ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Διπλωματική Εργασία

Μεταπτυχιακού Διπλώματος Ειδίκευσης

«*Αυτόματη παραγωγή συγκρίσεων προϊόντων από κριτικές χρηστών*»

Ιωάννα Λάζαρη

Επιβλέπων: Ίων Ανδρουτσόπουλος

ΑΘΗΝΑ, ΙΟΥΝΙΟΣ 2011

Περιεχόμενα

ΠΕΡΙΛΗΨΗ 4

1. Εισαγωγή 6

1.1 Αντικείμενο της εργασίας 6

1.2 Διάρθρωση της εργασίας 7

2. Δεδομένα 9

2.1 Συλλογή κριτικών 9

2.2 Σύνολα δεδομένων 10

3. Εξαγωγή λέξεων-κλειδιών 12

3.1 Hu και Liu 12

3.2 Blair-Goldensohn κ.ά. 16

3.2 Εξαγωγή λέξεων-κλειδιών με LDA 18

3.2.1 Latent Dirichlet Allocation 18

3.2.2 Brody και Elhadad 19

3.2.3 Άλλες μέθοδοι εξαγωγής λέξεων-κλειδιών με LDA 21

3.3 Εξαγωγή λέξεων-κλειδιών στο σύστημα της παρούσας εργασίας 22

4.1 Προγενέστερες μέθοδοι εξόρυξης γνώμης 26

4.1.1 Ο αλγόριθμος των Wilson, Wiebe και Hoffman 26

4.1.2 Άλλοι αλγόριθμοι εξαγωγής γνώμης 31

4.2 Εξόρυξη γνώμης στο σύστημα της παρούσας εργασίας..... 32

4.2.1 Εξόρυξη γνώμης με γλωσσικά μοντέλα..... 33

4.2.2 Εξόρυξη γνώμης με αφελή ταξινομητή Bayes..... 36

4.2.3 Εξόρυξη γνώμης με μηχανή διανυσμάτων υποστήριξης 38

5 Τελική παρουσίαση των προϊόντων 41

6 Συμπεράσματα και μελλοντικές κατευθύνσεις 49

ΒΙΒΛΙΟΓΡΑΦΙΑ..... 51

ΠΕΡΙΛΗΨΗ

Με την ανάπτυξη του παγκόσμιου ιστού υπάρχει πλέον μια πληθώρα ιστοσελίδων από τις οποίες οι καταναλωτές μπορούν να ενημερωθούν για προϊόντα που τους ενδιαφέρουν. Ιδιαίτερα χρήσιμες είναι ιστοσελίδες που περιλαμβάνουν σχόλια ή κριτικές για συγκεκριμένα προϊόντα. Όμως, ο αριθμός αυτών των ιστοσελίδων είναι τόσο μεγάλος που οι καταναλωτές είναι αναγκασμένοι να ξοδεύουν πολύ χρόνο για να εντοπίζουν και να συνθέτουν τις πληροφορίες που τους ενδιαφέρουν.

Για να αντιμετωπιστεί αυτό το πρόβλημα, δημιουργήθηκε ένα σύστημα παραγωγής συγκριτικών παρουσιάσεων δύο προϊόντων. Ο χρήστης εισάγει τα ονόματα δύο προϊόντων και το σύστημα παράγει αυτόματα μια συγκριτική παρουσίαση των κυριότερων χαρακτηριστικών τους, στηριζόμενο σε κείμενα κριτικών, τα οποία συλλέγονται από ιστοσελίδες ηλεκτρονικών καταστημάτων και ιστοσελίδες που φιλοξενούν αποκλειστικά κριτικές χρηστών.

Τα δύο βασικά βήματα που απαιτούνται για τον δημιουργία των συγκριτικών παρουσιάσεων είναι ο εντοπισμός των κύριων χαρακτηριστικών των προϊόντων και ο προσδιορισμός της γνώμης του κοινού (θετική ή αρνητική) για τα χαρακτηριστικά αυτά. Για κάθε βήμα, μελετήθηκε η σχετική βιβλιογραφία και αναπτύχθηκαν κατάλληλοι αλγόριθμοι οι οποίοι αξιολογήθηκαν πειραματικά.

1. Εισαγωγή

1.1 Αντικείμενο της εργασίας

Τα τελευταία χρόνια, η ραγδαία ανάπτυξη του παγκόσμιου ιστού έχει επηρεάσει πολύ έντονα τη συμπεριφορά των καταναλωτών. Υπάρχει πλέον μια πληθώρα ιστοσελίδων, περιοχών συζητήσεων (fora) και ιστολογίων (blogs), στα οποία ο καταναλωτής μπορεί να ενημερωθεί για τα προϊόντα που τον ενδιαφέρουν, όχι μόνο ως προς τα τυπικά τους χαρακτηριστικά, αλλά και ως προς τη γνώμη άλλων καταναλωτών σχετικά με τα προϊόντα αυτά. Όμως, ο αριθμός των ιστοσελίδων που περιέχουν σχόλια και κριτικές για προϊόντα (π.χ. φορητοί υπολογιστές) είναι πολύ μεγάλος και οι καταναλωτές είναι αναγκασμένοι να ξοδεύουν πολύ χρόνο για να εντοπίζουν και να συνθέτουν τις πληροφορίες που τους ενδιαφέρουν.

Στην παρούσα εργασία προσπαθούμε να αντιμετωπίσουμε αυτό το πρόβλημα μέσω ενός συστήματος στο οποίο ο χρήστης εισάγει τα ονόματα δύο προϊόντων. Το σύστημα παράγει αυτόματα μια συγκριτική παρουσίαση των δύο προϊόντων, βασισμένη σε κριτικές άλλων χρηστών. Πιο συγκεκριμένα, το σύστημά μας εντοπίζει για κάθε προϊόν το σύνολο των κυριότερων χαρακτηριστικών του, δηλαδή τα χαρακτηριστικά που σχολιάζονται συχνά στις κριτικές, καθώς και τη γενική άποψη του κοινού για κάθε χαρακτηριστικό. Στη συνέχεια, παρουσιάζει τα κυριότερα χαρακτηριστικά των δύο προϊόντων και τη γνώμη του κοινού για αυτά, εστιάζοντας στα χαρακτηριστικά για τα οποία η γνώμη του κοινού είναι αντίθετη στα δύο προϊόντα. Για παράδειγμα, έστω πως εξετάζουμε δύο βιντεοκάμερες. Για την πρώτη, οι κριτικές αναφέρουν πως έχει πολύ καλό zoom και καλό σύστημα ήχου, ενώ για τη δεύτερη πως έχει μέτριο zoom και καλό σύστημα ήχου. Στην τελική μας παρουσίαση,

θα εστιάσουμε περισσότερο στο zoom των δύο μοντέλων, το οποίο φαίνεται να είναι η βασική διαφορά τους και όχι τόσο στο σύστημα ήχου, το οποίο παίρνει καλές κριτικές και στις δύο κάμερες.

Το σύστημά μας αντλεί κριτικές χρηστών από ιστοσελίδες ηλεκτρονικών καταστημάτων και ιστοσελίδες που ασχολούνται αποκλειστικά με κριτικές αντικειμένων. Τα στάδια που ακολουθούμε για την παραγωγή της συγκριτικής παρουσίασης των δύο προϊόντων είναι τα ακόλουθα:

1. Συλλογή σχετικών κειμένων κριτικών από διάφορες ιστοσελίδες.
2. Προσδιορισμός των κυριότερων χαρακτηριστικών κάθε προϊόντος.
3. Προσδιορισμός της άποψης του κοινού για κάθε ένα από τα χαρακτηριστικά του προηγούμενου σταδίου.
4. Εντοπισμός των χαρακτηριστικών για τα οποία η άποψη του κοινού είναι αντίθετη στα δύο προϊόντα. Επίσης, εντοπίζουμε χαρακτηριστικά τα οποία αναφέρονται έντονα στις κριτικές ενός προϊόντος και είναι άξια αναφοράς, ακόμη κι αν δεν υπάρχουν καθόλου στις κριτικές του άλλου προϊόντος.
5. Παρουσίαση των αποτελεσμάτων μέσω γραφικής διεπαφής.

1.2 Διάρθρωση της εργασίας

Η εργασία αποτελείται από πέντε ακόμη κεφάλαια:

- Στο κεφάλαιο 2 περιγράφεται η διαδικασία συλλογής κριτικών από διάφορες ιστοσελίδες, καθώς και τα σύνολα δεδομένων που χρησιμοποιήθηκαν στα διάφορα στάδια της εργασίας.
- Το κεφάλαιο 3 ασχολείται με την εξαγωγή λέξεων – κλειδιών που αντιστοιχούν στα κύρια χαρακτηριστικά των προϊόντων. Παρουσιάζονται προγενέστεροι αλγόριθμοι, καθώς και η δική μας μέθοδος.
- Το κεφάλαιο 4 ασχολείται με τον προσδιορισμό της γνώμης του κοινού για κάθε χαρακτηριστικό. Παρουσιάζονται

προγενέστερες μέθοδοι εξόρυξης γνώμης, καθώς και η δική μας.

- Στο κεφάλαιο 5 παρουσιάζεται η διεπαφή χρήστη του συστήματος της εργασίας, μέσω της οποίας γίνεται η συγκριτική παρουσίαση των προϊόντων.
- Το κεφάλαιο 6 περιλαμβάνει τα συμπεράσματα που προέκυψαν από την εργασία, καθώς και προτάσεις για μελλοντική δουλειά.

