ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΤΕΧΝΟ – ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Τμήμα Βιομηχανικής Διοίκησης και Τεχνολογίας

Διπλωματική Εργασία
Μεταπτυχιακού Διπλώματος Ειδίκευσης
Τεχνο-Οικονομικά Συστήματα

ΘΕΜΑ: «Ανάλυση τεχνικών και μεθοδολογιών εξόρυξης γνώμης (opinion mining) και διαχείρισης φήμης σε πραγματικό χρόνο (on-line reputation management), σε περιβάλλοντα κοινωνικής δικτύωσης»

ΡΑΠΑΝΑΚΗΣ ΣΤΑΜΑΤΗΣ Α.Μ.: 032002046

Επιβλέπων: Δρ. Ιωάννης Χαραλαμπίδης
Διδάσκων μαθήματος Ηλεκτρονικές Συναλλαγές

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την .....

Δρ. Ιωαννης Χαραλαμπίδης Επίκουρος Καθηγητής Πανεπιστημίου Αιγαίου	Δ. Ασκούνης Αναπληρωτής καθηγητής Ε.Μ.Π.	Ι. Ψαρράς Καθηγητής Ε.Μ.Π. .....

Αθήνα, Οκτώβριος 2011

Ευχαριστώ θέρμα τον Δρ. Γ. Χαραλαμπίδη για την υποστήριξη του κατά την διάρκεια εκπόνησης της παρούσας διπλωματής εργασίας.

ΠΕΡΙΛΗΨΗ

Το βασικό αντικείμενο της εργασίας είναι η περιγραφή των τεχνικών και μεθοδολογιών εξόρυξης γνώμης (opinion mining) και διαχείρισης φήμης σε πραγματικό χρόνο (on-line reputation management) σε περιβάλλοντα κοινωνικής δικτύωσης. Θα διερευνήσουμε τους τρόπους με τους οποίους οι παραπάνω έννοιες βρίσκουν εφαρμογή σε περιβάλλοντα κοινωνικής δικτύωσης. Για τον σκοπό αυτό θα κάνουμε μια εκτενή ανασκόπηση της σχετικής βιβλιογραφίας και θα επιχειρήσουμε να παρουσιάσουμε τις επιμέρους προσεγγίσεις. Στα πλαίσια της εργασίας θα αναφερθούμε σε συστήματα διαβούλευσης πολιτικής και σε ταξινομητές πολιτικής γνώμής.

Ακολουθεί η περιγραφή μια σειράς από υποπροβλήματα που σχετίζονται με τα κοινωνικά δίκτυα, όπως η διαχείριση φήμης και η αναζήτηση απόψεων σε ιστολόγια. Καθένα από αυτά παρουσιάζει ιδιαιτερότητες και προσεγγίζεται μέσα από ένα ενιαίο πρίσμα. Οι επιδόσεις των αλγορίθμων κατάταξης των εγγράφων και αναζήτησης απόψεων σε ιστολόγια θα παρουσιαστούν με στοιχεία από αντίστοιχους διαγωνισμούς. Επιπλέον θα γίνουν συγκρίσεις των ταξινομητών πολιτικής γνώμης. Στη συνέχεια θα παρουσιάσουμε μια σειρά από ευρέως διαδεδομένα προϊόντα που χρησιμοποιούνται για τους παραπάνω σκοπούς. Τέλος θα προσπαθήσουμε να δώσουμε απάντηση σε μια σειρά καίριων ερωτημάτων και να προτείνουμε λύσεις που καλύπτουν τους τομείς που αναλύσαμε.

Λέξεις κλειδιά: Εξόρυξη γνώμης, διαχείριση φήμης, κοινωνική δικτύωση, ανάλυση συναισθήματος, πολικότητα κειμένου, συστήματα διαβούλευσης πολιτικής.

ABSTRACT

The main objective of this work is to describe the techniques and methodologies of opinion mining and online reputation management in social network environments. We will investigate the ways in which these concepts can be applied to social networking environments. To this end we will do an extensive review of the literature and attempt to classify the different approaches. As part of this work we will refer to opinion mining systems for Politics and in political opinion classifiers.

In the following we will present a series of sub problems related to the social networks, including reputation management and topic identification in blogs. Each of these areas is unique and is approached through a single prism. The performance of the ranking algorithms of search of documents and views in blogs will be presented with data from blog's topic indentification competetitions. Furthermore, comparisons of political opinion classifiers will also take place. Next we will present a series of widely available products used for these purposes. Finally, we will try to answer a series of key questions and propose solutions that cover the areas analyzed.

Key words: Opinion mining, reputation management, social networking, sentimental analysis, text polarity, policy consultation systems.

Πίνακας Περιεχομένων

Πίνακας Σχημάτων

1 Εισαγωγή

Τα τελευταία χρόνια έχει αλλάξει ο τρόπος με τον οποίο οι χρήστες του διαδικτύου διαχειρίζονται την πληροφορία. Οι χρήστες δεν περιορίζονται στην προβολή του διαδικτυακού περιεχομένου, αλλά μπορούν να το σημειώσουν, να το σχολιάσουν και γενικότερα να το εμπλουτίσουν. Με τον τρόπο αυτό, από απλοί αναγνώστες γίνονται και οι ίδιοι συντάκτες. Από καταναλωτές περιεχομένου, μπορούν εύκολα να γίνουν δημιουργοί περιεχομένου. Προχωρώντας ένα βήμα παραπέρα, δεν περιορίζονται στην επισήμανση σελίδων και στην αξιολόγηση της υπάρχουσας πληροφορίας. Μπορούν να μοιράζονται τα νέα τους, την γνώση και τις ιδέες τους με το σύνολο της διαδικτυακής κοινότητας.

Υπάρχουν πολλά μέσα δημοσίευσης περιεχομένου με τα οποία οι χρήστες μπορούν να εκφραστούν στο διαδίκτυο. Παραδείγματα τέτοιων μέσων αποτελούν τα ιστολόγια, ομάδες συζητήσεων αλλά και τα κοινωνικά δίκτυα. Οι χρήστες μπορούν να δημοσιεύουν κείμενα, να κάνουν σχόλια και να λαμβάνουν τις απαντήσεις άλλων χρηστών. Οι πληροφορίες αυτές καλύπτουν πολλές πτυχές της κοινωνικής τους ζωής. Από θέματα πολιτικής και υγείας μέχρι ταξίδια και κριτικές για μια υπηρεσία. Η δημοφιλία των μέσων δημοσίευσης περιεχομένου καθιστά σαφές ότι τα δεδομένα που αντιστοιχούν σε γνώμες χρηστών θα αποτελέσουν ένα σημαντικό κομμάτι των δεδομένων κειμένου στο διαδίκτυο.

Η ολοένα και αυξανόμενη παραγωγή περιεχομένου στο διαδίκτυο αποτελεί πρόκληση για την ικανότητα των αντίστοιχων συστημάτων να καταγράψουν την γνώμη και τις διαθέσεις των χρηστών σε μεγάλη κλίμακα. Στα πλαίσια αυτά, η **εξόρυξη γνώμης** (opinion mining) συγκεντρώνει όλο και περισσότερο το ενδιαφέρον της κοινότητας που ασχολείται με τον κλάδο της ανάκτησης πληροφορίας (information retrieval). Είναι ένας διεπιστημονικός τομέας που αφορά την επιστήμη της ανάκτησης πληροφορίας, της εξόρυξης δεδομένων, της μηχανικής μάθησης, καθώς και της στατιστικής και υπολογιστικής γλωσσολογίας.

Οι τεχνικές εξόρυξης γνώμης χρησιμοποιούνται κυρίως για την ανάλυση των κριτικών, εμπειριών του κοινού σχετικά με ένα προϊόν ή μια υπηρεσία. Επιπλέον, με την εφαρμογή τους επιτυγχάνεται η καταγραφή της τάσης διαμόρφωσης των απόψεων σε ένα θέμα και η εύρεση των κυρίαρχων συνιστωσών του. Αυτό συμβάλλει στο να αναδεικνύονται περισσότερες πτυχές του. Η γνώση αυτή αξιοποιείται μεταξύ άλλων από τα συστήματα διαβούλευσης πολιτικής. Τα συστήματα αυτά επιτρέπουν την συμμετοχή των πολιτών στην λήψη πολιτικών αποφάσεων και την διαχείριση των προτάσεων της δικτυακής κοινότητας γενικότερα. Αντίστοιχα, θα δούμε ότι οι τεχνικές εξόρυξης γνώμης χρησιμοποιούνται με σκοπό την έρευνα και την ανάλυση της εταιρικής φήμης μέσα από τα συστήματα διαχείρισης φήμης.

1.1 Αντικείμενο της Διπλωματικής εργασίας

Το αντικείμενο της διπλωματικής εργασίας είναι:

- Βιβλιογραφική επισκόπηση του πεδίου της εξόρυξης γνώμης από κείμενα.
- Περιγραφή της αρχιτεκτονικής των συστημάτων εξόρυξης γνώμης.
- Σύγκριση των τεχνικών εξόρυξης γνώμης που χρησιμοποιούνται, δίνοντας έμφαση στο πεδίο των ιστολογίων.
- Επισκόπηση των συστημάτων διαβούλευσης πολιτικής κάνοντας ιδιαίτερη αναφορά στις προκλήσεις της ταξινόμησης πολιτικού κειμένου.

- Παρουσίαση των συστημάτων διαχείρισης φήμης και παράθεση των πιο γνωστών λογισμικών εξόρυξης γνώμης και διαχείρισης φήμης.

1.2 Δομή της Διπλωματικής εργασίας

Στο **κεφάλαιο 2** γίνεται μια βιβλιογραφική αναφορά στην έννοια της εξόρυξης γνώμης και της ανάλυσης συναισθήματος. Αρχικά περιγράφονται οι ανάγκες που οδήγησαν στην ανάπτυξη του τομέα τα τελευταία χρόνια. Παρουσιάζονται οι κυριότερες δραστηριότητες που σχετίζονται με την εξόρυξη γνώμης και τα κίνητρα πίσω από αυτές. Επίσης προσδιορίζεται το πεδίο σε σχέση με το ευρύτερο πεδίο της εξόρυξης δεδομένων και τα συναφή πεδία της ανάκτησης πληροφορίας και της υπολογιστικής γλωσσολογίας. Γίνεται μια γενική περιγραφή των διαδικασιών εξόρυξης πληροφορίας από κείμενα. Το κεφάλαιο συνεχίζει με αναφορά στα βήματα της εξόρυξης γνώμης και στις διαφορετικές προσεγγίσεις με τις οποίες μπορεί αυτή να επιτευχθεί. Περιγράφεται η κάθε κατηγορία με παράθεση παραδειγμάτων και αναφορά στις αντίστοιχες επιστημονικές δημοσιεύσεις.

Στο **κεφάλαιο 3** γίνεται αρχικά μια περιγραφή του περιβάλλοντος των ιστολογίων ως το πιο προσφιλές μέσο δημοσίευσης περιεχομένου από τους χρήστες του διαδικτύου. Η δικτυακή κοινότητα των ιστολογίων, που περιλαμβάνει και τα μίκρο-ιστολόγια, αποτελεί μια πρόκληση για την κοινότητα της εξόρυξη γνώμης. Ιδιαίτερη έμφαση δίνεται στην αναζήτηση απόψεων σε ιστολόγια. Παρουσιάζονται τα αποτελέσματα διαδοχικών διαγωνισμών που αφορούν την αναζήτηση πληροφοριών σε ιστολόγια, την εύρεση θετικών και αρνητικών απόψεων προσομοιώνοντας ρεαλιστικά σεναρια αναζήτησης. Σε αυτά περιγράφεται η διαδικασία ανάπτυξης ενός συστήματος το οποίο θα εντοπίζει τις καταχωρήσεις ενός ιστολογίου που εκφράζουν κάποια άποψη σχετικά με ένα δεδομένο θέμα – στόχο, καθώς και η εύρεση θετικών/ αρνητικών απόψεων για το θέμα αυτό.

Στο **κεφάλαιο 4** ασχολούμαστε με τα συστήματα διαβούλευσης πολιτικής και τον σχεδιασμό ταξινομητών πολιτικού κειμένου. Διερευνούμε την εξόρυξη κειμένου και την εφαρμογή τεχνικών μηχανικής μάθησης για την καταγραφή της γνώμης του κοινού που αφορά ζητήματα πολιτικής. Περιγράφεται η αρχιτεκτονική αυτών των συστημάτων. Το πολιτικό κείμενο έχει ιδιαίτερα χαρακτηριστικά τα οποία και αναλύονται. Σε αυτό το πλαίσιο εξετάζονται οι ταξινομητές πολιτικής γνώμης. Στη συνέχεια προσδιορίζονται εκείνες οι τεχνικές εξόρυξης γνώμης που είναι οι πλέον κατάλληλες για το πολιτικό κείμενο.

Στο **κεφάλαιο 5** γίνεται μια εισαγωγή στα συστήματα διαχείρισης δικτυακής φήμης. Τα συστήματα αυτά ενσωματώνουν λειτουργίες που χρησιμοποιούν τις τεχνικές εξόρυξης γνώμης που περιγράψαμε στα προηγούμενα κεφάλαια. Αρχικά εξετάζουμε την σημασία και την χρήση του λογισμικού δικτυακής φήμης διεθνώς. Στη συνέχεια εξετάζεται η διείσδυση των υπηρεσιών διαχείρισης φήμης στον ελληνικό χώρο. Στο τέλος κάνουμε μια συγκριτική παρουσίαση λογισμικού εξόρυξης γνώμης και διαχείρισης φήμης, δίνοντας έμφαση στις λύσεις που προσφέρουν ελληνικές εταιρίες.

Στο **κεφάλαιο 6** θα καταθέσουμε τα συμπεράσματα μας σχετικά με την χρήση των κατάλληλων τεχνικών εξόρυξης γνώμης, ανάλογα με το πεδίο της εφαρμογής. Θα συνοψίσουμε τα ευρήματα που προέκυψαν από την επισκόπηση της βιβλιογραφίας και θα καταθέσουμε μια σειρά προτάσεων για τις ανοιχτές προκλήσεις του χώρου.

2 Ανάλυση του πεδίου

Η πληροφορία υπο μορφή κειμένου μπορεί να ταξινομηθεί δύο κατηγορίες, γεγονότα και γνώμες. Τα γεγονότα είναι αντικειμενικές δηλώσεις σχετικά με πρόσωπα και πράγματα στον κόσμο. Οι γνώμες (απόψεις) είναι υποκειμενικές δηλώσεις που αντανακλούν τα συναισθήματα ή τις αντιλήψεις των ανθρώπων σχετικά με πρόσωπα και πράγματα. Το μεγαλύτερο μέρος της υπάρχουσας έρευνας στην επεξεργασία κειμένου έχει επικεντρωθεί στην εξόρυξη και ανάκτηση πραγματικής πληροφορίας (γεγονότων). Είναι ενδεικτικό ότι οι περισσότερες μηχανές αναζήτησης μπορούν να ανακτούν μέσω λέξεων – κλειδιών γεγονότα, ενώ δεν μπορούν να ανακτήσουν απόψεις χρηστών, καθότι αφενός μεν αυτές είναι δύσκολο να αναπαρασταθούν από μεμονωμένες λέξεις κλειδιά και αφετέρου οι αλγόριθμοι κατάταξης των αποτελεσμάτων (search ranking strategies) δεν είναι κατάλληλοι για ανάκτηση/ εξόρυξη απόψεων.

Ο όγκος του διαδικτυακού περιεχομένου που αφορά γνώμες χρηστών ολοένα και αυξάνεται. Ο παγκόσμιος ιστός έχει αλλάξει δραματικά τον τρόπο με τον οποίο οι άνθρωποι εκφράζουν τις απόψεις τους. Μπορούν να γράψουν κριτικές προϊόντων σε σελίδες ηλεκτρονικού εμπορίου αλλά και να εκφράσουν τις απόψεις τους επάνω σε οποιοδήποτε ζήτημα σε πληθώρα μέσων. Τα συνηθέστερα μέσα είναι οι διαδικτυακές ομάδες συζήτησης (discussion groups), τα ιστολόγια (blogs) και οι χώροι δημόσιας συζήτησης (forums). Η μελέτη και η εξόρυξη πληροφορίας από περιεχόμενο που δημιουργείται από την συνδρομή των χρηστών (user-generated) αποτελεί μια πρόκληση καθότι στο διαδίκτυο υπάρχουν πολλές διαφορετικές πηγές και κάθε μια από αυτές περιέχει μεγάλη ποσότητα πληροφοριών. Σε πολλές περιπτώσεις οι γνώμες πρέπει να εξαχθούν μέσα από μακροσκελείς δημοσιεύσεις σε forum και ιστολόγια. Είναι πολύ δύσκολο για έναν αναγνώστη να αναζητήσει γνώμες για ενα θέμα, να συγκεντρώσει χειρωνακτικά όλες τις πηγές, να τις μελετήσει, να γράψει μια περίληψη για αυτές και στο τέλος να τις παρουσιάσει σε μια αξιοποιήσιμη μορφή. Για τους σκοπούς αυτούς έχει αναπτυχθεί τα τελευταία χρόνια ο κλάδος της εξόρυξης γνώμης.

Η **εξόρυξη γνώμης** (opinion mining) περιγράφεται ως το πρόβλημα της αναγνώρισης μιας γνώμης σε ενα συγκεκριμένο θέμα και της εκτίμησης του προσανατολισμού αυτής της γνώμης [1]. Ο *προσανατολισμός* ή *πολικότητα* (polarity) μιας γνώμης είναι το κατά πόσο μια γνώμη εκφράζει θετική ή αρνητική στάση πάνω στο θέμα που σχολιάζει. Η αξιολόγηση γίνεται βάση μιας νοητής κλίμακας συναισθημάτων, από αρνητικά σε ουδέτερα έως θετικά συναισθήματα. Η εξόρυξη γνώμης αποτελεί το υπόβαθρο πάνω στο οποίο βασίζονται άλλες εργασίες ανάλυσης υποκειμενικότητας. Προσφέρει μια σε βάθος εικόνα των απόψεων που εκφράζονται σε κείμενα και επιτρέπει την παραπέρα επεξεργασία των δεδομένων με σκοπό να προσδιορίσει την επικρατούσα γνώμη ή να επισημάνει αντικρουόμενες απόψεις. Η ποιότητα των αποτελεσμάτων της εξόρυξης γνώμης είναι ζωτικής σημασίας για την επιτυχία όλων των μετέπειτα εργασιών που βασίζονται σε αυτήν, γεγονός που την καθιστά ένα σημαντικό πρόβλημα.

Η εξόρυξη γνώμης συναντάται συχνά και με τον όρο **ανάλυση συναισθήματος** (sentimental analysis). Ο όρος *εξόρυξη γνώμης* και ο όρος *ανάλυση συναισθήματος* διαφέρουν λίγο σαν έννοιες, γεγονός που οφείλεται στο ότι αρχικά μελετήθηκαν σε διαφορετικά πεδία. Ο όρος εξόρυξη γνώμης προέρχεται από την κοινότητα της ανάκτησης πληροφορίας (information retrieval) και στοχεύει στην εξαγωγή και μετέπειτα επεξεργασία των απόψεων των χρηστών σχετικά με προϊόντα, ταινίες και υπηρεσίες. Η ανάλυση συναισθήματος, από την άλλη, είχε αρχικά διατυπωθεί σαν ένα πρόβλημα επεξεργασίας φυσικής γλώσσας για την ανάκτηση των απόψεων που εκφράζονται σε διάφορες

δημοσιεύσεις. Παρόλα αυτά, αυτοί οι δύο όροι έχουν παρόμοια έννοια και εμπίπτουν στο πεδίο της *ανάλυσης υποκειμενικότητας* (subjectivity analysis). Στην εργασία αυτή θα χρησιμοποιούμε αυτούς τους δύο όρους εναλλακτικά, θεωρώντας ότι έχουν την ίδια έννοια.

Στο κεφάλαιο αυτό παρουσιάζονται οι κυριότερες προσεγγίσεις και μεθοδολογίες που συναντώνται στη βιβλιογραφία στην εξόρυξη γνώμης από τον παγκόσμιο ιστό.

2.1 Αναζήτηση απόψεων

2.1.1 Εννοιολογική Προσέγγιση

Οι γνώμες μέσα σε ένα κείμενο μπορεί να εκφράζονται για κάποια οντότητα (όπως προϊόν, γεγονός, θέμα συζήτησης, πρόσωπο) είτε άμεσα με εκφράσεις που δηλώνουν συναίσθημα (π.χ. «Η ταινία ήταν υπέροχη»), είτε έμμεσα με σύγκριση της οντότητας με κάποια άλλη (π.χ. «Το αυτοκίνητο Α είναι καλύτερο από το Β»). Στόχος του ερευνητικού πεδίου της εξόρυξης γνώμης είναι η εφαρμογή εξειδικευμένων τεχνικών ανακάλυψης γνώσης κατά τη διαδικασία της αναζήτησης, ώστε ο χρήστης να είναι σε θέση να ανακτά απόψεις ή περίληψη αυτών των απόψεων για μία οντότητα. Ερωτήματα όπως «*Ποια είναι η άποψη του κοινού για τα κινητά τηλέφωνα Χ;*» ή «*Ποια τηλέφωνα είναι καλύτερα, τα Χ ή τα Υ;*» θα μπορούν να απαντηθούν μέσα από μηχανές αναζήτησης που εφαρμόζουν τεχνικές εξόρυξης πληροφορίας από κείμενο.

Τυπικά ερωτήματα χρηστών έχουν ως στόχο την ανάκτηση της άποψης ενός ατόμου για κάποιο χαρακτηριστικό μιας οντότητας, την συλλογή θετικών/ αρνητικών απόψεων για κάποια οντότητα (ή κάποιο μεμονωμένο χαρακτηριστικό της), την πληροφόρηση για το πώς οι απόψεις πάνω σε μία συγκεκριμένη οντότητα μεταβλήθηκαν χρονικά (π.χ. απόψεις για κάποιο πολιτικό πρόσωπο) ή το εάν υπερτερεί κάποιο αντικείμενο σε σχέση με ένα άλλο. Απ' αυτούς τους τύπους ερωτημάτων μόνο ο πρώτος μπορεί να απαντηθεί ικανοποιητικά από τις συνηθισμένες μηχανές αναζήτησης με κατάλληλη επιλογή λέξεων – κλειδιών ως είσοδο, αφού η άποψη ενός ατόμου για κάτι τις περισσότερες φορές περιγράφεται σε ένα κείμενο. Για την απάντηση ερωτημάτων του δεύτερου τύπου (συλλογή θετικών/ αρνητικών απόψεων) χρειάζεται η ανάκτηση πολλών απόψεων και η εξαγωγή της συναισθηματικής τους κατεύθυνσης (θετική/ αρνητική άποψη). Μετά από εξάλειψη τυχόντων παραπλανητικών απόψεων (spam) το τελικό αποτέλεσμα που δίνεται στο χρήστη μπορεί να είναι είτε μία περίληψη αυτών είτε οι ίδιες οι απόψεις ταξινομημένες σε θετικές και αρνητικές.

2.1.2 Μοντελοποίηση Απόψεων

Μία άποψη (opinion) είναι η εκδήλωση μίας στάσης, συμπεριφοράς ή επεφημίας απέναντι σε κάποιο αντικείμενο (object) από το υποκείμενο που την κατέχει (opinion holder). Το υποκείμενο αυτό μπορεί να είναι κάποιο πρόσωπο ή κάποιος φορέας (π.χ. οργανισμός, επιχείρηση) [2].

Ορισμός 1: Ένα αντικείμενο (object) είναι μια οντότητα Ο, η οποία μπορεί να είναι κάποιο προϊόν, άτομο, γεγονός, οργανισμός ή θέμα. Η οντότητα Ο αναπαρίσταται ως μία ιεραρχία από τα συστατικά της μέρη (components), τα συστατικά μέρη αυτών των μερών κ.τ.λ.. Κάθε συστατικό μίας οντότητας περιγράφεται από το δικό του σύνολο ιδιοτήτων (attributes). Η οντότητα Ο είναι ο κόμβος ρίζα αυτής της ιεραρχικής δομής, ο οποίος σχετίζεται επίσης με ένα σύνολο χαρακτηριστικών. Για συντομία κάθε συστατικό ενός αντικειμένου μαζί με τις ιδιότητές του θα αναφέρεται ως χαρακτηριστικό (feature). Μία άποψη μπορεί να ένα ή

περισσότερα χαρακτηριστικά. Το ίδιο το αντικείμενο Ο (κόμβος ρίζα της ιεραρχίας) είναι και αυτό ένα χαρακτηριστικό.

Ορισμός 2: Κάθε αντικείμενο Ο μπορεί να αναπαρασταθεί με ένα πεπερασμένο σύνολο χαρακτηριστικών, $F = \{f_1, f_2, \ldots, f_n\}$. Κάθε χαρακτηριστικό μπορεί να εκφραστεί με ένα πεπερασμένο σύνολο συνώνυμων λέξεων ή φράσεων. Ένα σύνολο $W = \{W_1, W_2, \ldots, W_n\}$ περιέχει όλα τα σύνολα συνώνυμων για κάθε χαρακτηριστικό.

Σύμφωνα με τα παραπάνω, σε μία κριτική ενός προϊόντος κάποιο υποκείμενο (opinion holder) j σχολιάζει το υποσύνολο $S_j \subseteq F$ των χαρακτηριστικών του Ο. Για κάθε χαρακτηριστικό $f_k \in S_j$ που σχολιάζεται, το υποκείμενο χρησιμοποιεί κάποια λέξη ή φράση από το W_k για να περιγράψει το χαρακτηριστικό και εκφράζει μία θετική, αρνητική ή ουδέτερη άποψη για αυτό.

Digital_camera_1:
  Feature: **picture quality**
    Positive: 253
      <individual review sentences>
    Negative: 6
      <individual review sentences>
  Feature: **size**
    Positive: 134
      <individual review sentences>
    Negative: 10
      <individual review sentences>
...

Σχήμα 1: Ιεραρχική αναπαράσταση μιας οντότητας

2.2 Διαδικασία εξόρυξης γνώμης

Η εξόρυξη γνώσης από κείμενα είναι ένα δύσκολο εγχείρημα και αυτό διότι τα δεδομένα ενός κειμένου έχουν αρκετές ιδιαιτερότητες σε σχέση με τα άλλα δεδομένα [3]. Αναφορικά, οι κυριότερες που αναφέρονται στη βιβλιογραφία είναι οι ακόλουθες:

- *Λεξιλογική και σημασιολογική αμφισημία*: Στο γραπτό λόγο χρησιμοποιούνται συχνά προσωπικές αντωνυμίες που είναι δύσκολο να εντοπιστούν το ποια ουσιαστικά αντικαθιστούν και με ποιά επίθετα σχετίζονται. Επίσης, χρησιμοποιούνται συχνά συνώνυμες λέξεις ή λέξεις με πολλαπλό νόημα. Ο όρος σημασιολογική αμφισημία αναφέρεται στο γεγονός ότι η ίδια φράση μπορεί να έχει διαφορετική σημασία ανάλογα με το περιεχόμενο μέσα στο οποίο συναντάται.

- *Σχετική εξάρτηση*: Σε μια γλώσσα μια έννοια ή μια ενέργεια είναι ένας συνδυασμός λέξεων και φράσεων.

- *Θόρυβος στα δεδομένα*: Για παράδειγμα, ορθογραφικά λάθη, συντακτικά και γραμματικά λάθη που προκύπτουν από την χρήση του προφορικού λόγου σε γραπτό υλικό.

- *Πολλές διαστάσεις του μοντέλου*: Κατά τη μοντελοποίηση ενός κειμένου κάθε λέξη του μπορεί να είναι και μια διάσταση του μοντέλου που χρησιμοποιείται για εκπαίδευση και μάθηση. Το διάνυσμα που αναπαριστά ένα κείμενο ανάλογα με την απουσία ή παρουσία κάποιας λέξης, συνήθως είναι και πολύ αραιό.

Οι έρευνες στην περιοχή της εξόρυξης γνώμης συνήθως ακολουθούν μια προσέγγιση που αποτελείται από δύο βήματα: την αναγνώριση των θεμάτων και των προτάσεων και την κατηγοριοποίηση των προτάσεων και των κειμένων. Στο πρώτο βήμα, αρχικά πρέπει να προσδιοριστούν τα θέματα που αναφέρονται στα κείμενα που αποτελούν τα δεδομένα εισόδου. Στην συνέχεια πρέπει να συγκεντρωθούν οι προτάσεις που περιέχουν κάποιο συναίσθημα, θετικό ή αρνητικό. Έπειτα, οι προτάσεις αυτές συσχετίζονται με τα θέματα που έχουν προσδιοριστεί. Έτσι, στο παρακάτω απόσπασμα κειμένου[1]:

Οι πανεπιστημιακοί σημειώνουν ότι το κράτος οφείλει στα πανεπιστήμια τα απαιτούμενα μέσα ώστε να συνεχίσουν να εκπληρώνουν τα καθήκοντα τους, τα οποία είναι η εκπαίδευση, η παραγωγή και διακίνηση ιδεών. Υποστηρίζουν ότι "στη σημερινή κρίση, κάθε φορολογικός πόρος που δίδεται στα πανεπιστήμια πρέπει να προάγει την προκοπή της χώρας" και υπογραμμίζουν ότι "η κουλτούρα της αφθονίας και της ευκολίας, που στηρίχθηκε στα δανεικά, πρέπει να ανατραπεί. Παράλληλα, στέλνουν μήνυμα στους συναδέλφους τους, γράφοντας ότι οι πανεπιστημιακοί οφείλουν να εκπληρώνουν τα έργα τους "με ευσυνειδησία και ευαισθησία στις ανάγκες των φοιτητών και της κοινωνίας ".

προσδιορίστηκαν τα παρακάτω θέματα:

- Χρηματοδότηση των Πανεπιστημίων
- Αποστολή και ρόλος του Πανεπιστημίου

Κάποια από τα σχόλια που ακολούθησαν αυτό το κείμενο ήταν τα παρακάτω:

1. *επαναξιολογηση του διδακτικου προσωπικου απο το μηδεν(οσοι εχουν δημοσιευσεις και εργο υπαρκτο σε διεθνες επιπεδο να παραμεινουν οι αλλοι ασεπ και επιλεγονται οι καλυτεροι)…*

2. *τα παραπανω ειναι κοινος τοπος και πιστευω δομουν ενα ισχυρο πανεπιστημιο.*

3. *Η κυβέρνηση και η τροικα θέλουν να μειώσουν το κόστος για την παιδεία και συνεπώς τόσο οι 20 όσο και όσα σχολιάζετε είναι εκτός τόπου και χρόνου.*

4. *Εκτος των άλλων, ειδικά το θέμα της σύνδεσης με τις εταιρείες είναι η κορυφαία πρόταση. Που νομίζετε ότι βρίσκεστε στη Silicon Valley ή σε μια χώρα με οικονομία που απογειώνεται όπως η φαντασία των περισσοτέρων στη χώρα αυτή.*

Παρατηρούμε ότι τα σχόλια των χρηστών συχνά περιέχουν ασυνταξίες, ορθογραφικά λάθη, λέξεις από άλλες γλώσσες κλπ. Απο τα 4 σχόλια που απομονώθηκαν, όλα εκφράζουν συναίσθημα με εξαίρεση το πρώτο, όπου ο χρήστης απλά κάνει μια πρόταση. Η πρόταση 3 αφορά το θέμα «*Χρηματοδότηση των Πανεπιστημίων*», η πρόταση 4 το θέμα «*Αποστολή και ρόλος του Πανεπιστημίου*» ενώ η πρόταση 2 θα μπορούσε να ταξινομηθεί και στις δύο κατηγορίες.

Το πρώτο σχόλιο, παρότι εκφράζει μια γνώμη που σχετίζεται με την χρηματοδότηση των Πανεπιστημίων, δεν εκφράζει θετικό ή αρνητικό συναίσθημα για την δημοσίευση. Επιπλέον

[1] Πηγή: http://www.antinews.gr/2011/01/31/83033/

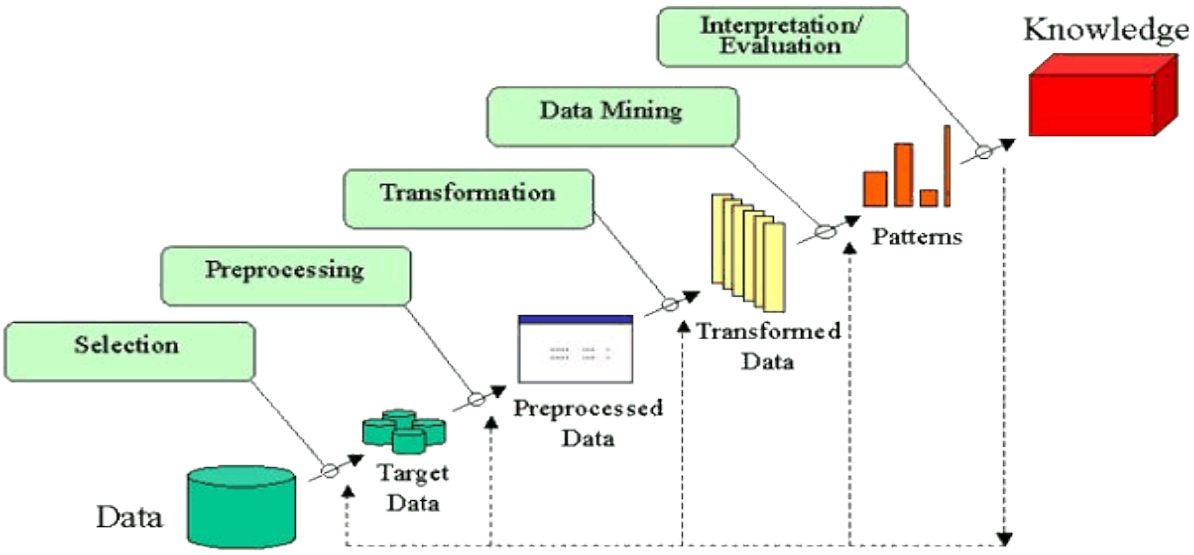
υπήρχαν αρκετά σχόλια που εξέφραζαν κάποιο συναίσθημα (αρνητικό ή θετικο) χωρίς να περιέχουν λέξεις/ φράσεις με αρνητικό ή θετικό προσανατολισμό. Σε αυτό το στάδιο, ο διαχωρισμός των προτάσεων σε εκείνες που διατυπώνονται με λέξεις που περιέχουν σαφή συναισθηματικό προσανατολισμό σε σχέση με εκείνες τις προτάσεις που εκφράζουν έμμεσα την συναισθηματική κατάσταση του φορέα της άποψης, έχει αποδειχθεί ότι βελτιώνει την ακρίβεια στο επόμενο βήμα [1]. Στο δεύτερο βήμα, το πρόβλημα της κατηγοριοποίησης αντιμετωπίζεται σαν ένα θέμα ταξινόμησης με διάκριση των σχολίων σε θετικά ή αρνητικά. Ωστόσο μπορεί να δημιουργηθούν συμπληρωματικές κατηγορίες με σκοπό να γίνει η ανάλυση πιο λεπτομερής. Έτσι μερικές εργασίες περιλαμβάνουν κατηγορίες για *ουδέτερες* ή *άσχετες* γνώμες. Με τον τρόπο αυτό επιτυγχάνεται η εστίαση του ταξινομητή σε γνώμες που παρουσιάζουν ουσιαστικό ενδιαφέρον για παραπέρα επεξεργασία.

Συνοψίζοντας, η εξόρυξη γνώμης θα μπορούσε να θεωρηθεί σαν ενα πρόβλημα ταξινόμησης που κάνει διάκριση μεταξύ διαφόρων κατηγοριών συναισθημάτων (θετικά, αρνητικά ή ουδέτερα). Αυτή η διάκριση ισχύει και στις μεθόδους που κατατάσσουν τα συναισθήματα σε μια αριθμητική κλίμακα, όπου οι διάφορες κατηγορίες συναισθημάτων ορίζονται μέσα σε ένα διάστημα τιμών αυτής της κλίμακας.

2.2.1 Συστήματα εξόρυξης γνώσης από κείμενα

Τα κριτήρια για να ένα αποτελεσματικό σύστημα εξόρυξης γνώσης από κείμενα ορίστηκαν ως εξής [4]:

- Το σύστημα θα πρέπει να λειτουργεί αποτελεσματικά σε μεγάλες συλλογές δεδομένων (κλιμάκωση).
- Θα πρέπει να χρησιμοποιεί περισσότερο αλγορίθμους παρά ευρετικές και χειρονακτικές μεθόδους.
- Θα πρέπει να εξάγει τα υποδείγματα (patterns) παρά την επιπρόσθετη προσθήκη εγγραφών.



