlink Networks. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-2011).* 2011.

Mohammad, Saif M. and Peter D. Turney. *Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.* 2010.

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

Narayanan, Ramanathan, Bing Liu, and Alok Choudhary. *Sentiment analysis of conditional sentences. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP- 2009).* 2009.

Nigam, Kamal, Andrew K. McCallum, Sebastian Thrun, and Tom Mitchell. *Text classification from labeled and unlabeled documents using EM. Machine Learning*, 2000.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM 2010).* 2010.

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).* 2002.

Pang, Bo and Lillian Lee. *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of Meeting of the Association for Computational Linguistics (ACL-2004)*. 2004.

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

Pang B, and L. Lee. *Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1–135*, 2008.

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.

Qu, Lizhen, Georgiana Ifrim, and Gerhard Weikum. *The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns. In Proceedings of the International Conference on Computational Linguistics (COLING-2010).* 2010.

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.

Rasch, G., *Probabilistic Models for Some Intelligence and Attainment Tests, (Copenhagen, Danish Institute for Educational Research), with foreward and after word by B.D. Wright. The University of Chicago Press, Chicago*, 1960/1980.

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.

Sadikov, Eldar, Aditya Parameswaran, and Petros Venetis. *Blogs as predictors of movie success. In Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM-2009).* 2009.

Scaffidi, Christopher, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng, and Chun Jin. *Red Opal: product-feature scoring from reviews. In Proceedings of Twelfth ACM Conference on Electronic Commerce (EC-2007).* 2007.

Snyder, Benjamin and Regina Barzilay. *Multiple aspect ranking using the good grief algorithm. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL / HLT-2007).* 2007.

Somasundaran, Swapna, Galileo Namata, Lise Getoor, and Janyce Wiebe. *Opinion graphs for polarity and discourse classification. In Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing.* 2009.

Stone, Philip. *The general inquirer: A computer approach to content analysis. Journal of Regional Science*, 1968.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. *Lexicon- based methods for sentiment analysis. Computational Linguistics*, 2011.

Tan, Songbo, Gaowei Wu, Huifeng Tang, and Xueqi Cheng. *A novel scheme for domain-transfer problem in the context of sentiment analysis. In Proceeding of the ACM conference on Information and knowledge management (CIKM-2007).* 2007.

Tikves, S., Banerjee, S., Temkit, H., Gokalp, S., Davulcu, H., Sen, A., Corman, S., Woodward, M., Nair, S., Rohmaniyah, I., Amin,A., *A system for ranking organizations using social scale analysis, Soc. Netw. Anal. Min.*, 2012.

Titov, Ivan and Ryan McDonald. *A joint model of text and aspect ratings for sentiment summarization. In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2008).* 2008.

Tsur, Oren, Dmitry Davidov, and Ari Rappoport. *A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM-2010).* 2010.

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 (ACL-2002).* 2002.

Wiebe, Janyce, Rebecca F. Bruce, and Thomas P. O'Hara. *Development and use of a gold-standard data set for subjectivity classifications. In Proceedings of the Association for Computational Linguistics (ACL-1999).* 1999.

Wiebe, Janyce. *Learning subjective adjectives from corpora. In Proceedings of National Conf. on Artificial Intelligence (AAAI-2000).* 2000.

Wilson, Theresa, Janyce Wiebe, 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.

Wu, Guangyu, Derek Greene, Barry Smyth, and Pádraig Cunningham. *Distortion as a validation criterion in the identification of suspicious reviews. In Proceedings of Social Media Analytics.* 2010.

Wu, Qion, Songbo Tan, and Xueqi Cheng. *Graph ranking for sentiment transfer. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (ACL-IJCNLP-2009).* 2009.

Xia, Rui and Chengqing Zong. *A POS-based ensemble model for cross-domain sentiment classification. In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP- 2010).* 2011.

Yang, Hui, Luo Si, and Jamie Callan. *Knowledge transfer and opinion detection in the TREC2006 blog track. In Proceedings of TREC.* 2006

Yu, Hong and Vasileios Hatzivassiloglou. *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, Jianxing, Zheng-Jun Zha, Meng Wang, and Tat-Seng Chua. *Aspect ranking: identifying important product aspects from online consumer reviews. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.* 2011.

Zhang, Wenbin and Steven Skiena. *Trading strategies to exploit blog and news sentiment. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2020).* 2010.

Zhang, Zhu and Balaji Varadarajan. *Utility scoring of product reviews. In Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2006).* 2006.

Zhuang, Li, Feng Jing, and Xiaoyan Zhu. *Movie review mining and summarization. In Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2006).* 2006.

Zhang, Wei and Clement Yu. *UIC at TREC 2007 Blog Report,* 2007.