2. Δεδομένα

2.1 Συλλογή κριτικών

Όπως ήδη έχουμε αναφέρει, το σύστημα της εργασίας παρέχει στο χρήστη τη δυνατότητα να επιλέξει δύο προϊόντα, για τα οποία στη συνέχεια θα παραχθεί μια συγκριτική παρουσίαση, βασισμένη σε κριτικές που έχουν γράψει γι' αυτά διάφοροι χρήστες σε ιστοτόπους. Συνεπώς, απαιτείται μια αυτοματοποιημένη διαδικασία, η οποία να εξάγει από διάφορους ιστοτόπους ένα σύνολο κειμένων κριτικών για κάθε προϊόν. Τα επόμενα στάδια επεξεργασίας του συστήματός μας αξιοποιούν τα κείμενα αυτά.

Χρησιμοποιούμε έναν προσαρμοσμένο στις ανάγκες μας Web Crawler. Οι Web Crawlers είναι προγράμματα τα οποία επισκέπτονται μια λίστα από ιστοτόπους, τους οποίους ορίζει ο χρήστης, και δημιουργούν ένα τοπικό αντίγραφό τους. Επίσης, μπορεί να παρέχουν κι άλλες επιλογές, όπως να αποθηκεύουν το περιεχόμενο της σελίδας απαλλαγμένο από ετικέτες HTML ή να επισκέπτονται και τις σελίδες στις οποίες οδηγούν οι σύνδεσμοι της τρέχουσας σελίδας διαδοχικά, μέχρι κάποιο προεπιλεγμένο βάθος.

Στην περίπτωσή μας, χρησιμοποιούμε έναν Web Crawler ώστε να δημιουργήσουμε τοπικά αντίγραφα ιστοσελίδων που γνωρίζουμε πως περιέχουν κριτικές χρηστών. Βασιστήκαμε στον ανοιχτού λογισμικού crawler4j, τον οποίο τροποποιήσαμε κατάλληλα.[1] Οι ιστοσελίδες τις οποίες ο crawler επισκέπτεται είναι οι ιστοσελίδες των ηλεκτρονικών καταστημάτων Amazon και Walmart, καθώς και ιστοσελίδες των ιστοτόπων Epinions και Viewpoints, που φιλοξενούν αποκλειστικά

[1] Βλ. http://code.google.com/p/crawler4j/. (Η υλοποίηση είναι σε Java)

κριτικές προϊόντων.[2] Οι ιστοσελίδες που αποθηκεύει ο crawler υφίστανται κατάλληλη επεξεργασία, ώστε να αποθηκεύονται μόνο τα τμήματα που περιέχουν τις κριτικές των χρηστών.

2.2 Σύνολα δεδομένων

Για τη διεξαγωγή των πειραμάτων μας, χρειάστηκε να συγκεντρώσουμε ένα σύνολο κειμένων κριτικών. Πιο συγκεκριμένα, για την εκπαίδευση και αξιολόγηση του αλγορίθμου εξαγωγής λέξεων-κλειδιών, δημιουργήσαμε 2 σύνολα δεδομένων (datasets), ένα 2.000 κειμένων που αφορούν φορητούς υπολογιστές (laptops) και ένα 500 κειμένων για φωτογραφικές μηχανές. Ονομάζουμε αυτά τα δύο σύνολα *Laptops1* και *Cameras*.

Για την εκπαίδευση και αξιολόγηση του αλγορίθμου εξόρυξης γνώμης, χρησιμοποιήσαμε το σύστημα βαθμολόγησης των προϊόντων (π.χ. αστεράκια) που παρέχουν οι ιστότοποι που αναφέραμε. Πιο συγκεκριμένα, κάναμε την παραδοχή πως αν ένας χρήστης-κριτής έχει δώσει σε ένα προϊόν το μέγιστο δυνατό βαθμό (π.χ. 5/5 αστεράκια), τότε το περιεχόμενο του κειμένου της κριτικής του είναι μόνο θετικό, δηλαδή όλες οι προτάσεις της κριτικής εκφράζουν αποκλειστικά θετικά σχόλια. Αντίστοιχα, αν μια κριτική έχει το μικρότερο δυνατό βαθμό (π.χ. 0/5 αστεράκια), θεωρούμε πως το περιεχόμενό της είναι μόνο αρνητικό, δηλαδή όλες οι προτάσεις της κριτικής εκφράζουν αποκλειστικά αρνητικά σχόλια. Πρόκειται, φυσικά, για παραδοχές που ενδέχεται να μην ισχύουν πάντα. Οδηγούν παρ' όλα αυτά σε χρήσιμα, όπως αποδεικνύουν τα πειράματα της εργασίας, σύνολα δεδομένων, χωρίς να απαιτείται πρόσθετη χειρωνακτική επισημείωση των κειμένων εκπαίδευσης και

[2] Βλ. http://www.amazon.com/, http://www.walmart.com/, http://www.epinions.com/, http://www.viewpoints.com/.

αξιολόγησης (π.χ. επισημείωση προτάσεων ως θετικών ή αρνητικών). Έτσι, για τη διεξαγωγή των πειραμάτων, τα οποία θα περιγραφούν στο κεφάλαιο 4, δημιουργήσαμε ένα σύνολο δεδομένων, το οποίο ονομάζουμε *Laptops2*, το οποίο περιέχει 950 εντελώς αρνητικές (0/5 αστεράκια) και 1200 εντελώς θετικές κριτικές (5/5 αστεράκια) χρηστών για φορητούς υπολογιστές.

3. Εξαγωγή λέξεων-κλειδιών

Στο κεφάλαιο αυτό περιγράφουμε προγενέστερες μεθόδους εξαγωγής λέξεων-κλειδιών (π.χ. «πληκτρολόγιο», «επεξεργαστής») που ονοματίζουν χαρακτηριστικά (aspects) των προϊόντων. Οι λέξεις - κλειδιά εξάγονται αυτόματα από κείμενα κριτικών. Περιγράφουμε αρχικά δύο ενδεικτικές προγενέστερες μεθόδους και κατόπιν μια μέθοδο βασισμένη στο μοντέλο Latent Dirichlet Allocation (LDA), που χρησιμοποιήθηκε και στο σύστημα αυτής της εργασίας.

3.1 Hu και Liu

Οι Hu και Liu [4] πρότειναν έναν αλγόριθμο εξαγωγής λέξεων-κλειδιών από ένα σύνολο κριτικών. Οι λέξεις-κλειδιά που εξάγει είναι μεμονωμένα ουσιαστικά ή ονοματικές φράσεις με μήκος το πολύ 3 λέξεις. Όταν εξάγονται φράσεις, θα έπρεπε ορθότερα να μιλάμε για φράσεις-κλειδιά, αλλά χρησιμοποιούμε πάλι τον όρο λέξεις-κλειδιά χάριν απλότητας.

Τα βήματα που ακολουθεί ο αλγόριθμος είναι τα παρακάτω:

1. Σε κάθε πρόταση των κριτικών αναγνωρίζονται τα μέρη του λόγου (POS tagging) και εντοπίζονται τα ονοματικά και τα ρηματικά τμήματα (noun groups, verb groups) της πρότασης. Για παράδειγμα, η πρόταση:

> "*I am absolutely in awe of this camera.*"

επισημειώνεται (με XML) ως εξής:

```
<S>
    <NG>
    <W C='PRP' L='SS' t='w' S='Y'> I </W>
    </NG>
    <VG>
    <W C='VBP'> am </W> <W C='RB'> absolutely </W>
    </VG>
    <W C='IN'> in </W>
    <NG>
    <W C='NN' > awe </W>
    </NG>
    <W C='IN' > of </W>
    <NG>
    <W C='DT' > this </W>
    <W C='NN' > camera </W>
    </NG>
    <W C='.' > . </W>
</S> .
```

όπου για παράδειγμα η επισημείωση <W C='NN'>…<W> …</W>δηλώνει ένα ουσιαστικό και η επισημείωση <NG>…</NG> ένα ονοματικό τμήμα.

2. Δημιουργείται ένα αρχείο με μία γραμμή για κάθε μια πρόταση των κριτικών. Η γραμμή περιέχει το σύνολο των ουσιαστικών και των ονοματικών φράσεων που προέκυψαν από την αντίστοιχη κριτική στο προηγούμενο βήμα. Από το σύνολο αυτό αφαιρούνται οι συχνές λέξεις (stop words). Ύστερα αφαιρούνται οι καταλήξεις των λέξεων (stemming) και γίνεται μερική διόρθωση των ορθογραφικών λαθών. Το αρχείο αυτό χρησιμοποιείται στο επόμενο βήμα.
3. Στη συνέχεια, κάνοντας την παρατήρηση πως όταν οι χρήστες σχολιάζουν ένα συγκεκριμένο χαρακτηριστικό ενός προϊόντος

χρησιμοποιούν συνήθως παρεμφερές λεξιλόγιο, γίνεται η παραδοχή πως τα ουσιαστικά ή οι ονοματικές φράσεις που αντιστοιχούν σε χαρακτηριστικά θα εμφανίζονται με μεγάλη συχνότητα στο σύνολο των προτάσεων των κριτικών. Για το λόγο αυτό, χρησιμοποιείται ο association rule miner CBA [5], ο οποίος δημιουργεί πρότυπα (patterns) από το σύνολο των ουσιαστικών και των ονοματικών φράσεων των γραμμών του αρχείου του προηγούμενου βήματος και στη συνέχεια εξετάζει με πόσες προτάσεις (γραμμές) του αρχείου ταιριάζει το κάθε πρότυπο. Αν το πλήθος των προτάσεων με τις οποίες ταιριάζει ένα πρότυπο ξεπερνά το 1% του συνόλου των προτάσεων, το πρότυπο αυτό αποτελεί υποψήφια λέξη-κλειδί.