Σχήμα 2: Τα στάδια της εξόρυξης γνώσης

Η διαδικασία της εξόρυξης γνώσης από κείμενα αποτελείται από συγκεκριμένα βήματα, όπως φαίνονται στο παρακάτω σχεδιάγραμμα.

Αναλυτικότερα:

- Βήμα 1^{ο} : *Συλλογή των εγγράφων σχετικών με το πρόβλημα*. Το βήμα αυτό εξετάζει η επιστήμη της ανάκτησης πληροφοριών (information retrieval). Το πρόβλημα που πρέπει να επιλυθεί σ' αυτή τη φάση είναι ο προσδιορισμός των εγγράφων που πρέπει να ανακτηθούν, δοθέντων κάποιων λέξεων/φράσεων αναζήτησης.

- Βήμα 2^{ο} : *Προεπεξεργασία των εγγράφων*. Το βήμα αυτό περιλαμβάνει όλες τις διαδικασίες μετασχηματισμού των αρχικών εγγράφων που ανακτώνται. Ένα κείμενο θεωρείται συνήθως ένας «σάκος με λέξεις» (bag of words) και μπορεί να μοντελοποιηθεί ως ένα πολυδιάστασο διάνυσμα. Κατά τη φάση της προεπεξεργασίας εξαλείφονται εκείνα τα χαρακτηριστικά του κειμένου που θα μπορούσαν να εισάγουν θόρυβο κατά τη διαδικασία της μοντελοποίησής του και της διαδικασίας εξόρυξης. Στο στάδιο αυτό μπορεί να λάβει χώρα η γλωσσική προεπεξεργασία (tokenization), δηλαδή η απαλοιφή από ένα κείμενο περιττών συμβόλων (%,!, HTML TAGS, ;, κ.τ.λ.). Επίσης μπορεί να λάβει χώρα και η διαδικασία της ανάθεσης των όρων στις γραμματικές τους κατηγορίες (part-of speech tagging) ή η διαδικασία της λημματοποίησης (lemmatization). Διαδεδομένη είναι και η μέθοδος αφαίρεσης των stopwords (λέξεις με μεγάλη συχνότητα σε ένα κείμενο, όπως άρθρα και σύνδεσμοι). Τέλος, μία κοινή μορφή επεξεργασίας των όρων ενός κειμένου είναι η αφαίρεση των καταλήξεων των λέξεων και η αντικατάστασή τους με τις ρίζες τους (stemming) [5]. Η επιλογή των διαδικασιών προεπεξεργασίας εξαρτάται από το στόχο και το πεδίο της διαδικασίας της μάθησης, αφού η μελέτη κάποιων χαρακτηριστικών μπορεί να είναι σκόπιμο να εξεταστεί ανάλογα με το πεδίο της εφαρμογής (π.χ. κατά την εξόρυξη γνώμης από blogs σύμβολα όπως τα θαυμαστικά ή λέξεις με έμφαση (tag BOLD) είναι σημαντικά χαρακτηριστικά).

- Βήμα 3^{ο} : *Μοντελοποίηση κειμένου και εξαγωγή χαρακτηριστικών*. Το βήμα αυτό είναι ίσως από τα πιο καθοριστικά για την ακρίβεια του παραγόμενου μοντέλου. Όπως αναφέρθηκε, ένα κείμενο μπορεί να αναπαρασταθεί ως ένας «σάκος με λέξεις» (bag of words). Η αναπαράσταση αυτή βασίζεται στη λογική του ότι κάθε κείμενο είναι μία συλλογή όρων ανεξαρτήτως σειράς και ότι το περιβάλλον ενός όρου δεν επηρεάζει το νόημά του. Έτσι, απλοποιεί την αναπαράσταση του κειμένου σε ένα διάνυσμα, όπου κάθε διάσταση είναι η παρουσία/απουσία μιας λέξης (Boolean model). Επίσης, κάθε διάσταση μπορεί να είναι η συχνότητα εμφάνισης μίας λέξης ή η κανονικοποιημένη συχνότητα εμφάνισης μίας λέξης (tf*idf), [5].

- Βήμα 4^{ο} : *Εφαρμογή διαδικασιών μάθησης από κείμενα*. Η μάθηση από κείμενα μπορεί να εφαρμοστεί στις εξής λειτουργίες:

    - **Ταξίνομηση**. Είναι η διαδικασία κατά την οποία τα κείμενα ταξινομούνται σε ένα προκαθορισμένο αριθμό κατηγοριών. Κατά τη διαδικασία της κατηγοριοποίησης, ένας ταξινομητής εκπαιδεύεται με ένα σύνολο εγγράφων, τα οποία έχουν συγκεκριμένα χαρακτηριστικά και στα οποία έχουν προστεθεί ετικέτες με την κατηγορία στην οποία ανήκουν. Στη συνέχεια ο ταξινομητής βρίσκει ένα μοντέλο για κάθε κατηγορία, το οποίο εκφράζεται ως συνάρτηση των χαρακτηριστικών των εγγράφων και χρησιμοποιεί το μοντέλο αυτό για να αναθέσει τα νέα έγγραφα που θα δεχτεί ως είσοδο σε μία κατηγορία.

- **Ομαδοποίηση**. Είναι η λειτουργία κατά την οποία ένα σύνολο εγγράφων διαιρείται σε ομάδες με βάση κάποιο μέτρο ομοιότητας. Στην ομαδοποίηση, οι κατηγορίες-ομάδες στις οποίες θα ανατεθούν τα έγγραφα δεν είναι προκαθορισμένες όπως στην κατηγοριοποίηση, αλλά ανακαλύπτονται κατά τη διαδικασία επεξεργασίας των κειμένων. Τα έγγραφα που ανήκουν σε μία ομάδα θα πρέπει να είναι παρόμοια μεταξύ τους περισσότερο απ' ότι με έγγραφα άλλων ομάδων.

- **Αναζήτηση και ανάκτηση πληροφοριών**. Κατά τη διαδικασία αυτή, δοθέντος ενός ερωτήματος (σε κείμενο), γίνεται αναζήτηση μέσα στο σύνολο των κειμένων για την εύρεση σχετικής πληροφορίας. Κείμενα με μη σχετική πληροφορία απομακρύνονται, και παρουσιάζονται αυτά που πληρούν τα κριτήρια της αναζήτησης.

- **Δημιουργία περίληψης των πληροφοριών.** Η λειτουργία αυτή περιλαμβάνει τη μείωση της ποσότητας ενός κειμένου, αποδίδοντας όμως το περιεχόμενο και το βασικό νόημα του.

2.3 Αλγόριθμοι ταξινόμησης κειμένων

Για την ταξινόμηση κειμένων χρησιμοποιούνται κυρίως οι παρακάτω μέθοδοι μάθησης [6]:

- Ταξινομητές που βασίζονται σε πιθανότητες, όπως ο αλγόριθμος Naive Bayes.
- Δένδρα αποφάσεων, όπως ο αλγόριθμος C4.5.
- Μέθοδοι παλινδρόμησης, όπως η ευθεία των ελάχιστων τετραγώνων Linear Least Squares Fit – LLSF.
- Online μέθοδοι μάθησης, όπως το νευρωνικό δίκτυο Perceptron.
- Η μέθοδος του Rocchio που χρησιμοποιείται στην ανάκτηση πληροφοριών.
- Νευρωνικά δίκτυα.
- Lazy learners, όπως ο αλγόριθμος των k-πλησιέστερων γειτόνων (kNN).
- Support Vector Machines (SVM).
- Μέθοδοι μάθησης συνόλου (ensemble learning methods).

Σύμφωνα με εμπειρικές έρευνες για τη σύγκριση των παραπάνω μεθόδων, οι αλγόριθμοι Naive Bayes και Support Vector Machines είναι οι αποτελεσματικότεροι στο πεδίο αυτό. Παρακάτω παρουσιάζονται οι δύο αυτοί βασικοί αλγόριθμοι μηχανικής μάθησης.

2.3.1 Ο αλγόριθμος Naive Bayes

Ο ταξινομητής Naive Bayes βασίζεται στην απλή υπόθεση ότι οι τιμές των χαρακτηριστικών είναι υπό συνθήκη ανεξάρτητες, δεδομένης της τιμής της εξαρτημένης μεταβλητής. Υπάρχουν αρκετές παραλλαγές στις εφαρμογές του αλγόριθμου αυτού. Ο McCallum [7] συνόψισε δύο βασικά μοντέλα Naive Bayes για την ταξινόμηση κειμένου, το πολυμεταβλητό μοντέλο του Bernoulli και το πολυωνυμικό μοντέλο.

Το πολυμεταβλητό μοντέλο του Bernoulli χρησιμοποιεί χαρακτηριστικά με τιμές boolean (παρουσία ή απουσία λέξεων) ενώ το πολυωνυμικό μοντέλο χρησιμοποιεί χαρακτηριστικά με τιμές μη-μηδενικούς ακεραίους (συχνότητα εμφάνισης λέξεων). Και τα δύο μοντέλα υποθέτουν την υπό συνθήκη ανεξαρτησία των χαρακτηριστικών και δε λαμβάνουν υπόψη τη σειρά των λέξεων.

Το πολυμεταβλητό μοντέλο του Bernoulli αποκαλείται επίσης μοντέλο δυαδικής ανεξαρτησίας (Binary Independence Model). Δοθέντος ενός συνόλου εγγράφων εκπαίδευσης D με λεξιλόγιο $V = w_1, w_2, \dots, w_m$ ένα έγγραφο αναπαρίσταται σαν ένα δυαδικό

διάνυσμα χαρακτηριστικών (λέξεων) με μήκος m: $d = (w_1, w_2, \dots, w_m)$. Κάθε χαρακτηριστικό - λέξη w_j έχει τιμή "1" αν η λέξη υπάρχει στο κείμενο και "0" αν δεν υπάρχει. Το μοντέλο αυτό δε λαμβάνει υπόψη τη συχνότητα εμφάνισης των λέξεων και το μήκος του εγγράφου. Δοθέντος ενός προβλήματος ταξινόμησης με l κλάσεις, η κλάση του εγγράφου καθορίζεται από την εξίσωση:

$$argmax_{lp(c_l)p(d|c_l)} = P(c_l) \prod_i P(w_i|c_l)^{w_i}(1 - P(w_i|c_l))^{1-w_i}$$

Το $P(w_i|c_l)$ υπολογίζεται από την εξίσωση

$$P(w_i|c_l) = \frac{n_c + mp}{n+m}$$

όπου n ο συνολικός αριθμός των εγγράφων και n_c ο συνολικός αριθμός των εγγράφων στα οποία συναντάται η λέξη w_i. Κάποιες φορές μία λέξη μπορεί να μη συναντάται σε μία κατηγορία, οπότε $n_c = 0$. Απαιτείται η χρήση τεχνικών εξομάλυνσης (smoothing), διότι μπορεί να υπάρχουν λέξεις που εμφανίζονται στα έγγραφα ελέγχου, αλλά όχι στα δεδομένα εκπαίδευσης. Εδώ χρησιμοποιούμε την m-εκτίμηση σε περίπτωση που $n_c = 0$. Για χαρακτηριστικά με boolean τιμές, $p = \frac{1}{2}$ και το m συνήθως τίθεται ίσο με 2.

Στο πολυωνυμικό μοντέλο, η τιμή κάθε χαρακτηριστικού-λέξης w_i είναι η συχνότητα εμφάνισής του στο έγγραφο. Αν συμβολίσουμε το συνολικό μήκος όλων των εγγράφων που ανήκουν στην κλάση c_l ως $l(c_l)$ και το μέγεθος του λεξιλογίου ως $|V|$, τότε το $P(w_i|c_l)$ υπολογίζεται από την εξίσωση

$$P(w_i|c_l) = \frac{freq(w_i) + 1}{l(c_l) + |V|}$$

και

$$P(c_l|d) = P(c_l) \prod_i P(w_i|c_l)^{w_i}$$

Η μέθοδος αυτή χρησιμοποιεί την εξομάλυνση Laplace. Στο πολυωνυμικό μοντέλο, ο εκπαιδευμένος ταξινομητής θα παραμείνει ίδιος αν αλλάξουμε τη σειρά των λέξεων μέσα σε ένα έγγραφο και ενώσουμε όλα τα παραδείγματα εγγράφων μίας κλάσης σε ένα μόνο παράδειγμα. Επομένως, το μήκος του κάθε ενός εγγράφου δε σχετίζεται με τους υπολογισμούς πιθανοτήτων.

Σύμφωνα με προηγούμενες έρευνες [7] το πολυμεταβλητό μοντέλο του Bernoulli είναι αποτελεσματικό όταν εφαρμόζεται σε σύνολα δεδομένων με μικρό λεξιλόγιο, ενώ το πολυωνυμικό μοντέλο είναι πιο αποτελεσματικό όταν έχουμε λεξιλόγιο μεγάλου μεγέθους. Άρα το πολυωνυμικό μοντέλο είναι πιο δημοφιλές σε εφαρμογές κατηγοριοποίησης κειμένου. Για ένα πρόβλημα δυαδικής ταξινόμησης, οι προβλέψεις και για τα δύο μοντέλα καθορίζονται από τον ακόλουθο λόγο κλάσης PR.

$$PR = \frac{P(c_1|d)}{P(c_2|d)}$$

Το έγγραφο d ανήκει στην κλάση c_1 αν $PR > 1$ και στη c_2 διαφορετικά. Μπορούμε να δούμε το λόγο PR ως ένα μέτρο εμπιστοσύνης (confidence measure) για τις προβλέψεις του

αλγόριθμου Naive Bayes. Μετά το λογαριθμικό μετασχηματισμό, μία θετική τιμή του $logPR$ αντιπροσωπεύει την εμπιστοσύνη ότι το παράδειγμα ανήκει στην κλάση c_1, ενώ μία αρνητική τιμή την εμπιστοσύνη ότι ανήκει στην κλάση c_2. Όσο μεγαλύτερη είναι η τιμή $|logPR|$, τόσο πιο σίγουρος είναι ο ταξινομητής για την πρόβλεψη. Το PR μπορεί να χρησιμοποιηθεί σαν μέθοδος για να βαθμολογήσουμε τα αποτελέσματα της πρόβλεψης.

2.3.2 Ο αλγόριθμος Support Vector Machines (SVM)

Ο αλγόριθμος SVM ανήκει στις επιβλεπόμενες μεθόδους μηχανικής μάθησης και προσπαθεί να μεγιστοποιήσει τη γενίκευση, ώστε να αντιμετωπίσει το πρόβλημα της υπερπροσαρμογής (overfitting).

Δοθέντων των δεδομένων εκπαίδευσης $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)$, ο SVM προσπαθεί να μεγιστοποιήσει το περιθώριο του ορίου απόφασης, βρίσκοντας το μέγιστο της συνάρτησης

$$W(a) = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i}^{l}\sum_{j}^{l} a_i a_j y_i y_j K(x_i, x_j)$$

με περιορισμούς

$$\sum_{i=1}^{l} a_i y_i = 0, a_i \geq 0, i = 1,2, \dots, l$$

Τα παραδείγματα εγγράφων που βρίσκονται στο περιθώριο έχουν μη μηδενικές τιμές a_i και ονομάζονται Support Vectors (SV). Τα υπόλοιπα έγγραφα έχουν μηδενικές τιμές a_i και θεωρείται ότι δε συνεισφέρουν στην ταξινόμηση.

Στον παραπάνω τύπο, ο όρος $K(x_i, x_j)$ είναι η συνάρτηση kernel (kernel function). Παρόλο που η SVM μπορεί να χειριστεί μη γραμμικά όρια με τη συνάρτηση kernel, μελέτες έχουν δείξει ότι η γραμμική συνάρτηση kernel είναι αποτελεσματική για την εργασία της κατηγοριοποίησης κειμένου ενώ η πολυωνυμική συνάρτηση kernel δε βελτιώνει αισθητά την απόδοση (Leopold και Kindermann, 2002). Επομένως εδώ χρησιμοποιούμε την απλή γραμμική συνάρτηση kernel, όρος $K(x_i, x_j) = x_i \cdot x_j$.

Δοθέντος ενός παραδείγματος ελέγχου x, η γραμμική συνάρτηση απόφασης είναι η εξής:

$$f(x) = w \cdot x + b$$

όπου

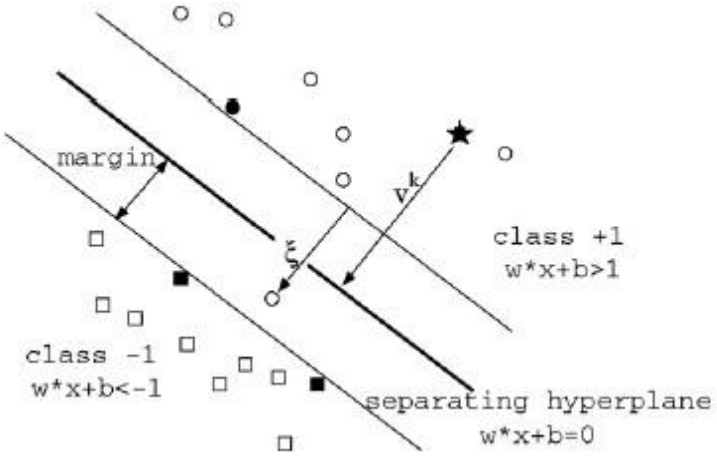
$$w = \sum_{i=1}^{l} a_i y_i x_i$$

και

$$b = y_i - w \cdot x_i$$

Η εξίσωση για την απόφαση ταξινόμησης είναι $D = sign(f(x))$.

Η τιμή της συνάρτησης απόφασης, δηλαδή η έξοδος του αλγορίθμου SVM για κάθε πρόβλεψη, μπορεί να θεωρηθεί σαν ένα είδος κριτηρίου εμπιστοσύνης της πρόβλεψης. Όσο μεγαλύτερη είναι η απόλυτη τιμή, τόσο πιο μακριά είναι το σημείο από το όριο της απόφασης, επομένως ο ταξινομητής είναι πιο «σίγουρος» για την πρόβλεψη.



Σχήμα 3: Γεωμετρική αναπαράσταση του τρόπου λειτουργίας των SVM

2.4 Κατηγορίες αλγορίθμων εξόρυξης γνώμης

Το πρόβλημα της αναζήτησης απόψεων σε δεδομένα κείμενα, προσεγγίζεται ερευνητικά τόσο σε επίπεδο ολόκληρου του κειμένου (γενικό συναίσθημα σε ένα κείμενο για κάποιο αντικείμενο), όσο και σε προτασιακό επίπεδο (εξαγωγή προτάσεων του κειμένου που εκφράζουν κάποια άποψη και ταξινόμηση αυτών σε θετικές και αρνητικές).

Κατά την πρώτη προσέγγιση (document level approach) γίνεται η παραδοχή ότι ένα κείμενο αναφέρεται σε ένα μόνο αντικείμενο και περιέχει την άποψη ενός μόνο υποκειμένου. Εντέλει, ταξινομείται ως θετικό, αρνητικό ή ουδέτερο σε σχέση με το αντικείμενο του ενδιαφέροντος. Κατά τη δεύτερη (sentence level approach) γίνεται η παραδοχή ότι μία πρόταση περιέχει μόνο μία άποψη. Μια τρίτη προσέγγιση βασίζεται στη γενική ιδέα ότι μια άποψη μπορεί να αναφέρεται τόσο σε μία οντότητα ολόκληρη, όσο και σε κάθε επιμέρους χαρακτηριστικό της. Συνεπώς, μία πιο λεπτομερής προσέγγιση της εξόρυξης απόψεων είναι αυτή σε επίπεδο χαρακτηριστικού (feature level approach). Η προσέγγιση αυτή περιλαμβάνει τη διαδικασία της αναγνώρισης των χαρακτηριστικών τα οποία έχει σχολιάσει ο χρήστης, τον σημασιολογικό προσδιορισμό των απόψεων και τέλος την ομαδοποίηση των συνώνυμων χαρακτηριστικών. Αποτέλεσμα αυτής της προσέγγισης μπορεί να είναι και μία περίληψη των απόψεων για κάθε σχολιασμένο χαρακτηριστικό.

Οι αλγόριθμοι που χρησιμοποιούνται για τους σκοπούς της εξόρυξης γνώμης μπορούν να χωριστούν σε τρεις κατηγορίες. Προσεγγίσεις που βασίζονται στην μηχανική μάθηση, προσεγγίσεις λεξικού και στατιστικές προσεγγίσεις. Ωστόσο απο τις πρώτες ημέρες της εξόρυξης γνώμης, η μηχανική μάθηση αποτελεί το πιο συχνά χρησιμοποιούμενο εργαλείο.

2.4.1 Προσεγγίσεις βασισμένες στην μηχανική μάθηση

Η προσέγγιση που βασίζεται στην μηχανική μάθηση αποτελεί μια προηγμένη λύση για το πρόβλημα της ταξινόμησης που μπορεί σε γενικές γραμμές να περιγραφεί σαν μια διαδικασία δύο βημάτων: 1) εκπαίδευση του μοντέλου από ενα σώμα δεδομένων εκπαίδευσης (εποπτευόμενα, μη-εποπτευόμενα) και 2) ταξινόμηση των πραγματικών δεδομένων με βάση το εκπαιδευμένο μοντέλο.

Υποθέτουμε ότι τα δεδομένα εκπαίδευσης είναι έγγραφα που αναπαρίστανται σε έναν χώρο D του οποίου οι διαστάσεις είναι τα χαρακτηριστικά των εγγράφων (όπως συχνότητα λέξεων, διγράμματα κλπ). Επιπλέον, σε αυτά τα έγγραφα έχει ανατεθεί μια τιμή που εκφράζει ένα συναίσθημα, από έναν χώρο S.

Για τα δεδομένα εκπαίδευσης $\{(D_i \in D, S_i \in S)\}$, πρέπει να βρεθεί μια τιμή g:

$$g: D \to S,\ g(D_i) = arg\max_S f(D_i, S_i) \qquad (1)$$

Η παραπάνω εξίσωση λέει ότι δοσμένου ενός συνόλου από ζεύγη εγγράφων D_i και τιμών συναισθημάτων S_i, θέλουμε να βρούμε μια εξίσωση g που να αντιστοιχεί έγγραφα σε τιμές συναισθημάτων σύμφωνα με τις καλύτερες προβλέψεις μιας συνάρτησης βαθμολόγησης f. Αυτή η συνάρτηση παίρνει σαν είσοδο έγγραφα και τιμές συναισθημάτων και προσδιορίζει το συναίσθημα του εγγράφου με μια πιθανότητα. Χωρίς απώλεια της γενικότητας, η διαδικασία εκπαίδευσης μπορεί να θεωρηθεί σαν μια εκτίμηση της συνάρτησης βαθμολόγησης f.

Η διαδικασία της μηχανικής μάθησης περιλαμβάνει τα παρακάτω βήματα. Πρώτον δημιουργείται ένα σύνολο δεδομένων εκπαίδευσης το οποίο μπορεί να είναι είτε σχολιασμένο με τιμές συναισθημάτων (εποπτευόμενη μάθηση) ή όχι (μη – εποπτευόμενη μάθηση). Δεύτερον, κάθε έγγραφο αναπαρίσταται σαν ενα διάνυσμα χαρακτηριστικών. Τρίτον, ένας ταξινομητής εκπαιδεύεται προκειμένου να διακρίνει τις τιμές των συναισθημάτων, αναλύοντας τα σχετικά χαρακτηριστικά των εγγράφων. Τέλος, ο ταξινομητής αυτό χρησιμοποιείται για να προβλέψει τιμές συναισθημάτων για νέα έγγραφα.

Στο πεδίο της εξόρυξης γνώμης έχουν συμβάλει εκατοντάδες δημοσιεύσεις τα περασμένα χρόνια. Γενικά, έχουν χρησιμοποιηθεί τόσο τεχνικές επιβλεπόμενης, όσο και μη επιβλεπόμενης μάθησης. Στις μεθόδους επιβλεπόμενης μάθησης εξετάζονται μεγαλύτερα δομικά στοιχεία (πρόταση, παράγραφος, κείμενο) για την εξόρυξη συναισθήματος. Στην εργασία των Pang et.al. [8] τα κείμενα μοντελοποιήθηκαν ως α) διανύσματα συχνότητας λέξεων και β) απουσίας ή παρουσίας λέξεων και εφαρμόστηκαν τρεις αλγόριθμοι μηχανικής μάθησης: Ο Naive Bayes, ένας αλγόριθμος βασισμένος σε SVM και ένας βασισμένος στο μέγεθος της εντροπίας. Καλύτερη απόδοση είχε αυτός που χρησιμοποιούσε SVM, με κείμενο που μοντελοποιήθηκε ως διάνυσμα παρουσίας ή απουσίας όρων. Πάντως, οι περισσότερες μεθοδολογίες επιβλεπόμενης μάθησης χρησιμοποιούνται κυρίως για ταξινόμηση συναισθήματος σε επίπεδο κειμένου.

Η τρέχουσα δημοτικότητα της προσέγγισης της μηχανικής μάθησης προέρχεται από την εργασία των Pang και Lee [8]. Οι συγγραφείς πρότειναν και αξιολόγησαν τρεις εποπτευόμενες μεθόδους ταξινόμησης: *απλοϊκός ταξινομητής Bayes* (Naïve Bayes - NB), *μέγιστης εντροπίας* (Maximum Entropy - ME) και *Μηχανές Διανυσμάτων Υποστήριξης* (Support Vector Machines - SVM). Σύμφωνα με την εκτίμησή τους, η μέθοδος των SVM είχε την καλύτερη επίδοση ενώ αυτή του NB είχε την χειρότερη από τις τρεις (με τις διαφορές ωστόσο μεταξύ τους να είναι μικρές). Σε κάθε περίπτωση και οι τρεις αλγόριθμοι ξεπέρασαν

την τυχαία επιλογή, παρουσιάζοντας μια μέση ακρίβεια της τάξης του 80%. Οι Dave et al. επέκτειναν το έργο των Pang και Lee δίνοντας έμφαση στην επιλογή των χαρακτηριστικών [9]. Χρησιμοποίησαν επίσης εξομάλυνση Laplace για τον NB, το οποίο αύξησε την ακρίβεια στο 87% (σε ένα συγκεκριμένο σύνολο δεδομένων). Ωστόσο, ο SVM επέτυχε παρόμοια ακρίβεια, υπολείποντας του NB μόνο όταν χρησιμοποιήθηκαν μεμονωμένες λέξεις (unigrams).

Στην ανάλυση συναισθήματος προκύπτει η ανάγκη για μια πιο λεπτομερή κατάταξη των απόψεων των κειμένων από το να ταξινομηθούν αυτές σαν θετικές ή αρνητικές. Για τους σκοπούς αυτούς εφαρμόζονται άλλες μέθοδοι ταξινόμησης εκτός από την παραδοσιακή τεχνική των SVM. Παρότι η τεχνική αυτή είναι αποδοτική για δυαδική ταξινόμηση, η ανάλυση συναισθήματος απαιτεί πιο εξεζητημένες λύσεις που στοχεύουν σε ταξινόμηση σε περισσότερες κατηγορίες.

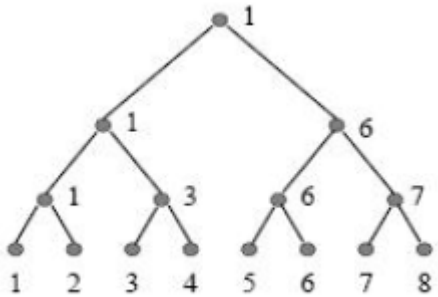
Οι δύο βασικές στρατηγικές που ακολουθούνται σε multi-class προβλήματα ταξινόμησης είναι:

1. One versus all
2. One versus one

Στην προσέγγιση One Vs All (OVA), για ένα πρόβλημα διαχωρισμού *L* κλάσεων, εκπαιδεύονται *L* SVMs με τα δεδομένα του συνόλου εκπαίδευσης. Κάθε μία SVM διαχωρίζει μια κλάση από όλες τις υπόλοιπες. Για την ταξινόμηση ενός νέου δείγματος δοκιμής ακολουθείται η στρατηγική winner-takes-all.

Στην τεχνική One Vs One (OVO) εκπαιδεύονται $\frac{L(L-1)}{2}$ SVMs. Κάθε μία SVM διαχωρίζει ένα ζευγάρι κλάσεων. ΄Επειτα, για κάθε νέο δείγμα γίνεται σύγκριση κάθε κλάσης με καθεμία από τις υπόλοιπες *L* – 1 κλάσεις ξεχωριστά. Για κάθε σύγκριση μεταξύ δύο κλάσεων η επικρατούσα κλάση παίρνει μια ψήφο. Το δείγμα ταξινομείται στην κλάση που συγκεντρώνει τελικά τις περισσότερες ψήφους [10]. Η προσέγγιση αυτή αναφέρεται ως max-wins voting SVM.

Μια άλλη προσέγγιση της τεχνικής One Vs One (OVO), που διαφοροποιείται από την προηγούμενη ως προς την ταξινόμηση νέων δειγμάτων, χρησιμοποιεί ένα δυαδικό δέντρο όπως αυτό του σχήματος. Η ταξινόμηση γίνεται ως εξής : Μετά από κάθε σύγκριση μεταξύ ενός ζεύγους κλάσεων προκύπτει η «νικήτρια» κλάση. Οι επικρατούσες κλάσεις ενός επιπέδου του δυαδικού δέντρου ανεβαίνουν στο επόμενο επίπεδο όπου και γίνεται ένας καινούργιος γύρος συγκρίσεων. Στο τέλος, η επικρατούσα κλάση θα εμφανιστεί στην κορυφή του δέντρου και θα αποτελεί την κλάση του τρέχοντος δείγματος δοκιμής.



Σχήμα 4: Δομή δυαδικού δέντρου για πρόβλημα 8 κλάσεων. Οι αριθμοί 1-8 αντιπροσωπεύουν τις κλάσεις. Η επικρατούσα κλάση θα εμφανιστεί στην κορυφή του δέντρου

Εκτός όμως από τις παραπάνω στρατηγικές, χρησιμοποιείται και μια ακόμη που βασίζεται σε κωδικές λέξεις (codewords). Η βασική ιδέα της τεχνικής αυτής συνοψίζεται ως εξής: Κάθε ετικέτα κλάσης αντιστοιχίζεται σε μια δυαδική κωδική λέξη από *d* bits και εκπαιδεύονται αντίστοιχα *d* SVMs, μία για κάθε bit, τα οποία χωρίζουν το '0' ή το '1' για κάθε bit στην ακολουθία. Έτσι προκύπτει για κάθε νέο δείγμα μια δυαδική ακολουθία από *d* bits [11]. Το νέο δείγμα αντιστοιχίζεται έπειτα στην κλάση της οποίας η κωδική λέξη ταυτίζεται ή είναι πιο κοντά στην ακολουθία bits που προέκυψε από τις SVMs. Σημαντικό ρόλο στην τεχνική αυτή παίζει η επιλογή πλαισίου κωδικοποίησης και το μήκος *d* των codewords που θα χρησιμοποιηθούν. Γενικά, όσο μεγαλύτερη απόσταση Hamming έχουν οι κωδικές λέξεις των κλάσεων ανά δύο μεταξύ τους τόσο πιο πιθανή είναι και η σωστή πρόβλεψη της κλάσης ενός νέου δείγματος, ακόμα κι αν κάποιοι από τους *d* ταξινομητές παρουσιάζουν σφάλμα κατά τη διάρκεια ταξινόμησης του.

Οι τεχνικές SVM αναπτύχθηκαν για να επιλύσουν προβλήματα ταξινόμησης, ωστόσο έχουν επεκταθεί και σε εφαρμογές παλινδρόμησης (regression), οπότε η διαδικασία ονομάζεται Support Vector Regression (SVR) [12]. Το μοντέλο που παράγεται από την τεχνική SVM εξαρτάται μόνο από ένα υποσύνολο των δεδομένων εκπαίδευσης, επειδή η συνάρτηση κόστους για την κατασκευή του μοντέλου δεν λαμβάνει υπόψιν της σημεία που βρίσκονται πέρα από ένα περιθώριο. Κατά αναλογία, το μοντέλο που παράγεται από την τεχνική SVR εξαρτάται μόνο από ένα υποσύνολο των δεδομένων εκπαίδευσης, επειδή η συνάρτηση κόστους για την κατασκευή του μοντέλου αγνοεί τα δεδομένα εκπαίδευσης που βρίσκονται κοντά στην πρόβλεψη του μοντέλου, εντός ενός ορίου *ε*.

Οι Pang και Lee πρότειναν για multi-class προβλήματα ταξινόμησης την χρήση της τεχνικής SVM με την προσέγγιση One Vs All (OVA) σε συνδυασμό με την τεχνική SVR και με *μετρική επισήμανσης* (metric labeling) [13]. Η μετρική επισήμανσης είναι μια ειδική περίπτωση μιας εκ των υστέρων βελτιστοποίησης της ανάθεσης μιας κλάσης (σε μια κατηγορία) σε σχέση με την προηγούμενη ανάθεση. Η νέα ανάθεση ελαχιστοποιεί το άθροισμα των αποστάσεων μεταξύ των ετικετών κοντινών σημείων, έτσι ώστε παρόμοιες κλάσεις να τοποθετούνται πιο κοντά σε μια κλίμακα αξιολόγησης. Τα αποτελέσματα έδειξαν με σαφήνεια ότι ο συνδυασμός των τεχνικών SVM με άλλες μη εποπτευόμενες μεθόδους ταξινόμησης επιτυγχάνει μεγαλύτερη ακρίβεια. Μια μεταγενέστερη εργασία για την υποστήριξη ή αντίθεση απόψεων, στο πλαίσιο πολιτικών κειμένων, μελέτησε περαιτέρω την τεχνική SVM μέσα από την μοντελοποίηση σχέσεων και συμφωνιών μεταξύ των συντακτών [14].

Η απόδοση των μεθόδων μηχανική μάθησης εξαρτάται σε μεγάλο βαθμό από την ποιότητα και την ποσότητα των δεδομένων εκπαίδευσης, τα οποία είναι λιγοστά σε σχέση με την ποσότητα των δεδομένων που δεν έχουν επισημανθεί. Οι Goldberg και Zhu πρότειναν μια ημι-εποπτευόμενη τεχνική μάθησης που εφαρμόζεται σε έναν γράφο αποτελούμενο από επισημασμένα και μη επισημασμένα δεδομένα [15]. Οι συγγραφείς αναπαριστούν τα έγγραφα με έναν γράφο, όπου οι κορυφές αντιστοιχούν σε έγγραφα και όπου οι ακμές του συνδέουν παρόμοια έγγραφα χρησιμοποιώντας μια μετρική απόστασης που υπολογίζεται απευθείας από τα χαρακτηριστικά των εγγράφων. Αυτές οι υποθέσεις είναι παρόμοιες με την μετρική επισήμανσης, με την διαφορά ότι χρησιμοποιούνται από πριν και ως εκ τούτου επιτρέπουν την χρήση ακόμα και σε μη επισημασμένα δεδομένα για σκοπούς εκπαίδευσης. Παρότι η προσέγγιση τους παρουσιάσε καλύτερη επίδοση από την τεχνική SVR, οι συγγραφείς αναφέρουν ότι είναι ευαίσθητη στην επιλογή του μέτρου ομοιότητας και ότι δεν μπορεί να επωφεληθεί από την χρήση πρόσθετων επισημασμένων δεδομένων.

Στις μελέτες που αναφέρθηκαν παραπάνω ο υπολογισμός των τεχνικών θεωρήθηκε ότι έγινε σε επίπεδο εγγράφου (document level), επιδεικνύοντας έτσι μια «κατά μέσο όρο» ακρίβεια σε ετερογενή σχόλια τα οποία αναφέρουν πολλαπλές πτυχές του θέματος με κάθε

μια να εκφράζει διαφορετικά συναισθήματα. Αυτό θέτει το πρόβλημα της ταξινόμησης συμφραζομένων συναισθημάτων, το οποίο απαιτεί όχι μόνο αλγορίθμους που λειτουργούν σε επίπεδο πρότασης αλλά και που συμπεριλαμβάνουν και το πλαίσιο της κάθε πρότασης στην ανάλυση τους (Wilson et al) [16]. Οι Pang και Lee (2005) έδειξαν ότι η τεχνική SVR, παρότι είναι λιγότερο ακριβής από την τεχνική SVM, παράγει επισημάνσεις που είναι πιο κοντά στις πραγματικές [8]. Τα στοιχεία αυτά υποστηρίζουν τον ισχυρισμό ότι με την χρήση μιας «βαθμωτής» συνάρτησης στην SVR, παρόμοιοι όροι λαμβάνουν παρόμοιες επισημάνσεις.