4. Ύστερα, για κάθε υποψήφια λέξη-κλειδί που αποτελείται από περισσότερες από μια λέξεις (φράση-κλειδί) ελέγχουμε τις προτάσεις στις οποίες εμφανίζεται. Αν η απόσταση μεταξύ των λέξεων που αποτελούν την υποψήφια λέξη-κλειδί είναι μικρότερη από 3 σε πάνω από 2 προτάσεις, τότε θεωρείται βέβαιο πως η υποψήφια λέξη-κλειδί που εξετάζεται αποτελεί όντως λέξη-κλειδί. Με τον τρόπο αυτό γίνεται εξαγωγή λέξεων-κλειδιών οι οποίες αποτελούνται από περισσότερες από δύο λέξεις.
5. Για κάθε μια από τις υπόλοιπες υποψήφιες λέξεις-κλειδιά, μετράμε τον αριθμό των προτάσεων στις οποίες περιέχεται και στις οποίες η υποψήφια λέξη-κλειδί δεν αποτελεί μέρος κάποιας από τις άλλες λέξεις-κλειδιά που προέκυψαν από το προηγούμενο βήμα. Αν το πλήθος αυτό είναι μεγαλύτερο ή ίσο με 3, τότε προκύπτει μια νέα λέξη-κλειδί. Για παράδειγμα, έστω πως η υποψήφια λέξη-κλειδί «manual» εμφανίζεται σε 10 προτάσεις και πως από το προηγούμενο βήμα έχει προκύψει η λέξη-κλειδί «manual settings», που εμφανίζεται σε 6 προτάσεις. Συνεπώς, η «manual» περιέχεται σε 4 προτάσεις χωρίς να είναι τμήμα κάποιας άλλης λέξης-κλειδιού. Αυτό σημαίνει πως είναι όντως λέξη-κλειδί, όπως και η «manual settings». Το αποτέλεσμά μας είναι το επιθυμητό, καθώς

οι δύο αυτές λέξεις-κλειδιά έχουν διαφορετική σημασία (manual = εγχειρίδιο, manual settings = χειροκίνητες ρυθμίσεις). Με τον τρόπο αυτό γίνεται εξαγωγή λέξεων-κλειδιών οι οποίες αποτελούνται από μια μόνο λέξη.

6. Στην κριτική ενός χρήστη, κοντά σε μια λέξη-κλειδί που ονοματίζει ένα χαρακτηριστικό ενός προϊόντος θα υπάρχουν και λέξεις που θα εκφράζουν την άποψη του χρήστη για το χαρακτηριστικό αυτό. Επομένως, χρησιμοποιώντας τις λέξεις-κλειδιά που προέκυψαν από τα προηγούμενα βήματα μπορεί να γίνει και εξαγωγή λέξεων που εκφράζουν άποψη. Για παράδειγμα, στη φράση *«This camera takes incredible pictures»* η λέξη-κλειδί *«pictures»* συνοδεύεται από τη λέξη *«incredible»*, που εκφράζει άποψη. Έτσι εξάγονται επίθετα που είναι δίπλα στις λέξεις-κλειδιά. Κατόπιν, αφού αφαιρεθούν οι καταλήξεις των επιθέτων που προκύπτουν και διορθωθούν τυχόν ορθογραφικά λάθη, προκύπτει μια λίστα ριζών (stems) επιθέτων που εκφράζουν γνώμη.

Οι χρήστες πολλές φορές μπορεί να χρησιμοποιήσουν το ίδιο επίθετο για να σχολιάσουν παραπάνω από ένα χαρακτηριστικά, για παράδειγμα *«excellent software», «excellent picture quality»*. Εξετάζοντας τις προτάσεις που περιέχουν επίθετα του προηγούμενου βήματος είναι, επομένως, δυνατόν να εντοπισθούν πρόσθετα ουσιαστικά και ονοματικά τμήματα που βρίσκονται δίπλα στα επίθετα και αποτελούν πρόσθετες λέξεις-κλειδιά.

Ο παραπάνω αλγόριθμος ήταν από τους πρώτους που χρησιμοποιήθηκαν για εξαγωγή λέξεων-κλειδιών από ένα σύνολο κριτικών. Σύμφωνα με τους Brody και Elhadad [3] δίνει ικανοποιητικά αποτελέσματα κυρίως σε ό,τι αφορά την ανάκτηση λέξεων-κλειδιών που εμφανίζονται συχνά ως ουσιαστικά ή ονοματικές φράσεις. Δεν έχει όμως καλά αποτελέσματα στην ανάκτηση λέξεων-

κλειδιών που εμφανίζονται με μικρότερη συχνότητα ή περιγράφουν πιο αόριστα χαρακτηριστικά (π.χ. την ατμόσφαιρα ενός εστιατορίου). Μια παρόμοια προσέγγιση ακολούθησαν και οι Popescu και Etzioni [6] .

3.2 Blair-Goldensohn κ.ά.

Μια πιο πρόσφατη προσέγγιση είναι ο αλγόριθμος των Blair-Goldensohn κ.ά. [7], ο οποίος εστιάζει σε κριτικές χρηστών για υπηρεσίες, όπως ξενοδοχεία και εστιατόρια και συνδυάζει τεχνικές που χρησιμοποιεί ο προηγούμενος αλγόριθμος με τεχνικές μηχανικής μάθησης. Ο αλγόριθμος χωρίζεται σε δύο τμήματα: τη δυναμική εξαγωγή λέξεων-κλειδιών και τη στατική εξαγωγή λέξεων-κλειδιών.

Επίσης, χρησιμοποιεί ένα λεξικό όρων συναισθήματος (sentiment lexicon), το οποίο δημιουργήθηκε από ένα χειρωνακτικά κατασκευασμένο σύνολο λέξεων με γνωστή αρνητική ή θετική πολικότητα. Σε κάθε μια από τις αρχικές λέξεις έχει αντιστοιχηθεί ένα score, το οποίο υποδεικνύει το βαθμό βεβαιότητας για την πολικότητά της. Στη συνέχεια, χρησιμοποιώντας τις συνώνυμες και αντίθετες λέξεις του Wordnet [11] επεκτείνεται το αρχικό λεξικό προσθέτοντας νέες λέξεις με τα αντίστοιχα score.

Τα βήματα που ακολουθεί ο αλγόριθμός για τη δυναμική εξαγωγή λέξεων-κλειδιών είναι:

1. Εξάγει λέξεις-κλειδιά με τον αλγόριθμο των Hu και Liu.
2. Εξάγει λέξεις-κλειδιά οι οποίες ταιριάζουν με χειρωνακτικά κατασκευασμένα συντακτικά πρότυπα. Για παράδειγμα, ένα συντακτικό πρότυπο είναι «επίθετο + ουσιαστικό» (π.χ. *«great tacos»*).
3. Από τις λέξεις-κλειδιά που έχουν προκύψει αφαιρούνται αυτές που αποτελούνται εξολοκλήρου από πολύ συχνές λέξεις (stopwords) ή έχουν μικρή συχνότητα εμφάνισης σε προτάσεις.

4. Για κάθε λέξη-κλειδί, αθροίζονται τα βάρη των όρων του λεξικού συναισθήματος με τους οποίους συνυπάρχει η λέξη-κλειδί σε συντακτικά πρότυπα στα κείμενα που εξετάζονται. Αν το άθροισμα είναι κάτω από ένα κατώφλι, η λέξη-κλειδί απορρίπτεται.

Η παραπάνω μέθοδος εντοπίζει με επιτυχία λέξεις-κλειδιά οι οποίες εμφανίζονται με μεγάλη συχνότητα. Υπάρχει όμως η περίπτωση να υπάρχουν χαρακτηριστικά που να μπορούν να περιγραφούν από ένα ευρύ σύνολο λέξεων. Για παράδειγμα, στον τομέα των εστιατορίων, η λέξη «κοτόπουλο» ή η λέξη «αστακός» ουσιαστικά είναι υπο-περιπτώσεις του χαρακτηριστικού «φαγητό» (που είναι και λέξη-κλειδί), συνεπώς θα πρέπει αυτό να το λάβουμε υπόψη μας, γιατί αλλιώς είναι πιθανό να παράγουμε υπερειδικευμένες λέξεις-κλειδιά, που όμως δεν θα είναι χρήσιμες.

Για να επιλυθεί το πρόβλημα αυτό, χρησιμοποιήθηκε ο ταξινομητής Maximum Entropy (ΜΕ), ο οποίος εκπαιδεύτηκε ξεχωριστά σε δύο τομείς, εστιατόρια και ξενοδοχεία, ώστε να μπορεί να προβλέπει αν κάποια πρόταση αναφέρεται σε κάποιο (και ποιο) ευρύτερο χαρακτηριστικό (π.χ. φαγητό, διακόσμηση). Τα χαρακτηριστικά είχαν οριστεί εκ των προτέρων για κάθε έναν από τους δύο τομείς και χρησιμοποιήθηκαν ως οι κατηγορίες του ταξινομητή. Το σώμα εκπαίδευσης περιελάμβανε προτάσεις κριτικών στις οποίες άνθρωποι είχαν επισημειώσει τις σωστές κατηγορίες (χαρακτηριστικά). Η παραπάνω διαδικασία αποτελεί τη στατική εξαγωγή λέξεων-κλειδιών.

Συνδυάζοντας τα αποτελέσματα της δυναμικής και της στατικής εξαγωγής παράγεται η τελική λίστα λέξεων-κλειδιών. Με βάση τα αποτελέσματα που παρουσίασαν οι συγγραφείς, ο αλγόριθμος επιτυγχάνει μεγάλο ποσοστό ορθότητας (accuracy) και επιπλέον μας δίνει τη δυνατότητα να ομαδοποιήσουμε τις λέξεις-κλειδιά της δυναμικής εξαγωγής σε κατηγορίες της στατικής εξαγωγής.

3.2 Εξαγωγή λέξεων-κλειδιών με LDA

Το μοντέλο Latent Dirichlet Allocation (LDA) ή επεκτάσεις του έχουν χρησιμοποιηθεί, μεταξύ άλλων εφαρμογών, και σε αλγορίθμους εντοπισμού λέξεων-κλειδιών σε σύνολα κριτικών. Παρακάτω περιγράφουμε το μοντέλο LDA, καθώς και πώς χρησιμοποιείται σε έναν αλγόριθμο εξαγωγής λέξεων-κλειδιών.

3.2.1 *Latent Dirichlet Allocation*

Το LDA [1] είναι ένα πιθανοτικό μοντέλο κειμένων. Θεωρεί πως τα κείμενα αποτελούνται από κρυφές (latent) θεματικές ενότητες (topics), οι οποίες περιέχουν λέξεις. Οι λέξεις που εμφανίζονται (σε μια συλλογή κειμένων) συχνότερα σε μια θεματική ενότητα είναι πιο αντιπροσωπευτικές της ενότητας αυτής. Παρακάτω φαίνεται ένα παράδειγμα θεματικών ενοτήτων και των αντίστοιχων πιο αντιπροσωπευτικών λέξεων, που παράγονται από ένα σώμα κειμένων για φορητούς υπολογιστές (laptops). Τα ονόματα των θεματικών ενοτήτων είναι δικά μας (οι θεματικές ενότητες του LDA δεν έχουν συγκεκριμένα ονόματα).

Θεματική Ενότητα	Αντιπροσωπευτικές λέξεις
customer support	excellent, repair, technician, fix, disappointing, rude, break
Sony	quality, laptop, love, use, purchase
battery	life, lasts, hours, problem

Το μοντέλο LDA θεωρεί ότι κάθε έγγραφο d παράγεται με την ακόλουθη διαδικασία, όπως περιγράφεται από τους Blei κ.ά. [1]:

1. Επιλέγουμε θ~Dir(α), όπου Dir η κατανομή Dirichlet.
2. Για κάθε λέξη w_n από τις N λέξεις του εγγράφου:
    a. Επιλέγουμε μια θεματική ενότητα z_n~Multinomial(θ).

b. Επιλέγουμε μια λέξη w_n με κατανομή p(w_n| z_n, β), μια επίσης πολυωνυμική κατανομή η οποία εξαρτάται από τη θεματική ενότητα z_n.

Η παράμετρος θ είναι εν γένει διαφορετική ανά έγγραφο και καθορίζει ουσιαστικά την κατανομή των θεματικών ενοτήτων στο κάθε έγγραφο. Το πλήθος των δυνατών θεματικών ενοτήτων θεωρείται γνωστό. Υπάρχουν δε αλγόριθμοι [1] με τους οποίους εκτιμώνται (κατά το στάδιο της εκπαίδευσης του μοντέλου) οι παράμετροι α και β από ένα σώμα κειμένων. Οι πιο αντιπροσωπευτικές λέξεις μιας θεματικής ενότητας z είναι εκείνες με τις υψηλότερες p(w_n| z, β). Δοθέντος επίσης ενός νέου εγγράφου, μετά την εκπαίδευση του μοντέλου LDA, υπάρχουν αλγόριθμοι που εκτιμούν την κατανομή των θεματικών ενοτήτων του εγγράφου.

Το μοντέλο LDA, όπως περιγράφηκε παραπάνω, δε δίνει το επιθυμητό αποτέλεσμα εάν χρησιμοποιηθεί για την εξαγωγή των πιο αντιπροσωπευτικών λέξεων-κλειδιών κάθε θεματικής ενότητας από μεγάλες συλλογές κειμένων κριτικών. Μερικές από τις θεματικές ενότητες που παράγονται από ένα σύνολο κριτικών για MP3 players αντιστοιχούν σε συγκεκριμένα μοντέλα (π.χ. iPod και Sony Walkman), ενώ το ζητούμενο είναι να δημιουργηθούν θεματικές ενότητες που να αντιστοιχούν π.χ. στην ποιότητα του ήχου, τις λειτουργίες της συσκευής κλπ.

3.2.2 Brody και Elhadad

Σε μια πρόσφατη προσέγγιση εξαγωγής λέξεων-κλειδιών (Brody και Elhadad [3]) χρησιμοποιήθηκε το μοντέλο LDA σε επίπεδο προτάσεων. Δηλαδή κάθε πρόταση θεωρείται ως ένα ξεχωριστό κείμενο. Η προσέγγιση αυτή έχει να κάνει με τη φύση των κειμένων των κριτικών χρηστών. Όταν το μοντέλο LDA χρησιμοποιείται για ένα σύνολο ομοειδών προϊόντων (για παράδειγμα laptops) σε

επίπεδο κειμένων, αυτό που κυρίως κάνει τις κριτικές να διαφέρουν μεταξύ τους είναι χαρακτηριστικά τα οποία δε μπορούν να εκτιμηθούν, όπως η εταιρεία κατασκευής. Ενώ, όταν το μοντέλο LDA χρησιμοποιείται σε επίπεδο προτάσεων, οι προτάσεις διαφέρουν σε χαρακτηριστικά (π.χ. μνήμη, επεξεργαστής) που αντιστοιχούν στις επιθυμητές θεματικές ενότητες.

Για την εύρεση του αριθμού των θεματικών ενοτήτων που θα χρησιμοποιηθούν, ακολουθήθηκε μια διαδικασία επικύρωσης cluster (cluster validation [17]), στην οποία κάθε cluster αντιστοιχεί σε μια θεματική ενότητα. Επίσης, κάθε πρόταση θεωρούμε πως ανήκει σε ένα μόνο cluster, σε εκείνο που αντιστοιχεί στην πιθανότερη θεματική της ενότητα, όπως εκτιμάται από το μοντέλο LDA. Πιο αναλυτικά, για κάθε υποψήφιο αριθμό θεματικών ενοτήτων, και με βάση τις κατανομές που προέκυψαν από την εφαρμογή του μοντέλου LDA σε ένα σύνολο προτάσεων *D,* δημιουργούμε τον πίνακα C_k, του οποίου ένα στοιχείο (i,j) είναι 1 αν οι προτάσεις d_i, d_j περιέχονται στο ίδιο cluster. Επίσης, με τον ίδιο τρόπο δημιουργούμε και τον πίνακα R_k βασισμένο σε τυχαία ανάθεση των προτάσεων σε clusters. Στη συνέχεια, γίνεται επιλογή ενός υποσύνολου *D′* των παραπάνω προτάσεων, για το οποίο δημιουργούμε τους αντίστοιχους πίνακες C'_k και R'_k, όπως παραπάνω. Με βάση τα παραπάνω αποτελέσματα και τη συνάρτηση σταθερότητας:

$$F(C, C') = \frac{\sum_{i,j} 1\{C_{i,j} = C'_{i,j} = 1, d_i, d_j \in D'\}}{\sum_{i,j} 1\{C_{i,j} = 1, d_i, d_j \in D'\}}$$

υπολογίζεται το score:

$$score = F(C',C) - F(R',R).$$

Δεδομένου του αριθμού clusters (θεματικών ενοτήτων), ο παραπάνω υπολογισμός επαναλαμβάνεται *n* φορές και υπολογίζεται ο μέσος όρος των scores των επαναλήψεων. Η παραπάνω διαδικασία

επαναλαμβάνεται για αριθμό clusters (θεματικών ενοτήτων) από 10 έως 20 και επιλέγεται ο αριθμός clusters με το μεγαλύτερο μέσο score.

Στη συνέχεια, για κάθε λέξη *w* και θεματική ενότητα *t* υπολογίζεται το παρακάτω score το οποίο μετράει τη συσχέτιση της λέξης με αυτή την ενότητα. Πρόκειται ουσιαστικά για μια παραλλαγή του pointwise mutual information [18].

$$Score_t\,(w) = \mathrm{p(w,\,t)} \cdot \log \frac{p(w,t)}{p(w)\cdot p(t)},$$

Οι πιθανότητες p(w), p(t), p(w,t) υπολογίζονται χρησιμοποιώντας τις εκτιμήσεις του μοντέλου LDA.