Εκτός από την επιλογή των αλγορίθμων και των δεδομένων επιλογής, η επίδοση των προσεγγίσεων μηχανικής μάθησης εξαρτάται σε μεγάλο βαθμό και από την επιλογή των χαρακτηριστικών (features). Ο πιο απλός τρόπος είναι να κωδικοποιηθεί κάθε στοιχείο του συνόλου από την παρουσία ή την απουσία του στο έγγραφο. Στην περίπτωση που το ρόλο των χαρακτηριστικών παίζουν λέξεις του κειμένου, αυτό θα παράγει μια αναπαράσταση του εγγράφου ως ένα δυαδικό διάνυσμα. Επεκτείνοντας αυτήν την αναπαράσταση, μπορούμε να χρησιμοποιήσουμε τις σχετικές συχνότητες των εμφανίσεων των λέξεων. Ωστόσο δεν είναι όλες οι λέξεις εξίσου αντιπροσωπευτικές και επομένως χρήσιμες για τους σκοπούς της ανάλυσης υποκειμενικότητας. Το γεγονός αυτό μας δίνει την ευκαιρία να κάνουμε τη μαθησιακή διαδικασία πιο αποτελεσματική με την μείωση των διαστάσεων του D (συνάρτηση 1). Οι Osherenko et al. αποδεικνύουν ότι είναι δυνατή η χρήση μόνο ενός μικρού συνόλου από τις πιο συναισθηματικές λέξεις, σχεδόν χωρίς καμία επίπτωση στην απόδοση του ταξινομητή του [17]. Η άμεση χρήση των τιμών συναισθημάτων από τέτοια λεξικά έχει δείξει μικρή ή και καθόλου αύξηση της ακρίβειας. Ως εκ τούτου, οι μελέτες συνήθως χρησιμοποιούν τις συχνότητες των λέξεων. Για παράδειγμα, οι Devitt και Ahmad εντοπίζουν τις λέξεις που εκφράζουν συναίσθημα σε ένα έγγραφο με την χρήση του λεξικού Senti-WordNet, αλλά στη συνέχεια χρησιμοποιουν τις συχνότητες εμφάνισης τους για το έργο της κατάταξης [18]. Αυτή η προσέγγιση είναι επίσης δημοφιλής με μεθόδους λεξικού, τις οποίες θα περιγράψουμε στην συνέχεια.

2.4.2 Προσεγγίσεις βασισμένες σε λεξικό

Η προσέγγιση λεξικού βασίζεται σε ένα προϋπάρχον λεξικό που περιέχει λέξεις συνοδευόμενες από τον προσανατολισμό τους, όπως τα General Inquirer, WordNet-Affect ή το SentiWordNet που αποτελεί και το πιο δημοφιλές λεξικό σήμερα [19].

Οι υφιστάμενες εργασίες εκμεταλλεύονται αυτά τα λεξικά κυρίως για την αναγνώριση των λέξεων που εκφράζουν κάποιο συναίσθημα, παρόλο που κάποιες πρόσφατες μελέτες έδειξαν ότι είναι δυνατή η απευθείας χρήση τιμών διαβάθμισης συναισθήματος, παρέχοντας μια τιμή συναισθήματος σε μια συνεχή κλίμακα [20]. Στην περίπτωση αυτή ο προσανατολισμός μιας πρότασης ή ενός κειμένου καθορίζεται συνήθως από τον μέσο όρο των τιμών συναισθήματος των ανεξάρτητων λέξεων. Για παράδειγμα, οι περισσότερες από τις μεθόδους λεξικού αθροίζουν τις τιμές συναισθήματος μιας πρότασης ή ενός εγγράφου και υπολογίζουν την τιμή συναισθήματος που προκύπτει χρησιμοποιώντας αλγορίθμους βασισμένους σε κανόνες [21]. Πιο εξελιγμένα εργαλεία όπως ο *Αναλυτής Συναισθήματος* που εισήχθει από τον Yi et al. [22] ή η γλωσσική προσέγγιση των Thet et al [23], εξάγουν τιμές συναισθήματος για ορισμένα θέματα χρησιμοποιώντας προηγμένες μεθόδους που εκμεταλλεύονται ειδικά χαρακτηριστικά του τομέα, καθώς και υποδείγματα προτάσεων που εκφράζουν γνώμη και ανάλυση μερών του λόγου. Οι δύο παραπάνω προσεγγίσεις οδηγούν σε καλύτερες επιδόσεις έχοντας όμως αυξημένη υπολογιστική πολυπλοκότητα.

Οι μη επιβλεπόμενες μέθοδοι για την εξόρυξη γνώμης επικεντρώνονται στην εξαγωγή και ταξινόμηση λέξεων ή προτάσεων από το κείμενο, οι οποίες θεωρούνται ως ατομικές

μονάδες που εκφράζουν συναίσθημα (Turney [24], Riloff και Wiebe [25], Hatzivassiloglou και Wiebe [26]). Το γενικό συναίσθημα ενός κειμένου είναι τελικά το άθροισμα θετικών και αρνητικών συναισθημάτων των επιμέρους μονάδων. Επειδή, σε αυτές τις μεθόδους δεν χρησιμοποιούνται δεδομένα εκπαίδευσης, γίνεται χρήση εξωτερικών βάσεων γνώσης (π.χ.WordNet), οι οποίες επιστρέφουν το σημασιολογικό περιεχόμενο και τη συναισθηματική κατεύθυνση μίας λέξης.

Χρησιμοποιούμε τώρα έναν τύπο που ορίζει την πιο γενική περίπτωση της ανάθεσης γνώμης σε ένα έγγραφο με την χρήση ενός λεξικού:

$$S(D) = \frac{\sum_{w \in D} S_w \cdot weight(w) \cdot modifier(w)}{\sum weight(w)} \qquad (2)$$

Στην παραπάνω εξίσωση, το S_w αναπαριστά την τιμή συναισθήματος του λεξικού για την λέξη w, η οποία αθροίζεται σε σχέση με κάποια συνάρτηση στάθμισης *weight()* και ενός τελεστή μετασχηματισμού *modifier()* ο οποίος χειρίζεται την άρνηση, την ένταση των λέξεων και άλλες περιπτώσεις που επηρεάζουν την τιμή συναισθήματος. Οι συναρτήσεις στάθμισης μπορεί να ορισθούν στατικά για κάθε πρόταση ή να υπολογισθούν δυναμικά, λαμβάνοντας υπόψιν τις θέσεις των λέξεων. Συνήθως οι συναρτήσεις στάθμισης αναπαριστούν ένα πλαίσιο γύρω από την λέξη που αντιστοιχεί στο θέμα, λαμβάνοντας υπόψιν τις τιμές συναισθήματος των λέξεων που είναι άμεσοι γείτονες με αυτήν. Για παράδειγμα, μια συνάρτηση στάθμισης μπορεί να έχει την τιμή 1 για δύο ή περισσότερες λέξεις που περικλείουν την λέξη που αντιστοιχεί στο θέμα και τιμή 0 σε διαφορετική περίπτωση. Πιο εξεζητημένες μέθοδοι μπορεί επίσης να χρησιμοποιηθούν, όπως η επεξεργασία φυσικής γλώσσας (NLP) που μπορεί να οδηγήσει σε εναν δυναμικό υπολογισμό της συνάρτησης στάθμισης για κάθε πρόταση, λαμβάνοντας υπόψιν την συγκεκριμένη δομή της.

Η χρήση λεξικού μπορεί να συνδυασθεί και με μεθόδους μηχανικής μάθησης, όπως έχει ήδη αναφερθεί. Σημειώνουμε ότι το να βασιζόμαστε στις τιμές συναισθήματος ενός λεξικού δεν είναι πάντα εφικτό, καθώς το λεξικό μπορεί να μην ενδείκνυται για χρήση σε συγκεκριμένα σύνολα δεδομένων (που περιέχουν τεχνικούς όρους και αφορούν ένα συγκεκριμένο πεδίο). Επιπλέον, οι μέθοδοι λεξικού δεν μπορούν να προσαρμόσουν τις τιμές συναισθήματος σε ένα συγκεκριμένο πλαίσιο συμφραζομένων. Προκύπτει ότι οι λέξεις μπορεί να έχουν διαφορετικές τιμές συναισθήματος ανάλογα με τον έαν χρησιμοποιούνται σε διαφορετικό πλαίσιο [20]. Αντίθετα με τις προσεγγίσεις που βασίζονται σε λεξικό, οι μέθοδοι μηχανικής μάθησης προσαρμόζονται στο σώμα του κειμένου με το οποίο έχουν εκπαιδευτεί.

2.4.3 Στατιστικές προσεγγίσεις

Η στατιστική προσέγγιση στοχεύει στο να ξεπεράσει τα προβλήματα που αναφέρθηκαν παραπάνω. Για παράδειγμα, οι Farni και Klenner πρότειναν να εξάγουν τις πολικότητες χρησιμοποιώντας την εμφάνιση των επιθέτων σε ένα κείμενο [20]. Σε αυτήν την περίπτωση, η προσαρμοστικότητα επιτυγχάνεται μέσω της κατασκευής ενός λεξικού που βασίζεται στο κείμενο. Όσον αφορά το πρόβλημα της μη εμφάνισης ορισμένων λέξεων, η στατιστική μέθοδος με βάση το κείμενο προτείνει να ξεπεραστεί με την χρήση ενός σώματος κειμένου που είναι αρκετά μεγάλο. Για τον σκοπό αυτό, είναι δυνατόν να χρησιμοποιηθεί ολόκληρο το σύνολο των εγγράφων του διαδικτύου που έχουν ενταχθεί σε ένα ευρετήριο σαν ένα σώμα κειμένου για την κατασκευή ενός λεξικού [24]. Χρησιμοποιούνται στην μη – επιβλεπόμενη ταξινόμηση συναισθήματος.

Μπορούμε να προσδιορίσουμε τον προσανατολισμό μιας λέξης με την μελέτη των συχνοτήτων με την οποία αυτή η λέξη εμφανίζεται σε ένα μεγάλο σώμα κειμένων που έχουν επισημειωθεί. Εάν η λέξη εμφανίζεται πιο συχνά σε κείμενα με θετική (ή αρνητική) έννοια, τότε έχει έναν θετικό (ή αρνητικό) προσανατολισμό. Ίσες εμφανίσεις δείχνουν ουδέτερες λέξεις. Αν και υπολογιστικά είναι αποδοτική, η βασική μέθοδος απαιτεί ένα μεγάλο επισημειωμένο σώμα κειμένου, το οποία είναι ένας περιοριστικός παράγοντας.

Οι πιο εξελιγμένες μέθοδοι βασίζονται στην παρατήρηση ότι οι λέξεις που εκφράζουν παρόμοια γνώμη εμφανίζονται πιο συχνά μαζί σε ένα σώμα κειμένου. Αντίστοιχα, αν δύο λέξεις εμφανίζονται συχνά μαζί στο ίδιο πλαίσιο, είναι πιθανό να έχουν τον ίδιο προσανατολισμό. Συνεπώς, ο προσανατολισμός μιας άγνωστης λέξης μπορεί να καθοριστεί υπολογίζοντας την σχετική συχνότητα συνύπαρξης της με μια άλλη λέξη, η οποία διατηρεί τον προσανατολισμό της (για παράδειγμα η λέξη «καλό»). Για να επιτύχουν αυτόν τον σκοπό, ο Peter Turney πρότεινε την χρήση του κριτηρίου της *σημειακής αμοιβαίας πληροφορίας* (ΣΑΠ – Pointwise Mutual Information) [24]. Η σημειακή αμοιβαία πληροφορία είναι ένα μέτρο ομοιότητας που υπολογίζει κατά πόσον δύο λέξεις συνεμφανίζονται πάντα μαζί ή όχι σε τμήματα κειμένου. Το κάθε τμήμα μπορεί να είναι μια πρόταση, μια παράγραφος κ.τ.λ.

Η σημειακή αμοιβαία πληροφορία δύο λέξεων w_1 και w_2 υπολογίζεται ως εξής:

$$PMI(w_1, w_2) = log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

Όπου:

$\boldsymbol{P(w_1, w_2)}$: η πιθανότητα να βρεθούν στο ίδιο τμήμα οι λέξεις w_1, w_2.
$\boldsymbol{P(w_1)}$: η πιθανότητα να βρεθεί σε ένα τμήμα η λέξη w_1.
$\boldsymbol{P(w_2)}$: η πιθανότητα να βρεθεί σε ένα τμήμα η λέξη w_2.

Ο προσανατολισμός μιας λέξης x (PMI-IR) υπολογίζεται σαν την διαφορά μεταξύ των τιμών ΣΑΠ που έχουν υπολογιστεί μεταξύ δύο λιστών λέξεων: μια λίστα με θετικές λέξεις, *pWords*, όπως η λέξη «εξαιρετικά» και μια λίστα με αρνητικές λέξεις, *nWords*, όπως η λέξη «φτωχή»:

$$PMI - IR(x) = \sum_{p \in pWords} PMI(x, p) - \sum_{n \in nWords} PMI(x, n)$$

Δηλαδή, το εάν μία φράση έχει θετική ή αρνητική συναισθηματική χροιά είναι η διαφορά της πιθανότητας να συναντάται συχνά, σε κείμενα, κοντά στη λέξη “τέλειος” μείον τη διαφορά να συναντάται συχνά, σε κείμενα, κοντά στη λέξη “φτωχός”. Οι παραπάνω πιθανότητες μπορούν να εκτιμηθούν χρησιμοποιώντας ένα μεγάλο σώμα κειμένων (ή ιστιοσελίδων). Οι Chaovalit et al. [27] χρησιμοποίησαν την συλλογή εγγράφων της μηχανής αναζήτησης AltaVista. Συγκεκριμένα, μετράται το πλήθος χτυπημάτων (hits) κατά την αναζήτηση κάθε όρου με τη μηχανή αναζήτησης της AltaVista. Οι μηχανές αναζήτησης επιτρέπουν τον υπολογισμό της πιθανότητας να βρεθούν στο ίδιο τμήμα όχι μόνο δύο λέξεις, αλλά και ολόκληρες φράσεις, το οποίο είναι ένα χρήσιμο χαρακτηριστικό γνώρισμα. Για τον υπολογισμό του συναισθηματικού προσανατολισμού ενός κειμένου υπολογίζεται ο μέσος όρος όλων των φράσεων του. Αν αυτός είναι θετικός τότε η κριτική ταξινομείται ως θετική κριτική, αλλιώς ως αρνητική.

Μια άλλη σημαντική εργασία στην κατηγορία αυτή είναι αυτή του Dave et. al. [24]. Μία κριτική ταξινομείται ανάλογα με το άθροισμα ενός σκορ που έχουν λάβει διάφορα

χαρακτηριστικά της. Πρώτα επιλέγεται ένα σύνολο χαρακτηριστικών $F = \{f_1, f_2, f_3, \dots\}$. Η συνάρτηση υπολογισμού του σκορ ενός χαρακτηριστικού είναι η ακόλουθη:

$$score(f_i) = \frac{P(f_i|C) - P(f_i|C')}{P(f_i|C) + P(f_i|C')}$$

Στην παραπάνω συνάρτηση, C είναι μία κλάση (θετική κριτική) και C' το συμπλήρωμά της (αρνητική κριτική). Το σκορ είναι το πηλίκο της πιθανότητας εμφάνισης του χαρακτηριστικού σε μία θετική κριτική μείον την πιθανότητα εμφάνισης του χαρακτηριστικού σε μία αρνητική κριτική προς το άθροισμα της πιθανότητας εμφάνισης του χαρακτηριστικού σε μία αρνητική κριτική συν την πιθανότητα εμφάνισης του χαρακτηριστικού σε μία θετική κριτική. Η κλάση της κριτικής καθορίζεται από τον τύπο:

$$class(d_j) = \begin{cases} C & eval(d_j) > 0 \\ C' & eval(d_j) < 0 \end{cases}$$

Όπου:

$$eval(d_j) = \sum_i score(f_i)$$

Δηλαδή κατατάσσεται ως θετική (κλάση C), εάν το άθροισμα των σκορ των χαρακτηριστικών της είναι θετικό.

Η χρήση των στατιστικών μεθόδων στον υπολογισμό του προσανατολισμού μιας γνώμης έχει βρει ενδιαφέρουσα εφαρμογή στην εργασία του Ben He et al. [28], οι οποίοι προτείνουν την χρήση ένός λεξικού γνώμης σε συνδυασμό με μεθόδους ανάκτησης πληροφορίας με σκοπό να ανακτήσουν γνώμες που έχουν εκφραστεί σε blogs. Η προσέγγιση τους κατασκευάζει πρώτα ένα λεξικό εξάγοντας συχνούς όρους από ολόκληρη την συλλογή, οι οποίοι στην συνέχεια ταξινομούνται ανάλογα με την συχνότητα τους μεταξύ κειμένων που εκφράζουν άποψη. Ο προσανατολισμός ενός εγγράφου υπολογίζεται σαν ένα σκόρ συνάφειας σε ένα ερώτημα που συντίθεται από τους κορυφαίους όρους αυτού του λεξικού. Τέλος, το σκορ της σχετικότητας γνώμης συνδυάζεται με το σκόρ της σχετικότητας θέματος, παρέχοντας μια κατάταξη των κειμένων που εκφράζουν μια γνώμη για αυτό το θέμα.

Παρόμοια με την στατιστική είναι και η σημασιολογική προσέγγιση. Παρέχει τιμές συναισθήματος άμεσα (όπως η στατιστική προσέγγιση) με την εξαίρεση ότι στηρίζεται σε διαφορετικές αρχές για τον υπολογισμό της ομοιότητας μεταξύ των λέξεων. Η βασική αρχή όλων των προσεγγίσεων σε αυτήν την κατηγορία είναι ότι λέξεις που παρουσιάζουν σημασιολογική εγγύτητα θα πρέπει να λάβουν παρόμοιες τιμές συναισθήματος. Το λεξικό WordNet παρέχει διάφορα είδη σημασιολογικών σχέσεων μεταξύ των λέξεων και μπορεί να χρησιμοποιηθεί για τον υπολογισμό του προσανατολισμού λέξεων. Η δυνατότητα για αποσαφήνιση των εννοιών των λέξεων με την χρήση του Senti -WordNet μπορεί να υπηρετήσει σαν ένας τρόπος για να συμπεριληφθεί το πλαίσιο αυτών των λέξεων στο έργο της εξόρυξης γνώμης. Παρόμοια με τις στατιστικές μεθόδους, χρησιμοποιούνται ως σημείο εκκίνησης δύο σύνολα με λέξεις με θετικό και αρνητικό προσανατολισμό για την κατασκευή ενός λεξικού.

Οι Kamps et al πρότειναν την χρήση της ελάχιστης απόστασης μιας σχέσης "συνωνύμων", επιδεικνύοντας μια ικανοποιητική συμφωνία (70%) με ενα σχολιασμένο λεξικό [29]. Ένας άλλος δημοφιλής τρόπος της χρήσης του Senti – WordNet είναι να ληφθεί μια λίστα με λέξεις

που εκφράζουν συναίσθημα επεκτείνοντας με επαναλήψεις το αρχικό σύνολο με συνώνυμα και αντώνυμα (Kim και Hovy) [30]. Ο προσανατολισμός μιας άγνωστης λέξης καθορίζεται από την σχετική καταμέτρηση των θετικών και των αρνητικών συνωνύμων αυτής της λέξης (Kim και Hovy, 2004) [30]. Διαφορετικά, άγνωστες λέξεις μπορεί να απορρίπτονται (Hu και Liu, 2004a) [31]. Ωστόσο, είναι σημαντικό να γνωρίζουμε ότι από την στιγμή που η σχετικότητα του συνωνύμου μειώνεται με το μήκος της διαδρομής μεταξύ των συνωνύμων και της αρχικής λέξης, έτσι θα πρέπει να συμβαίνει και με την τιμή της πολικότητας. Αν και όπως επισημάνθηκε από τον Godbole et al [32], θα πρέπει αρχικά να εξετάσουμε μόνο τα μονοπάτια που περνούν μέσα από τις λέξεις που έχουν τον ίδιο προσανατολισμό.

2.4.4 Άλλες προσεγγίσεις

Στην εργασία «*Αυτόματη παραγωγή συγκρίσεων προϊόντων από κριτικές χρηστών*» έχουν χρησιμοποιηθεί γλωσσικά μοντέλα για σκοπούς εξόρυξης γνώμης [33]. Ένα γλωσσικό μοντέλο αναθέτει σε κάθε ακολουθία m λέξεων μια πιθανότητα $P(w_1, \dots, w_m)$, η οποία δείχνει πόσο πιθανό είναι να εμφανιστεί αυτή η ακολουθία λέξεων σε ένα κείμενο της κατηγορίας των κειμένων για την οποία έχει εκπαιδευτεί το μοντέλο. Η πιθανότητα $P(w_1, \dots, w_m)$ υπολογίζεται με τον παρακάτω τύπο χρησιμοποιώντας πιθανότητες εμφάνισης n – γραμμάτων, οι οποίες εκτιμούνται από σύνολα κειμένων εκπαίδευσης. Θεωρούμε ότι στην αρχή κάθε ακολουθίας λέξεων υπάρχουν n – 1 ψευδο-λέξεις.

$$P(w_1, \dots, w_m) = \prod_{i=1}^{m} P(w_i | w_1, \dots, w_{i-1}) \approx \prod_{i=1}^{m} P\left(w_i | w_{i-(n-1)}, \dots, w_{i-1}\right)$$

2.5 Ταξινόμηση απόψεων σε επίπεδο προτάσεων

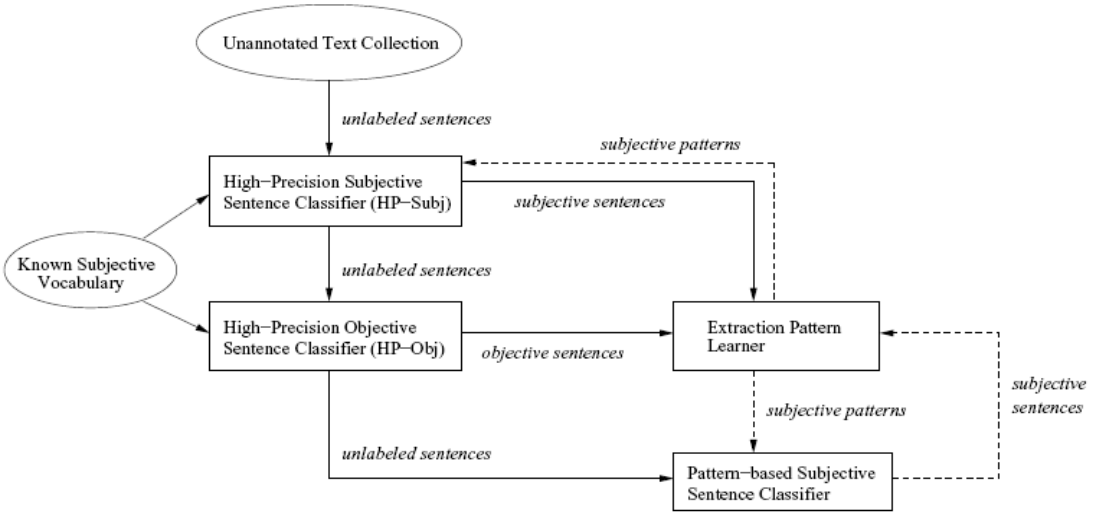
Είδαμε ότι η ταξινόμηση απόψεων σε επίπεδι εγγράφου γίνεται κυρίως με αλγορίθμους μηχανικής μάθησης. Στο σημείο αυτό θα αναφέρουμε αλγορίθμους που προσπαθούν να καθορίσουν αν μια πρόταση ενός κειμένου εκφράζει άποψη (subjective) ή όχι (objective). Αυτό μας βοηθάει στο να εξάγουμε από ένα κείμενο μόνο εκείνες τις απόψεις που εκφράζουν άποψη και να αγνοήσουμε τις υπόλοιπες. Έτσι μπορούμε να ετσιάσουμε μόνο στις προτάσεις που μας ενδιαφέρουν.

2.5.1 Αναγνώριξη προτασεων που εκφράζουν άποψη

Όλες οι μέθοδοι που αφορούν την επίλυση αυτού του προβλήματος χρησιμοποιούν κάποια τεχνική μηχανικής μάθησης. Η μεταβλητή στόχος μπορεί να πάρει δύο τιμές, objective (αντικειμενική/ ουδέτερη πρόταση) και subjective (υποκειμενική/ πρόταση που εκφράζει άποψη).

Οι Riloff και Wiebe εφάρμοσαν μία επαναληπτική διαδικασία (bootstrapping approach) για την εκμάθηση εξαγωγής προτάσεων που περιέχουν υποκειμενικότητα [34]. Χρησιμοποίησαν αρχικά δύο ταξινομητές υψηλής ακρίβειας (HP- Subj και HP-Obj) για την ανάκτηση κάποιων ουδέτερων και υποκειμενικών προτάσεων. Οι προτάσεις που ανακτήθηκαν απ' αυτούς χρησιμοποιήθηκαν ως σύνολο εκπαίδευσης σε κάποιον άλλο αλγόριθμο, ο οποίος μπορούσε να μάθει να εξάγει ακολουθίες λέξεων (extraction of patterns) που σχετίζονταν με την υποκειμενικότητα. Συνήθως, αυτές οι ακολουθίες περιοριζόνταν σε καθορισμένες συντακτικές μορφές φράσεων, π.χ. <subj> passive-verb. Τα πρότυπα που εξαγόταν μπορούσαν στη συνέχεια να χρησιμοποιηθούν για την εξαγωγή περισσότερων

υποκειμενικών προτάσεων. Η διαδικασία που ακολουθήθηκε παρουσιάζεται στο παρακάτω διάγραμμα:



Σχήμα 5: Μεθοδολογία των Rilloff και Wiebe

Παρακάτω παρουσιάζονται κάποια πρότυπα που εκφράζουν την υποκειμενικότητα μίας πρότασης.

SYNTACTIC FORM	EXAMPLE PATTERN
<subj> passive-verb	<subj> was satisfied
<subj> active-verb	<subj> complained
<subj> active-verb dobj	<subj> dealt blow
<subj> verb infinitive	<subj> appear to be
<subj> aux noun	<subj> has position
active-verb <dobj>	endorsed <dobj>
infinitive <dobj>	to condemn <dobj>
verb infinitive <dobj>	get to know <dobj>
noun aux <dobj>	fact is <dobj>
noun prep <np>	opinion on <np>
active-verb prep <np>	agrees with <np>
passive-verb prep <np>	was worried about <np>
infinitive prep <np>	to resort to <np>

Σχήμα 6:Τα πρότυπα που εκφράζουν την υποκειμενικότητα μίας πρότασης

Σε άλλες εργασίες που εκπονήθηκαν για την εξαγωγή υποκειμενικών προτάσεων χρησιμοποιήθηκαν ταξινομητές όπως ο Naive Bayes και ο Multiple Naive Bayes.

2.5.2 Αναγνώριξη του προσανατολισμού των απόψεων

Οι Yu και Hazivassiloglou σε εργασία τους σχετικά με την αναγνώριση του προσανατολισμού μίας άποψης σε προτασιακό επίπεδο, χρησιμοποίησαν αρχικά μία σειρά από κλασικούς ταξινομητές για την αναγνώριση των υποκειμενικών προτάσεων [35]. Πιο

συγκεκριμένα έκαναν χρήση τριών αλγορίθμων μηχανικής μάθησης: του Naive Bayes, πολλαπλών ταξινομητών Naive Bayes και μιας μεθοδολογίας μέτρησης ομοιότητας προτάσεων (SIMFINDER) για την ανακάλυψη υποκειμενικών προτάσεων. Για την εξόρυξη του προσανατολισμού των προτάσεων έκαναν χρήση μίας μεθόδου παρόμοιας με αυτή του Turney [24], αλλά για τον υπολογισμό του σημασιολογικού προσανατολισμού χρησιμοποιήθηκαν περισσότερες λέξεις πέρα από τις λέξεις excellent και poor.

$$L(W_i, POS_j) = log\left(\frac{\frac{Freq(W_{all}, POS_j, ADJ_p) + \varepsilon}{Freq(W_{all}, POS_j, ADJ_p)}}{\frac{Freq(W_{all}, POS_j, ADJ_n) + \varepsilon)}{Freq(W_{all}, POS_j, ADJ_n)}}\right)$$

Στον παραπάνω τύπο, $Freq(W_{all}, POS_j, ADJ_p)$ είναι η συχνότητα συνύπαρξης όλων των λέξεων που είναι μέρη του λόγου POS_j με κάποιο προκαθορισμένο θετικό επίθετο ADJ_p και η σταθερά ε είναι συντελεστής εξομάλυνσης. Για την κατηγοριοποίηση κάθε λέξης λαμβάνεται υπόψη ο μέσος LLR όλων των λέξεων στην πρόταση και χρησιμοποιούνται κατώφλια για να ληφθεί μία απόφαση για το αν είναι θετική, αρνητική ή ουδέτερη.

2.6 Ανακάλυψη του σημασιολογικού προσανατολισμού λέξεων και φράσεων

Κατά την ανακάλυψη του σημασιολογικού προσανατολισμού των απόψεων προέκυψε η αναγκαιότητα του καθορισμού του σημασιολογικού προσανατολισμού λέξεων ή φράσεων. Κάποιες λέξεις (κυρίως επίθετα και επιρρήματα) είναι πάντα θετικά ή αρνητικά σε κάθε περιεχόμενο (π.χ. όμορφος (+), θαυμάσιος (+), καλός (+), απαίσιος (-), cost someone a leg and a arm(-)). Ωστόσο, η θετική ή αρνητική σημασία κάποιων λέξεων μπορεί να εξαρτάται πολλές φορές και από το περιεχόμενο στο οποίο αναφέρονται (π.χ. μεγάλο ταξίδι (-), μεγάλη επιτυχία (+)).

Για την κατάρτιση θετικών και αρνητικών λιστών λέξεων και φράσεων χρησιμοποιούνται κυρίως τρεις προσεγγίσεις:

- Χειροκίνητη κατάρτιση της λίστας (Manual approach).
- Κατάρτιση λίστας με τη χρήση λεξικού (knowledge based approach).
- Κατάρτιση λίστας από κειμενικό υλικό (corpus-based approach).

2.6.1 Κατάρτιση λίστα από κειμενικό υλικό (Corpus – Based Approach)

Η μεθοδολογία αυτή βασίζεται στη συχνότητα συνεμφάνισης ακολουθιών λέξεων σε μεγάλες κειμενικές βάσεις δεδομένων [26]. Στην κατηγορία αυτή ανήκουν και οι μέθοδοι που χρησιμοποίησαν ο Turney [24] (μέθοδος PMI) και οι Yu et al [35] (μέθοδος LLR) για τον προσδιορισμό της συναισθηματικής χροιάς μίας λέξης ή φράσης. Η διαφορά ανάμεσα στις μεθόδους των δύο παραπάνω ερευνών είναι ότι στη δεύτερη περίπτωση μελετήθηκε η συνύπαρξη λέξεων με περισσότερα θετικά και αρνητικά επίθετα. Γίνεται αντιληπτό ότι αυτές οι μέθοδοι μπορούν να εφαρμοστούν μόνο στην περίπτωση που το λεξιλόγιο εμπίπτει σε κάποιον συγκεκριμένο τομέα, ώστε το αρχικό δείγμα λέξεων που θα χρησιμοποιηθεί να είναι μονοσήμαντα θετικά ή αρνητικά προσδιορισμένο. Σε άλλες εργασίες χρησιμοποιήθηκαν

περιορισμοί ή σύνδεσμοι για να ταξινομηθούν λέξεις που εκφράζουν άποψη. Αυτή η μέθοδος προκύπτει από την παρατήρηση ότι επίθετα που συνδέονται με τη λέξη «ΚΑΙ» είναι συνήθως συνώνυμα [36]. Η σύνδεση με άλλους συνδέσμους (ή, αλλά, ούτε... ούτε) έχει αντίστοιχες ιδιότητες. Μετά τον εντοπισμό συνδέσμων μπορεί να εφαρμοστούν τεχνικές μηχανικής μάθησης (log-linear model), με σκοπό να καθοριστεί εάν δύο συνδεδεμένα επίθετα εκφράζουν απόψεις κοινής συναισθηματικής χροιάς ή όχι. Τεχνικές ομαδοποίησης χρησιμοποιούνται για την παραγωγή δύο ομάδων λέξεων (θετικών και αρνητικών).

2.6.2 Κατάρτιση της λίστας με τη χρήση κάποιου λεξικού συνωνύμων/ αντωνύμων (dictionary – based approach)

Σε πάρα πολλές εργασίες συναντάται η χρήση του on-line λεξικού WordNet για την κατάρτιση λιστών λέξεων και φράσεων θετικού ή αρνητικού σημασιολογικού προσανατολισμού προτάσεων. Το WordNet παρέχει πληροφορία για τις συνώνυμες λέξεις κάποιας αναζητούμενης καθώς και για άλλα μέρη του λόγου που σχετίζονται με τη λέξη αυτή (για κάθε ουσιαστικό, παρέχονται επίθετα και επιρρήματα που έχουν την ίδια ρίζα/stem).

Η διαδικασία της κατάρτισης της λίστας ξεκινάει με ένα μικρό πλήθος αρχικών λέξεων θετικής και αρνητικής χροιάς. Το WordNet χρησιμοποιείται για αναζήτηση συνωνύμων και αντωνύμων των λέξεων, πληροφορία για τις συνώνυμες λέξεις κάποιας αναζητούμενης καθώς και για άλλα μέρη του λόγου που σχετίζονται με τη λέξη αυτή (για κάθε ουσιαστικό, παρέχονται επίθετα και επιρρήματα που έχουν την ίδια ρίζα/stem). Η διαδικασία λήγει με ανθρώπινη παρατήρηση των αποτελεσμάτων.

Σε κάποιες εργασίες χρησιμοποιήθηκε επιπλέον πληροφορία από το WordNet, όπως οι αναλυτικές περιγραφές της σημασίας των λέξεων ή και τεχνικές μηχανικής μάθησης. Το μειονέκτημα της προσέγγισης αυτής για την κατάρτιση θετικών και αρνητικών λέξεων είναι ότι δε λαμβάνεται υπόψη η ιδιότητα κάποιων λέξεων να αλλάζουν συναισθηματική χροιά ανάλογα με το περιεχόμενο στο οποίο αναφέρονται και επομένως δεν μπορούν να προσδιοριστούν μονοσήμαντα.

3 Εξόρυξη γνώμης από Ιστολόγια

Τα τελεταία χρόνια γίνεται πολύς λόγος για τα ιστολόγια (blogs) και τον τρόπο με τον οποίο επηρεάζουν τα μέσα μαζικής ενημέρωσης και αλλάζουν τον τρόπο που οι άνθρωποι επικοινωνούν και μοιράζονται γνώσεις. Ένας μεγάλος αριθμός ακαδημαϊκών ερευνών σχετίζονται με τα ιστολόγια και καλύπτουν πολλές διαφορετικές πτυχές τους. Από εθνογραφικές και κοινωνιολογικές μελέτες μέχρι την ανάπτυξη τυπικών μαθηματικών μοντέλων ροής πληροφοριών μέσω των ιστολογίων. Σήμερα ο χρήστης μπορεί να δημοσιεύει την άποψή του για τις ειδησεις που διαβάζει προσαρτώντας την στη σελίδα ειδήσεων και να την καταστήσει έτσι διαθέσιμη στους υπόλοιπους χρήστες. Με την χρήση ενός ιστολογίου έχει την δυνατότητα να δημιουργήσει μια προσωπική σελίδα στην οποία μπορεί να αρθρογραφεί δημοσίως για όποιο θέμα επιθυμεί. Η δημιουργία της πληροφορίας παύει να είναι πλέον μονοπώλιο λίγων, αφού ο κάθε χρήστης μπορεί εάν το επιθυμεί να την δημιουργήσει.

Σήμερα παρατηρείται μεγάλη άνθιση τέτοιων σελίδων και μπορεί κάποιος να βρεί ιστολόγια σχεδόν για κάθε θέμα. Πολύς κόσμος προτιμά να ενημερώνεται μέσω ιστολογίων αντί να καταφεύγει στις συμβατικές πηγές πληροφόρησης. Στις πληροφορίες που "ανεβαίνουν" στο ιστολόγιο δεν υπάρχει άμεσος έλεγχος από κάποιον ιεραρχικά ανώτερο χρήστη ή οργανισμό, γεγονός που τις καθιστά μη εύκολα ελεγχόμενες. Η εξέλιξη των ιστολογίων έχει επηρρεάσει τα παραδοσιακά ΜΜΕ, τις αγοραστικές τάσεις των καταναλωτών και έχει παίξει σημαντικό ρόλο στην αλλαγή των κοινωνικών ισσοροπιών σε πολλές χώρες. Η δυναμική αυτής της εξέλιξης ενισχύεται συνεχώς. Από την σκοπιά της επιστήμης υπολογιστών, η σφαίρα των ιστολογίων προσφέρει νέες προκλήσεις για τομείς όπως η υπολογιστική γλωσσολογία, η μηχανική μάθηση και πιο συγκεκριμένα η εξόρυξη γνώμης. Η εξόρυξη γνώμης σε ιστολόγια προσπαθεί με έναν αυτοματοποιημένο τρόπο να εντοπίσει και να αναλύσει τις απόψεις που προβάλλονται μέσα από αυτά.