Για κάθε θεματική ενότητα t, επιλέγεται το 75% των λέξεων με το μεγαλύτερο score. Το ποσοστό αυτό επιλέχθηκε, καθώς δίνει ως αποτέλεσμα ένα σχετικά μικρό σύνολο λέξεων-κλειδιών (100-200). Έτσι, παράγεται ένα σύνολο από θεματικές ενότητες με τις αντίστοιχες λέξεις-κλειδιά.

3.2.3 Άλλες μέθοδοι εξαγωγής λέξεων-κλειδιών με LDA

Έχουν προταθεί και άλλοι αλγόριθμοι εξαγωγής λέξεων-κλειδιών οι οποίοι βασίζονται σε επεκτάσεις του LDA (Titov & McDonald 2008, Zhao κ.ά. 2010). Για παράδειγμα, η επέκταση του LDA των Titov & McDonald (2008) έχει δύο επίπεδα θεματικών ενοτήτων, τις καθολικές και τις τοπικές. Στις καθολικές θεματικές ενότητες αντιστοιχούν χαρακτηριστικά που στην πλειοψηφία τους δεν γίνεται να εκτιμηθούν, όπως το μοντέλο του προϊόντος, ενώ στις τοπικές θεματικές ενότητες αντιστοιχούν χαρακτηριστικά τα οποία μπορούν να εκτιμηθούν, για παράδειγμα η ποιότητα ήχου. Κατά συνέπεια, οι τελικές λέξεις-κλειδιά παράγονται από τις τοπικές θεματικές ενότητες.

Ο αλγόριθμος των Titov και McDonald (2008) είναι αρκετά πολύπλοκος και δεν παράγει καλύτερα αποτελέσματα από τον αλγόριθμο των Brody και Elhadad (σύμφωνα με την ανάλυση των τελευταίων). Επίσης, επειδή θεωρούμε πως μέθοδοι όπως των Hu κ.ά. και Blair-Goldensohn κ.ά. (βλ. προηγούμενες ενότητες) που απαιτούν τη χειρωνακτική κατασκευή κανόνων ή/και πόρους όπως το WordNet είναι δύσκολο να μεταφερθούν σε νέα γνωστικά πεδία (domains) ή νέες γλώσσες, επιλέξαμε τη μέθοδο των Brody και Elhadad ως βάση για την εξαγωγή λέξεων-κλειδιών στο σύστημα της παρούσας εργασίας.

3.3 Εξαγωγή λέξεων-κλειδιών στο σύστημα της παρούσας εργασίας

Για την εξαγωγή των λέξεων-κλειδιών που θα χρησιμοποιηθούν στο σύστημα της παρούσας εργασίας, αρχικά ακολουθούμε την προσέγγιση των Brody κ.ά. και εκπαιδεύουμε το μοντέλο LDA σε επίπεδο προτάσεων, ώστε οι θεματικές ενότητες που θα προκύψουν να αντιστοιχούν σε χαρακτηριστικά των προϊόντων που μπορούν να εκτιμηθούν. Ύστερα, εξετάζοντας τις πιο αντιπροσωπευτικές λέξεις κάθε θεματικής ενότητας, επιδιώκουμε να βρούμε αυτόματα και το όνομα του χαρακτηριστικού που αντιστοιχεί στη θεματική ενότητα, το οποίο θέλουμε τελικά να εξαγάγουμε.

Πιο συγκεκριμένα, στα πειράματα αυτής της εργασίας εκπαιδεύσαμε το LDA στις προτάσεις του συνόλου δεδομένου Laptops1.[3] Θέσαμε τον αριθμό των θεματικών ενοτήτων ίσο με 14, καθώς μελετώντας το σύνολο δεδομένων παρατηρήσαμε πως αυτός είναι (περί-

[3] Χρησιμοποιήσαμε τους αλγορίθμους LDA του MALLET (βλ. http:// http://mallet.cs.umass.edu/) (Η υλοποίησή τους είναι σε γλώσσα Java).

που) ο ιδανικός αριθμός χαρακτηριστικών. Προέκυψαν 14 θεματικές ενότητες με τις αντίστοιχες αντιπροσωπευτικές λέξεις.

Ύστερα για κάθε πρόταση κάθε κειμένου (στα πειράματα, του συνόλου Laptops1) αναγνωρίζουμε τα μέρη του λόγου (POS tagging) και κρατάμε ως υποψήφιες λέξεις-κλειδιά τα ουσιαστικά.[4]

Για κάθε υποψήφια λέξη-κλειδί η οποία εμφανίζεται τουλάχιστον σε 15 προτάσεις, υπολογίζουμε (χρησιμοποιώντας το μοντέλο LDA) την πιθανότητα να ανήκει σε κάθε μια από τις θεματικές ενότητες. Αν η πιθανότητα αυτή είναι μεγαλύτερη από 10^{-3} σε λιγότερες από 3 θεματικές ενότητες, η υποψήφια λέξη-κλειδί θεωρείται ότι είναι πράγματι λέξη-κλειδί, διαφορετικά την απορρίπτουμε, γιατί συνδέεται με πολλές θεματικές ενότητες.

Τέλος, για κάθε λέξη-κλειδί που προέκυψε, ελέγχουμε αν υπάρχει κάποια άλλη λέξη-κλειδί με την οποία να συνεμφανίζονται τουλάχιστον στο 60% των προτάσεων στις οποίες περιέχεται η πρώτη λέξη-κλειδί Αν αυτό συμβαίνει, οι δύο λέξεις αφαιρούνται από τη λίστα των λέξεων-κλειδιών και συνθέτουν μια νέα μεγαλύτερη (π.χ. «battery life»).

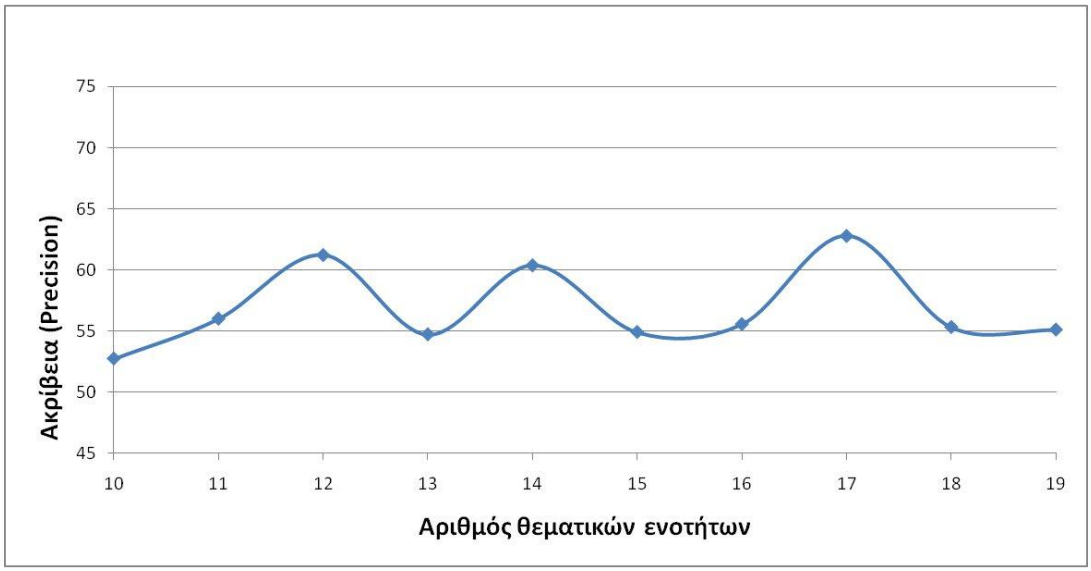
Παρακάτω παρουσιάζονται τα αποτελέσματα του αλγορίθμου για διάφορες κατηγορίες προϊόντων:

Κατηγορία	Λέξεις-κλειδιά
Laptops	processor, keyboard, webcam, operating system, video, graphics, memory card, ram, support, battery life, usb, performance, drive, windows, quality, speed, webcam, touchpad, price

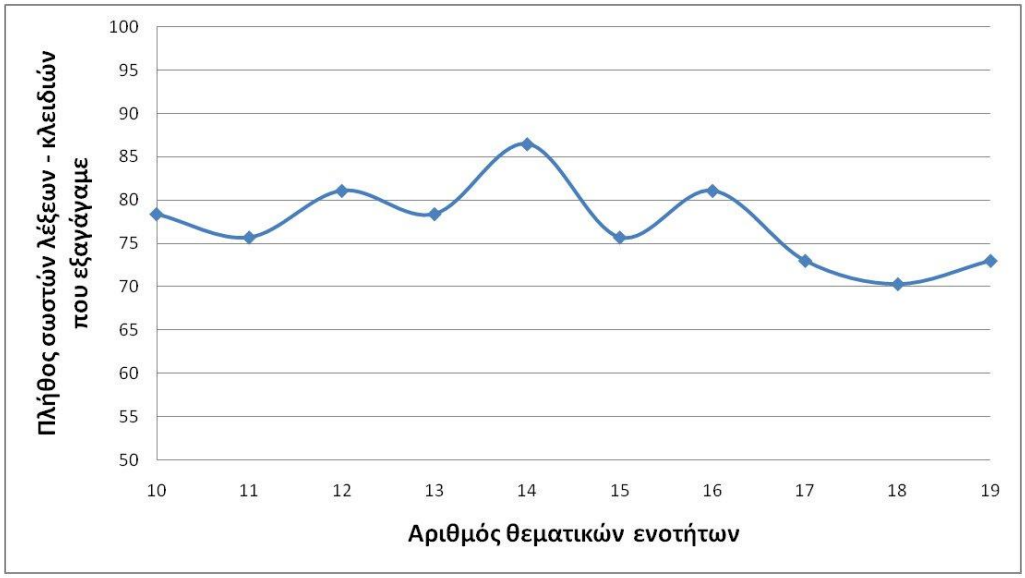
[4] Χρησιμοποιήσαμε τον Stanford tagger (βλ.http://nlp.stanford.edu/software/tagger.shtml).

Φωτογραφικές Μηχανές	zoom, pocket size, touch screen, computer, memory, battery, buttons, settings, lcd, stabilization, megapixels, lens, colors, price, menu, resolution, usb

Για να αξιολογήσουμε τα αποτελέσματά μας, υπολογίσαμε την τιμή της ακρίβειας (precision) (το ποσοστό των λέξεων–κλειδιών που εξαγάγαμε οι οποίες είναι όντως λέξεις-κλειδιά) και του πλήθους των σωστών λέξεων-κλειδιών που εξαγάγαμε για διαφορετικά πλήθη θεματικών ενοτήτων. Ο υπολογισμός τους έγινε χειρωνακτικά και παρακάτω φαίνονται τα διαγράμματα που προκύπτουν για τα σύνολα δεδομένων μας:



Διάγραμμα 1: Ακρίβεια (precision) εξαγωγής λέξεων-κλειδιών για διαφορετικά πλήθη θεματικών ενοτήτων



Διάγραμμα 2: Πλήθος σωστών λέξεων - κλειδιών που εξαγάγαμε για διαφορετικά πλήθη θεματικών ενοτήτων στο σύνολο δεδομένων Laptops 1.

Όπως παρατηρούμε, η μέγιστη τιμή του Precision επιτυγχάνεται αν επιλέξουμε 17 θεματικές ενότητες, για το πλήθος αυτό όμως, το πλήθος των σωστών λέξεων που εξαγάγαμε είναι αρκετά μικρό, συνεπώς θεωρούμε πως η βέλτιστη τιμή πλήθους θεματικών ενοτήτων είναι 14, όπου επιτυγχάνεται υψηλή τιμή και για τα δύο μέτρα που εξετάζουμε.

4. Εξόρυξη γνώμης

Στο κεφάλαιο αυτό παρουσιάζουμε ενδεικτικά μερικές προηγούμενες μεθόδους εξόρυξης γνώμης και κατόπιν εστιαζόμαστε στη μέθοδο που χρησιμοποιείται στο σύστημα της παρούσας εργασίας.

4.1 Προγενέστερες μέθοδοι εξόρυξης γνώμης

Στον τομέα της εξόρυξης γνώμης (opinion mining) έχουν παρουσιαστεί διάφορες προσεγγίσεις. Πολλές από αυτές [9, 10] χρησιμοποιούν υπολογιστικά λεξικά, όπως το WordNet [11]. Οι περισσότερες, όπως και η δική μας, προϋποθέτουν επίσης έναν αλγόριθμο εξαγωγής λέξεων-κλειδιών.

4.1.1 Ο αλγόριθμος των Wilson, Wiebe και Hoffman

Οι Wilson, Wiebe και Hoffman [12] προσπαθούν να προσδιορίσουν το συναίσθημα που εκφράζει ένας χρήστης σε κάθε πρόταση ενός κειμένου του, χρησιμοποιώντας τεχνικές μηχανικής μάθησης. Ο αλγόριθμός τους χωρίζεται σε δύο βασικά στάδια:

1. Έλεγχος αν η πρόταση είναι υποκειμενική, δηλαδή αν εκφράζει συναίσθημα ή όχι.
2. Προσδιορισμός της πολικότητας των υποκειμενικών προτάσεων. Μια πρόταση μπορεί να εκφράζει θετική γνώμη («The food was great»), αρνητική γνώμη («The decoration was awful»), θετική και αρνητική γνώμη μαζί («The sound quality was excellent, but the battery didn’t last over one hour.») ή ουδέτερη γνώμη («Jerome says the hospital feels no different than a hospital in the states»).

Ο αλγόριθμος χρησιμοποιεί λεξικό 8000 υποκειμενικών όρων (π.χ. «like», «love»). Η δημιουργία του λεξικού βασίστηκε σε μια αρχική λίστα υποκειμενικών όρων, που κατασκευάστηκε από τους Riloff και Wiebe [19]. Οι Riloff και Wiebe χρησιμοποίησαν τεχνικές μηχανικής μάθησης για να προσδιορίσουν για ένα σύνολο προτάσεων ποιες από αυτές είναι υποκειμενικές ή ουδέτερες. Στη συνέχεια, χρησιμοποιώντας και πάλι μεθόδους μηχανικής μάθησης και τα αποτελέσματα του προηγούμενου βήματος, έκαναν εξαγωγή προτύπων (patterns) υποκειμενικών προτάσεων. Για παράδειγμα από την υποκειμενική πρόταση «I love the design» παράγεται το πρότυπο «<x> love <y>». Από τα πρότυπα αυτά προήλθαν και οι όροι της αρχικής λίστας.

Κάθε όρος του λεξικού μπορεί να είναι ισχυρά υποκειμενικός ή ασθενώς υποκειμενικός, ανάλογα με το βαθμό βεβαιότητας που έχουμε για την πολικότητα του. Επίσης, περιλαμβάνεται και μια αρχική εκτίμηση της πολικότητας κάθε όρου, αν δηλαδή είναι θετική, αρνητική, και τα δύο ή ουδέτερη. Η πολικότητα ενός όρου είναι αρνητική ή θετική αν εκφράζει αρνητικό ή θετικό συναίσθημα αντίστοιχα. Αν ένας όρος μπορεί να χρησιμοποιηθεί για να εκφράσει και θετικό και αρνητικό συναίσθημα, η πολικότητά του είναι θετική και αρνητική. Ακόμη, στην αρχική λίστα όρων του λεξικού προστέθηκαν όροι από τις λίστες θετικών και αρνητικών λέξεων του General Inquirer (General Inquirer, 2000), οι οποίοι κρίθηκαν από τους συγγραφείς ως υποκειμενικοί. Τέλος, το λεξικό περιλαμβάνει και όρους που μπορεί να μην εκφράζουν κανένα συναίσθημα, μπορούν όμως να χρησιμοποιηθούν ως συνοδευτικοί κατά την έκφραση θετικού και αρνητικού συναισθήματος (π.χ. «feel», «look»). Οι όροι αυτοί έχουν ουδέτερη πολικότητα.

Για τη διεξαγωγή των πειραμάτων χρησιμοποιήθηκε ένα σύνολο κειμένων, στα οποία κάθε πρόταση είχε επισημειωθεί χειρωνακτικά ως προς την πολικότητά της.

Αρχικά, για να διαπιστωθεί κατά πόσο η αρχική εκτίμηση της πολικότητας ενός όρου που υπάρχει στο λεξικό συμφωνεί με την πολικότητα της πρότασης στην οποία εμφανίζεται, οι Wilson κ.ά. εκπαίδευσαν έναν ταξινομητή με την υπόθεση πως η πολικότητα κάθε πρότασης είναι αυτή του όρου του λεξικού που περιέχει. Από ό,τι όμως αποδείχτηκε, από τα αποτελέσματα των πειραμάτων, η αρχική εκτίμηση (που υπάρχει στο λεξικό) της πολικότητας ενός όρου τις περισσότερες φορές δεν συμφωνεί με την πολικότητα μιας πρότασης που περιέχει τον όρο. Αυτό συμβαίνει γιατί σε πολλές περιπτώσεις, όροι που μπορούν να εκφράσουν υποκειμενικότητα χρησιμοποιούνται σε ουδέτερες προτάσεις και αυτός είναι κυρίως ο λόγος που αρχικά γίνεται έλεγχος για την υποκειμενικότητα ή μη μιας πρότασης. Για να ξεπεραστεί το πρόβλημα αυτό, έγινε η εξής παραδοχή, η οποία εφαρμόστηκε στις προτάσεις που εξετάστηκαν στο στάδιο της εκπαίδευσης του συστήματος: Αν ένας όρος του λεξικού δεν περιέχεται σε κάποια υποκειμενική πρόταση, τότε είναι ουδέτερος. Αν όμως περιέχεται έστω και σε μια υποκειμενική πρόταση, τότε είναι υποκειμενικός και η πολικότητά του είναι αυτή της πρότασης. Αν ο όρος περιέχεται σε προτάσεις με θετική και με αρνητική πολικότητα, τότε η πολικότητά του είναι θετική ή αρνητική ανάλογα με τα συμφραζόμενα.

Και για τα δύο στάδια του αλγορίθμου χρησιμοποιήθηκε ο ταξινομητής AdaBoost (Schapire και Singer, 2000). Για το πρώτο στάδιο, του προσδιορισμού της υποκειμενικότητας, οι ιδιότητες που χρησιμοποιήθηκαν είναι οι παρακάτω:

- Ιδιότητες σε επίπεδο λέξεων (όρων λεξικού): Περιλαμβάνουν για κάθε λέξη που εμφανίζεται στο κείμενο και περιέχεται και στο λεξικό τις παρακάτω 5 ιδιότητες:
    - Την ίδια τη λέξη (αλφαριθμητικό).
    - Την προηγούμενή της λέξη (αλφαριθμητικό).
    - Την επόμενή της λέξη (αλφαριθμητικό).