3.1 Το περιβάλλον των ιστολογίων

Σε έναν κόσμο που αλλάζει συνεχώς η αύξηση της επιρροής της σφαίρας των blogs σχετικά με θέματα που αφορούν από τις επιχειρήσεις μέχρι την πολιτική συνεχίζει να αυξάνεται. Η στάση των συγγραφέων των blogs (bloggers), η οποία φαίνεται να μη διαφοροποιείται πολύ από παράγοντες όπως η ηλικία ή το φύλο, ή ακόμη η γεωγραφική περιοχή, φαίνεται να γίνεται όλο και πιο καταλυτική για τη διαμόρφωση μίας παγκοσμιοποιημένης κοινής γνώμης και συνδράμει αποφασιστικά στον εκδημοκρατισμό της έκφρασης.

3.1.1 Η δικτυακή κοινότητα των ιστολογίων

Το ιστολόγιο, γνωστό συχνά με την άκλιτη ονομασία μπλογκ (blog), είναι μορφή ιστοχώρου. Είναι λίστα καταχωρήσεων από την πιο πρόσφατη καταχώρηση στην παλαιότερη. Το περιεχόμενο των καταχωρήσεων μπορεί είναι οτιδήποτε, όπως νέα, κοινωνικός σχολιασμός, σχολιασμός των μέσων μαζικής ενημέρωσης και των διασημοτήτων, προσωπικά ημερολόγια και ειδικά θέματα όπως τεχνολογία, μόδα, αθλητικά, τέχνες και άλλα. Συνήθως δεν απαιτείται ενδελεχής επιμέλεια του κώδικα της ιστοσελίδας, μιας και συχνά είναι εγκατεστημένα αυτόματα συστήματα, που παρέχουν την δυνατότητα στο διαχειριστή του ιστολογίου να συντάξει μια καταχώρηση σε πολύ λίγα βήματα. Με βάση το περιεχόμενο του ένα ιστολόγιο μπορεί να κατηγοριοποιηθεί σε προσωπικό ημερολόγιο (journal), ενημερωτικό ιστολόγιο (news blog) και σε άλλες κατηγορίες. Στο επίκεντρο του

ενδιαφέροντος βρίσκονται τα ιστολόγια που λειτουργούν σαν δικτυακά ημερολόγια, δίνοντας τη δυνατότητα να δούμε το περιβάλλον των ανθρώπων, τα ενδιαφέροντά τους, τις απόψεις τους και τα συναισθήματά τους.

Η επέκταση της χρήσης των ιστολογίων, οφείλεται κατά πολύ στο γεγονός ότι κάθε χρήστης του διαδικτύου μπορεί εύκολα και δωρεάν να ξεκινήσει το δικό του ιστολόγιο μέσω πολλών φορέων που προσφέρουν συστήματα τα οποία στηρίζονται σε λογισμικό και έχουν μετατρέψει την σύνταξη των ιστολογίων σε πολύ απλή διαδικασία. Η διαφορά-κλειδί μεταξύ του περιεχομένου άλλων ιστοσελίδων και των ιστολογίων είναι ότι τα τελευταία αντιπροσωπεύουν μεμονωμένα άτομα. Οι κυριότερες διαφορές μεταξύ του περιεχομένου των ιστολογίων και του περιεχομένου άλλων ιστοσελίδων προκύπτουν από το εξής: ένα ιστολόγιο, στις περισσότερες περιπτώσεις, είναι σαν μία αντιπροσώπευση της ζωής ενός ανθρώπου και επιδεικνύει τη δυναμική και αλληλεπιδραστική συμπεριφορά που είναι χαρακτηριστική στους ανθρώπους. Οι προσωπικές ιστοσελίδες (home pages) σχετίζονται επίσης με άτομα, αλλά δεν έχουν την ίδια δυναμική συμπεριφορά που έχουν τα ιστολόγια, που μοιάζει με την ανθρώπινη.

3.1.2 Μίκρο – ιστολόγια (microblogs)

Τα μικρο-ιστολόγια είναι μια μορφή ιστολογίων που δίνουν τη δυνατότητα στους χρήστες να δημοσιεύουν συνοπτικά κείμενα, τα οποία δεν υπερβαίνουν συνήθως τους 140 - 200 χαρακτήρες, καθώς και εικόνες ή εξωτερικούς συνδέσμους (video κλπ). Μέσω αυτών των μικρών δημοσιεύσεων οι χρήστες μπορούν να γνωστοποιούν την τρέχουσα κατάσταση τους, την τοποθεσία τους καθώς και άλλες πληροφορίες. Τις δημοσιεύσεις αυτές μπορούν να τις διαχειριστούν οι χρήστες είτε διαδικτυακά, είτε μέσω αποστολής τους με sms ή e-mail, είτε μέσω προγραμμάτων άμεσης αλληλογραφίας (instant messaging clients). Δίνεται, επίσης η δυνατότητα στους χρήστες να ενσωματώνουν τα μικρο-ιστολόγια τους και σε άλλα blogs ή σελίδες μέσω μικρών εφαρμογών (widgets).

Το Twitter[2] είναι η πιο διαδεδομένη υπηρεσία μίκρο-ιστολογίου, η οποία επιτρέπει στους χρήστες του να γράφουν σύντομα μηνύματα και να διαβάζουν τα μηνύματα άλλων χρηστών της υπηρεσίας (γνωστά ως tweets). Οι χρήστες επικοινωνούν μεταξύ τους απαντώντας σε ερώτησεις του τύπου: "Τί κάνεις τώρα;". Το Twitter έχει μετεξελιχθεί σε έναν χώρο όπου χρήστες από όλο τον κόσμο κοινοποιούν τις σκέψεις τους σε όσους έχουν επιλέξει. Από προεπιλογή είναι δημόσια ορατά σε όλους τους χρήστες μέσω της σχετικής υπηρεσίας αναζήτησης, ωστόσο οι αποστολείς μπορούν να περιορίσουν την λειτουργία αυτή θέτοντας μια λίστα από εγκεκριμένους χρήστες. Όλοι οι χρήστες μπορούν να στέλνουν και να λαμβάνουν μηνύματα μέσω της ιστοσελίδας του Twitter, από συμβατές εξωτερικές εφαρμογές αλλά και μέσω υπηρεσιών γραπτών μηνυμάτων (SMS) που διατίθενται σε πολλές χώρες. Από τη δημιουργία του το 2006 από τον Jack Dorsey, σήμερα περιγράφεται ώς το «SMS του Διαδικτύου».

3.1.3 Η γλώσσα των ιστολογίων

Τα περισσότερα ιστολόγια είναι ένα μέσο για τον χρήστη ώστε να εκφράζει τις σκέψεις και τους προβληματισμούς του ανεπιφύλακτα. Ως τέτοια, τα ιστολόγια επιδεικνύουν μια μοναδική ανεπίσημη γλώσσα που έχει ιδιότητες και μονολόγου και διαλόγου. Είναι καταχωρήσεις ημερολογίου και ανοιχτές προσκλήσεις για συζήτηση ταυτόχρονα. Η γλώσσα που χρησιμοποιείται σε πολλά ιστολόγια είναι μία μίξη προφορικού και γραπτού λόγου που αναμειγνύει χαρακτηριστικά και των δύο, όπως το είδος του λεξιλογίου (π.χ. χρήση jargon) ή η δομή των προτάσεων. Επιπρόσθετα, πολλά από τα προσωπικά ιστολόγια ανήκουν σε

[2] http://twitter.com/

εφήβους και νέους, με αποτέλεσμα ένα ακόμα πιο ανεπίσημο περιβάλλον, όπου συζητούνται ανοιχτά θέματα που θεωρούνται προσωπικά. Ένα ακόμη χαρακτηριστικό της γλώσσας των ιστολογίων είναι το υποκειμενικό ύφος που αποτελεί πρόσφορο έδαφος για μελέτες ανάλυσης συναισθήματος.

Αυτά τα χαρακτηριστικά της σφαίρας των ιστολογίων έχουν έμμεσο αντίκτυπο σε μια σειρά εργασιών πρόσβασης πληροφοριών. Η ανεπίσημη, και πολλές φορές μη σύμφωνη με τη γραμματική, γλώσσα που χρησιμοποιείται στα ιστολόγια τα καθιστά δύσκολο πεδίο για πολλά NLP εργαλεία όπως συντακτικούς αναλυτές ή λογισμικά εξαγωγής των μερών του λόγου. Αυτό με τη σειρά του επηρεάζει τις εργασίες που ωφελούνται από αυτά τα εργαλεία, όπως οι αναλυτές συναισθήματος.

3.1.4 Αναγνώριση των χαρακτηριστικών του συγγραφέα ενός ιστολογίου

Η επιστημονική κοινότητα καταβάλλει όλο και μεγαλύτερη προσπάθεια στην ανάπτυξη υπολογιστικών μεθόδων που θα αποκαλύπτουν το πρόσωπο που βρίσκεται πίσω από το ιστολόγιο. Ο προσδιορισμός της προσωπικότητας του χρήστη μπορεί να είναι ιδιαίτερα επωφελής σε μελέτες που αφορούν στην ανάλυση συναισθήματος και στην εξόρυξη γνώμης. Αυτό συμβαίνει, επειδή οι άνθρωποι διαφέρουν στην προσωπικότητα και στο πώς εκτιμούν διάφορα γεγονότα, άρα και στο πόσο έντονα εκφράζουν την εκτίμηση ή την αποδοκιμασία τους για αυτά. Καθοριστικός παράγοντας για τον προσδιορισμό της προσωπικότητας είναι η γλώσσα. Οι λεξιλογικές επιλογές κάθε συντάκτη μπορεί να αντικατοπτρίζουν δομικές διαφορές στην προσωπικότητά τους. Πρέπει, ωστόσο, να λαμβάνεται υπόψη το γεγονός ότι η χρήση διαφορετικής γλώσσας μπορεί να σχετίζεται με τοπικούς ιδιωματισμούς και διαλέκτους, γεγονός το οποίο καθιστά δυσκολότερη και πολυπλοκότερη τη μελέτη για τον προσδιορισμό της προσωπικότητας.

3.2 Αναζήτηση απόψεων σε ιστολόγια

Καθώς αυξάνεται η ποσότητα των πληροφοριών στα ιστολόγια, χρειάζονται καλύτεροι μηχανισμοί κατάταξης, όμοιοι με αυτούς που χρησιμοποιούνται στις μηχανές αναζήτησης ιστού.Οι χρήστες δεν μπορούν πλέον να εξετάζουν όλα τα αποτελέσματα και ενδιαφέρονται πρώτα για τα πιο σχετικά. Η εργασία της ανάκτησης ιστολογίων που αξιολογήθηκε στο διαγωνισμό TREC έδωσε μια ουσιαστική ώθηση σε αυτό το πεδίο.

Η σφαίρα των ιστολογίων διαφοροποιείται από το υπόλοιπο διαδίκτυο με την έννοια ότι αποτελείται από δυναμικό και χρονικά προσδιορισμένο περιεχόμενο. Η ανάκτηση τέτοιου περιεχομένου μπορεί να βελτιστοποιηθεί με τη χρήση εξειδικευμένων τεχνικών ανάκτησης. Τέτοιες τεχνικές αξιοποιούν τη χρήση δομημένης πληροφορίας (RSS) που παρέχεται από τα ιστολόγια, σε συνδυασμό με το παραδοσιακό περιεχόμενό τους σε HTML μορφή. Έτσι, ενώ οι συνήθεις μηχανές αναζήτησης συνεχίζουν να προσπαθούν να ανακτήσουν και να δεικτοδοτήσουν τα ιστολόγια με βάση λέξεις-κλειδιά που εμφανίζονται αυτούσιες στο HTML περιεχόμενό τους, ερευνάται η ανάπτυξη μιας σειράς από εξειδικευμένες μηχανές αναζήτησης και ανάλυσης τέτοιου περιεχομένου, που θα αποκρίνονται ακριβέστερα στα ερωτήματα του χρήστη.

Στο πεδίο της αναζήτησης απόψεων από ιστολόγια αναφέρεται και τμήμα του διαγωνισμού TREC. Ο διαγωνισμός TREC για ιστολόγια ξεκίνησε το 2006 με στόχο να εξερευνήσει την εργασία της αναζήτησης πληροφοριών στη σφαίρα των ιστολογίων. Τα εγχειρήματα της αναζήτησης απόψεων, εύρεσης θετικών ή αρνητικών στάσεων και εύρεσης ιστολογίων ήταν καλές προσομοιώσεις ρεαλιστικών σεναρίων αναζήτησης χρήστη. Δόθηκε η δυνατότητα στους συμμετέχοντες να αξιολογήσουν τις τεχνικές και τις προσεγγίσεις που ακολούθησαν οι

συμμετέχοντες, προσφέροντας μια πληρέστερη κατανόηση αυτών των εργασιών. Ο διαγωνισμός συστημάτων TREC διοργανώνεται από το Εθνικό Ινστιτούτο Προτύπων και Τεχνολογίας (National Institute of Standards and Technology, NITS) και το Υπουργείο Άμυνας των Η.Π.Α. για να στηρίξει την έρευνα στον τομέα της ανάκτησης πληροφοριών. Από το 2006 διενεργείται ειδικός διαγωνισμός που αφορά τα ιστολόγια, με στόχο να εξερευνηθεί η συμπεριφορά αναζήτησης πληροφοριών στη σφαίρα των ιστολογίων.

3.2.1 Διαγωνισμός TREC 2006 Blog track

Στο διαγωνισμό TREC 2006 Blog track, στόχος ήταν η ανάπτυξη ενός συστήματος το οποίο θα εντοπίζει τις καταχωρήσεις ενός ιστολογίου που εκφράζουν κάποια άποψη σχετικά με ένα δεδομένο θέμα – στόχο. Ο στόχος αυτός μπορεί να είναι μία οντότητα, για παράδειγμα ένα άτομο, μία τοποθεσία ή ένας οργανισμός, αλλά και μία ιδέα, ένα προϊόν ή ένα γεγονός. Συνοπτικά, τα συστήματα που θα αναπτυσσόταν έπρεπε να απαντήσουν στο ερώτημα «*Τι σκέφτονται οι άνθρωποι για το X*», όπου X ο δεδομένος στόχος. Δεν ήταν απαραίτητο ο τίτλος της καταχώρησης του ιστολογίου να περιέχει το όνομα του στόχου, αλλά απαιτούνταν να εκφράζεται κάποια άποψη για το στόχο αυτό είτε στην καταχώρηση είτε σε κάποιο από τα σχόλια.

Για τους σκοπούς του διαγωνισμού αυτού έπρεπε να δημιουργηθεί ένα σύνολο δεδομένων ελέγχου, δηλαδή μία συλλογή από ιστολόγια που θα αποτελούσαν μία ρεαλιστική αντιπροσώπευση της σφαίρας των ιστολογίων. Η συλλογή αυτή θα έπρεπε να περιέχει αρκετό αριθμό ιστολογίων ώστε να υπάρχουν σε αυτή κάποιες αναγνωρίσιμες ιδιότητες της σφαίρας των ιστολογίων, να καλύπτουν αρκετό χρονικό διάστημα και να περιέχουν spam μηνύματα. Η συλλογή αυτή δημιουργήθηκε από το Πανεπιστήμιο της Γλασκώβης και της δόθηκε η ονομασία Blog06. Τα θέματα-στόχοι για την εργασία της ανάκτησης πληροφοριών, 50 στο σύνολο, επιλέχθηκαν από ένα σύνολο ερωτημάτων που έγιναν σε εμπορικές μηχανές αναζήτησης σε ιστολόγια. Τα θέματα που δημιουργήθηκαν είχαν τρία πεδία, τον τίτλο, με βάση τους όρους αναζήτησης που δίνονται, την περιγραφή, που βασίστηκε σε μία εκτίμηση για το τι αναζητούσαν οι χρήστες και την αφήγηση, όπου δινόταν μια περιγραφή σχετικά με το ποιες απόψεις θεωρούνταν σχετικές με την αναζήτηση.

Η αξιολόγηση των συστημάτων οργανώθηκε από το NIST και περιελάμβανε τη βαθμολόγηση της σχετικότητας κάθε εγγράφου με το θέμα – στόχο από έναν αξιολογητή. Η αξιολόγηση περιελάμβανε δύο επίπεδα: στο πρώτο επίπεδο καθοριζόταν αν το περιεχόμενο της καταχώρησης που επέστρεφε το σύστημα δεν εξεταζόταν λόγω του ότι ήταν διεύθυνση με ακατάλληλο περιεχόμενο (-1), η καταχώρηση του ιστολογίου εξεταζόταν αλλά δεν περιείχε πληροφορίες για το θέμα-στόχο (0) ή η καταχώρηση περιείχε πληροφορίες για το θέμα-στόχο αλλά δεν εξέφραζε κάποια άποψη για αυτό. Στο δεύτερο στάδιο, αν η καταχώρηση στο ιστολόγιο εξέφραζε κάποια άποψη ή συναισθήματα για το θέμα-στόχο, κατατασσόταν ανάλογα με το αν περιείχε αρνητικές απόψεις για το θέμα, αν περιείχε και θετικές και αρνητικές απόψεις ή αν περιείχε μόνο θετικά σχόλια για το θέμα-στόχο. Οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση της εργασίας ανάκτησης δεδομένων ήταν η μέση ακρίβεια (mean average precision, MAP), η R-ακρίβεια (R- Precision, R-Prec), η δυαδική προτίμηση (binary Preference, bPref) και η ακρίβεια στα 10 έγγραφα (Precision at 10 documents, P@10).

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα της ανάκτησης απόψεων ενός γύρου από τις 14 ομάδες με την καλύτερη μέση ακρίβεια (MAP), ταξινομημένα σύμφωνα με τη μέση ακρίβεια. Τα T, TD και TDN συμβολίζουν αν χρησιμοποιήθηκαν σε αυτό το γύρο από το διαγωνιζόμενο το πεδίο του τίτλου, το πεδίο του τίτλου και της περιγραφής, ή τα πεδία του τίτλου, της περιγραφής και της αφήγησης, αντίστοιχα.

Group	Run	Topics	MAP	R-prec	bPref	P@10
Indiana Univ.	woqln2	TDN	**0.2052**	0.2881	**0.2934**	0.4680
Indiana Univ.	wxoqf2	TDN	0.2019	**0.2934**	0.2824	0.4500
Univ. of Maryland	ParTiDesDmt2	TD	0.1887	0.2421	0.2573	0.3780
Univ. of Illinois at Chicago	uicst	T	0.1885	0.2771	0.2693	**0.5120**
Tsinghua Univ.	THUBLOGMF	T	0.1798	0.2647	0.2563	0.3600
Univ. of Amsterdam	UAmsB06All	T	0.1795	0.2771	0.2625	0.4640
CMU (Callan)	blog06r2	T	0.1576	0.2455	0.2458	0.3580
Univ. of California, Santa Cruz	ucscauto	T	0.1549	0.2355	0.2264	0.4380
Fudan Univ.	mcwil2knl	TDN	0.1179	0.1860	0.1920	0.2940
Univ. of Pisa	pisaBlDes	TD	0.0873	0.1765	0.1620	0.3400
Univ. of Maryland B.C.	UABas11	T	0.0764	0.1307	0.1202	0.2140
Univ. of Arkansas at Little Rock	UALR06a260r2	T	0.0715	0.1393	0.1357	0.3320
Chinese Academy of Sciences	IIIS	T	0.0621	0.1134	0.1553	0.2000
National Institute of Informatics	NII1	T	0.0466	0.1030	0.0851	0.3140
Robert Gordon Univ.	rguOPN	T	0.0000	0.0004	0.0003	0.0000

Σχήμα 7:Μετρικές Αξιολόγησης ενός γύρου των διαγωνιζομένων (TREC-2006)

Στον επόμενο πίνακα παρουσιάζονται τα αποτελέσματα του καλύτερου γύρου κάθε διαγωνιζόμενου αναφορικά με τη σχετικότητα του θέματος. Λήφθηκαν υπόψη τα έγγραφα που βαθμολογήθηκαν με 1 και άνω σύμφωνα με την κλίμακα αξιολόγησης που παρουσιάστηκε παραπάνω. Οι περισσότεροι συμμετέχοντες προσέγγισαν το εγχείρημα της ανάκτησης απόψεων σαν μια διαδικασία δύο σταδίων. Στο πρώτο στάδιο τα έγγραφα βαθμολογούνταν μόνο με βάση τη σχετικότητά τους με το θέμα, με τη χρήση έτοιμων (off-the-shelf) συστημάτων ανάκτησης και σταθμισμένων μοντέλων. Στο δεύτερο στάδιο τα αποτελέσματα αναβαθμολογούνταν ή φιλτράρονταν με τη χρήση μίας ή περισσότερων ευριστικών μεθόδων για την ανίχνευση απόψεων στα έγγραφα που ανακτήθηκαν από το πρώτο στάδιο. Οι τεχνικές που αναφέρθηκαν από τους συμμετέχοντες για την αναγνώριση απόψεων στο περιεχόμενο των εγγράφων περιλαμβάνουν προσεγγίσεις βασισμένες σε λεξικό και προσεγγίσεις ταξινόμησης κειμένου.

Group	Run	Topics	MAP	R-Prec	bPref	P@10
Indiana Univ.	wxoqf2	TDN	**0.2983**	**0.3925**	0.4225	0.6500
Indiana Univ.	woqln2	TDN	0.2963	0.3892	**0.4272**	0.6720
Tsinghua Univ.	THUBLOGMF	T	0.2959	0.3816	0.4177	0.6080
Univ. of Maryland	ParTitDesDef	TD	0.2849	0.3490	0.3998	0.6200
Univ. of Amsterdam	UAmsB06All	T	0.263	0.3674	0.3849	0.6940
Univ. of Illinois at Chicago	uicst	T	0.237	0.3315	0.3415	0.6860
CMU (Callan)	blog06r2	T	0.2324	0.3470	0.3599	0.5480
Univ. of Illinois at Chicago	uicsr	T	0.2267	0.3278	0.3410	**0.7060**
Univ. of California, Santa Cruz	ucscauto	T	0.2203	0.3047	0.3312	0.6480
Fudan Univ.	mcwil2knl	TDN	0.1668	0.2589	0.2826	0.4400
Univ. of Pisa	pisaBlDes	TD	0.1327	0.2329	0.2328	0.5880
Univ. of Maryland B.C.	UABas11	T	0.1288	0.1805	0.1911	0.4520
Univ. of Arkansas at Little Rock	UALR06a500r4	T	0.1192	0.1950	0.1966	0.5180
Chinese Academy of Sciences	IIIS	T	0.1071	0.1903	0.2673	0.3400
National Institute of Informatics	NII1	T	0.0834	0.1522	0.1345	0.5640
Robert Gordon Univ.	rguOPN	T	0.0001	0.0010	0.0010	0.0060

Σχήμα 8:Μετρικές Αξιολόγησης του καλύτερου γύρου των διαγωνιζομένων (TREC-2006)

Στις προσεγγίσεις που βασίζονται σε λεξικό χρησιμοποιούνταν μία λίστα (με τον σημασιολογικό προσανατολισμό όρων) για να βαθμολογηθούν τα έγγραφα, ανάλογα με τη συχνότητα εμφάνισης των όρων του λεξικού σε αυτά. Σε κάποιες περιπτώσεις συνδυαζόταν

με πληροφορίες για την απόσταση μεταξύ των όρων αυτών και των λέξεων του ερωτήματος-στόχου μέσα στο έγγραφο. Οι αναφορές για την επιτυχία αυτής της προσέγγισης ποικίλλουν. Κάποιες ομάδες ανέφεραν υποβάθμιση των αποτελεσμάτων σε σχέση με τα βασικά αποτελέσματα ανάκτησης, ενώ άλλες ομάδες παρατήρησαν βελτίωση.

Στις προσεγγίσεις ταξινόμησης κειμένου, ένας ταξινομητής εκπαιδευόταν με δεδομένα που περιείχαν απόψεις για κάποιο θέμα (όπως ιστοσελίδες με κριτικές προϊόντων) και δεδομένα που δεν περιείχαν απόψεις (όπως δικτυακές εγκυκλοπαίδειες) και στη συνέχεια χρησιμοποιούνταν για να εκτιμήσει αν εκφραζόταν κάποια άποψη στο περιεχόμενο των ανακτημένων εγγράφων. Οι περισσότερες ομάδες που χρησιμοποίησαν αυτή την προσέγγιση προτίμησαν το αλγόριθμο SVM για την ταξινόμηση, αν και χρησιμοποιήθηκαν και άλλοι ταξινομητές. Η επιτυχία αυτής της μεθόδου ήταν περιορισμένη, πιθανόν εξαιτίας της διαφοράς μεταξύ των δεδομένων εκπαίδευσης και του περιεχομένου των ιστολογίων.

3.2.2 Διαγωνισμός TREC 2007 Blog track

Ο διαγωνισμός TREC για τα ιστολόγια συνεχίστηκε και την επόμενη χρονιά, το 2007, με την προσθήκη ενός νέου εγχειρήματος και ενός υποεγχειρήματος. Όμοια με το διαγωνισμό του 2006, υπήρχε ο στόχος της ανάκτησης απόψεων που αφορούσε τον εντοπισμό των καταχωρήσεων blog που εκφράζουν άποψη σχετικά με κάποιο δεδομένο θέμα – στόχο. Το νέο υποεγχείρημα που προστέθηκε ήταν η εύρεση θετικής ή αρνητικής στάσης σχετικά με το θέμα – στόχο σε μία καταχώρηση ιστολογίου, ενώ το δεύτερο εγχείρημα αφορούσε την εύρεση feed σχετικά με κάποιο θέμα. Το σύνολο δεδομένων που χρησιμοποιήθηκε ήταν η συλλογή Blog06 από το διαγωνισμό της προηγούμενης χρονιάς. Αναφορικά με το εγχείρημα της εύρεσης απόψεων για κάποιο θέμα – στόχο, η διαδικασία που ακολουθήθηκε, η κλίμακα αξιολόγησης καθώς και οι μετρικές αξιολόγησης ήταν ίδιες με αυτές του διαγωνισμού της προηγούμενης χρονιάς. Παρατηρήθηκε ότι η απόδοση των συστημάτων στην εργασία της ανάκτησης εγγράφων στο διαγωνισμό TREC 2007 ήταν υψηλότερη απ' ότι στο διαγωνισμό της προηγούμενης χρονιάς. Απαιτείται περισσότερη έρευνα για να διαπιστωθεί αν αυτό οφείλεται στο ότι τα θέματα αναζήτησης που χρησιμοποιήθηκαν ήταν ευκολότερα ή αν οι διαγωνιζόμενοι χρησιμοποίησαν πιο αποδοτικές μεθόδους ανάκτησης.

Group	Run	Automatic	Fields	MAP	R-prec	b-Bref	P@10
UIC (Zhang)	uic1c	yes	T	**0.4341**	**0.4529**	**0.4724**	**0.690**
UAmsterdam (deRijke)	uams07topic	yes	T	0.3453	0.3872	0.3953	0.562
IndianaU (Yang)	oqlr2fopt	yes	TDN	0.3350	0.3925	0.378	0.576
UGlasgow (Ounis)	uogBOPFProxW	yes	T	0.3264	0.3657	0.3497	0.552
DalianU (Yang)	DUTRun2	yes	T	0.3190	0.3671	0.3686	0.600
FudanU (Wu)	FDUTisdOpSVM	yes	T	0.3179	0.3467	0.3501	0.454
FIU (Netlab team)	FIUDDPH	yes	TD	0.3053	0.3498	0.3475	0.492
UNeuchatel (Savoy)	UniNEblog3	yes	TD	0.3049	0.3438	0.3266	0.516
CAS (Liu)	Relevant	yes	T	0.3041	0.3600	0.3779	0.446
UArkansas Littlerock (Bayrak)	UALR07BlogIU	yes	T	0.2911	0.3263	0.3134	0.580
UWaterloo (Olga)	UWopinion3	yes	T	0.2631	0.3344	0.298	0.496
CAS (NLPR-IACAS)	NLPRPTD2	yes	TD	0.2587	0.3088	0.2956	0.456
Zhejiangu (Qiu)	EAGLE1	yes	T	0.2561	0.3159	0.2867	0.428
BUPT (Weiran)	prisOpnBasic	yes	T	0.2466	0.3018	0.2835	0.456
KobeU (Eguchi)	KobePrMIR01	yes	T	0.2460	0.3011	0.2744	0.440
NTU (Chen)	NTUManualOp	no	T	0.2393	0.2659	0.2749	0.486
KobeU (Seki)	Ku	yes	T	0.1689	0.2417	0.219	0.254
RGU (Mukras)	rgu0	yes	T	0.1686	0.2266	0.2163	0.288
UBuffalo (Ruiz)	UB1	yes	TDN	0.1501	0.2001	0.1887	0.266
Wuhan (Lu)	NOOPWHU1	yes	T	0.0011	0.0071	0.0072	0.008

Σχήμα 9:Μετρικές Αξιολόγησης του καλύτερου γύρου των διαγωνιζομένων (TREC-2007)

Στον παραπάνω πίνακα παρουσιάζονται οι μετρικές αξιολόγησης του καλύτερου γύρου κάθε ομάδας που συμμετείχε, για το εγχείρημα της εύρεσης άποψης.

Το νέο υποεγχείρημα που προστέθηκε στο διαγωνισμό TREC 2007 ήταν η εύρεση θετικής ή αρνητικής στάσης για το θέμα-στόχο στο ανακτημένο έγγραφο, δηλαδή αν η άποψη που εκφραζόταν ήταν θετική, αρνητική ή ένας συνδυασμός και των δύο. Η μετρική που χρησιμοποιήθηκε για την αξιολόγηση των αποτελεσμάτων ήταν η R – accuracy που αντιπροσωπεύει το ποσοστό των ανακτημένων εγγράφων πάνω από το βαθμό R που ταξινομήθηκαν σωστά, όπου R ο αριθμός των εγγράφων που εκφράζουν άποψη για το συγκεκριμένο θέμα – στόχο. Ο παρακάτω πίνακας παρουσιάζει τα αποτελέσματα του καλύτερου γύρου για κάθε διαγωνιζόμενο με βάση την R- accuracy.

Group	Run	Fields	R-Acc	A@10	A@1000
UIC (Zhang)	uic75cpnm	T	**0.2295**	**0.3700**	**0.0493**
IndianaU (Yang)	oqlr2f2optP	TDN	0.1941	0.3080	0.0438
UAmsterdam (de Rijke)	uams07ipolt	T	0.1827	0.2640	0.0418
DalianU (Yang)	DUTRun2P	T	0.1721	0.3080	0.0406
Zhejiangu (Qiu)	EAGLE2P	T	0.1510	0.2380	0.0427
UGlasgow (Ounis)	uogBOPFPol	T	0.1460	0.2020	0.0397
NTU (Chen)	NTUManualOpP	T	0.1161	0.2300	0.0348
CAS (Liu)	DrapStmSub	T	0.0818	0.1060	0.0243
BUPT (Weiran)	prisPolC2	T	0.0726	0.2020	0.0124
UBuffalo (Ruiz)	pUB11	TDN	0.0671	0.1000	0.0195
Wuhan (Lu)	OTPSWHU102	T	0.0032	0.0040	0.0010

Σχήμα 10:Αποτελέσματα καλύτερου γύρου για κάθε διαγωνιζόμενο με βάση το R-accuracy (TREC-2007)

Οι συμμετέχοντες χρησιμοποίησαν διάφορες τεχνικές στις προσεγγίσεις τους κατά την ανάπτυξη των συστημάτων. Παρουσιάζουμε στη συνέχεια αυτές που βελτίωσαν τα αποτελέσματα της εύρεσης εγγράφων σχετικών με το θέμα αναζήτησης.

Ευρετηριοποίηση (Indexing)

Όλοι οι διαγωνιζόμενοι χρησιμοποίησαν ως δείκτη το στοιχείο του μόνιμου συνδέσμου (Permalink) προς μία καταχώρηση blog, με εξαίρεση την ομάδα του Πανεπιστημίου Waterloo που χρησιμοποίησε και τα τρία στοιχεία της συλλογής δεδομένων, το μόνιμο σύνδεσμο, τα feeds και την αρχική σελίδα (Homepage).

Ανάκτηση

Όπως και στο διαγωνισμό του 2006, οι περισσότερες ομάδες ακολούθησαν μία προσέγγιση δύο σταδίων για την ανάκτηση εγγράφων. Στο πρώτο στάδιο τα έγγραφα βαθμολογούνταν με βάση κάποια σταθμισμένα μοντέλα και στο δεύτερο στάδιο αναβαθμολογούνταν λαμβάνοντας υπόψη χαρακτηριστικά της εύρεσης απόψεων.

Χαρακτηριστικά Εύρεσης Απόψεων

Χρησιμοποιήθηκαν κυρίως δύο αποτελεσματικές τεχνικές για την εύρεση εγγράφων που περιείχαν απόψεις. Η πρώτη βασίστηκε στην αυτόματη δημιουργία ενός σταθμισμένου λεξικού με βάση την αξιολόγηση των αποτελεσμάτων σχετικότητας στο εγχείρημα εύρεσης

απόψεων του διαγωνισμού TREC 2006. Η βαρύτητα του κάθε όρου εκτιμά κατά πόσο αυτός εκφράζει κάποια άποψη. Στη συνέχεια το λεξικό αυτό δόθηκε σαν ερώτημα ώστε να δώσει μία εκτίμηση για τη θετική ή αρνητική στάση κάθε εγγράφου της συλλογής. Η δεύτερη προσέγγιση βασίστηκε σε μία λίστα όρων που εκφράζουν υποκειμενική άποψη. Η αναβαθμολόγηση του εγγράφου γινόταν ανάλογα με την εγγύτητα των όρων του ερωτήματος με τους όρους αυτής της λίστας.

Το δεύτερο και καινούριο εγχείρημα-στόχος που προστέθηκε στο TREC 2007 Blog track ήταν η αναζήτηση feed (Blog distillation – feed search). Συχνά οι χρήστες κάνουν αναζήτηση σε ιστολόγια που ασχολούνται με κάποιο συγκεκριμένο θέμα ώστε να εγγραφούν σε αυτά και να τα διαβάζουν σε τακτική βάση. Το ερώτημα στο οποίο θα απαντούσαν τα συστήματα που θα αναπτυσσόταν ήταν το εξής: «*Βρες ένα ιστολόγιο στο οποίο το Χ είναι βασικό και επαναλαμβανόμενο θέμα*». Για ένα δεδομένο θέμα-στόχο Χ, τα συστήματα θα έπρεπε να προτείνουν feeds τα οποία ασχολούνται συστηματικά με το Χ και πιθανότατα θα ενδιέφεραν το χρήστη ώστε να τα προσθέσει στο δικό του πρόγραμμα ανάγνωσης feed. Το σύνολο δεδομένων που χρησιμοποιήθηκε ήταν η συλλογή Blog06 του προηγούμενου διαγωνισμού, ενώ τα θέματα αναζήτησης προτάθηκαν από τους συμμετέχοντες. Για την εκτίμηση των αποτελεσμάτων, ζητήθηκε από τους αξιολογητές να διαβάσουν κάποια έγγραφα του feed και στη συνέχεια να κρίνουν αν το blog ασχολείται συστηματικά και κατά βάση με το ζητούμενο θέμα-στόχο. Οι μετρικές που χρησιμοποιήθηκαν ήταν η μέση ακρίβεια (mean average precision, MAP), η R- Precision, η δυαδική προτίμηση (binary Preference, bPref) και η ακρίβεια στα 10 έγγραφα (Precision at 10 documents, P@10). Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα του καλύτερου γύρου για κάθε ομάδα.