    - Την αρχική πολικότητα της λέξης σύμφωνα με το λεξικό (αλφαριθμητικό).
- Το βαθμό βεβαιότητάς μας για την πολικότητα της λέξης σύμφωνα με το λεξικό (αλφαριθμητικό). Ιδιότητες τροποποίησης λέξεων (όρων λεξικού): Περιλαμβάνουν για κάθε λέξη που εμφανίζεται στο κείμενο και περιέχεται και στο λεξικό τις παρακάτω 8 ιδιότητες:
    - Αν πριν από τη λέξη υπάρχει επίθετο (δυαδική).
    - Αν πριν από τη λέξη υπάρχει επίρρημα (δυαδική).
    - Αν πριν από τη λέξη υπάρχει κάποια λέξη με ειδικό βάρος (intensifier), σύμφωνα με τη λίστα λέξεων με ειδικό βάρος που έχουν δημιουργήσει οι συγγραφείς για τα πειράματά τους (δυαδική).
    - Αν η λέξη που εμφανίζεται στο κείμενο και περιέχεται και στο λεξικό είναι έχει ειδικό βάρος.
- Ιδιότητες δομής: Περιλαμβάνουν δυαδικές ιδιότητες που προκύπτουν από το δέντρο εξαρτήσεων (dependency tree) της πρότασης. Δηλώνουν αν η πρόταση έχει αντικείμενο, αν έχει συνδετικό ρήμα και αν έχει παθητικό ρήμα.
- Ιδιότητες επιπέδου πρότασης: Περιλαμβάνουν το πλήθος των ισχυρά υποκειμενικών και ασθενώς υποκειμενικών όρων της εξεταζόμενης πρότασης, της προηγούμενής της και της επόμενής της. Επίσης, περιλαμβάνουν το πλήθος των επιθέτων και των επιρρημάτων, και δυαδικές ιδιότητες που δηλώνουν την ύπαρξη αντωνυμιών, αριθμών και τροπικών βοηθητικών ρημάτων εκτός του «will».
- Ιδιότητες επιπέδου κειμένου: Περιλαμβάνουν μια μόνο ιδιότητα που είναι το θέμα του κειμένου (αλφαριθμητικό), το οποίο είναι γνωστό εκ των προτέρων.

Σύμφωνα με τα αποτελέσματα που παρουσίασαν οι Wilson κ.ά. [12], το ποσοστό ορθότητας (accuracy) toy ταξινομητή για διάφορα σύνολα δεδομένων είναι περίπου 75%.

Ο ταξινομητής που αναπτύχθηκε για το δεύτερο στάδιο του αλγορίθμου, δηλαδή για τον προσδιορισμό της πολικότητας, εξετάζει μόνο τις υποκειμενικές προτάσεις που προέκυψαν από το προηγούμενο βήμα. Η πολικότητα μιας πρότασης μπορεί να είναι θετική, αρνητική, αρνητική και θετική ή ουδέτερη. Ο ταξινομητής που χρησιμοποιείται είναι και πάλι ο AdaBoost. Οι ιδιότητες που χρησιμοποιούνται είναι οι παρακάτω:

- Ιδιότητες σε επίπεδο λέξεων (όρων λεξικού): Περιλαμβάνουν για κάθε λέξη του κειμένου η οποία περιέχεται στο λεξικό, τις εξής 2 ιδιότητες:
    - Την ίδια τη λέξη (αλφαριθμητικό).
    - Την αρχική πολικότητα της λέξης με βάση το λεξικό (αλφαριθμητικό). Επίσης περιλαμβάνονται και 8 ιδιότητες σχετικές με την πολικότητα της λέξης αυτής:
        - Η πρώτη ιδιότητα είναι δυαδική και δηλώνει αν στις προηγούμενες 4 λέξεις από την εξεταζόμενη λέξη, υπάρχει κάποια λέξη που δηλώνει άρνηση ή αν υπάρχει κάποια λέξη που δηλώνει άρνηση στους απογόνους της λέξης με βάση το δέντρο εξάρτησης.
        - Η δεύτερη ιδιότητα είναι δυαδική και δηλώνει αν το υποκείμενο της πρότασης έχει άρνηση
        - Η τρίτη ιδιότητα είναι αλφαριθμητικό και με βάση το δέντρο εξάρτησης δηλώνει την πολικότητα του γονέα της εξεταζόμενης λέξης.
        - Η τέταρτη ιδιότητα είναι αλφαριθμητικό και με βάση το δέντρο εξάρτησης και αν η σχέση μεταξύ της εξεταζόμενης λέξης και κάποιου από τα παιδιά της

είναι obj, adj, mod ή vmod, τότε η ιδιότητα παίρνει την τιμή της πολικότητας του παιδιού.

- Η πέμπτη ιδιότητα είναι αλφαριθμητικό και εξετάζει αν η εξεταζόμενη λέξη περιλαμβάνεται σε κάποια σύζευξη. Αν ναι, η ιδιότητα παίρνει την τιμή της πολικότητας του κόμβου – αδερφού, με βάση το δέντρο εξάρτησης.
- Η έκτη ιδιότητα είναι δυαδική και δηλώνει αν στις 4 λέξεις που προηγούνται από την εξεταζόμενη λέξη, υπάρχει κάποια λέξη που χρησιμοποιείται για να τροποποιήσει την πολικότητα
- Η έβδομη ιδιότητα είναι δυαδική και δηλώνει αν στις 4 λέξεις που προηγούνται από την εξεταζόμενη λέξη, υπάρχει κάποια λέξη που χρησιμοποιείται για να δώσει αρνητική πολικότητα στην πρόταση.
- Η όγδοη ιδιότητα είναι δυαδική και δηλώνει αν στις 4 λέξεις που προηγούνται από την εξεταζόμενη λέξη, υπάρχει κάποια λέξη που χρησιμοποιείται για να δώσει θετική πολικότητα στην πρόταση.

Με βάση τα αποτελέσματα που παρουσίασαν οι Wilson κ.ά. στο [12], η ορθότητα (accuracy) του ταξινομητή για διάφορα σύνολα δεδομένων κυμαίνεται από 61% έως 65%. Αν και μεταγενέστερες προσπάθειες επιτυγχάνουν υψηλότερο ποσοστό, ο αλγόριθμος είναι ενδεικτικός του τρόπου με τον οποίο μπορούμε να κάνουμε εξαγωγή γνώμης με μεθόδους μηχανικής μάθησης.

4.1.2 Άλλοι αλγόριθμοι εξαγωγής γνώμης

Ενδεικτικά αναφέρουμε επίσης τις μεθόδους εξόρυξης γνώμης των Godbole, Srinivasaiah και Skiena [10] και Titov και McDonald [13], τις οποίες επίσης μελετήσαμε στη διάρκεια της εργασίας αλλά δεν περιγράφουμε εδώ χάριν συντομίας.

4.2 Εξόρυξη γνώμης στο σύστημα της παρούσας εργασίας

Η ενότητα αυτή περιγράφει τις μεθόδους εξόρυξης γνώμης που δοκιμάστηκαν πειραματικά στη διάρκεια της παρούσας εργασίας, καθώς και τα αποτελέσματά τους. Σε όλες τις περιπτώσεις θεωρούμε εδώ γνωστά τα χαρακτηριστικά των προϊόντων και τις αντι-προσωπευτικές λέξεις-κλειδιά που τα περιγράφουν.

Στο στάδιο αυτής της ενότητας, ασχολούμαστε μόνο με προτάσεις (των κριτικών) που περιλαμβάνουν αντιπροσωπευτικές λέξεις-κλειδιά των χαρακτηριστικών. Για κάθε μία τέτοια πρόταση, επιδιώκουμε να προσδιορίσουμε αν εκφράζει θετική ή αρνητική γνώμη. Αν μια πρό-ταση εκφράζει θετική (αντίστοιχα αρνητική) γνώμη, θεωρούμε ότι η θετική (ή αντίστοιχα αρνητική) γνώμη αναφέρεται στα χαρακτη-ριστικά (συχνά μόνο ένα) που ονοματίζονται από τις λέξεις-κλειδιά της πρότασης.

Για τα πειράματα αυτής της ενότητας χρησιμοποιήσαμε 2.510 προτάσεις από κείμενα με (εντελώς) αρνητική πολικότητα (0/5 αστεράκια) του συνόλου Laptops2 και 2.610 προτάσεις από κείμενα με (εντελώς) θετική πολικότητα (5/5 αστεράκια) από το ίδιο σύνολο δεδομένων. Όπως έχουμε ήδη πει, θεωρούμε πως όλες οι προτάσεις που περιέχονται σε μια κριτική που έχει πάρει 0/5 αστεράκια έχουν αρνητική φόρτιση. Αντίστοιχα όσες προτάσεις περιέχονται σε κριτικές που έχουν λάβει 5/5 θεωρούνται ως θετικά φορτισμένες.