Group	Run	Fields	MAP	R-prec	b-Bref	P@10	MRR
CMU (Callan)	CMUfeedW	T	**0.3695**	**0.4245**	**0.3861**	**0.5356**	0.7537
UGlasgow (Ounis)	uogBDFeMNZP	T	0.2923	0.3654	0.3210	0.5311	0.7834
UMass (Allen)	UMaTDPCSwGR	TD	0.2741	0.3356	0.3027	**0.5356**	**0.8407**
KobeU (Seki)	kudsn	T	0.2420	0.3148	0.2714	0.4622	0.7605
DalianU (Yang)	DUTDRun4	TDN	0.2399	0.3126	0.2740	0.4378	0.7337
UTexas-Austin (Efron)	utblnrr	T	0.2197	0.3100	0.2649	0.4511	0.7245
UAmsterdam (deRijke)	uams07bdtblm	T	0.1605	0.2346	0.1820	0.3067	0.6320
UBerlin (Neubauer)	ADABoostM1	TDN	0.0176	0.0468	0.0330	0.0978	0.2881
WuhanU (Lu)	TDWHU200	T	0.0135	0.0419	0.0297	0.0578	0.1386

Σχέδιο 11: Αποτελέσματα του καλύτερου γύρου για κάθε διαγωνιζόμενο (TREC 2007)

Παρατηρήθηκε ότι συστήματα τα οποία μπορούσαν να εντοπίσουν και να αφαιρέσουν τα spam ιστολόγια είχαν καλύτερη απόδοση ανάκτησης. Οι συμμετέχοντες χρησιμοποίησαν διάφορες τεχνικές δεικτοδότησης και ανάκτησης για το συγκεκριμένο εγχείρημα. Αναφορικά με τη δεικτοδότηση, χρησιμοποιήθηκαν δύο είδη δεικτών. Κάποιες ομάδες δημιούργησαν δείκτη με βάση το στοιχείο Feeds της συλλογής Blog06 και κάποιες άλλες το στοιχείο Permalinks (μόνιμο σύνδεσμο). Η ομάδα που είχε την καλύτερη απόδοση πειραματίστηκε και με τους δύο τύπους δεικτών και ανέφερε ότι η δεικτοδότηση με βάση το στοιχείο Feeds είχε ως αποτέλεσμα μεγαλύτερη απόδοση στην ανάκτηση εγγράφων.

Για την ανάκτηση δεδομένων, κάποιες ομάδες συνέδεσαν αυτό το εγχείρημα με άλλα υπάρχοντα εγχειρήματα αναζήτησης. Για παράδειγμα, το Πανεπιστήμιο της Μασαχουσέτης αντιμετώπισε το στόχο της εύρεσης ιστολογίου σαν ένα πρόβλημα κατανεμημένης αναζήτησης. Οι περισσότερες ομάδες που χρησιμοποίησαν ευρετήριο στο στοιχείο Permalink, πρότειναν διάφορες τεχνικές για το άθροισμα των αποτελεσμάτων των καταχωρήσεων ιστολογίου σε ένα συνολικό αποτέλεσμα για το feed. Σε σχέση με τον αντίστοιχο διαγωνισμό του 2006, η απόδοση των συστημάτων για το πρώτο εγχείρημα, του

εντοπισμού απόψεων, ήταν καλύτερη στο TREC 2007. Κάποιοι διαγωνιζόμενοι πρότειναν νέες τεχνικές ανίχνευσης απόψεων που βελτίωσαν τα αποτελέσματά τους στην εύρεση εγγράφων σχετικών με το θέμα αναζήτησης. Οι αποδόσεις στο υποεγχείρημα της εύρεσης θετικής ή αρνητικής στάσης δεν ήταν αρκετά υψηλές, κάτι που δείχνει ότι το θέμα αυτό είναι ακόμη ανοιχτό πρόβλημα που απαιτεί περαιτέρω έρευνα. Το δεύτερο εγχείρημα είχε ως αποτέλεσμα την πρόταση κάποιων υποσχόμενων τεχνικών ανάκτησης.

3.2.3 Διαγωνισμός TREC 2008 Blog track

Στο διαγωνισμό που διοργανώθηκε το 2008 συνέχισαν να ερευνώνται τα εγχειρήματα της ανάκτησης απόψεων, εύρεσης θετικής ή αρνητικής στάσης και εύρεσης feed με τη χρήση της συλλογής Blog06. Επιπρόσθετα, υπήρξε ένα νέο εγχείρημα εύρεσης καταχωρήσεων σε ιστολόγια σχετικών με ένα ζητούμενο θέμα.

Οι διαγωνισμοί των δύο προηγούμενων ετών δείξανε ότι η απόδοση της ανάκτησης καταχωρήσεων που εκφέρουν άποψη για κάποιο θέμα εξαρτάται σε μεγάλο βαθμό από την απόδοση της εύρεσης εγγράφων σχετικών με αυτό το θέμα. Οι περισσότεροι συμμετέχοντες στους διαγωνισμούς των δύο προηγούμενων ετών χρησιμοποίησαν, όπως είδαμε, μία προσέγγιση δύο σταδίων για την ανάκτηση απόψεων: στο πρώτο στάδιο γινόταν η ανάκτηση εγγράφων σχετικά με το ζητούμενο θέμα και στο δεύτερο η αξιολόγηση του κατά πόσο το έγγραφο εκφράζει κάποια άποψη. Για να ερευνηθεί η σχέση αυτών των σταδίων προστέθηκε το τελευταίο εγχείρημα, ώστε να αξιολογηθούν οι τεχνικές που χρησιμοποιούνται για την ανάκτηση δεδομένων σχετικών με κάποιο θέμα. Η διαδικασία που ακολουθήθηκε, ο τρόπος αξιολόγησης και οι μετρικές ήταν όμοια με αυτά των προηγούμενων διαγωνισμών TREC. Στον παρακάτω πίνακα παρουσιάζονται οι μετρικές για τον καλύτερο γύρο κάθε ομάδας, στα εγχειρήματα της εύρεσης καταχωρήσεων σχετικών με κάποιο θέμα καθώς και της εύρεσης απόψεων.

			Topic-Relevance					Opinion-Finding				
Group	Run	Fields	MAP	R-prec	bPref	P@10	MRR	MAP	R-prec	bPref	P@10	MRR
KLE	KLEPsgFeedT	T	0.4954	0.5150	0.5364	0.7920	0.9058	0.4052	0.4366	0.4314	0.6440	0.8184
UAms_De_Rijke	uams08n1o1	T	0.4644	0.4867	0.5034	0.7620	0.8892	0.3797	0.4176	0.4117	0.6620	0.8052
UIC_IR_Group	uicirnoa	T	0.4403	0.4804	0.5062	0.7700	0.8667	0.3438	0.3956	0.3929	0.5880	0.7480
UniNE	UniNEBlog1	TD	0.4344	0.4608	0.4662	0.6440	0.8199	0.3565	0.3887	0.3677	0.5540	0.7605
UoGtr	uogBLProxCE	T	0.4219	0.4548	0.4481	0.7060	0.8228	0.3531	0.3840	0.3646	0.6100	0.7723
THUIR	THUrelTwpmf	T	0.4067	0.4565	0.4625	0.6940	0.8263	0.3313	0.3942	0.3749	0.5900	0.7487
BUPT_pris_	prisba	T	0.4065	0.4506	0.4561	0.6780	0.8290	0.3346	0.3876	0.3684	0.5580	0.7456
DUTIR	DUT08BRun1	T	0.3617	0.4188	0.4345	0.6540	0.7633	0.2974	0.3586	0.3598	0.5420	0.7204
iitkgp	IITKGPNOSPAM	T	0.3598	0.4090	0.4394	0.7400	0.8817	0.2988	0.3664	0.3642	0.5720	0.7955
IU-SLIS	wdoqsBase	T	0.3431	0.3918	0.4001	0.7280	0.8636	0.2818	0.3367	0.3215	0.5900	0.7551
UWaterlooEng	UWBase2	T	0.3309	0.3824	0.3875	0.6380	0.8127	0.2753	0.3391	0.3249	0.5160	0.7254
aic-dcu	DCUCDVPtbl	T	0.3303	0.3671	0.3601	0.6520	0.7783	0.2875	0.3280	0.3089	0.5560	0.7066
UTD_SLP_Lab	SplBaseTD	TD	0.3298	0.3751	0.3787	0.6380	0.7423	0.2682	0.3305	0.3133	0.5200	0.6618
UIUC	UIUCb08uwTtl	T	0.3240	0.3766	0.3771	0.6800	0.8223	0.2723	0.3336	0.3133	0.5540	0.7777
fub	FIUbasePL2c9	T	0.3199	0.3738	0.3601	0.6120	0.7351	0.2659	0.3206	0.2915	0.5020	0.6862
KobeU-Seki	ku	T	0.3035	0.3602	0.3531	0.5820	0.7053	0.2475	0.3051	0.2806	0.4960	0.6585
KU	kunlpKLtt	T	0.2791	0.3568	0.3487	0.5700	0.7784	0.2263	0.3042	0.2815	0.4520	0.6955
USI	run0	T	0.2567	0.3363	0.3289	0.4020	0.5472	0.2048	0.2604	0.2523	0.3060	0.4605
feup_irlab	feupB	T	0.2518	0.3190	0.3243	0.5800	0.7133	0.2006	0.2660	0.2573	0.4360	0.5745
york	york08bb2	T	0.2074	0.2923	0.2863	0.5540	0.7954	0.1700	0.2489	0.2343	0.4520	0.7308

Σχήμα 12: Μετρικές Αξιολόγησης του καλύτερου γύρου των διαγωνιζομένων (TREC 2008)

Οι αποδόσεις των συστημάτων τόσο στην ανάκτηση εγγράφων όσο και στην εύρεση απόψεων στο διαγωνισμό του 2006 ήταν σημαντικά χαμηλότερες από ότι στους διαγωνισμούς του 2007 και 2008, κάτι που πιθανόν να οφείλεται στο γεγονός ότι τα θέματα που χρησιμοποιήθηκαν στον TREC 2006 ήταν πιο δύσκολα. Οι πιο αποτελεσματικές προσεγγίσεις που χρησιμοποιήθηκαν στο εγχείρημα της εύρεσης απόψεων ήταν οι εξής:

- Η χρήση ενός SVM ταξινομητή για το διαχωρισμό υποκειμενικών και αντικειμενικών κειμένων ώστε να καθοριστεί αν το κείμενο που εκφέρει υποκειμενική άποψη είναι σχετικό με το θέμα.
- Η χρήση ενός λεξικού όρων για να καθοριστεί αν μια καταχώρηση ιστολογίου εκφράζει άποψη για το θέμα – στόχο, υπολογιζόντας το άθροισμα των βαρών των αντίστοιχων όρων που υπήρχαν στην καταχώρηση.

Ένα από τα συμπεράσματα του TREC 2007 Blog track, ήταν ότι το εγχείρημα της ανίχνευσης θετικής ή αρνητικής στάσης θα πρέπει να αποτελεί αναπόσπαστο κομμάτι της διαδικασίας εύρεσης άποψης. Αντί για εγχείρημα ταξινόμησης, όπου το σύστημα αναγνωρίζει τι είδους άποψη εκφράζει το έγγραφο, στον διαγωνισμό του 2008 ορίστηκε ξανά ώστε να προσομοιώνει ένα σενάριο αναζήτησης χρήστη, κατά το οποίο το σύστημα θα ανακτά τα έγγραφα που εκφράζουν είτε θετική είτε αρνητική άποψη και θα τα εμφανίζει στο χρήστη κατηγοριοποιημένα.

Στον παρακάτω πίνακα βλέπουμε τα αποτελέσματα του καλύτερου γύρου για κάθε ομάδα. Για να υπολογιστεί η συνολική απόδοση κάθε συστήματος σε αυτό το εγχείρημα, υπολογίστηκε ο μέσος όρος της μέσης ακρίβειας (MAP) για την εύρεση θετικών στάσεων και της μέσης ακρίβειας για την εύρεση αρνητικών τάσεων (τον συμβολίζουμε με Mix MAP).

Group	Run	Fields	Baseline	Mix			Positive			Negative		
				MAP	Δ MAP	P@10	MAP	Δ MAP	P@10	MAP	Δ MAP	P@10
IU-SLIS	top3dt1mP5	T	N/A	0.1677	N/A	0.2170	0.1752	N/A	0.2040	0.1601	N/A	0.2300
KLE	KLEPolarity	T	N/A	0.1662	N/A	0.2020	0.1828	N/A	0.2360	0.1496	N/A	0.1680
aic-dcu	DCUCDVPgpo	TD	baseline4	0.1547	9.70%	0.1900	0.1612	5.22%	0.2000	0.1483	15.14%	0.1800
KobeU-Seki	kup4	T	baseline4	0.1448	2.68%	0.1820	0.1566	2.22%	0.1980	0.1329	3.18%	0.1660
THUIR	THUpolTmfPNR	T	THUrelTwpmf	0.1353	7.16%	0.1870	0.1289	6.27%	0.1880	0.1417	7.92%	0.1860
UoGtr	uogPL41	T	baseline4	0.1348	-4.41%	0.1640	0.1394	-9.01%	0.1700	0.1301	1.01%	0.1580
UWaterlooEng	UWnb1Pol	T	baseline1	0.1278	0.71%	0.1780	0.1430	4.84%	0.2040	0.1126	-4.17%	0.1520
iitkgp	KGPPOL1	T	IITKGPTITLE1	0.1139	-6.15%	0.1990	0.1304	-1.95%	0.2300	0.0975	-11.12%	0.1680
UTD_SLP_Lab	NTrMM47P	TD	baseline4	0.1129	-19.94%	0.2130	0.1323	-13.64%	0.2220	0.0934	-27.48%	0.2040
UIC_IR_Group	uicpolrun1	T	N/A	0.1099	N/A	0.2400	0.1594	N/A	0.3000	0.0604	N/A	0.1800
UniNE	UniNEpolLR1	TD	UniNEBlog1	0.0775	-41.33%	0.1780	0.0882	-35.90%	0.2000	0.0667	-47.31%	0.1560
fub	FIUpBL3DFR	T	baseline3	0.0723	-45.26%	0.1610	0.0618	-55.09%	0.1760	0.0828	-34.60%	0.1460
SUNY_Buffalo	UBpol1	T	N/A	0.0661	N/A	0.0820	0.0752	N/A	0.1080	0.0570	N/A	0.0560
tno	tnobase1	D	baseline1	0.0449	-64.62%	0.0990	0.0544	-60.12%	0.1360	0.0353	-69.96%	0.0620
KU	kunlpKLttPs	T	kunlpKLtt	0.0416	-54.39%	0.1560	0.0542	-38.34%	0.1900	0.0291	-69.21%	0.1220
DUTIR	DUTIR08Run2P	T	DUT08BRun2	0.0301	-73.43%	0.1500	0.0352	-72.28%	0.1840	0.0250	-74.87%	0.1160

Σχήμα 13: Μετρικές Αξιολόγησης του καλύτερου γύρου των διαγωνιζομένων με Mix MAP (TREC-2008)

Με Δ MAP συμβολίζεται η διαφορά του Mix MAP του συγκεκριμένου γύρου και του Mix MAP στο στάδιο της ανάκτησης σχετικών εγγράφων. Μία σχετική αύξηση στην απόδοση υποδεικνύει ότι οι χρησιμοποιούμενες τεχνικές ανίχνευσης θετικής ή αρνητικής στάσης υπήρξαν χρήσιμες. Ωστόσο, στις περισσότερες περιπτώσεις παρατηρείται μία σχετική μείωση στην απόδοση, κάτι που δείχνει ότι οι τεχνικές που χρησιμοποιήθηκαν από τις περισσότερες ομάδες δεν ήταν αποτελεσματικές.

Το εγχείρημα της εύρεσης ιστολογίων, που εξετάστηκε πρώτη φορά στο διαγωνισμό του 2007, ασχολείται με ένα σενάριο αναζήτησης όπου ο χρήστης έχει ως στόχο να βρει ένα blog και να το προσθέσει στο πρόγραμμα ανάγνωσης feed. Η βαθμολόγηση έγινε από τους αξιολογητές με βάση την εξής κλίμακα: Spam (Το έγγραφο αποτελεί spam blog), Μη σχετικό (Δε θα έκανα εγγραφή σε αυτό το feed), Σχετικό (Το blog περιέχει αρκετές καταχωρήσεις σχετικές με το θέμα και πιθανόν να έκανα εγγραφή), Πολύ Σχετικό (Θα έκανα εγγραφή σε αυτό το ιστολόγιο).

Group	Run	Topic	nDCG	MAP	R-prec	bPref	P@10	MRR	MAP(2)
KLE	KLEDistFBB	TD	0.5443	0.2994	0.3508	0.3224	0.4560	0.7458	0.2852
CMU-LTI-DIR	cmuLDwikiSP	T	0.5170	0.3056	0.3646	0.3535	0.4340	0.8051	0.2750
uMass	UMassBlog3	TD	0.4969	0.2711	0.3286	0.3117	0.4240	0.7612	0.2772
UAms_De_Rijke	uams08bl	T	0.4904	0.2638	0.3137	0.3024	0.4200	0.7294	0.2547
SUNY_Buffalo	UBDist4	TDN	0.4824	0.2633	0.3160	0.3088	0.3820	0.7293	0.2449
UoGtr	uogTrBDfeNWD	T	0.4758	0.2521	0.3121	0.2932	0.4040	0.7425	0.2452
KobeU-Seki	kudb	T	0.4712	0.2422	0.2947	0.2903	0.3440	0.7469	0.2398
USI	BM25LenNorm	T	0.4663	0.2566	0.3144	0.2882	0.3960	0.7016	0.2282
WHU	PermMeWhu	T	0.4023	0.1898	0.2591	0.2451	0.3180	0.5554	0.1827
iitkgp	FEEDKGP1	TD	0.3613	0.1720	0.2484	0.2129	0.3220	0.5077	0.1826
feup_irlab	feupbase	T	0.3478	0.1413	0.1890	0.1690	0.2560	0.5970	0.1621
DUTIR	DUTIR08DRun4	TDN	0.3394	0.1632	0.2365	0.2059	0.2780	0.4298	0.1359

Σχήμα 14: Μετρικές Αξιολόγησης του καλύτερου γύρου των διαγωνιζομένων με nDCG (TREC-2008)

4 Εξόρυξη γνώμης σε συστήματα διαβούλευσης πολιτικής

Σε αυτό το κεφάλαιο διερευνούμε την εξόρυξη κειμένου και την εφαρμογή τεχνικών μηχανικής μάθησης για την καταγραφή της γνώμης του κοινού που αφορά ζητήματα πολιτικής. Η ανάπτυξη νέων μορφών διαδικτυακής επικοινωνίας, κυρίως μέσα από τα εργαλεία του Web 2.0 (ιστολόγια, σελίδες κοινωνικής δικτύωσης) αλλά και η μαζική χρήση αυτών των εργαλείων ειδικά απο τις νεότερες γενιές, έχουν δημιουργήσει νέες ευκαιρίες δημόσιας διαβούλευσης. Τα συστήματα διαβούλευσης πολιτικής αποτελούν μια προσπάθεια βελτίωσης της διεπαφής μεταξύ του πολιτικού συστήματος και της κοινωνίας, προκειμένου να εξασφαλιστεί καλύτερη επικοινωνία και συμμετοχή της δεύτερης στη φάση της διαμόρφωσης πολιτικής. Τα συστήματα αυτά επιτρέπουν στους κυβερνώντες να εκμεταλλευτούν σε επόμενες δράσεις τους τις απόψεις που εκφράστηκαν και γενικά να βελτιώσουν την κατανόηση των θεμάτων που συζητήθηκαν.

4.1 Εξόρυξη γνώμης στην πολιτική

Οι αποφάσεις των κυβερνήσεων είναι από τα θέματα που συζητούνται περισσότερο στις δικτυακές κοινότητες (forum, ιστολόγια και άλλες). Αυτό δεν οφείλεται μόνο στο ότι η νέα γενιά χρηστών του διαδικτύου έχει μεγαλύτερες δυνατότητες να μοιράζεται πληροφορίες και να συνεργάζεται δικτυακά αλλά και στο ότι οι κυβερνήσεις σε όλον τον κόσμο δημοσιεύουν μεγάλο μέρος των αποφάσεων και των νομοθετημάτων τους στο διαδίκτυο. Είναι προφανές ότι οι γνώμες των πολιτών έχουν μεγάλη σημασία στην πολιτική και οι νόμοι και οι νομοθεσίες μπορούν να αλλάξουν κάτω από την (αρνητική) πίεση της κοινής γνώμης. Οι πολιτικοί προσπαθούν να κατανοήσουν τι σκέφτονται οι ψηφοφόροι για εκκρεμή θέματα πολιτικής και για τις προτάσεις της κυβέρνησης. Ο κοινωνικός αντίκτυπος των προτάσεων τους θα πρέπει να καταγραφεί και να αναλυθεί περαιτέρω από τα συστήματα αυτά.

Για τους σκοπούς της συλλογής των απόψεων των πολιτών έχουν δημιουργηθεί κυβερνητικά κοινωνικά δίκτυα. Τα δίκτυα αυτά στοχεύουν στο να συλλέγουν τις απόψεις των πολιτών ώστε αυτές στη συνέχεια να λαμβάνονται υπόψιν στην διαδικασία λήψης κυβερνητικών αποφάσεων. Αρχικές προσεγγίσεις ήθελαν μια κεντρική πλατφόρμα, ελεγχόμενη από δημόσιους φορείς να αποτελεί το «θεσμοθετημένο» χώρο (αλλά και μονοπώλιο) της δημόσιας διαβούλευσης. Πιο ανοιχτές προσεγγίσεις συνειδητοποιούν ότι δημόσια διαβούλευση επί των πολιτικών συμβαίνει στην πραγματικότητα καθημερινά και συνεχώς, σε χιλιάδες τοποθεσίες στο διαδίκτυο, σε προσωπικα blogs, σε mailing lists, forums, ιστοσελίδες που μπορούν να ανήκουν σε μη κυβερνητικές οργανώσεις ή και σε απλούς πολίτες. Ζητούμενο λοιπόν δεν είναι μόνο η δημιουργία κεντρικά ελεγχόμενων πλατφορμών με αυξημένες λειτουργίες υποστήριξης της δημόσιας διαβούλευσης, αλλά εν γένει η βελτίωση της ικανότητας του πολιτικού συστήματος να «αισθάνεται» την άποψη της κοινωνίας, όπως αυτή εκφράζεται ελεύθερα και χωρίς κανέναν ενδοιασμό στο διαδίκτυο. Στο πλαίσιο αυτό, η δικτυακή διαβούλευση έχει αναδειχθεί ως ένα νέο μέσο επικοινωνίας και ανοιχτής συζήτησης στα θέματα σχεδιασμού και εφαρμογής δημόσιων πολιτικών.

4.1.1 Συστήματα διαβούλευσης πολιτικής

Τα συστήματα διαβούλευσης πολιτικής έχουν αρχίσει να αναπτύσσονται τα τελευταία χρόνια. Οι κυβερνήσεις σε πολλές χώρες του κόσμου έχουν αναδείξει τη διακτυακή διαβούλευση ως νέο μέσο επικοινωνίας και ανοιχτής συζήτησης. Πολίτες και κοινωνικοί φορείς συμμετέχουν εθελοντικά στην λήψη αποφάσεων σε διάφορα επίπεδα. Η ανάγκη για

συστήματα διαβούλευσης πολιτικής πηγάζει από την αδυναμία συντονισμού και συμμετοχής πολλών διαφορετικών συνιστωσών. Μέσω αυτών επιτυγχάνονται η όσο το δυνατόν λεπτομερέστερη και σε βάθος διερεύνηση όλων των πτυχών ενός σύνθετου προβλήματος. Τα συστήματα αυτά συμβάλλουν στην αποτύπωση και κατανόηση των πραγματικών διαδικασιών που εφαρμόζονται και των δυσλειτουργιών που συναντώνται.

Η ανοιχτή διαβούλευση για πολιτικές προτάσεις έχει αναδειχθεί ως ένα σημαντικό εργαλείο για την ανάπτυξη της διαφάνειας, την διερεύνηση και εφαρμογή καλύτερων πολιτικών και τη δημιουργία μιας πιο δυναμικης σχέσης μεταξύ πολίτη και κράτους. Οι φορείς που διεξάγουν τις διαβουλεύσεις εξοικειώνονται σταδιακά στο να είναι πιο ανοικτοί στους πολίτες, παρουσιάζοντας σχέδια δημόσιων πολιτικών πριν αυτά εφαρμοστούν και είναι σε θέση να αναμένουν χρήσιμες προτάσεις για την τελική τους διαμόρφωση. Οι συμμετέχοντες στην διαδικασία αποκτούν για πρώτη φορά το δικαίωμα να ενημερώνονται για την επικείμενη νομοθεσία πριν αυτή διαμορφωθεί οριστικά και με αυτόν τον τρόπο αναπτύσσουν σταδιακά την προσδοκία για μια γόνιμη συζήτηση και ότι οι απόψεις τους θα ληφθούν υπόψιν.

Στην Ελλάδα από τον Οκτώβριο του 2009 λειτουργεί μια πλατφόρμα ηλεκτρονικής, συμμετοχικής διαβούλευσης, το opengov.gr[3]. Οι διαβουλεύσεις στο opengov.gr στην συντριπτική τους πλειοψηφία αποτελούν προ – νομοθετικές διαβουλεύσεις, δηλαδή διαβουλεύσεις επί προτάσεων νόμων. Το opengov είναι αναμφίβολα μια ελληνική πρακτική ηλεκτρονικής δημόσιας διαβούλευσης που μπορεί να παρουσιασθει ως καινοτομία διεθνώς. Η λειτουργία του ξεκίνησε τον Οκτώβριο του 2009 και μέχρι σήμερα έχει συγκεντρώσει πάνω από 68.000 σχόλια και 4.5 εκατομμύρια επισκέψεις. Κάθε υπουργείο έχει μια δικτυακή πλατφόρμα διαβούλευσης η οποία έχει τα εξής χαρακτηριστικά:

- Πολύ απλό και φιλικό περιβάλλον χρήσης που μοιάζει με blog.

- Ο σχολιασμός γίνεται είτε ανά άρθρο, είτε ανα παράγραφο, είτε ανά ερώτημα, ανάλογα με το είδος της διαβούλευσης.

- Τα πιο πρόσφατα σχόλια των πολιτών εμφανίζονται πρώτα.

- Οι συμμετέχοντες έχουν την δυνατότητα να αξιολογήσουν τις απόψεις των άλλων θετικά ή αρνητικά με ένα κλικ.

- Σε κάθε διαβούλευση αναφέρεται ο σκοπός και η διάρκεια της και υπάρχουν και κάποια εργαλεία άμεσης διάχυσης της διαβούλευσης στα μέσα κοινωνικής δικτύωσης.

Τα προϊόντα της έκθεσης διαβούλευσης πρέπει να είναι ποιοτικά και αυτό προϋποθέτει την χρήση εργαλείων και κάποια μεθοδολογία για την αποδελτίωση και τη σύνταξη της έκθεσης αναφοράς. Δύναται να χρησιμοποιηθούν εργαλεία που κάνουν λεξικογραφική ανάλυση, όμως την ποιοτική δουλειά την κάνουν ομάδες ατόμων που δουλεύουν με μέθοδο. Η ομάδα ηλεκτρονικής διακυβέρνησης προτείνει μια μεθοδολογία 4 βημάτων προς τις ομάδες σε όλα τα Υπουργεία [37].

- Μελέτη του νομοσχεδίου άρθρο προς άρθρο για κατανόηση αντικειμένου και σκοπού της ρύθμισης και καθημερινή ανάγνωση των σχολίων των πολιτών για την απόκτηση γενικής αίσθησης επι των απόψεων.

- Καταγραφή πλαγιότιτλου και σύντομης περίληψης για κάθε σχόλιο.

[3] http://www.opengov.gr/home/

- Ομαδοποίηση πλαγιότιτλων και σχολίων σε κοινά θέματα, που αναδεικνύουν οι πολίες. Σε κάθε κοινό θέμα τίθεται τίτλος και ένα – δύο αυτούσια συμπεράσματα από τα αποσπάσματα από τα σχόλια των πολιτών.

Από την μελέτη και το άθροισμα προκύπτει μια αναφορά. Η πρακτική τροποποιείται ανάλογα με τον φορέα. Συνήθως επιλέγεται μια μεθοδολογία ποιοτικά προσανατολισμένη, δηλαδή κατάταξη των σχολίων σε θετικά, αρνητικά και ουδέτερα. Για τον σκοπό αυτό χρησιμοποιούνται τεχνικές εξόρυξης γνώμης που εστιάζουν στο πολιτικό κείμενο, όπως θα περιγραφούν στην συνέχεια.

4.1.2 Μηχανισμοί συλλογής γνώμης για πολιτικές αποφάσεις

Οι παραδοσιακοί μηχανισμοί καταγραφής της κοινής γνώμης χρησιμοποιούν την τεχνολογία της πληροφορικής εδώ και δεκαετίες. Δημοσκοπήσεις, ερωτηματολόγια και έρευνες της κοινής γνώμης βασίζονται στην ανάπτυξη ενός δικτύου υπολογιστών στο οποίο μεταβιβάζονται τα δεδομένα που έχουν συλλεχθεί, τυπικά από ερωτηματολόγια. Στη συνέχεια, προσωπικό από διάφορες ειδικότητες (στατιστικολόγοι, σύμβουλοι κλπ) αναλαμβάνει να αναδείξει και να αναλύσει τις διαφορετικές τάσεις που διαμορφώνονται με βάση τις μετρήσεις. Τα στοιχεία παρουσιάζονταν στους πολιτικούς μαζί με τα βασικότερα συμπεράσματα που εξήχθησαν.

Η συλλογή της γνώμης των χρηστών από το διαδίκτυο για πολιτικές αποφάσεις παρουσιάζει ορισμένα σημαντικά πλεονεκτήματα. Τα μέσα που χρησιμοποιούν οι παραδοσιακές μέθοδοι συλλογής της κοινής γνώμης (ερωτηματολόγια, τηλεφωνήματα) από την φύση τους περιορίζουν το εύρος των απαντήσεων. Επιπλέον, ορισμένες φορές οι ερωτήσεις τίθενται με έναν τέτοιο τρόπο που προδιαθέτουν τις απαντήσεις προς μια συγκεκριμένη κατεύθυνση (π.χ. απαντήσεις υπαγορευμένες από τη διατύπωση της ερώτησης). Αντίθετα, η γνώμη που εκφράζεται στο διαδίκτυο μπορεί να είναι περισσότερο αντικειμενική καθότι εκφράζεται αβίαστα και όχι υπό πίεση. Από την άλλη η εξαγωγή της και η ανάλυση της είναι σαφώς πιο δύσκολη από την δομημένη πληροφορία που τυπικά περιέχουν τα ερωτηματολόγια.

Οι ταξινομητές πολιτικές γνώμης έχουν εφαρμογή κυρίως σε δύο πεδία:

- Σε συστήματα ηλεκτρονικής θέσπιση κανόνων (e-Rulemaking)

Η ηλεκτρονική θέσπιση κανόνων είναι η χρήση των ψηφιακών τεχνολογιών από κυβερνητικές υπηρεσίες σε διαδικασίες θέσπισης κανόνων και λήψης αποφάσεων [38]. Στις Η.Π.Α. ο νόμος ορίζει μια διαδικασία γνωστή ως *Αναγγελία και παρατηρήσεις στην θέσπιση κανόνων* (Notice and Comment Rulemaking), η οποία απαιτεί από τους οργανισμούς που ασχολούνται με την θέσπιση κανόνων να αναζητήσουν σχόλια από το κοινό πριν από την θέσπιση νέων κανονισμών. Επιπλέον οι οργανισμοί πρέπει να αποδείξουν ότι ο τελικός κανονισμός που προτείνουν αντιμετωπίζει όλα τα ουσιαστικά θέματα που τέθηκαν από το κοινό. Τα συστήματα αυτά χρησιμοποιούν μεταξύ άλλων και ταξινομητές πολιτικής γνώμης, ενώ αποτελούν αναπόσπαστο κομμάτι της δημόσιας διοίκησης στις Ηνωμένες Πολιτείες και αλλού.

- Σε συστήματα ηλεκτρονικής διακυβέρνησης (e-Government).

Τα σύστήματα ηλεκτρονικής διακυβέρνησης αποσκοπούν στην βελτίωση των παρέχομενων υπηρεσιών από το κράτος με την αξιοποίση της τεχνολογίας, κυρίως της πληροφορικής. Μέσα από την χρήση τους αναπτύσσεται μια αμφίδρομη, άμεση και ευρεία σχέση μεταξύ πολιτικών συντελεστών, κοινωνικών ομάδων και κοινωνικών εταίρων. Η

χρήση τους δεν περιορίζεται ως το βασικό εργαλείο εφαρμογής της διοικητικής μεταρρύθμισης στον δημόσιο τομέα, παρά αποτελούν πλέον ένα μέσο χάραξης, εφαρμογής και ανατροφοδότησης της πολιτικής [39].

Μια γενικότερη πρόκληση για τα συστήματα αυτά είναι να καταγράφουν τον παλμό της κοινωνίας, όπως αυτός εκφράζεται μέσα από το διαδίκτυο. Σε αυτό το πλαίσιο, προτεραιότητα αποτελεί η αναθεώρηση του μοντέλου της συμμετοχής των πολιτών σε αυτά, παρέχοντας τους εύκολη πρόσβαση σε διάφορα μέσα έκφρασης. Η αύξηση του βαθμού της συμμετοχής των πολιτών και η διεύρυνση της χρήσης τους από πολλές κοινωνικές ομάδες είναι ένας βασικός δείκτης επιτυχίας αυτών των συστημάτων.

Για να επιτύχουν τους σκοπούς αυτούς αξιοποιούνται τεχνικές εξόρυξης γνώμης. Τα συστήματα ηλεκτρονικής διακυβέρνησης ενσωματώνουν διαδικασίες με τις οποίες μπορούν να αναλύσουν τον αντίκτυπο που έχουν οι κυβερνητικές αποφάσεις, καθώς και να προσδριορίσουν την γνώμη των πολιτών πάνω σε συγκεκριμένα ζητήματα. Η δημόσια διοίκηση θα αναβαθμιστεί και η κυβερνητική γραφειοκρατία θα καταπολεμηθεί αν αυτή η γνώμη μπορεί να συλλέγεται και να εξάγεται σε συστηματική βάση.

4.2 Δημιουργία μηχανισμών συλλογής γνώμης

Η περιγραφή της αρχιτεκτονικής συστημάτων πολιτικής γνώμης που ακολουθεί βασίζεται στην εργασία ”Public opinion mining for Govermental Decision” για ένα σύστημα εξόρυξης γνώμης που μπορεί να αποτελέσει μέρος μιας υποδομής ηλεκτρονικής διακυβέρνησης [39]. Κύριος σκοπός του συστήματος είναι η ταξινόμηση των απόψεων των χρηστών για διάφορα θέματα, καθώς το αν εκφράζουν θετική ή αρνητική άποψη σε αυτά. Η δυνατότητα αυτή είναι ιδιαίτερα χρήσιμη σε συστήματα ηλεκτρονικής διακυβέρνησης διότι επιτρέπει στους πολιτικούς να εντοπίζουν και να εστιάζουν π.χ. στις αρνητικές απόψεις και να τις λαμβάνουν υπόψιν τους.

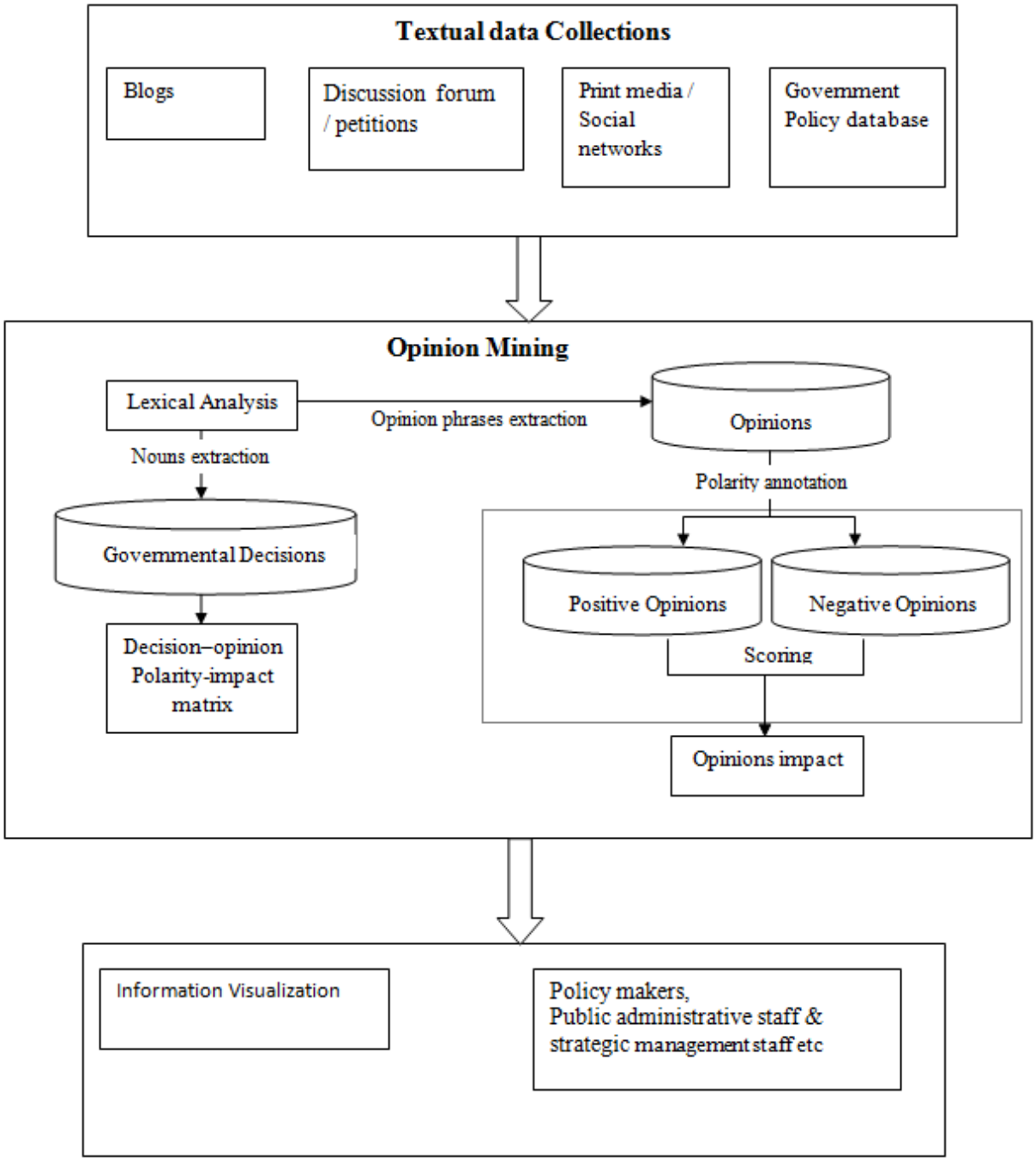
Συστήματα με αντίστοιχη αρχιτεκτονική χρησιμοποιούνται για παρόμοιους σκοπούς. Στην εργασία “Text Annotation for Political Science Research” οι συγγραφείς έχουν συντάξει μια λίστα με δημοσιεύσεις για εφαρμογές της εξόρυξης γνώμης και της μηχανικής μάθησης στο πεδίο της πολιτικής [40]. Έχουν αναπτυχθεί αντίστοιχα συστήματα που ταξινομούν τους λόγους των μελών του Κογκρέσου (δημοκρατικοί/ ρεπουμπλικάνοι), δημόσια σχόλια πολιτών, απόψεις που εμφανίζονται σε άρθρα εφημερίδων και τα οποία χρησιμοποιούν μια αντίστοιχη αρχιτεκτονική. Θεωρούμε λοιπόν την παρακάτω αρχιτεκτονική αντιπροσωπευτική για συστήματα που έχουν αναπτυχθεί στο πεδίο της πολιτικής.

4.2.1 Αρχιτεκτονική των συστημάτων συλλογής γνώμης

Τα συστήματα εξόρυξης γνώμης που χρησιμοποιούνται σε συστήματα διαβούλευσης πολιτικής αποτελούν τυπικά τμήμα ενός ευρύτερου συστήματος. Συνήθως το σύστημα αυτό είναι ένα σύστημα ηλεκτρονικής διακυβέρνησης. Η αρχιτεκτονική αυτών των συστημάτων δεν διαφέρει από την τυπική αρχιτεκτονική ενός συστήματος εξόρυξης γνώμης. Οι διαφορές εντοπίζονται κυρίως στις επιλογές των ταξινομητών και γενικότερα σε ορισμένους κανόνες κατάταξης (σε κατηγορίες) που ορίζουν οι επιμέρους σχεδιαστές [41].

Για τον προσδριορισμό των απόψεων των πολιτών μέσα από τις τοποθετήσεις τους που σχετίζονται με πολιτικά θέματα, αρχικά μεταφορτώνονται τα περιεχόμενα των δημοσιεύσεων των χρηστών. Αυτά επεξεργάζονται με σκοπό να προσδιορισθεί (από το υπόλοιπο κείμενο) εκείνο που περιέχει τις απόψεις των χρηστών. Αρχικά το περιεχόμενο των HTML σελίδων αναλύεται (HTML parsing) με σκοπό να απομακρυνθούν μη - κειμενικά στοιχεία. Τέτοια

στοιχεία περιλαμβάνουν εικόνες, γραφικές αναπαραστάσεις κειμένου, πλαίσια, αρχεία εντολών και βίντεο. Έπειτα εφαρμόζεται στο κύριο σώμα του κειμένου η διαδικασία του διαχωρισμού λεκτικών μονάδων (tokenization) με σκοπό να εξαχθούν τα λεξιλογικά στοιχεία του κειμένου. Στη συνέχεια, το κείμενο περνάει μέσα από έναν αναγνωριστή μέρους του λόγου (Part of Speech Tagger - POS) ο οποίος επισημειώνει τις λέξεις του κειμένου στην κατάλληλη γραμματική κατηγορία.



Σχήμα 15: Αρχιτεκτονική συστήματος εξόρυξης γνώμης

Στο επόμενο βήμα χρησιμοποιείται ένας συντακτικός αναλυτής (parser), μέσω του οποίου εξάγονται τα ουσιαστικά που εμφανίζονται στο κείμενο και τα επίθετα που αναφέρονται σε

αυτά. Ο συντακτικός αναλυτής προσδιορίζει τα ουσιαστικά στα οποία αναφέρονται τα επίθετα και δημιουργεί ζεύγη ουσιαστικών – επιθέτων. Με βάση αυτά τα ζεύγη μπορούμε να προσδιορίσουμε τόσο το θέμα του κειμένου όσο και την γνώμη των χρηστών. Το θέμα του κειμένου περιγράφεται από τα ουσιαστικά ενώ η γνώμη των χρηστών από τα επίθετα που αναφέρονται σε αυτά. Επομένως το πρόβλημα του προσδιορισμού του πώς οι χρήστες κρίνουν τις πολιτικές αποφάσεις ανάγεται στην εύρεση της συναισθηματικής αξίας των επιθέτων που χρησιμοποιούν στις δημοσιεύσεις τους.

Προτού προσδιοριστεί η συναισθηματική αξία ενός επιθέτου, αυτό αναζητείται στην οντολογία FrameNet [42]. Αυτή περιέχει πάνω από 10.000 λεκτικές μονάδες, εκ των οποίων οι 6.000 είναι πλήρως επισημασμένες, ιεραρχικά ορισμένες σε 800 σημασιολογικά πλαίσια. Το FrameNet αντιστοιχεί σε κάθε επίθετο ένα πλαίσιο (frame), μια περιγραφή του επιθέτου με σημασιολογικούς όρους. Τα επίθετα στα οποία δίνονται σημασιολογικές περιγραφές (πλαίσια) που θεωρούνται άσχετα με την έκφραση γνώμης, δεν λαμβάνονται υπόψιν. Στα υπόλοιπα προσδιορίζεται χειρονακτικά μια τιμή συναισθήματος. Έχει αποδειχθεί σε διάφορες μελέτες ότι ο χειρονακτικός προσδιορισμός της τιμής συναισθήματος σε επίθετα είναι ένας αντικειμενικός τρόπος προσδιορισμού, καθότι μελέτες έχουν δείξει ότι οι αξιολογητές συμφωνούν στις τιμές που ορίζουν [36]. Τα επίθετα θα χρησιμοποιηθούν για την εκπαίδευση ενός ταξινομητή, με τρόπο που θα περιγραφεί στην συνέχεια.

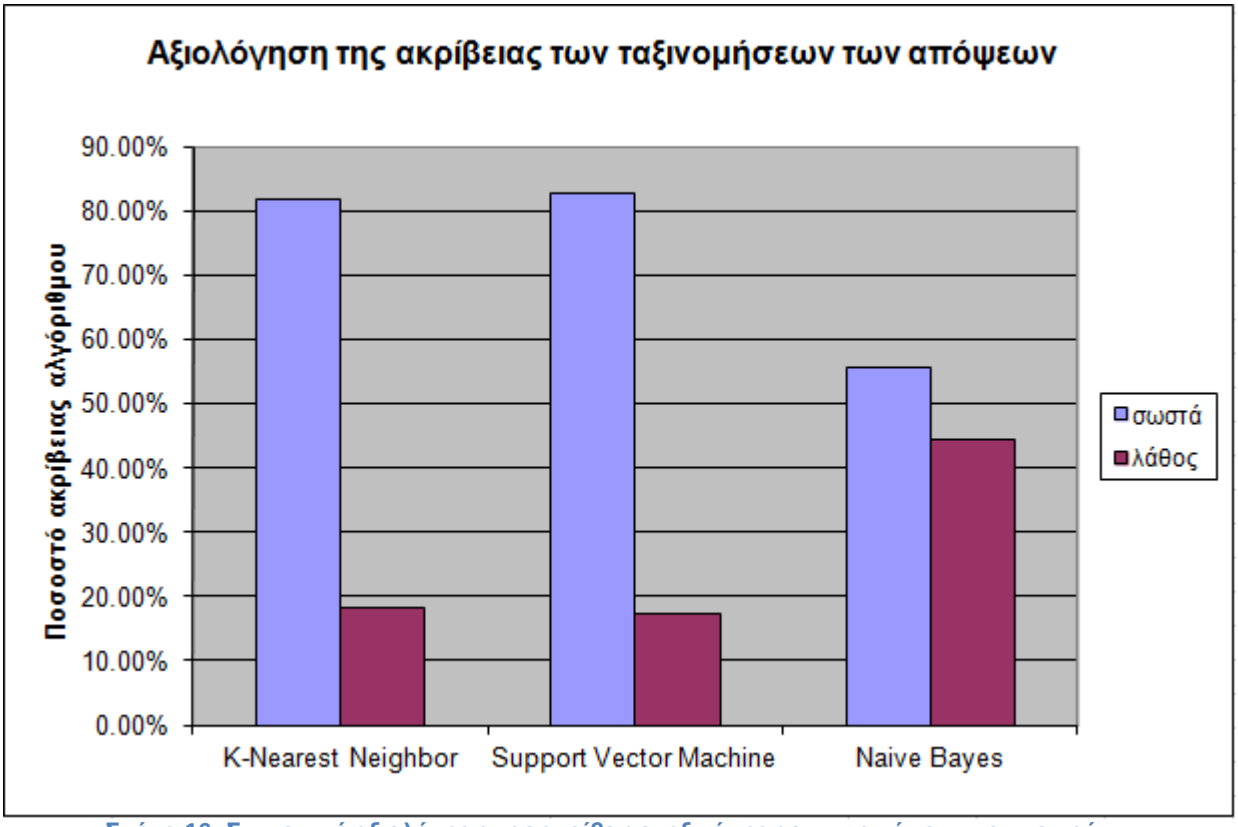
Σκοπός του ταξινομητή είναι να κατατάξει τις γνώμες – φράσεις σε σύνολα από υποστηρικτικές (θετικές) και αντιτιθέμενες (αρνητικές) απόψεις. Σε αυτό το πλαίσιο, εμπλουτίζεται η λίστα με τα επίθετα στα οποία έχουν ανατεθεί τιμές συναισθήματος με λέξεις από το WordNet. Το WordNet είναι μια λεξιλογική ιεραρχία που οργανώνει έννοιες σε σύνολα συνωνύμων και τις συνδέει με βάση τις σημασιολογικές τους σχέσεις [43]. Η λίστα αυτή επεκτείνεται σε δύο στάδια. Στο πρώτο στάδιο προστίθενται στην λίστα επίθετα που είναι συνώνυμα με αυτά που ήδη περιέχει. Στα επίθετα αυτά δίνεται η ίδια τιμή συναισθήματος με τα συνώνυμα τους. Στο δεύτερο στάδιο προστίθενται στην λίστα επίθετα που είναι αντώνυμα με αυτά που ήδη περιέχει. Αντίστοιχα, αυτά λαμβάνουν αντίθετη τιμή συναισθήματος από αυτήν που έχουν τα συνώνυμα τους. Το σύνολο που προκύπτει χωρίζεται σε δεδομένα εκπαίδευσης και ελέγχου. Το σύνολο εκπαίδευσης χρησιμοποιείται για να εκπαιδευτεί ο ταξινομητής ώστε να αναγνωρίζει αυτόματα την τιμή συναισθήματος των φράσεων που εξάγονται από το σύστημα. Η ακρίβεια του ταξινομητή αξιολογείται με βάση το σύνολο ελέγχου και εκτελώντας επαναληπτικές ταξινομήσεις.

4.2.2 Επιδόσεις ταξινομητών

Προκειμένου να συγκριθούν οι επιδόσεις διαφόρων ταξινομητών, στην εργασία ”Public opinion mining for Govermental Decision” [39] εκτελέστηκαν πειράματα από ένα σύνολο πραγματικών πολιτικών απόψεων που συνέλλεξαν από το forum anatheorisi[4], ένα forum που εστιάζει σε θέματα πολιτικής. Αφού επεξεργάστηκαν αυτά τα σχόλια όπως περιγράφτηκε στην προηγούμενη παράγραφο, εξήχθησαν 652 διαφορετικές φράσεις που εκφράζουν άποψη. Έπειτα, εκπαίδευσαν τρεις διαφορετικούς ταξινομητές με την χρήση της πλατφόρμας Weka. Αυτοί ήταν ο SVN, ο Κοντινότερος Γείτονας (K-Nearest Neighbor) και ο απλοϊκός ταξινομητής Bayes.

Το παρακάτω γράφημα δείχνει το μέσο όρο των περιπτώσεων που προβλέφθηκαν σωστά και λανθασμένα από τους τρεις ταξινομητές για κάθε σύνολο δεδομένων. Πιο συγκεκριμένα, η εικόνα προσδιορίζει το ποσοστό των φράσεων γνώμης που προσδιορίσθηκαν σωστά και λανθασμένα από τους ταξινομητές σαν θετικές ή αρνητικές.

[4] http://anatheorisi.org/



Σχήμα 16: Συγκριτική αξιολόγηση της ακρίβειας ταξινόμησης των απόψεων του κοινού

Με βάση τα αποτελέσματα των πειραμάτων μπορούμε να συμπεράνουμε ότι η διαδικασία που περιγράφτηκε είναι αρκετά αποτελεσματική στον αυτόματο εντοπισμό της στάσης του κοινού απέναντι στις κυβερνητικές αποφάσεις.

4.3 Ταξινομητές πολιτικού κειμένου

Ένας τύπος κειμένου που έχει συγκεντρώσει το ενδιαφέρον τα τελευταία χρόνια είναι το πολιτικό κείμενο. Πρόκειται για κείμενο που περιγράφει τον πολιτικό λόγο σε επίσημη γλώσσα. Αυτός ο τύπος κειμένου συναντάται σε συνεδριάσεις του κοινοβουλίου, εξαγγελίες πολιτικών και άρθρα πολιτικών εφημερίδων. Έχει συνταχθεί από πολιτικούς, αναλυτές και εμπειρογνώμονες. Οι ταξινομητές πολιτικού κειμένου έχουν σαν στόχο να ταξινομήσουν ορθά πολιτικά κείμενα ανάλογα με το εάν υποστηρίζουν ή αντιτίθενται σε μια συγκεκριμένη πολιτική άποψη που τίθεται υπο συζήτηση. Με την χρήση των τεχνικών εξόρυξης γνώμης είναι δυνατή η εύρεση και κατηγοριοποίηση πολιτικών απόψεων καθώς και η παρακολούθηση της εξέλιξης τους στον χρόνο.

Τυπικές πηγές πολιτικού κειμένου αποτελούν άρθρα εφημερίδων στο διαδίκτυο, πολιτικά ιστολόγια αλλά και υλικό που είναι διαθέσιμο από τα κοινοβούλια πολλών χωρών. Στους δικτυακούς τόπους των κοινοβουλίων, όπως και στο ελληνικό, υπάρχουν σε ηλεκτρονική μορφή ομιλίες και πρακτικά συνεδριάσεων. Η αξιοποίηση αυτού του υλικού απαιτεί την δημιουργία των κατάλληλων εργαλείων. Αποδέκτες αυτών των εργαλείων είναι οι πολίτες καθώς και όσοι ασχολούνται με την ιστορική και δημοσιογραφική έρευνα. Αντίστοιχα εργαλεία για πολιτικό κείμενο χρησιμοποιούνται από συστήματα ηλεκτρονικής διακυβέρνησης και συστήματα ηλεκτρονικής θέσπισης κανόνων (e-Rulemaking) [47].

Στη συνέχεια θα παρουσιάσουμε τα χαρακτηριστικά του πολιτικού κειμένου και θα το συγκρίνουμε με άλλους τύπους κειμένων. Με βάση αυτά θα περιγράψουμε τους ταξινομητές πολιτικού κειμένου και θα αναφέρουμε παραδείγματα χρήσης τους.

4.3.1 Χαρακτηριστικά πολιτικού κειμένου

Η επίσημη, «ευπρεπής» γλώσσα η οποία χαρακτηρίζει το πολιτικό κείμενο, οι ορολογίες καθώς και η χρήση λέξεων και φράσεων που εμπεριέχουν έναν ιδιαίτερο συναισθηματικό προσανατολισμό είναι ορισμένα από τα χαρακτηριστικά του πολιτικού κειμένου. Πολλές λέξεις έχουν διαφορετική σημασία όταν χρησιμοποιούνται στον πολιτικό λόγο, όπως για παράδειγμα οι λέξεις «ταμπακιέρα», «τζάκι». Η «ξύλινη γλώσσα» που συναντάμε σε πολιτικά κείμενα (με λέξεις όπως «οπορτουνισμός», «ρεφορμισμός») διαφέρει σημαντικά από το στύλ της γλώσσας που συναντούμε σε αντίστοιχες συζητήσεις πολιτικού περιεχομένου στο διαδίκτυο. Σχετικές μελέτες έχουν προσδιορίσει τρια κύρια στυλ έκφρασης πολιτικής γνώμης, καθένα από τα οποία συναντάται σε:

- Κοινοβουλευτικές συζητήσεις
- Πολιτικές συζητήσεις σε δικτυακές κοινότητες
- Απαντήσεις ηλεκτρονικού ταχυδρομείου σε θέματα πολιτικής

Προκειμένου να προσδιορισθούν οι ιδιαιτερότητες του πολιτικού κειμένου και οι επιδόσεις των ταξινομητών γενικού σκοπού σε αυτό, έγιναν συγκρίσεις με κείμενα κριτικών ταινιών και κείμενα ειδήσεων [51]. Στα πειράματα αυτά χρησιμοποιήθηκε η μέθοδος λεξικού για τον προσδιορισμό της συναισθηματικής αξίας. Τα αποτελέσματα έδειξαν ότι η ταξινόμηση πολιτικού κειμένου είναι δύσκολη για δύο κυρίως λόγους.

1. Η συναισθηματική αξία των λέξεων είναι χαμηλή

2. Τα ουσιαστικά (που εξαρτώνται από το θέμα) είναι περισσότερο χρήσιμα στον προσδιορισμό της γνώμης σε σχέση με τα επίθετα

Τα περισσότερα κείμενα με πολιτικό περιεχόμενο, κυρίως όσα προέρχονται από το κοινοβούλιο, είναι προσεκτικά μελετημένοι λόγοι. Προκειμένου να μην παραβιαστούν οι κανόνες της ευπρέπειας σε πολλές περιπτώσεις περιορίζεται σε σημαντικό βαθμό η συναισθηματική γλώσσα. Οι νομοθέτες δεσμεύονται από τους σιωπηρούς κανόνες και τις συμβάσεις της κοινοβουλευτικής συζήτησης και χρησιμοποιούν λιγότερα επίθετα και χαρακτηρισμούς στον λόγο τους. Έτσι πολιτικό κείμενο (αυτού του είδους) παρουσίαζει μικρή συναισθηματική αξία, δηλαδή μπορεί πιο δύσκολα να ταξινομηθεί σαν κείμενο με θετική ή αρνητική τιμή συναισθήματος. Αντίθετα, στα κείμενα που εκφράζουν κριτικές ταινιών η συναισθηματική αξία των προτάσεων είναι μεγάλη. Αυτό συμβαίνει επειδή περιέχουν περισσότερα επίθετα, τα οποία τυπικά έχουν μια υψηλή συναισθηματική αξία. Τα κείμενα που απλά παραθέτουν γεγονότα χωρίς σχολιασμό, όπως ορισμένα κείμενα ειδήσεων, έχουν ακόμα μικρότερη συναισθηματική αξία από το πολιτικό κείμενο. Σε γενικές γραμμές, όσα περισσότερα επίθετα περιέχει ένα κείμενο τόσο μεγαλύτερη είναι η συναισθηματική του αξία.

Η ανάλυση μιας πρότασης στα επιμέρους μέρη του λόγου από τα οποία αποτελείται έχει δείξει ότι κάθε μέρος του λόγου έχει διαφορετική επίδραση στην αναγνώριση του συναισθήματος που εκφράζει η πρόταση. Όπως αναφέρθηκε, τα επίθετα θεωρούνται οτι έχουν μεγαλύτερη αξία συναισθήματος. Για να εξακριβωθεί αν αυτός ο κανόνας ισχύει και στο πολιτικό κείμενο, χρησιμοποιήθηκε η τεχνική της ανάλυσης μερών του λόγου. Πιο συγκεκριμένα, χρησιμοποιήθηκε η μέθοδος της επιλογής χαρακτηριστικών (feature selection) σε συνδυασμό με την στατιστική επεξεργασία των μερών του λόγου προκειμένου να επιλεγούν οι λέξεις που συμβάλλουν περισσότερο στην ταξινόμηση και έτσι να

προσδιοριστούν ποιά μέρη του λόγου είναι περισσότερο σημαντικά. Έγινε μια σύγκριση αναφοράς σε σχέση με τα κείμενα που κριτικάρουν ταινίες. Ενώ στις κριτικές ταινιών τα επίθετα είναι πιο ενδεικτικά του εάν μια κριτική είναι θετική, αρνητική ή ουδέτερη, αυτό δεν συνέβη με τα πολιτικά κείμενα. Εκεί τα ουσιαστικά ήταν περισσότερο ενδεικτικά, τα οποία μάλιστα εξαρτιόνταν από το θέμα του κειμένου. Αυτό το αποτέλεσμα δεν είναι δύσκολο να ερμηνευθεί στην πράξη. Η επιλογή του θέματος του πολιτικού κειμένου (όπως αυτή προσδιορίζεται από τα ουσιαστικά) είναι μια έκφραση πολιτικής γνώμης από μόνη της. Όροι όπως «ανάπτυξη», «απασχόληση», «ποινικοποίηση», «αποδόμηση» που συναντάμε σε πολιτικά κείμενα, εκφράζουν έμμεσα και το συναίσθημα του συντάκτη.

4.3.2 Βελτίωση των ταξινομητών πολιτικής γνώμης

Για σκοπούς ταξινόμησης πολιτικού κειμένου χρησιμοποιούνται στην πράξη παραλλαγές των αλγορίθμων SVM και του απλού ταξινομητή Bayes. Μια κλασική εργασία αξιολόγησης των ταξινομητών πολιτικής γνώμης είναι ο καθορισμός του εάν ένα κείμενο εκφράζει απόψεις της δεξίας πολιτικής πτέρυγας ή της αριστεράς. Πειράματα που πραγματοποιήθηκαν με κείμενα από ιστολόγια και χρήση των ταξινομητών SVM και του απλού ταξινομητή Bayes (NB) έδειξαν ότι ο δεύτερος ξεπερνάει τον πρώτο στα ποσοστά ορθής ταξινόμησης, επιτυγχάνοντας ποσοστά της τάξης του 80%, με τυπική απόκλιση 2,39, ενώ ο SVM ταξινομεί ορθά το 75% των εγγράφων με τυπική απόκλιση 2,64 [44]. Παρόμοια αποτελέσματα επιβεβαιώνονται και από άλλους ερευνητές.

Σε γενικές γραμμές, ένας ταξινομητής αποδίδει καλύτερα όσο πιο κοντά είναι η μορφή του κειμένου (στο οποίο εφαρμόζεται) με το κείμενο της εκπαίδευσης του. Ωστόσο ούτε η μορφή του κειμένου μπορεί να προσδιοριστεί εκ των προτέρων αλλά ούτε και είναι θεμιτό να δημιουργηθεί ένας ταξινομητής που θα αποδίδει ικανοποιητικά μόνο σε συγκεκριμένους τύπους κειμένων. Οι ταξινομητές πολιτικού κειμένου είναι ευαίσθητοι στον ορισμό των κλάσεων του συνόλου εκπαίδευσης [44]. Μια μη ισορροπημένη σύνθεση των κλάσεων ταξινόμησης εισάγει μια μεροληψία στα αποτελέσματα. Η κλάση της πλειοψηφίας είναι πιο πιθανό να ταξινομηθεί ορθά από την κλάση με τα λιγότερα δεδομένα εκπαίδευσης. Όσο τα δεδομένα εκπαίδευσης μειώνονται, αυξάνεται η επίδραση της μεροληψίας λόγω της μη ισορροπημένης σύνθεσης των δεδομένων εκπαίδευσης.

Αναφέρθηκε ότι ο πολιτικός λόγος περιέχει όρους με ιδιαίτερη σημασία, καθώς και λέξεις που δεν χρησιμοποιούνται συχνά στην καθημερινή ομιλία. Τα υπάρχοντα λεξικά θα μπορούσαν να εμπλουτισθούν με τιμές συναισθήματος για κάποιους από αυτούς τους όρους. Άλλοι ερευνητές έχουν προτείνει την βελτίωση των ταξινομητών πολιτικής γνώμης μέσα από την αναγνώριση της ευρύτερης ιδεολογίας την οποία ασπάζεται ο ομιλητής. Η αναγνώριση της ιδεολογίας του ομιλητή αποκαλύπτει από μόνη της πολλά στοιχεία για τις απόψεις του. Σε περιπτώσεις που η ταυτότητα του συντάκτη δεν είναι γνωστή εκ των προτέρων, όπως συμβαίνει με τους κοινοβουλευτικούς λόγους, αυτή θα μπορούσε να προσδιορισθεί από τα κοινωνικά δίκτυα στα οποία ανήκει ο συντάκτης [50].

Στο πλαίσιο της προσπάθειας αναγνώρισης της ιδεολογίας του συντάκτη του πολιτικού κειμένου, ορισμένοι ερευνητές προτείνουν την αξιοποίηση των δικτυακών παραπομπών των υπερσυνδέσμων των κειμένων (hyperlinks cocitation) με σκοπό την εκτίμηση του πολιτικού προσανατολισμού των εγγράφων [45]. Η ιδέα τους προέρχεται από το κριτήριο της σημειακής αμοιβαίας πληροφορίας (Pointwise Mutual Information). Με το κριτήριο αυτό μπορεί να υπολογιστεί κατά πόσο δύο λέξεις εμφανίζονται σε ένα κείμενο μαζί ή όχι. Δύο λέξεις που εμφανίζονται γειτονικά σε κείμενα έχουν μεγάλη πιθανότητα να έχουν την ίδια συναισθηματική αξία. Επέκτειναν αυτήν την ιδέα και σε επίπεδο κειμένων και όχι μόνο λέξεων. Προσδιορίζουν τον *πολιτιστικό προσανατολισμό* (cultural orientation) ενός εγγράφου σαν τον βαθμό στον οποίο αυτό αναφέρεται σε μια συγκεκριμένη κοινότητα. Κάνουν την

παραδοχή ότι κείμενα προοδευτικά είναι απίθανο να σχετίζονται και να εμφανίζονται σε μέσα που υπάρχουν πολλά συντηρητικά κείμενα. Έτσι, προοδευτικά κείμενα θα συναντώνται σε πηγές που ασπάζονται προοδευτικές απόψεις και με τον τρόπο αυτό μπορεί έμμεσα να προσδιορισθεί και η ιδεολογία του συντάκτη.

Μια άλλη προσπάθεια προσδιορισμού της ιδεολογίας θα μπορούσε να γίνει με βάση τις θέσεις που έχει εκφράσει ένας χρήστης για άλλα θέματα. Παρακολουθώντας τις απόψεις του σε μια σειρά θεμάτων – κλειδιά, μπορεί να προσδιορισθεί έμμεσα ο ιδεολογικός χώρος στον οποίο ανήκει. Για παράδειγμα, οι Poole και Rosental [46] βρήκαν ότι στην μεταπολεμική ιστορία του αμερικανικού κονγκρέσου, το 85% των γερουσιαστών μπορεί να προσδιοριστεί με βάση τις ψήφους τους σε μια σειρά θεμάτων, όπως φαίνεται στην παρακάτω εικόνα.

Σχήμα 17: Η σχέση μεταξύ ιδεολογίας και γνώμης σε διάφορα θέματα για τα μέλη του αμερικανικού κογκρέσου

Αν έχουμε μια συλλογή κειμένων του προσώπου του οποίου θέλουμε να προσδιορίσουμε την ιδεολογία, τότε με βάση αυτήν, εφαρμόζοντας τεχνικές μηχανικής μάθησης, μπορούμε να ταξινομήσουμε τις απόψεις του για κάθε ένα από τα θέματα σε μια από τις δύο κατηγορίες. Βρίσκοντας τις απόψεις του σε μια σειρά θεμάτων – κλειδιών, θα έχουμε καταφέρει να προσδιορίσουμε την ιδεολογία του.

Η διαδικασία της ταξινόμησης πολιτικών κειμένων μπορεί να επιβαρυνθεί αν γίνουν λάθη σε διάφορα στάδια της προετοιμασίας. Για παράδειγμα, τα σύνολα ταξινόμησης θα πρέπει να είναι ορισμένα με σαφήνεια και τα δεδομένα θα πρέπει να είναι αντιπροσωπευτικά και ομοιόμορφα κατανεμημένα. Ωστόσο στην ταξινόμηση πολιτικού κειμένου αυτό δεν είναι προφανές. Τα πολιτικά κείμενα εμπεριέχουν αοριστία και ασάφεια. Ακόμα και οι άνθρωποι δυσκολεύονται να συμφωνήσουν μεταξύ τους στην κατηγορία στην οποία ανήκει ένα πολιτικό κείμενο. Τα σύνολα κειμένων που επιλέγονται κάποιες φορές μπορεί να είναι βολικά για τους ερευνητές, αλλά εισάγουν μια μεροληψία στο αποτέλεσμα. Τέλος, η συνεργασία των επιστημόνων της πληροφορικής (που σχεδιάζουν τον πολιτικό ταξινομητή) και των πολιτικών επιστημόνων (που επιλέγουν τα χαρακτηριστικά των δεδομένων) μπορεί να μην είναι αρμονική, κυρίως λόγω έλλειψης αλληλοκατανόησης.

5 Διαχείριση δικτυακής φήμης

Η διαχείριση φήμης σε πραγματικό χρόνο (on-line reputation management - ORM) είναι η πρακτική της συνεχούς έρευνας και ανάλυσης της φήμης όπως αυτή αναπαράγεται σε όλους τους τύπους των ηλεκτρονικών μέσων ενημέρωσης και ιδιαίτερα του διαδικτύου. Η διαχείριση της εταιρικής φήμης (ORM) σχετίζεται με όλα τα εργαλεία παρακολούθησης όπως επίσης και με τις τεχνικές που χρησιμοποιούνται για να καθοδηγούν και να ελέγχουν αυτές τις συζητήσεις και τις γνώμες που μερικές φορές μπορεί να γίνουν πολύ επικίνδυνες για την υπόσταση μιας επιχείρησης. Στο κεφάλαιο αυτό θα παρουσιάσουμε κάποια προϊόντα που χρησιμοποιούν τεχνικές εξόρυξης γνώμης για σκοπούς διαχείρισης φήμης. Περιγράφονται επίσης οι τρόποι διαχείρισης της δικτυακής φήμης καθώς και η παρουσία αντίστοιχων υπηρεσιών στην Ελλάδα.

5.1 Δικτυακή φήμη

Η διαχείριση διαδικτυακής φήμης αναφέρεται σε εκείνες τις διαδικασίες που απαιτούνται για την σωστή προώθηση της εικόνας μίας επιχείρησης, προϊόντος ή υπηρεσίας μέσω του διαδικτύου. Με την άνοδο των κοινωνικών δικτύων και των ιστολογίων, η φήμη ενός προϊόντος ή επιχείρησης διαμορφώνεται και από τις συζητήσεις των χρηστών στο διαδίκτυο. Η διαχείριση διαδικτυακής φήμης αποσκοπεί στο να αναπτύξει εκείνες τις διαδικασίες ώστε κάθε ενδιαφερόμενος να γίνει κυρίαρχος της διαδικτυακής του φήμης και να μειωθεί ο αντίκτυπος από τυχόν αρνητικές αναφορές.

Ο όρος προέκυψε από την αναγνώριση της σημασίας της για τις επιχειρήσεις: η αντίληψη για κάποιον ή για κάτι μπορεί εύκολα να επηρεαστεί μέσα από μια έρευνα στο διαδίκτυο. Δεδομένου ότι ο όγκος των πληροφοριών που δημιουργείται από τους χρήστες στο διαδίκτυο αυξάνεται γεωμετρικά, τα αποτελέσματα αναζήτησης στο διαδίκτυο άρχισαν να επηρεάζονται από το περιεχόμενο των χρηστών. Έτσι η επιθυμία να αλλαχθούν κάποια από αυτά τα αποτελέσματα (αρνητικού περιεχομένου προφανώς) ήταν φυσικό επακόλουθο.

Οι εταιρίες πρέπει να παρακολουθούν συστηματικά την «εικόνα» που αναπτύσσεται στο διαδίκτυο για την επωνυμία τους. Η επωνυμία μιας επιχείρησης αποτελεί περιουσιακό της στοιχείο και η διατήρηση μιας θετικής αντίληψης γύρω από αυτήν είναι κάτι παραπάνω από σημαντική. Με την βοήθεια των κατάλληλων εργαλείων, οι ενδιαφερόμενοι μπορούν να προσδιορίσουν την εικόνα που επικρατεί γύρω από την επωνυμία τους. Στη συνέχεια θα πρέπει να αναπτύξουν την εικόνα που θέλουν να αντιλαμβάνονται οι αγοραστές καθώς και μια στρατηγική για την οικοδόμηση της. Νέες ερωτήσεις προκύπτουν για τον ενδιαφερόμενο. Προσφέρει η επωνυμία αξιοπιστία στο χώρο της αγοράς; Έχει δοθεί αρκετή σημασία στο τι αναφέρεται στα ιστολόγια; Στοχεύει σε μια ηγετική θέση στην αγορά; Έχει δημιουργήσει καινοτόμα εργαλεία για τη βιομηχανία του; Έχει φροντίσει για καλές διασυνδέσεις με τη δημιουργία ενός δικτύου επαφών στους επαγγελματικούς ή/και κοινωνικούς ιστοτόπους;

Οι συνηθέστερες απειλές για την δικτυακή φήμη μιας εταιρίας είναι η αρνητική κάλυψη από τα μέσα μαζική ενημέρωσης, τα παράπονα πελατών σε ιστοτόπους και η αποκάλυψη πληροφοριών για την εκ των έσω λειτουργία της επιχείρησης από πρώην υπαλλήλους. Η διασπορά αρνητικών ειδήσεων από δυσαρεστημένους πελάτες, νυν/πρώην εργαζόμενους μπορεί να αποτελέσει σημαντικό πλήγμα για την αξιοπιστία της επιχείρησης. Θα πρέπει να εντοπιστούν οι πηγές της αρνητικής δημοσιότητας και τον καθορισμό των επόμενων βημάτων ώστε να δωθούν απαντήσεις στα αρνητικά δημοσιεύματα. Αυτά μπορεί να

περιλαμβάνουν από απλές απαντήσεις στα αντίστοιχα μέσα προβολής μέχρι και την αναζήτηση των πηγών τους για προσφυγή στην νομική οδό σε πιο σοβαρές περιπτώσεις.

Η διαδικτυακή φήμη μπορεί να προσδιοριστεί και από την αντίληψη που έχει κανείς στο διαδίκτυο με βάση τις δημοσιεύσεις του και γενικότερα το ψηφιακό του αποτύπωμα. Ψηφιακά αποτυπώματα είναι οι πληροφορίες που συσσωρεύονται μέσω όλων των περιεχομένων που διαμοιράζονται στο διαδίκτυο και της σχετικής ανατροφοδότησης που προσφέρεται. Ελέγχοντας κάποιος το ψηφιακό του αποτύπωμα μπορεί να διαχειριστεί τη φήμη του στον ιστό.

Κοινές τεχνικές για διαδικτυακές δημόσιες σχέσεις αποτελούν η δημιουργία νέων περιεχομένων με θετικά σχόλια, η συμμετοχή στη σφαίρα του κοινωνικού διαδικτύου (μέσω φόρουμ, ιστολογιών, κοινωνικής δικτύωσης), η προώθηση των υπαρχόντων θετικών περιεχομένων και η οικοδόμηση ενός γενικά δυνατού κοινωνικού προφίλ. Πιο δύσκολες, αλλά παρόλα αυτά σχετικές τεχνικές, μπορούν να περιλαμβάνουν την αφαίρεση αρνητικού περιεχομένου, όπου αυτό είναι εφικτό. Μια πιθανή αναφορά στα κοινωνικά μέσα ενημέρωσης μπορεί να αποτελέσει απειλή, αν υπάρχει αρνητική κριτική, αλλά μπορεί, επίσης, να ωφελήσει, εάν αντιστραφούν αυτές οι επικρίσεις.

Το eBay ήταν μία από τις πρώτες διαδικτυακές εταιρείες που ασχολήθηκε με πληροφορίες που δημιουργούσαν οι καταναλωτές. Με τη χρήση των βαθμολογιών των διάφορων χρηστών, οι αγοραστές και οι πωλητές αποκτούσαν φήμη, γεγονός που βοηθούσε τους επόμενους χρήστες να αποφασίζουν σχετικά με τις αγοραπωλησίες τους. Η πρώτη εταιρεία, που δημιουργήθηκε και απεθυνόταν και σε πρόσωπα, όχι μόνο επιχειρήσεις, ήταν η ReputationDefender[5]. Το Παγκόσμιο Οικονομικό Φόρουμ αναγνώρισε την αυξανόμενη σημασία αυτού του τομέα, χαρίζοντας στη ReputationDefender τον τίτλο ως ενός από τους 31 Πρωτοπόρους Τεχνολογίας για το 2011. Τυπικές πηγές αναζήτησης για τους σκοπούς της διαχείρισης φήμης είναι τα κοινωνικά δίκτυα, οι ιστοσελίδες με απόψεις καταναλωτών ή σχόλια χρηστών, συνεργατικές ιστοσελίδες (όπως η Wikipedia), σελίδες ειδήσεων, forum συζητήσεων και ιστολόγια.

5.2 Δικτυακή φήμη και ελληνική πραγματικότητα

Υπάρχουν πολλές ελληνικές εταιρίες οι οποίες παρέχουν υπηρεσίες διαχείρισης φήμης. Ο τομέας αυτός έχει αρχίσει να αναπτύσσεται ραγδαία τα τελευταία χρόνια. Μια απλή αναζήτηση στο διαδίκτυο με σχετικούς όρους θα αποκαλύψει πολλές μικρές εταιρίες. Ωστόσο ο αριθμός τους δεν είναι ενδεικτικός της διείσδυσης που έχει αυτός ο τομέας. Η αγορά ακόμα είναι μικρή και τα σχετικά στοιχεία ελάχιστα. Οι εταιρίες αυτές απευθύνονται κυρίως σε επιχειρήσεις. Ένας κλάδος στον οποίο υπάρχει αυξανόμενο ενδιαφέρον για τέτοιες υπηρεσίες στην Ελλάδα είναι ο τουριστικός κλάδος. Πολλά ξενοδοχεία θέλουν να μάθουν τα σχόλια που γράφουν οι τουρίστες που καταλύουν σε αυτά, για τις υπηρεσίες που τους προσφέρουν. Επίσης, ο κλάδος της εστίασης έχει δείξει αντίστοιχο ενδιαφέρον, με μεγάλα εστιατόρια και αλυσίδες να παρακολουθούν το τι γράφεται για αυτούς.

Σε γενικές γραμμές όμως, η διαχείριση δικτυακής φήμης στην Ελλάδα είναι περιορισμένη καθότι οι περισσότερες εταιρίες δεν δείχνουν να έχουν αντιληφθεί την σημασία της. Ένας απλός τρόπος απόδειξης αυτού του γεγονότος είναι μια αναζήτηση σε μια μηχανή αναζήτησης (Google) για μεγάλες επώνυμες εταιρίες της Ελλάδας. Δοκιμάζοντας την επωνυμία Jumbo (αλυσίδα καταστημάτων με παιδικά παιχνίδια) το τρίτο αποτέλεσμα αναφέρεται σε καταγγελίες των εργαζομένων στην αλυσίδα, με τον τίτλο «*Το αίσχος των*

[5] http://www.reputation.com/

Jumbo». Επίσης, σε μια αναζήτηση για τα εστιατόρια γρήγορα φαγητού Goody's, το τέταρτο αποτέλεσμα είναι ένα βίντεο με τίτλο «*GOODYS Αγρινίου και σκουλίκι στην σαλάτα!!!Αισχος*». Αντίστοιχες αρνητικές δημοσιεύσεις μπορούν να βρεθούν για πολλές άλλες γνωστές εταιρίες.

Παρατηρούμε ότι εταιρίες που δαπανούν πολύ μεγάλα ποσά για να διαφημιστούν και να χτίσουν μια δυνατή επωνυμία στα παραδοσιακά μέσα ενημέρωσης, δε δίνουν καμία σημασία στην εικόνα τους στο διαδίκτυο. Ωστόσο, με την αλματώδη ανάπτυξη του διαδικτύου στην Ελλάδα, καθίσταται απαραίτητη η προστασία της διαδικτυακής υπόστασης των εταιριών. Μπορεί στο διαδίκτυο να είναι δύσκολο μια εταιρία να χτίσει μια καλή σχέση με τους πελάτες της, όμως όταν το καταφέρει αυτό, μπορεί να γίνει η κυρίαρχος στο χώρο της. Από την άλλη πλευρά, είναι πολύ εύκολο να καταποντιστεί η εταιρική της εικόνα από ένα απλό σχόλιο σε κάποιο blog ή από κάποιο σχόλιο σε forum. Ο λόγος είναι πως στον Ιστό οι πληροφορίες εξαπλώνονται ταχύτατα και οι καταναλωτές δηλώνουν άμεσα την γνώμη τους (αρνητική ή θετική) για ένα προϊόν. Επομένως, είναι ιδιαίτερα σημαντικό για μια εταιρία να καταλάβει την δυναμική που έχει η διαχείριση διαδικτυακής φήμης.

5.3 Λογισμικό διαχείρισης δικτυακής φήμης

Ο χειρισμός της δικτυακής φήμης μπορεί να γίνει είτε με την ανάθεση του σχετικού καθήκοντος σε μια εξειδικευμένη εταιρία (search marketeers) είτε με την απ' ευθείας χρήση κατάλληλου λογισμικού για την παρακολούθηση του παγκόσμιου ιστού. Στην ενότητα αυτή θα αναφέρουμε κάποια δημοφιλή εργαλεία που διατίθενται για αυτόν τον σκοπό.

5.3.1 Google Alerts

Το Google Alerts αποτελεί μια υπηρεσία παρακολούθησης περιεχομένου, που προσφέρει η μηχανή αναζήτησης της εταιρείας Google. Η διαφορά με την παραδοσιακή αναζήτηση είναι ότι ειδοποιεί αυτόματα τους χρήστες, όταν νέο περιεχόμενο (από ειδήσεις, διαδίκτυο, blogs, βίντεο, ομάδες συζήτησης) ταυτίζεται με μια σειρά από όρους αναζήτησης, που επιλέγονται από το χρήστη. Η ενημέρωση μπορεί να πραγματοποιηθεί μέσω ηλεκτρονικού ταχυδρομείου, μέσω web feed ή μέσω εμφάνισης στη σελίδα iGoogle του χρήστη. Τα αποτελέσματα, που παρέχονται, είναι μόνο από αναζητήσεις μέσω της μηχανής αναζήτησης Google. Επί του παρόντος, υπάρχουν έξι διαφορετικοί τύποι καταχωρήσεων. Ο χρήστης μπορεί να αναζητήσει στην κατηγορία για τα πάντα, όπου δε γίνεται συγκεκριμένος έλεγχος, στην κατηγορία των νέων, όπου το περιεχόμενο αναζητείται στα δέκα πρώτα αποτελέσματα του Google News και στην κατηγορία του Ιστού, όπου ξεχωρίζουν οι νέες ιστοσελίδες, που εμφανίζονται στην εικοσάδα των αποτελεσμάτων του Google Web Search. Επιπλέον, υπάρχουν η κατηγορία των blogs, δηλαδή από τη δεκάδα των αποτελεσμάτων του Google Blog Search, η κατηγορία των βίντεο από Google Video Search, και η κατηγορία των ομάδων από Google Groups.

Οι χρήστες έχουν τη δυνατότητα να καθορίζουν τη συχνότητα των ελέγχων για τα νέα αποτελέσματα. Οι τρεις επιλογές, που είναι διαθέσιμες, είναι «μία φορά την ημέρα», «μια φορά την εβδομάδα» ή «όπως συμβαίνει». Αυτό δε συνεπάγεται αναγκαστικά ότι τόσο συχνά θα έχει ο χρήστης ειδοποιήσεις. Θα πρέπει να υπάρχει και νέο υλικό, που να έχει ταυτιστεί με τους όρους της αναζήτησης. Τέλος, οι ειδοποιήσεις που λαμβάνονται είναι διαθέσιμες σε μορφή απλού κειμένου, αλλά και HTML.

5.3.2 reputationdefender

Το ReputationDefender είναι μια υπηρεσία που έχει σχεδιαστεί ειδικά για να προσφέρει τις υπηρεσίες του σε εταιρίες. Συνδυάζει προηγμένες τεχνολογίες και ζωντανή υποστήριξη από ένα άλλο έμπειρο εργαλείο, τους Συμβούλους Φήμης. Το εργαλείο βοηθάει τη διαδικτυακή φήμη του πελάτη αυξάνοντας τα θετικά περιεχόμενα και καταπολεμώντας άμεσα τα λανθασμένα ή παραπλανητικά αποτελέσματα για την επωνυμία του πελάτη. Για τη σωστή και γρήγορη λειτουργία του χρησιμοποιεί πέντε επιμέρους εργαλεία. Το VisibilityPlus αναζητεί δημοσιεύσεις σχετικές με όποιο προσωπικό προφίλ, έχει δημιουργήσει ο πελάτης. Το Profile Optimizer βελτιώνει τα διάφορα προφίλ του χρήστη, ώστε να μεγιστοποιηθεί η ικανότητά τους να εμφανίζονται σε αποτελέσματα αναζήτησης άλλων χρηστών. Το ImageMaker κατασκευάζει βιογραφία του πελάτη και μετά από έγκρισή του, τη δημοσιεύει ελεύθερα στο διαδίκτυο. Τέλος, το ConnectEdge και το PR Control προωθούν τα ήδη υπάρχοντα θετικά περιεχόμενα στο διαδικτυακό χώρο.

5.3.3 Trackur

Μια πολύ καλή υπηρεσία που ξεχωρίζει λόγω της ταχύτητας της και κάνει αναζήτηση σε κοινωνικά δίκτυα εύκολη υπόθεση είναι το Trackur [47]. Το εργαλείο είναι εξαιρετικά εύκολο στην εγκατάσταση και τη χρήση του και προσφέρει πολλές επιλογές. Ο χρήστης μπορεί να το χρησιμοποιήσει δωρεάν για μια ορισμένη περίοδο. Η αναζήτηση γίνεται με την εισαγωγή λέξεων – κλειδιών που θέλει να παρακολουθήσει, τα αποτελέσματα φιλτράρονται και επιστρέφονται οι ζητούμενες αναφορές. Το αποτέλεσμα μπορεί να κρατηθεί για μελλοντική παρακολούθηση ή ο χρήστης να επισκεφτεί άμεσα την πηγή για να πάρει τα κατάλληλα μέτρα. Ακόμη, υπάρχει η δυνατότητα να διαμοιραστεί το αποτέλεσμα με άλλο χρήστη. Μια σημαντική λειτουργία είναι η βαθμολόγηση των πηγών, δηλαδή ο χρήστης δε γνωρίζει μόνο ποια είναι η πηγή που αναφέρει το όνομά του ή την επωνυμία της εταιρίας του αλλά και κατά πόσο αυτή η πηγή επηρεάζει την κοινωνία του διαδικτύου. Η βαθμολόγηση των πηγών γίνεται από τους ίδιους τους χρήστες σε μια κλίμακα επί τις εκατό. Πέρα από αυτό, κάθε στοιχείο έχει κατάλληλη ετικέτα σχετικά με το αν είναι θετικό, αρνητικό ή ουδέτερο. Έτσι ο χρήστης ξέρει ποιος λέει θετικά πράγματα για αυτόν και ποιος επιτίθεται στη φήμη του. Τα βασικά χαρακτηριστικά που κάνουν το εργαλείο ιδιαίτερο είναι μεταξύ άλλων η απευθείας εισαγωγή των αποτελεσμάτων σε κάποια βάση δεδομένων του χρήστη.

5.3.4 Search.twitter.com

Ένα ακόμη εργαλείο που παρατίθεται είναι μια πρόσφατη εφαρμογή της ήδη επιτυχημένης ιστοσελίδας search.twitter.com. Η μέχρι τότε εφαρμογή αφορούσε την ενημέρωση σχετικά με αξιοσημείωτα νέα και πρόσωπα που ενδιέφεραν τον χρήστη προσωπικά, αλλά η ανάγκη για πιο γενικές πληροφορίες οδήγησε στη δημιουργία του search.twitter.com. Εξυπηρετεί ανάγκες σχετικά με την αναζήτηση, το φιλτράρισμα και της αλληλεπίδρασης με τον όγκο των ειδήσεων και των πληροφοριών που διαβιβάζονται στο twitter κάθε δευτερόλεπτο. Βοηθάει με τις υπηρεσίες του να φιλτράρεται όλος ο όγκος των πληροφοριών σε πραγματικό χρόνο. Ο χρήστης μπορεί να τροποποιήσει τα ερωτήματά του χρησιμοποιώντας διάφορους τελεστές σύνθετης αναζήτησης, ενώ όλα τα αποτελέσματα αναζήτησης είναι διαθέσιμα μέσω API (του standard Atom και του JSON) προς χρήση αναγνωστών.

5.3.5 Qualia

Η εταιρία Qualia με το προϊόν Aino προσφέρει καινοτόμες υπηρεσίες διαχείρισης φήμης [48]. Δημιουργήθηκε στην Αθήνα το 2004 σαν υποτμήμα του Ινστιτούτου Γλώσσας και Επεξεργασίας Λόγου. Αποτελείται από ειδικούς σε τεχνικά και επιστημονικά πεδία ευρείας κλίμακας, που προσπαθούν να παράγουν καινοτόμες υπηρεσίες. Θεωρώντας ότι ο

υπερβολικός φόρτος πληροφοριών μπορεί να είναι αρνητικός για άτομα ή εταιρίες, προσφέρουν σύγχρονες τεχνολογίες για να βοηθήσουν τους πελάτες τους να κατανοήσουν τι λένε για το πρόσωπό τους ή τις επιχειρήσεις τους. Το προϊόν Aino οροσφέρει παρακολούθηση της διαδικτυακής φήμης και έρευνα σε τηλεόραση, ραδιόφωνο και στο διαδίκτυο, συνήθως κεντρικοποιημένα ως προς την επωνυμία του πελάτη. Αναζητητούν τη δημόσια εικόντα της επιχείρησης και εντοπίζουν συσχετίσεις μεταξύ κοινού και προϊόντος, επικοινωνίας και Μ.Μ.Ε., επωνυμιών και εμπειριών. Η διαφορά σε σχέση με τα προηγούμενα προϊόντα είναι ότι δεν παρακολουθεί μόνο δικτυακές πηγές. Προσφέρεται σαν αυτόνομο προϊόν ή με την μορφή μιας συνδρομητικής υπηρεσίας.

5.3.6 onlinereputation

Η ελληνική αυτή ιστοσελίδα ακολουθεί τρία βασικά βήματα για τη διαχείριση της διαδικτυακής φήμης [49]. Σε πρώτο επίπεδο, παρακολουθείται και καταγράφεται η όποια αναφορά στο διαδίκτυο για την επωνυμία του πελάτη. Ξεπερνώντας την ταχύτητα και την πολυπλοκότητα του διαδικτύου, με ευέλικτους τρόπους ανακτάται όποιο σχόλιο έχει γίνει σε κοινωνικό μέσο, ιστολόγιο ή φόρουμ. Στη δεύτερη φάση γίνεται η ανάλυση των πληροφοριών, που καταγράφτηκαν ώστε να ενημερωθεί ο πελάτης για το πώς επηρεάζεται η επωνυμία του και κατ' επέκταση η φήμη του από αυτές. Πραγματοποιείται διαχωρισμός των αποτελεσμάτων σε θετικά και αρνητικά, ενώ στη συνέχεια θέτονται επίπεδα επικινδυνότητας σύμφωνα με τη δυναμική των πληροφοριών που συλλέχθηκαν. Η καινοτομία του συστήματος βρίσκεται στο τρίτο βήμα. Αφού έχουν εντοπιστεί τα κακόβουλα σχόλια, το onlinereputation επεμβαίνει με συμμετοχή στο διαδικτυακό διάλογο ώστε να βελτιωθεί η φήμη. Αν είναι κατασκευασμένα ώστε να βλάψουν τη φήμη του πελάτη, τότε η επέμβαση είναι άμεση. Ειδάλλως, η επέμβαση είναι απλά επηρεάζοντας τις διαδικτυακές κατευθύνσεις για την επωνυμία του.

5.3.7 iSieve

Η εταιρία i-sieve Technologies Ltd παρέχει υπηρεσίες αυτόματης ανάλυσης και χαρακτηρισμού περιεχομένου στο διαδίκτυο, βασισμένη σε καινοτόμα τεχνολογία διήθησης πληροφορίας [50]. Αντικείμενο της i-sieve είναι η εφαρμογή μεθόδων τεχνητής νοημοσύνης στην ανάλυση ιστοσελίδων και δικτυακού περιεχομένου. Οι υπηρεσίες που παρέχει η εταιρία περιλαμβάνουν την ανάλυση της φήμης μιας μάρκας (brand reputation), την επίδραση μιας εκστρατείας μάρκετινγκ ενός προϊόντος (marketing campaign impact) και γενικότερα ολοκληρωμένες μετρήσεις έκθεσης σε ηλεκτρονικά μέσα για εταιρίες και τα προϊόντα τους. Ο παγκόσμιος ιστός ανιχνεύεται με σκοπό να προσδιοριστούν οι παρατηρήσεις των χρηστών σε πηγές όπως κοινωνικά δίκτυα, ιστολόγια (blogs), forums, chat-rooms καθώς και βίντεο. Πιο συγκεκριμένα, ο πελάτης θέτει κάποια ερωτήματα (π.χ. σχολιασμός μιας εταιρίας στο web, μέτρηση θετικών και αρνητικών απόψεων για μια ταινία). Σκοπός είναι να εξαχθούν σχόλια που σχετίζονται με το ερώτημα που έχει τεθεί. Τα σχόλια αυτά θα χαρακτηριστούν ώς θετικά, αρνητικά ή «άσχετα αλλά ενδιαφέροντα».

Το πρώτο στάδιο σε κάθε έργο είναι η εκπαίδευση του αλγορίθμου μηχανικής μάθησης του συστήματος με σκοπό αυτό να μπορεί να απαντήσει συγκεκριμένα στο ερώτημα που έχουν τεθεί. Η διαδικασία αυτή είναι χειρονακτική. Η ομάδα της i-sieve παραθέτει παραδείγματα σχετικών σχολίων στο λογισμικό τα οποία στη συνέχεια αναλύονται, ώστε αυτό να είναι σε θέση να διαχωρίζει τα σημαντικά σχόλια από τα ασήμαντα. Τα παραδείγματα που χρησιμοποιούνται για εκπαίδευση είναι ομαδοποιημένα με βάση την θεματολογία τους. Για παράδειγμα, μπορεί να υπάρξει μια ομάδα παρατηρήσεων σχετικά με την μουσική που χρησιμοποιείται σε μια διαφήμιση, ένα άλλο σύνολο σχετικά με την αποτελεσματικότητα του προϊόντος και ούτω καθεξής.

Αυτή η διαδικασία μάθησης διαρκεί συνήθως μεταξύ 2 ημερών και μίας εβδομάδας. Μόλις οι υπολογιστές έχουν μάθει τι πρέπει να αναγνωρίζουν, η υπόλοιπη διαδικασία αυτοματοποιείται. Συνήθως αναζητούνται στον παγκόσμιο ιστό περισσότερο από μισό εκατομμύριο πηγές και ταξινομούνται αναλόγως. Η ακρίβεια που επιτυγχάνει το σύστημα είναι της τάξης του 90%. Όταν το λογισμικό παραμετροποιηθεί, συντονισθεί και ξεκινήσει να αναζητά πηγές από τον παγκόσμιο ιστό, γίνονται κάποιοι έλεγχοι σε κάθε γύρο αποτελεσμάτων. Αν το σύστημα βρει πηγές που σχετίζονται αλλά δεν ταιριάζουν με το προκαθορισμένο ερώτημα, ειδοποιεί την ομάδα και αυτή είναι σε θέση να καθορίσει μια νέα συστάδα, ακόμα και εάν αυτό δεν είχε αρχικά ζητηθεί. Έτσι μπορεί να προκύψει χρήσιμη γνώση που να σχετίζεται με το προϊόν/ υπηρεσία και που δεν είχε προσδιορισθεί αρχικά.

Με την χρήση αυτοματοποιημένων συστημάτων εντοπισμού ιστοσελίδων (web crawlers) και web spiders είναι σε θέση να προσδιοριστούν τα σχετικά με το ερώτημα σχόλια από το διαδίκτυο. Ο αριθμός των αναφορών σε ένα συγκεκριμένο σχόλιο καθώς και ο αριθμός των φορών που το σχόλιο αυτό επαναλαμβάνεται, αποτελεί έναν παράγοντα της σχετικής σημασίας και επιρροής του. Η επιρροή ενός ορισμένου προσώπου ή σχολίου περιγράφεται με τον όρο buzz factor. Ο εντοπισμός των κυριότερων πηγών επιρροής είναι μέσα στους στόχους της ανάλυσης.

Η διαδικασία αποτελείται από 4 βήματα:

1. **Εκπαίδευση του αλγορίθμου** Filter X. Η εκπαίδευση γίνεται από τους ειδικούς της εταιρίας οι οποίοι εντοπίζουν σχετικά παραδείγματα από σχόλια χρηστών στον παγκόσμιο ιστό. Η ομάδα ομαδοποιεί (clusters) τα σχόλια και δημιουργεί αντίστοιχα δεδομένα εκπαίδευσης.

2. **Συλλογή δεδομένων** από τον παγκόσμιο ιστό. Το στάδιο αυτό περιλαμβάνει την χρήση web crawlers και web spiders για την εξαγωγή των σχολίων από μια πληθώρα πηγών (κείμενο, βίντεο και ήχο).

3. Το **φιλτράρισμα** του περιεχομένου (content filtering) επιτυγχάνεται από τον αλγόριθμο FilterX. Ο FilterX είναι ένας αλγόριθμος μηχανικής μάθησης ο οποιός χρησιμοποιώντας τεχνικές επεξεργασίας φυσικής γλώσσας ταξινομεί τα σχόλια. Χρησιμοποιείται ανεξάρτητα και σε άλλα προϊόντα όπως φίλτρα κατά της πορνογραφίας. Ο αλγόριθμος FilterX δημιουργεί μια οντολογία για κάθε ερώτημα που του τίθεται. Με βάση αυτήν την οντολογία ταξινομεί τα επιμέρους σχόλια. Η χρήση οντολογίας αντί για προκαθορισμένες κατηγορίες και λίστες που χρησιμοποιούν άλλα προγράμματα, επιτρέπει χρήση του σε κάθε γλώσσα και τύπο περιεχομένου. Η εταιρία έχει αναπτύξει παράλληλα και το σύστημα BuzzSense, σύστημα στοχευμένης αναζήτησης πληροφορίας στο Διαδίκτυο με ειδίκευση στην επιχειρηματική πληροφόρηση (business intelligence).

4. **Ανάλυση των ευρημάτων**. Ακολουθεί η ανάλυση των ευρημάτων και η ερμηνεία των αποτελεσμάτων από τους ειδικούς της iSieve. Η τελική αναφορά παρουσιάζει μεταξύ άλλων κάθε σχόλιο ομαδοποιημένο, την πηγή του (url), τον παράγοντα επιρροής του (buzz factor) και τυχόν παρατηρήσεις την ομάδας της i-sieve.

5.4 Λογισμικό εξόρυξης γνώμης

Στο σημείο αυτό θα παρουσιάσουμε λογισμικό εξόρυξης γνώμης το οποίο διαφέρει από τα προηγούμενα στο ότι δεν εστιάζει στην διαχείριση φήμης, δεν παρέχει δηλαδή ολοκληρωμένα εργαλεία για την διαχείριση της. Μας ενδιαφέρει περισσότερο η κατάταξη των γνωμών σε κατηγορίες και η ανάλυση τους, παρά η απλή εύρεση θετικών/ αρνητικών δημοσιευμάτων σε ένα θέμα. Περιγράφουμε ορισμένα από τα πιο γνωστά προϊόντα εξόρυξης γνώμης, παραθέτοντας πληροφορίες για τους αλγόριθμους που χρησιμοποιούν, όπου αυτές είναι διαθέσιμες.

5.4.1 MOBI

Το MOBI (Mass Opinion Business Intelligence) είναι ένα προϊόν εξόρυξης γνώμης της εταιρείας Wise Window [51]. Συνδυάζει τεχνικές τεχνητής νοημοσύνης, επεξεργασίας φυσικής γλώσσας, έναν πατενταρισμένο web crawler και μια προηγμένη πιθανολογική μηχανή για να παρακολουθεί σε πραγματικό χρόνο συζητήσεις που γίνονται στο διαδίκτυο. Διατρέχει πολλές δημοφιλείς σελίδες όπως το Twitter, facebook, YouTube, TripAdvisor, IMDB, καταστήματα, ιστολόγια, forum κλπ. Σκοπός είναι να αναλύσει συντακτικά (parse) ότι λέγεται για ένα πρόσωπο, προϊόν ή επωνυμία και να εξάγει το εννοιολογικό πλαίσιο που αναδύεται. Η ανάλυση αυτών των μοτίβων μπορεί να αποκαλύψει πολύτιμη γνώση στους ενδιαφερόμενους και να τους προσδώσει έτσι ένα ανταγωνιστικό πλεονέκτημα σε σχέση με τον ανταγωνισμό.

Βασίζεται σε μια στατιστική τεχνική γνωστή σαν ανάλυση συστάδων (cluster analysis). Η ανάλυση συστάδων χρησιμοποιείται εκτεταμένα στην επιστημονική έρευνα, όπου υπάρχει ανάγκη ταξινόμησης και κατάταξης των αντικειμένων μελέτης σε ομάδες. Έχει σκοπό να διαχωρίσει το σύνολο των παρατηρήσεων σε φυσικές ομάδες, έτσι ώστε τα μέλη κάθε ομάδας να είναι όσο το δυνατό πιο όμοια μεταξύ τους, ενώ τα μέλη διαφορετικών ομάδων να είναι όσο το δυνατό πιο ανόμοια. Στην περίπτωση του προϊόντος αυτό σημαίνει ανάγνωση κειμένου, εντοπισμός λέξεων και ομαδοποίηση τους. Από την ανάλυση ομάδων προκύπτουν ομάδες λέξεων. Η σημασία τους υπολογίζεται ανάλογα με την συχνότητα εμφάνισής τους σε ένα κείμενο. Το κλειδί για την επιτυχή ανάλυση συστάδων είναι να περιέχουν στενά σχετιζόμενο περιεχόμενο. Όσο πιο ισχυρές είναι οι σχέσεις ανάμεσα σε λέξεις του ίδιου εννοιολογικού πλαισίου, τόσο πιο εύκολο είναι να εντοπιστούν τα πρότυπα και οι αλλαγές σε αυτά ή ακόμα και ο σχηματισμός νέων προτύπων. Για τον λόγο αυτό το MOBI επεξεργάζεται τις κατηγορίες που προκύπτουν κάθετα, δηλαδή σε ευρύτερες κατηγορίες όπως μουσική, ταινίες, ηλεκτρονικά, αυτοκίνητα και άλλες.

Ο προσδιορισμός των σχέσων μεταξύ των λέξεων σε κάθε μία από αυτές τις κάθετες κατηγορίες απαιτεί μια *ταξονομία*. Η ταξινομία είναι ένα σχήμα ταξινόμησης στο οποίο προσδιορίζονται η ιεραρχία των σχέσεων και των γνωρισμάτων. Για παράδειγμα, η ταξινομία ηλεκτρονικών συσκεών, θα μπορούσε να περιλαμβάνει μία ιεραρχία από κατηγορίες προιόντων (ψηφιακές φωτογραφικές μηχανές, τηλεοράσεις, βιντεοκάμερες, ηλεκτρονικούς υπολογιστές), τύπους προιόντων σε κάθε κατηγορία (laptop, επιτραπέζιους, netbooks, ταμπλέτες), μάρκες προιόντων (Apple, Samsung, Acer, Dell, HP) και χαρακτηριστικά (μέγεθος, ταχύτητα επεξεργαστή, μέγεθος μνήμης και χωρητικότητα δίσκου). Το MOBI εκπαιδεύεται στο να κάνει συσχετίσεις μεταξύ των ομάδων και των χαρακτηριστικών που προκύπτουν, παρά στο να εκτελεί απλά αναζητήσεις. Οι τεχνικοί της εταιρείας Wise Window εισάγουν στο MOBI μία ταξινομία που είναι διαθέσιμη δημόσια (π.χ. το Amazon για ηλεκτρονικές συσκευές, το IMDB για ταινίες και το TripAdvisor για ταξίδια) αντί να επιβαρύνονται οι πελάτες τους με αυτό το καθήκον. Το MOBI φιλτράρει αυτές τις ταξινομίες

ανά τακτά χρονικά διαστήματα, αναλύοντας τις αντιστοιχίες μεταξύ των δεδομένων των χρηστών με σκοπό να διαπιστώσει με ποιον τρόπο τα προιόντα, οι άνθρωποι ή οι επωνυμίες προϊόντων γίνονται αντικείμενο συζήτησης.

5.4.2 SAS Text Analytics

Το SAS Text Analytics βασίζεται στη στατιστική ανάλυση και την στατιστική μοντελοποίηση κειμένων [52]. Για την αξιολόγηση των συναισθημάτων που εκφράζονται σε σύνολα κειμένων, το λογισμικό παρέχει προεπιλεγμένες παραμέτρους που έχουν προσδιοριστεί εκ των προτέρων, καθώς και τη δυνατότητα ο χρήστης να αλλάζει τις ρυθμίσεις του προγράμματος για να προσδιορίσει καλύτερα τα στοιχεία κλειδιά. Λειτουργεί προσδιορίζοντας γλωσσικούς κανόνες. Αφήνει τους ειδικούς σε θέματα περιεχομένου να προσδιορίσουν τα θέματα που θα εξεταστούν κατά τη διάρκεια της ανασκόπησης των συναισθημάτων μέσω boolean συντελεστών συμφραζομένων. Παρέχει τη δυνατότητα να συνδυαστούν η στατιστική αυστηρότητα και η πείρα των ειδικών για να προσδιοριστούν τα χαρακτηριστικά γνωρίσματα και τα στοιχεία εκείνα τα οποία όταν συνδυαστούν, θα δώσουν πιο ακριβείς αξιολογήσεις συναισθημάτων. Επιπλέον παρέχει τη δυνατότητα να ανατεθούν βάρη στα ταιριάσματα που θα προκύψουν, ώστε να προσδιοριστεί το βέλτιστο ταίριασμα για οποιοδήποτε αρχείο ή πηγή. Υποστηρίζει πολύπλοκους γλωσσικούς κανόνες για να συνδυαστούν όροι, εκφράσεις ή μέρη του λόγου, να προσδιοριστεί η απόσταση από έννοιες ή να εντοπίσει την ύπαρξη εννοιών.

Μερικά ακόμα χαρακτηριστικά του SAS Text Analytics είναι ότι περιλαμβάνει σύστημα διαχείρισης ώστε να δημιουργήθουν κανόνες συναισθημάτων. Δίνει τη δυνατότητα να χρησιμοποιηθούν φράσεις που θα εκφράσουν έννοιες σε συνδυασμό με κανόνες άλγεβρας Boole και άλλα γλωσσικά εργαλεία. Επιτρέπει βελτιώσεις ή/και αλλαγές παράγωντας πολλά μοντέλα. Παρέχει τη δυνατότητα να συνταχθούν και να ελεγχθούν οι αλλαγές που έγιναν στα μοντέλα. Παρακολουθεί τα αποτελέσματα ανά τακτά χρονικά διαστήματα και στη συνέχεια ενημερώνει τα μοντέλα για τυχόν αλλαγές. Ο χρήστης του SAS Text Analytics μπορεί να πραγματοποιηεί συγκρίσεις μεταξύ διαφορετικών συνόλων εκπαίδευσης ώστε να επιτευχθούν συνεχείς βελτιώσεις των αξιολογήσεων συναισθημάτων.

Το λογισμικό χρησιμοποιεί επιπλέον μία μηχανή αναζήτησης που λειτουργεί πολυνηματικά και με κατανεμημένο τρόπο ώστε να μεγιστοποιήσει την υποστήριξη αναζητήσεων μεγάλης κλίμακας. Χρησιμοποιεί τεχνογνωσία από την επιστήμη της γλωσσολογίας για να εξάγει URL. Οι αναζητήσεις στο διαδίκτο μπορούν να διακοπούν και να ανακεφαλαιωθούν. Παρέχει αυτόματη διαγραφή διπλοτύπων. Προσαρμόζεται για να υποστηρίξει τους επιθυμητούς περιορισμούς κατά την αναζήτηση (π.χ. σε συγκεκριμένα format αρχείων, σε συγκεκριμένους servers, να περιορίσουν το πεδίο αναζήτησης ή να προσδιορίσουν σε βάθος αναζήτηση). Οι αναζητήσεις μπορούν να προσδιοριστούν ώστε να αλλάζουν τμηματικά με βάση προηγούμενες αναζητήσεις.

5.4.3 BuzzMetrics

Το προϊόν BuzzMetrics είναι πλήρως προσαρμόσιμο ώστε να επιτρέπει τους επιχειρηματίες να παρακολουθούν και να αναλύουν χωρίς κόπο οτιδήποτε λέγεται διαδικτυακά για τα προιόντα τους ή την εταιρία τους μέσα από μία ευρεία γκάμα από μέσα ενημέρωσης που προέρχονται από απλούς χρήστες [53]. Το BuzzMetrics αναλύει περιεχόμενο και μηνύματα από περισσότερα από 100 εκατομμύρια ιστολόγια, ομάδες χρηστών και κοινωνικά δίκτυα παγκοσμίως. Τρία είδη ανάλυσης επιτρέπουν γρήγορες αναζητήσεις μέσα στον τεράστιο όγκο των διαθέσιμων συζητήσεων στον παγκόσμιο ιστό και

τη σύνταξη αναφορών ή την δημιουργία ειδοποιήσεων. Η κάλυψη των μέσων που παρέχει περιλαμβάνει:

- Ευρείας γκάμας πηγές μέσων ενημέρωσης που έχουν δημιουργηθεί από απλούς χρήστες (Consumer-generated media, CGM). Αυτά τα μέσα ενημέρωσης περιλαμβάνουν ιστολόγια, ομάδες συζητήσεων και κοινωνικά δίκτυα.
- Δυνατότητες υποστήριξης πολλών γλωσσών.
- Κάλυψη video και μικρο-ιστολόγια: Twitter, YouTube.
- Πιο παραδοσιακές πηγές μέσων ενημέρωσης: Διαδικτυακό περιεχόμενο εφημερίδων, περιοδικών και άλλες.
- Τμηματοποίηση των περιοχών από τις οποίες προέρχονται οι πηγές των δεδομένων.

Το BuzzMetrics παραδίδει πάνω από 30 κατανοητές αναφορές (συμπεριλαμβανομένων και των αυτόματα παραγόμενων αναλύσεων συναισθήματος και τοπικής διαχείρισης) που μπορούν να ενημερωθούν κατόπιν συγκεκριμένου χρονοδιαγράμματος ή να παραχθούν άμεσα. Παραδίδει στο χρήστη βαθμολογίες συναισθημάτων χρησιμοποιώντας λεξικά ανάλυσης συναισθήματος ευρείας χρήσης με σκοπό τον προσδιορισμό των τάσεων και την αξιολόγηση επιδόσεων. Επιδιώκει να προσδιορίσει τον τόνο μίας συζήτησης ώστε να εκτιμήσει το κατά πόσο είναι θετική, αρνητική ή ουδέτερη. Επιπλέον παρακολουθεί και εντοπίζει σημαντικά άρθρα, ιστολόγια, video και δημοσιεύσεις που σχετίζονται περισσότερο με το προιόν που ενδιαφέρει τον πελάτη.

5.4.4 BlogPulse

Το BlogPulse δουλεύει όπως μία τυπική μηχανή αναζήτησης, μόνο που έχει προγραμματιστεί να αναζητεί δεδομένα μόνο από ιστολόγια [54]. Είναι ένα αυτόματο σύστημα αναζήτησης των τάσεων της αγοράς μέσα από ιστολόγια. Αναλύει και παραδίδει αναφορές καθημερινά από την μπλογκόσφαιρα. Το BlogPulse παρέχει τέσσερεις βασικές υπηρεσίες:

- **Υπηρεσία αναζήτησης**. Η αναζήτηση που εκτελεί το BlogPulse επιτρέπει να εκτελεστεί η αναζήτηση με βάση λέξεις κλειδιά ή ακόμα και με URI.
- **Υπηρεσία εντοπισμού συζητήσεων**. Καθώς η μπλογκόσφαιρα εξελίσσεται, οι συζητήσεις γίνονται με τρόπο κατανεμημένο και αποκεντροποιημένο σε εκατομμύρια ιστολόγια και άλλους ιστότοπους. Μία βασική αρχή του εντοπισμού συζητήσεων είναι να εντοπίζει τις αναρτήσεις εκείνες ανάμεσα στις συζητήσεις που επηρεάζουν περισσότερο την κοινή γνώμη.
- **Υπηρεσία εντοπισμού των τάσεων της αγοράς**. Επιτρέπει στον χρήστη να δημιουργεί γραφήματα που εντοπίζουν τάσεις ανά τακτά χρονικά διαστήματα βασιζόμενος σε λέξεις κλειδιά, φράσεις ή συνδέσμους. Επιπλέον, δίνει τη δυνατότητα να συγκρίνει μεμονωμένα τους συνδέσμους που έδωσε σαν αποτέλεσμα μία αναζήτηση ή να χρησιμοποιήσει και τα τρία παραπάνω πεδία (λέξεις κλειδιά, φράσεις και συνδέσμους) για να συγκρίνει τους συνδέσμους μεταξύ τους. Η υπηρεσία εντοπισμού τάσεων της αγοράς δίνει ένα γράφημα των αναρτήσεων που σχετίζονται με λέξεις κλειδιά ή URI. Μπορεί επίσης να συγκρίνει διάφορα αντικείμενα μέσα σε ένα γράφημα. Περιλαμβάνει επίσης και τάσεις υψηλού ενδιαφέροντος που αφορούν ειδήσεις και άλλες συζητήσεις που ενδιαφέρουν τον πελάτη.

- **Υπηρεσία δημιουργίας προφίλ**. Δίνει πιο εις βάθος πληροφορίες για συγκεκριμένα ιστολόγια, συμπεριλαμβανομένης μίας ανασκόπησης σε αναρτήσεις, αναφορές, τάσεις, πηγές και συσχέτισης με άλλα ιστολόγια.

5.4.5 Sentiment

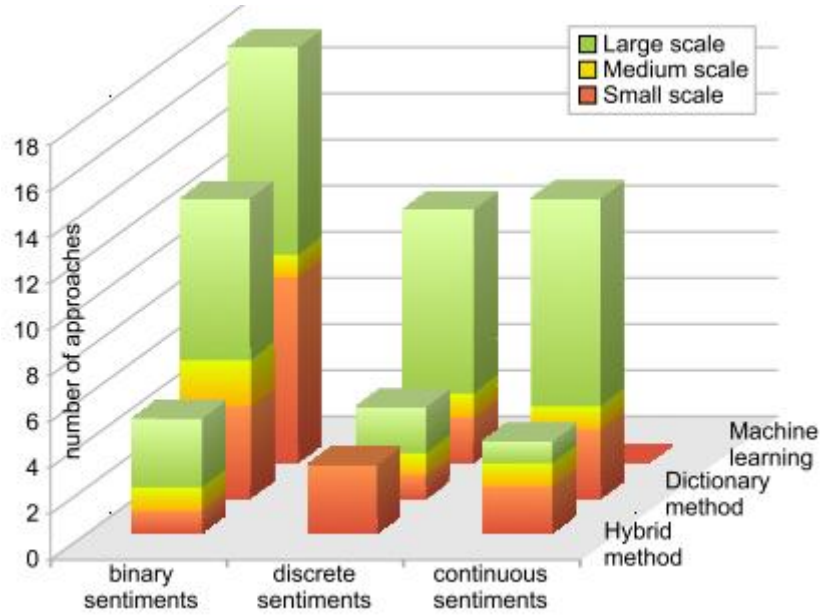
Η Infonic είναι μία Αγγλική εταιρεία που ήταν γνωστή στο παρελθόν ως Corpora [55]. Το προιόν τους για ανάλυση συναισθήματος με την ονομασία Sentiment, χρησιμοποιείται στον Dow Jones και σε μια μηχανή αναζήτησης του πρακτορείου ειδήσεων Reuters (Reuters NewsScope Sentiment Engine). Το Sentiment χρησιμοποιεί τεχνολογία επεξεργασίας φυσικής γλώσσας για την ανάλυση ειδήσεων σε ηλεκτρονική μορφή. Η ανάλυση δίνει σαν αποτέλεσμα μια αξιολόγηση (θετική/ αρνητική/ ουδέτερη). Το λογισμικό επιτρέπει στους επιχειρηματίες την αξιολόγηση ειδήσεων που αφορούν χιλιάδες εταιρίες, εύκολα και γρήγορα. Στα πλεονεκτήματα του συγκαταλέγεται το ότι ομαδοποιεί τα αποτελέσματα (ειδήσεις κλπ) σε πραγματικό χρόνο. Αυτό επιτρέπει στους επενδυτές να λαμβάνουν γρήγορα υπόψιν τους τα νέα διεθνή γεγονότα και έτσι να προσαρμόζουν αντίστοιχα τις αποφάσεις τους, γεγονός που τους προσδίδει ένα ανταγωνιστικό πλεονέκτημα.

6. Συμπεράσματα

Τα τελευταία χρόνια έχουμε γίνει μάρτυρες ενός αυξανόμενου ενδιαφέροντος για την επεξεργασία και ανάλυση αδόμητων δεδομένων, όπως το κείμενο που δημοσιεύουν οι χρήστες στο διαδίκτυο. Ο πλούτος της πληροφορίας αυτού του περιεχομένου καθιστά το έργο της επεξεργασίας και ανάλυσης απαραίτητο για να προκειμένου να αξιοποιηθούν όλες αυτές οι διαθέσιμες πληροφορίες και να παραχθεί νέα γνώση.

Στην εργασία αυτή κάναμε μια βιβλιογραφική επισκόπηση μιας ειδικής κατηγορίας των αλγορίθμων εξόρυξης γνώσης, της εξόρυξης γνώμης. Πρόκειται για έναν τομέα που άρχισε να αναπτύσσεται τα τελευταία χρόνια και τέθηκε στο επίκεντρο του ενδιαφέροντος λόγω των πρακτικών εφαρμογών του και της προσδοκίας να ανακαλύψη χρήσιμα και αξιοποιήσιμα μοτίβα (pattern) από αδόμητα δεδομένα του διαδικτύου. Εξετάσαμε τις πιο γνωστές προσεγγίσεις που αφορούν τα προβλήματα της εξόρυξης γνώμης. Αυτά έχουν αναδειχθεί ως ένας σημαντικός τομέας της εξόρυξης δεδομένων από τον παγκόσμιο ιστό και οι τάσεις δείχνουν μια αυξανόμενη συμμετοχή της ερευνητικές κοινότητας, μαζί με μια προσπάθεια για πιο εξελιγμένους και ισχυρούς αλγόριθμους.

Οι αλγόριθμοι αυτοί ανήκουν σε τρεις κύριους τύπους. Στο σχήμα 18 απεικονίζουμε την κατανομή των δημοσιεύσεων μεταξύ των πιο δημοφιλών τύπων αλγορίθμων (μηχανικής μάθησης, προσέγγιση βασισμένη σε λεξικό και υβριδικές προσεγγίσεις) και της αναπαράστασης των τιμών συναισθήματος.



Σχήμα 18: Ο αριθμός των αλγορίθμων ανάλογα με την αναπαράσταση του συναισθήματος, την αλγοριθμική προσέγγιση και την επεκτασιμότητα της μεθόδου

Όπως έχουμε αναφέρει στο κεφάλαιο 2, οι τιμές συναισθήματος μπορεί να είναι δυαδικές (θετικές/ αρνητικές), διακριτές (π.χ. -1, 0, 1) ή συνεχείς σε μια κλίμακα (π.χ. από 1 – 10). Παρατηρούμε ότι η πλειονότητα των δημοσιεύσεων χρησιμοποιούν μεθόδους βασισμένες σε μηχανική μάθηση. Ακολουθούν οι προσεγγίσεις βασισμένες σε λεξικό. Στην κατηγορία αυτή συμπεριελήφθησαν και οι στατιστικές και σημασιολογικές προσεγγίσεις. Οι υβριδικές

μέθοδοι που συνδυάζουν τις παραπάνω προσεγγίσεις (συνήθως σαν ένας συνδυασμός των προσεγγίσεων λεξικού με εργαλεία επεξεργασία φυσικής γλώσσας) δεν είναι ακόμα δημοφιλείς, πιθανότατα λόγω της πολυπλοκότητας τους.

Όσον αφορά τον τρόπο αναπαράστασης των τιμών συναισθήματος, οι περισσότερες έρευνες χρησιμοποιούν την δυαδική αναπαράσταση. Ωστόσο, οι άλλοι δύο τρόποι αναπαράστασης έχουν κερδίσει δημοτικότητα, δεδομένου ότι προσφέρουν πιο λεπτομερή ανάλυση και κατηγοριοποίηση. Ο σχετικά χαμηλός αριθμός δημοσιεύσεων που χρησιμοποιεί διακριτές τιμές συναισθήματος στις υβριδικές μεθόδους και σε εκείνες που βασίζονται σε λεξικό, μπορεί να εξηγηθεί από την διαθεσιμότητα των συνεχών αναπαραστάσεων των τιμών συναισθήματος που προσφέρουν μεγαλύτερη ακρίβεια. Αυτές οι μελέτες χρησιμοποιούν είτε δυαδική είτε συνεχή αναπαράσταση, ανάλογα με τον σκοπό τους. Από την άλλη, η συνεχής αναπαράσταση συναισθήματος δεν προτιμάται από τους αλγορίθμους ταξινόμησης, κάνοντας την μια σπάνια επιλογή για τις προσεγγίσεις που βασίζονται στην μηχανική μάθηση.

Τα χρώματα σε κάθε στήλη της εικόνας αντιστοιχούν στον αριθμό των αλγορίθμων που είναι σε θέση να χρησιμοποιηθούν σε μεγάλα, μεσαίας και μικρής κλίμακας σύνολα δεδομένων. Αυτό σχετίζεται άμεσα με την πολυπλοκότητα των προτεινόμενων αλγορίθμων. Έτσι, οι αλγόριθμοι που μπορούν να αξιοποιηθούν μόνο σε εποπτευόμενη λειτουργία, δεν μπορούν να κλιμακώσουν με την αύξηση του μεγέθους των συνόλων δεδομένων. Από το γράφημα φαίνεται ότι υπάρχουν κυρίως δύο προσεγγίσεις που είναι ενδεδειγμένες για μεγάλα σύνολα δεδομένων. Αυτές είναι οι προσεγγίσεις που βασίζονται σε λεξικό (σε μια συνεχή κλίμακα τιμών συναισθήματος) και προσεγγίσεις που βασίζονται στην μηχανική μάθηση με δυαδικές και διακριτές τιμές συναισθήματος. Οι μέθοδοι λεξικού έχουν την ικανότητα της μη – εποπτευόμενης ταξινόμησης βασισμένης σε κανόνες, η οποία είναι υπολογιστικά αποδοτική. Από την αλλη πλευρά, οι μέθοδοι που βασίζονται στην μηχανική μάθηση επιτυγχάνουν ανώτερα αποτελέσματα και καλύτερη προσαρμογή στο εκάστοτε πεδίο, έχοντας όμως πληρώσει το τίμημα της φάσης εκπαίδευσης.

6.1 Σύγκριση μεθόδων

Όπως μπορεί να φανεί και από το σχήμα 18, οι προσεγγίσεις λεξικού και μηχανικής μάθησης έχουν προσελκύσει την προσοχή της ερευνητικής κοινότητας. Εξελίσσονται παράλληλα για τουλάχιστον μια δεκαετία και υπάρχουν διαθέσιμες αρκετές μελέτες που συγκρίνουν την αποτελεσματικότητά τους σε διάφορα σύνολα δεδομένων. Στη συνέχεια θα αναφέρουμε τις πιο ενδιαφέρουσες μελέτες σύγκρισης τους και θα σχολιάσουμε συνοπτικά τα αποτελέσματα τους. Στον παρακάτω πίνακα υπάρχει μια πλήρης λίστα της αξιολόγησης των επιδόσεων διαφόρων μελετών.

Οι Chaovalit et al. [27] πραγματοποίησαν μία αξιολόγηση μεταξύ ενός ταξινομητή βασισμένου σε ν-γράμματα (n-gram) και μεθόδων στατιστικής προσέγγισης σε ένα σύνολο δεδομένων από κριτικές ταινιών. Η μελέτη έδειξε ότι η ακρίβεια της μεθόδου μηχανικής μάθησης κυμαίνεται από 66% μέχρι 85%, κάνοντάς την συγκρίσιμη με το ποσοστό 77% της ακρίβειας που επετεύχθη με την μη – επιβλεπόμενη προσέγγιση λεξικού. Οι Gindl et al. [56] συνέκριναν την ακρίβεια μεταξύ διαφόρων μεθόδων λεξικού με μεθόδων μηχανικής μάθησης σε σύνολα δεδομένων από τον παγκόσμιο ιστό (Amazon, IMDb και TripAdvisor). Τα αποτελέσματα κατέδειξαν την ανωτερότητα των μεθόδων μηχανικής μάθησης σε σχέση με τις μεθόδους βασισμένες σε λεξικό και στα τρία σύνολα δεδομένων. Τα καλύτερα αποτελέσματα επετεύχθησαν με την μέθοδο της μέγιστης εντροπίας, της οποίας η ακρίβεια ήταν στα επίπεδα του 80%.

Δημοσίευση	Σύνολο δεδομένων	Ακρίβεια αλγορίθμου συναισθήματος (Precision, %)
Dave et al (2003)	AZ, CN	SVM (85.8 - 87.2) NB (81.9 - 87.0)
Hu and Liu (2004a)	AZ, CN	Semantic (84.0)
Turney (2002)	EP	PMI (56.8)
Turney and Littman (2003)	HM	SO-LSA (67.7 - 88.9) PMI (61.8 - 71.0)
	GI	SO-LSA (65.3 - 82.0) PMI (61.3 - 68.7)
Kamps et al (2004)	GI	Semantic (76.7)
Read and Caroll (2009)	GI	PMI (71.7) Semantic Space (83.8) Similarity (67.6)
	SemEval*	PMI (46.4) Semantic Space (44.4) Similarity (53.1)
	IMDB	PMI (68.7) Semantic Space (66.7) Similarity (60.8)
Gindl and Liegl (2008), average	AZ (N/A)	Dictionary (59.5 - 62.4) NB (66.0) ME (83.8)
	TA (N/A)	Dictionary (70.9 - 76.4) NB (72.4) ME (78.9)
	IMDB	Dictionary (61.8 - 64.9) NB (58.5) ME (82.3)
Pang et al (2002)	IMDB	NB (81.5) ME (81.0) SVM (82.9)
Chaovalit and Zhou (2005)	IMDB	N-Gram (66.0 - 85.0) PMI (77.0)
Goldberg and Zhu (2006)	IMDB	SVR (50.0 - 59.2) Graph (36.6 - 54.6)
Annett and Kondrak (2008)	IMDB	NB (77.5) SVM (77.4) ADTree (69.3)
Thet et al (2009)	IMDB	Dictionary (81.0)
Ku et al (2007)	NTCIR	Statistics (66.4)
Choi et al (2009)	NTCIR	Dictionary + Clustering (~70.0)
Osherenko and Andr´e (2007)	SAL*	SVM + Dictionary (34.5)
Yu and Hatzivassiloglou (2003)	TREC	Statistics (68.0 - 90.0)
Ku et al (2005)	TREC	Dictionary (62.0)
Missen and Boughanem (2009)	TREC	Semantic (MAP 28.0, P@10 64.0)
Yi et al (2003)	N/A	Dictionary (87.0 Reviews, 91.0 - 93.0 News)
Gamon (2004)	N/A	SVM (69.0 nearest classes, 85.0 farthest classes)
Kim and Hovy (2004)	N/A	Semantic (67.0 - 81.0)
Thomas et al (2006)	N/A	Multiple SVM (71.0)
Nadeau et al (2006)	N/A	LR (35.0 - 50.0) NB + Dictionary (38.0)
Chen et al (2006)	N/A	DT (71.7) SVM (84.6) NB (77.5)
Devitt and Ahmad (2007)	N/A	Semantic (50.0 - 58.0, f-measure)
Shimada and Endo (2008)	N/A	SVM OVA (58.4) ME (57.1) SVR (57.4) SIM (55.7)
Hare et al (2009)	N/A	MNB (75.1) SVM (74.4)
Zhu et al (2009)	N/A	Dictionary (69.0)
Bodendorf (2009)	N/A	SVM OVA (69.0)
Melville (2009)	N/A	NB + Dictionary (63.0 - 91.0)
Prabowo and Thelwall (2009)	N/A	SVM-only (87.3) SVM + RuleBased + Dictionary + Statistics (91.0)
Feng et al (2009)	N/A	Dictionary (65.0)
Go et al (2009)	TS	NB (82.7) ME (83.0) SVM (82.2)
Bifet and Frank (2010)	TS	MNB (82.5) SGD (78.6), Hoeffding tree (69.4)
	N/A	MNB (86.1) SGD (86.3) Hoeffding tree (84.8)
Pak and Paroubek (2010)	N/A	MNB (70.0) at recall value 60.0
Dave et al (2003)	AZ, CN	SVM (85.8 - 87.2) NB (81.9 - 87.0)

Πίνακας 19: Ακρίβεια της εξόρυξης γνώμης για διαφορετικές εφαρμογές και με βάση τα δεδομένα που αναφέρθηκαν από τους συγγραφείς. Τα σύνολα δεδομένων που δεν είναι δημόσια διαθέσιμα αναφέρονται ως N/A

Μία ακόμη σύγκριση μεταξύ των δημοφιλέστερων τύπων αλγορίθμων για εξόρυξη γνώμης έγινε από τους Annett και Kondrak [57], δείχνοντας ότι ορισμένοι σημασιολογικοί αλγόριθμοι έχουν συγκρίσιμα αποτελέσματα με τις μεθόδους μηχανικής μάθησης σε ότι αφορά την ακρίβεια, παρόλο που δεν απαιτούν μια υπολογιστικά απαιτητική φάση εκπαίδευσης. Πιο συγκεκριμένα, μια προσέγγιση λεξικού που χρησιμοποιεί το WordNet πέτυχε ακρίβεια παρόμοια με αυτήν των δένδρων αποφάσεων (60.4% έναντι 67.4%). Σε κάθε περίπτωση, αυτοί οι αλγόριθμοι δεν υποκαθιστούν, παρά αλληλοσυμπληρώνουν ο ένας τον άλλον.

Όπως αποδείχθηκε από τους Prabowo και Thelwall [58], μόνο ένας συνδυασμός από διαφορετικών ειδών ταξινομητές είναι δυνατόν να επιτύχει σταθερή απόδοση. Προκειμένου να αναπτυχθεί η υβριδική προσέγγιση, συνδύασαν διάφορους ταξινομητές βασισμένους σε κανόνες, με μία στατιστική μέθοδο και έναν SVM ταξινομητή. Με αυτόν τον τρόπο, πέτυχαν μία απόδοση από 83% έως 91%, ανάλογα με το σύνολο δεδομένων. Ωστόσο, δεν έχει παρουσιαστεί ακόμη στη διεθνή βιβλιογραφία μία συστηματική συγκριτική μελέτη η οποία να υλοποιεί και να αξιολογεί όλες τις σχετικές μεθοδολογίες μέσα από το ίδιο ερευνητικό πλαίσιο εργασίας. Επομένως, τα αποτελέσματα που παραθέτουν τις επιδόσεις των αλγορίθμων στον παραπάνω πίνακα δεν είναι άμεσα συγκρίσιμα, καθώς οι διάφορες μελέτες χρησιμοποιούν το δικό τους πλαίσιο αξιολόγησης και μεθοδολογίες έλεγχου.

6.2 Προτάσεις για το μέλλον

Η εξόρυξη και ανάλυση γνώμης είναι ένα απαιτητικό και διεπιστημονικό εγχείρημα. Απαιτεί να συγκεράσουν τις προσπάθειές τους ερευνητές από διαφορετικά πεδία. Μία τυπική λύση σε αυτήν την περιοχή προϋποθέτει γρήγορη και επεκτάσιμη ανάκτηση πληροφορίας, επεξεργασία κειμένου και ταυτοποίηση του θέματος, έτσι ώστε να είναι δυνατή η εκτέλεση αλγορίθμων μάθησης που υποστηρίζονται από εργαλεία επεξεργασίας φυσικής γλώσσας.

Οι αλγόριθμοι της εξόρυξης γνώμης έχουν με την πάροδο του χρόνου βελτιωθεί και ως προς την επίδοση και ως προς τη δυνατότητα ανάλυσης. Οι πρώτοι αλγόριθμοι που προτάθηκαν στην βιβλιογραφία ήταν αποτελεσματικοί στο να κάνουν διαχωρισμό ανάμεσα σε δύο ή τρεις κατηγορίες συναισθημάτων. Η μετάβαση σε περισσότερες κατηγορίες γνώμης απαίτησε τον επανασχεδιασμό των εφαρμοζόμενων μεθόδων μηχανικής μάθησης. Οι συνεχόμενες τιμές συναισθήματος μπορούν να αποκτηθούν μόνο με τη χρησιμοποίηση λεξιλογικών μεθόδων. Με βάση αυτό, στο μέλλον είναι πιθανόν η αυξανόμενη απαίτηση για ποιότητα των συναισθημάτων να απαιτήσει την ανάπτυξη νέων μεθόδων, οι οποίες θα διαθέτουν χαρακτηριστικά τόσο από μεθόδους μηχανικής μάθησης, όσο και από λεξιλογικές μεθόδους.

Προκειμένου να ενοποιηθούν οι πρόσφατες εξελίξεις στην Εξόρυξη Γνώμης, είναι απαραίτητο να αναπτυχθεί μια κοινά αποδεκτή κλίμακα για την αναπαράσταση των γνωμών. Για την ανάλυση συναισθήματος η επιλογή της συνεχόμενης κλίμακας στο πεδίο τιμών [-1,1] θεωρείται μια αποδεκτή επιλογή, καθώς μπορεί εύκολα να συμπεριλάβει και τις διακριτές κατηγορίες γνώμης (-1,0,1). Ωστόσο, για αντιτιθέμενες γνώμες δεν υπάρχει τέτοια προφανής επιλογή. Χρειάζεται να αναπαραστήσουμε διαφορές στις γνώμες που δεν μπορούν να αντιστοιχιστούν επακριβώς σε πραγματικούς αριθμούς. Για παράδειγμα, το ζεύγος “η γάτα είναι μαύρη – είναι μία άσπρη γάτα” το οποίο έχει την προφανή αντίφαση, δεν μπορεί να αναπαρασταθεί με τα +1/-1, καθώς το σύνολο που περιλαμβάνει μόνο δύο χρώματα (μαύρο, άσπρο) δεν είναι πλήρες, καθώς μπορεί επίσης να υπάρχουν και άλλα χρώματα.

Η εργασία επίσης αποκαλύπτει την ανάγκη να διευθετηθούν τα προβλήματα της συνάθροισης, διαχείρισης και ανάλυσης γνωμών σε μεγάλη κλίμακα. Αυτό θα επιτρέψει τον

σχεδιασμό συστημάτων διαβούλευσης πολιτικής και ηλεκτρονικής διακυβέρνησης με μεγαλύτερο βαθμό αποτελεσματικότητας. Για τους σκοπούς αυτούς θα πρέπει να υιοθετηθούν πρακτικές που χρησιμοποιούνται στην παραδοσιακή διαχείριση δεδομένων. Προβλήματα όπως η έλλειψη κατάλληλων επισημασμένων συνόλων δεδομένων, ώστε να συγκριθούν οι αλγόριθμοι εξόρυξης γνώμης, θα πρέπει επίσης να αντιμετωπιστούν. Στο ίδιο πλαίσιο, θα πρέπει να εισαχθεί μια κοινή κλίμακα βαθμολόγησης. Προκειμένου να επιτελεστεί σημαντική πρόοδος σε αυτή την κατεύθυνση χρειάζεται να εισαχθεί ένα κατάλληλο πλαίσιο εργασίας και να οριστούν με τυπικό τρόπο να σχετιζόμενα προβλήματα. Προβλέπεται ότι στα επόμενα χρόνια θα εισαχθεί ένα συνεργατικό πλαίσιο έρευνας το οποίο θα προωθήσει την αιχμή της τεχνολογίας και θα θέσει νέους στόχους για εξέλιξη.

7. Βιβλιογραφία

1. *Mining Subjective Data on the Web.* **Mikalai Tsytsarau, Themis Palpanas.** Trento : Ingegneria e Scienza dell'Informazione, University of Trento., 2010.

2. *A Holistic Lexicon-Based Approach to Opinion Mining.* **Xiaowen Ding, Bing Liu, Philip S. Yu.** Illinois : University of Illinois at Chicago, 2008.

3. *A State Of The Art Opinion Mining And Its Application Domains.* **Haji Binali, Vidyasagar Potdar, Chen Wu.** Australia : Curtin University of Technology, 2010.

4. **Sharp, Mark.** Text Mining. [Online] Rutgers University, School of Communication, Information and Library Studies, December 11, 2001. http://comminfo.rutgers.edu/~msharp/text_mining.htm.

5. **Yu, Bei.** *An evaluation of text classification methods for literary study.* Illinois : University of Illinois at Urbana-Champaign, 2006.

6. *Machine Learning in Automated Text Categorization.* **Sebastiani, Fabrizio.** Pisa, Italy : Consiglio Nazionale delle Ricerche, 2002.

7. *Support Vector Machine Active Learning with Applications to Text Classification.* **Simon Tong, Daphne Koller.** Stanford : Computer Science Department, Stanford University, 1998.

8. *Thumbs up? Sentiment Classification using Machine Learning Techniques.* **Bo Pang, Lillian Lee, Shivakumar Vaithyanathan.** Philadelphia : Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing, 2002.

9. *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.* **Dave et al, Dave, K., Lawrence, S., Pennock.** Budapest : International Conference on the World Wide Web, 2003.

10. *Half-Against-Half Multi-class Support Vector Machines.* **Hansheng Lei, Venu Govindaraju.** New York : Department of Computer Science and Engineering, State University of New York at Bu®alo, 2004.

11. *Multi-class Classification with Error Correcting Codes.* **Jorg Kindermann, Edda Leopold, Gerhard Paass.** s.l. : German National Research Center for Information Technology, 2000.

12. *Support Vector Machines for Classification and Regression.* **R.Gunn, Steve.** Southampton : University of Southampton, 1998.

13. *Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales.* **Bo Pang, Lillian Lee.** Ann Arbor : Proceedings of ACL-05, 43rd Meeting of the Association for Computational Linguistics, 2005.

14. *Get out the vote: determining support or opposition from congressional floor-debate.* **Matt Thomas, Bo Pang, Lillian Lee.** Stroudsburg : EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006.

15. *Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization.* **Goldberg A, Zhu X.** New York : TextGraphs Workshop On Graph Based Methods For Natural Language Processing, 2006.

16. *Recognizing contextual polarity in phrase-level sentiment analysis.* **Wilson et al, Theresa Wilson, Janyce Wiebe, Paul Hoffmann.** New York : Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language (HLT/EMNLP), 2005.

17. *Affect Sensing using Lexical Means: Comparison of a Corpus with Movie Reviews and a Corpus with Natural Language Dialogues.* **Osherenko, Alexander.** Augsburg : University of Augsburg, Germany, 2007.

18. *Sentiment polarity identification in financial news: A cohesion-based approach.* **Devitt A, Ahmad K.** Prague : 45th Annual Meeting of the Association of Computational Linguistics, 2007.

19. **Cogley, James.** *Sensing Sentiment in On-Line Recommendation Texts and Ratings.* Dublin : School of Computer Science and Statistics, 2010.

20. *Old Wine or Warm Beer: Target-Specific Sentiment Analysis of Adjectives. In: Proceedings of the Symposium on Affective Language in Human and Machine.* **Fahrni A, Klenner M.** Aberdeen, Scotland : Proceedings of the Symposium on Affective Language in Human and Machine, 2008.

21. *Semi-Supervised Learning Literature Survey.* **Zhu, Xiaojin.** Madison : University of Wisconsin, 2007.

22. *Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques.* **Yi et al, Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, Wayne Niblack.** Melbourne, Florida : Proceedings of the third IEEE International Conference on Data Mining (ICDM), 2003.

23. *Sentiment analysis of movie reviews on discussion boards using a linguistic approach.* **Tun Thura Thet, Jin-Cheon Na, Christopher S.G. Khoo, Subbaraj Shakthikumar.** Hong Kong : Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, 2009.

24. *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews.* **PD, Turney.** Morristown : Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002.

25. **Riloff E, Wiebe J, Phillips W.** *Exploiting subjectivity classification to improve information extraction.* s.l. : The MIT Press, 2005.

26. **Hatzivassiloglou, V., Wiebe, J.** *Effects of Adjective Orientation and Gradability on Sentence Subjectivity.* Saarbrücken, Germany : 18th International Conference on Computational Linguistics (COLING-2000), 2000.

27. *Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches.* **Pimwadee Chaovalit, Lina Zhou.** Maryland : Department of Information Systems, University of Maryland , 2005.

28. *Ranking opinionated blog posts using OpinionFinder.* **Ben He, Craig Macdonald, Iadh Ounis.** Singapore : Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008.

29. *Using WordNet to measure semantic orientation of adjectives.* **Jaap Kamps, Maarten Marx, R. ort. Mokken, Maarten de Rijke.** Lisbon : Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, 2004.

30. *Determining the sentiment of opinions.* **Kim SM, Hovy E.** Morristown : Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, 2004.

31. *Mining and summarizing customer reviews. In Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining (KDD).* **Liu, Minqing Hu and Bing.** Seattle, USA : Department of Computer Science, University of Illinois, 2004.

32. *Large-scale sentiment analysis for news and blogs.* **Godbole N, Srinivasaiah M, Skiena S.** Boulder, Colorado : Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2007.

33. **Λαζάρη, Ιωάννα.** *Αυτόματη παραγωγή συγκρίσεων προϊόντων από κριτικές χρηστών.* Αθήνα : Οικονομικό Πανεπιστήμιο Αθηνών, 2011.

34. *Learning Subjective Nouns Using Extraction Pattern Bootstrapping.* **Ellen Riloff, Janyce Wiebe, Theresa Wilson.** Edmonton, Canada : CoNLL-2003: Seventh Conference on Natural Language Learning (CoNLL), 2003.

35. *Towards answering opinion questions: Separating facts from Opinions and Identifying the Polarity of Opinion Sentences.* **Hong Yu, Vasileios Hatzivassiloglou.** Columbia : Department of Computer Science, Columbia University, 2003.

36. *Predicting the semantic orientation of adjectives.* **V. Hatzivassiloglou, K. McKeown.** Madrid : ACL EACL, 1997.

37. **Χρήστος, Παπαδημητρίου.** *Καινοτομίες στη Διαβούλευση: Η συμμετοχή του πολίτη στη λήψη αποφάσεων.* Αθήνα : Υπουργείο Διοικητικής Μεταρρύθμισης και Ηλεκτρονικής Διακυβέρνησης, Ιούλιος, 2011.

38. *Collective Text Analysis for eRulemaking.* **Namhee Kwon, Stuart W. Shulman, Eduard Hovy.** San Diego : 7th Annual International Conference on Digital Government Research, 2006.

39. *Public Opinion Mining for Governmental Decisions.* **George Stylios, Dimitris Christodoulakis, Jeries Besharat, Maria-Alexandra Vonitsanou, Ioanis Kotrotsos, Athanasia Koumpouri, Sofia Stamou.** 2, Patra, Greece : University of Patras, 2010, Τόμ. 8.

40. *Text Annotation for Political Science Research.* **Claire Cardie.** Ithaca : Cornell University, 2008.

41. *Decision Support for e-Governance: A text mining approach.* **G. Koteswara Rao, Shubhamoy Dey.** 3, Indore, India : International Journal of Managing Information Technology (IJMIT), August 2011, Τόμ. 3.

42. *Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text.* **Hovy, Soo-Min Kim and Eduard.** Sydney : Proceedings of Workshop on Sentiment and Subjectivity in Text, 2006.

43. *WordNet: An online electronic lexical database.* **Ch., Fellbaum.** s.l. : MIT Press, 1998.

44. *Mining sentiment classification from political web logs.* **Durant, K. T. and Smith M. D.** s.l. : Proceedings of the workshop on Web Mining and Web Usage Analysis of the the ACM SIGKDD international conference on Knowledge Discovery and Data Mining, 2006.

45. *Using cocitation information to estimate political orientation in web documents.* **Efron, Miles.** Austin : University of Texas, 2005.

46. *Congress: A Political-Economic History of Roll-Call Voting.* **Pool, Keith T., Howard Rosental.** Oxford : Oxford University Press, 1997.

47. Trackur. *http://www.trackur.com/.* [Ηλεκτρονικό]

48. Qualia. [Ηλεκτρονικό] http://www.qualia.gr/.

49. Onlinereputation. [Ηλεκτρονικό] http://www.onlinereputation.gr/.

50. iSieve. [Ηλεκτρονικό] http://i-sieve.com/.

51. MOBI. [Ηλεκτρονικό] http://www.wisewindow.com/about-us/blog/item/8-mobi-mass-opinion-business-intelligence.

52. SAS. *SAS Text Analytics.* [Ηλεκτρονικό] http://www.sas.com/text-analytics/.

53. Buzz Metrics. [Ηλεκτρονικό] http://www.nielsen-online.com/products_buzz.jsp?section=pro_buzz.

54. BlogPulse. [Ηλεκτρονικό] http://www.blogpulse.com/.

55. Infonic. [Ηλεκτρονικό] http://www.infonic.com/product_sentiment.php.

56. *Evaluation of Different Sentiment Detection Methods for Polarity Classification on Web-Based Reviews.* **Gindl, S., Liegl, J.** Patras : 18th European Conference on Artificial Intelligence (ECAI-2008), 2008.

57. *A comparison of sentiment analysis techniques: Polarizing movie blogs.* **Annett M, Kondrak G.** Canada : Proceedings of the Canadian Society for computational studies of intelligence, 2008.

58. *Sentiment analysis: A combined approach.* **Prabowo R, Thelwall M.** s.l. : Journal of Infometrics, 2009.

59. *A survey on sentiment detection on reviews.* **Huifeng Tang, Songbo Tan, Xueqi Cheng.** Expert Systems with Applications, An International Journal, Beijing, China : Elsevier, 2009, Τόμ. 36.

60. *A preliminary Investigation into Sentiment Analysis of Informal Political Discourse.* **Tony Mullen, Robert Malouf.** s.l. : Proceedings of the AAAI Workshop on Analysis of Weblogs, 2006.

61. *Political Leaning Categorization by Exploring Subjectivities in Political Blogs.* **Maojin Jiang, Shlomo Argamon.** Illinois Institute of Technology, Chicago : Proceedings, International Conference on Data Mining (DMIN 2008), pp. 647-653, 2008.

62. *Collective Text Analysis for eRulemaking.* **Kwon, N., Shulman, S.W., and Hovy, E.H.** San Diego, CA. : Proceedings of the Sixth National Conference on Digital Government research, 2006.

63. *Research of Data Mining in Government Transparent Decision-making.* **Guoshi Guan, Liuyang Zhou, and Pengcheng Tang.** China : Academy Publisher, 2009.

64. *Classifying Party Affiliation from Political Speech.* **Bei Yu, Daniel Diermeier and Stefan Kaufmann.** s.l. : Journal of Informatino Technology & Politics, The Haworth Press, 2008, Τόμ. 5(1).

65. *The design of OPTIMIST, an Opinion Mining System for Portuguese Politics.* **Mário J. Silva, Paula Carvalho, Luís Sarmento, Pedro Magalhães and Eugénio Oliveira.** Portugal : Proceedings of EPIA'2009, Aveiro, 2009.

66. *Exploring the characteristics of opinion expressions for political opinion classification.* **Yu, B., Kaufmann, S., & Diermeier, D.** s.l. : Proceedings of the 9th International Conference on Digital Government Research, p82-91, 2008.

67. *Building a Sentiment Summarizer for Local Service Reviews.* **Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald,.** s.l. : WWW Workshop NLP Challenges, 2008.

68. *Opinion mining of customer feedback data on the web.* **Dongjoo Lee Seoul National University, Seoul, Republic of Korea.** New York : Proceeding ICUIMC '08 Proceedings of the 2nd international conference on Ubiquitous information management and communication , 2008. 978-1-59593-993-7.