Προκειμένου να αποδείξουμε την ορθότητα της παραπάνω παραδοχής, δηλαδή πως η πολικότητα των προτάσεων που χρησιμοποιούμε συμφωνεί με τη βαθμολογία των κριτικών από τις οποίες προέρχονται οι προτάσεις, ελέγξαμε ένα δείγμα 100 προτά-σεων από το σύνολο που αναφέραμε παραπάνω. Από αυτές το 86% έχει ταξινομηθεί σωστά, δηλαδή οι προτάσεις έχουν όντως την πολικότητα που υποθέτουμε πως έχουν (θετική ή αρνητική), 4% των

προτάσεων έχουν αντίθετη πολικότητα από αυτή που υποθέτουμε πως έχουν και 10% των προτάσεων έχουν ουδέτερη πολικότητα.

4.2.1.Εξόρυξη γνώμης με γλωσσικά μοντέλα

Η πρώτη προσέγγιση που ακολουθήσαμε για τον προσδιορισμό της γνώμης (θετική ή αρνητική) του χρήστη σε κάθε πρόταση χρησιμοποιεί γλωσσικά μοντέλα.

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.

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

Στην περίπτωσή μας, θεωρήσαμε πως πολλοί διαφορετικοί χρήστες όταν γράφουν θετικές (ή αντίστοιχα αρνητικές) κριτικές, χρησιμοποιούν πολλές φορές ίδιες ή παρόμοιες ακολουθίες λέξεων. Για παράδειγμα, σε πολλά διαφορετικά κείμενα θετικών κριτικών είναι πολύ πιθανό να συναντήσει κανείς ακολουθίες λέξεων, όπως «I like» ή « is amazing».

Η μέθοδος που χρησιμοποιήσαμε για την κατάταξη μιας πρότασης στη θετική ή αρνητική κατηγορία (θετική ή αρνητική γνώμη) περιγράφεται παρακάτω:

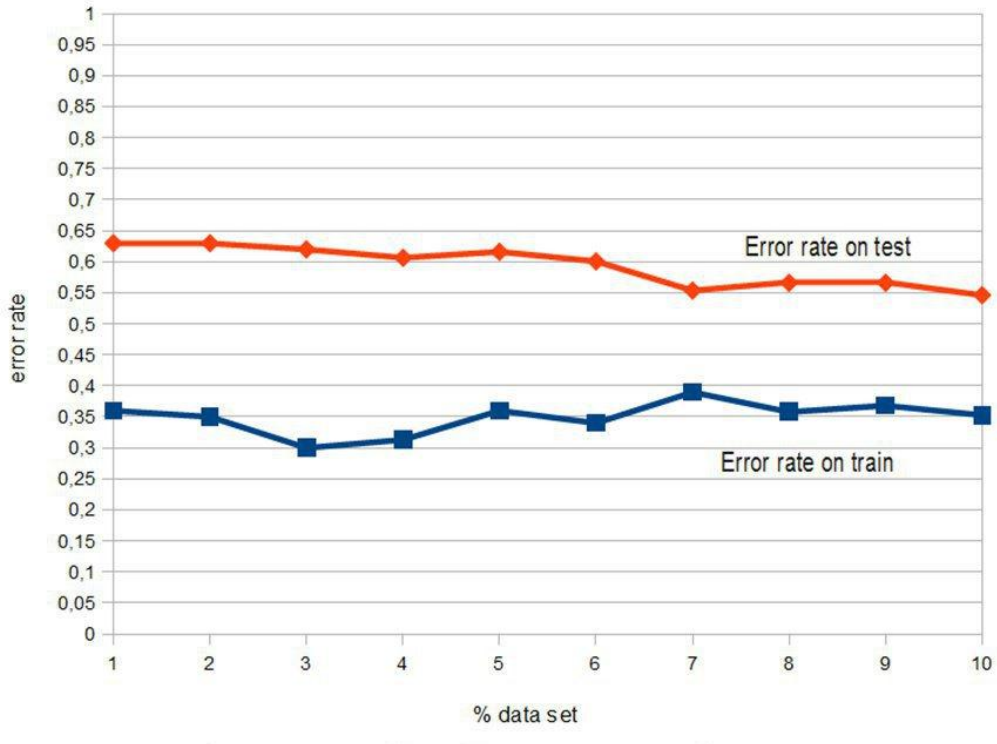
1. Εκπαιδεύσαμε ένα γλωσσικό μοντέλο στο σύνολο των αρνητικά φορτισμένων προτάσεων του συνόλου εκπαίδευσης. (Ο τρόπος

χωρισμού των δεδομένων σε σύνολα εκπαίδευσης και αξιολόγησης περιγράφεται παρακάτω.)

2. Εκπαιδεύσαμε ένα γλωσσικό μοντέλο στο σύνολο των θετικά φορτισμένων προτάσεων του συνόλου εκπαίδευσης.
3. Για κάθε νέα πρόταση, της οποίας την κατηγορία έπρεπε να μαντέψουμε, υπολογίσαμε την πιθανότητα που επιστρέφουν τα δύο εκπαιδευμένα μοντέλα.
4. Κατατάξαμε την πρόταση στην κατηγορία (θετική ή αρνητική) της οποίας το μοντέλο επέστρεψε την μεγαλύτερη πιθανότητα.

Για την εκπαίδευση των γλωσσικών μοντέλων χρησιμοποιήσαμε το εργαλείο SRILM με 2-grams.[5]

Παρακάτω, φαίνονται οι καμπύλες μάθησης που προέκυψαν:



Διάγραμμα 3: Καμπύλες μάθησης για εξόρυξη γνώμης με χρήση γλωσσικών μοντέλων

[5] Βλ. http://www-speech.sri.com/projects/srilm/.

Το ποσοστό λάθους (error rate) είναι το ποσοστό των προτάσεων κάποιου συνόλου (βλ. παρακάτω) που κατατάξαμε λανθασμένα. Για τα πειράματά μας χωρίσαμε το σύνολο δεδομένων σε 11 ίσα τμήματα (περίπου 460 προτάσεις σε κάθε τμήμα). Ένα από αυτά το χρησιμοποιήσαμε για την αξιολόγηση του συστήματος (test data) και τα υπόλοιπα 10 για την εκπαίδευση του (train data). Οι πιθανότητες των γλωσσικών μοντέλων εκτιμούνται από τα δεδομένα εκπαίδευσης. Αρχικά το σύστημά μας εκπαιδεύτηκε χρησιμοποιώντας το 10% των δεδομένων εκπαίδευσης (1 τμήμα) και σε κάθε νέο πείραμα προσθέταμε άλλο ένα 10% των δεδομένων εκπαίδευσης (ένα ακόμα τμήμα).

Η καμπύλη «Error rate on test» του διαγράμματος μας δείχνει πόσο καλά τα πάει το σύστημά μας στα δεδομένα αξιολόγησης, δηλαδή σε δεδομένα διαφορετικά από αυτά που έχει συναντήσει κατά την εκπαίδευσή του. Η καμπύλη «Error rate on train» μας δείχνει πόσο καλά τα πάει το σύστημα στα ίδια δεδομένα στα οποία έχει εκπαιδευτεί, δηλαδή στα τμήματα που συνάντησε κατά την εκπαίδευσή του. Εν γένει ένα σύστημα μάθησης τα πηγαίνει καλύτερα στα δεδομένα στα οποία έχει εκπαιδευτεί και χειρότερα σε δεδομένα που δεν έχει ξανασυναντήσει. Επομένως η καμπύλη «Error rate on train» μπορεί να θωρηθεί διαισθητικά ένα κάτω φράγμα της καμπύλης «Error rate on test». Επίσης, συνήθως το ποσοστό λάθους σε νέα δεδομένα (error rate on test) μειώνεται όσο προστίθενται περισσότερα δεδομένα εκπαίδευσης, ενώ το ποσοστό λάθους στα ίδια τα δεδομένα εκπαίδευσης συνήθως αυξάνεται, επειδή γίνεται δυσκολότερο το σύστημα να κάνει υπερ-εφαρμογή (overfitting), διαισθητικά να απομνημονεύσει ιδιαιτερότητες, των δεδομένων εκπαίδευσης.

Στην περίπτωσή μας, το ποσοστό λάθους σε νέα δεδομένα παρουσιάζει σχετικά μικρή μείωση, ενώ παραμένει σχετικά μεγάλη η απόσταση από την καμπύλη του ποσοστού λάθους στα δεδομένα εκπαίδευσης. Τα στοχεία αυτά αποτελούν ενδείξεις ότι το ποσοστό λάθους σε νέα δεδομένα θα μπορούσε να μειωθεί περαιτέρω χρησιμοποιώντας περισσότερα δεδομένα εκπαίδευσης, αλλά η μείωση θα ήταν μάλλον μικρή.

4.2.2 Εξόρυξη γνώμης με αφελή ταξινομητή Bayes

Στην επόμενη προσέγγισή μας χρησιμοποιήσαμε λογισμικό που είχε αναπτυχθεί από τους Κοσμόπουλο, Ανδρουτσόπουλο και Παλιούρα [14] για τη διήθηση ανεπιθύμητης αλληλογραφίας, δηλαδή για την κατάταξη μηνυμάτων ηλεκτρονικού ταχυδρομείου σε δύο κατηγορίες (spam, ham). Το λογισμικό αυτό υποστηρίζει πολλές παραλλαγές του αφελούς ταξινομητή Bayes (Naïve Bayes) [20]. Χρησιμοποιήσαμε την πολυωνυμική (multinomial) μορφή με διάφορες βελτιώσεις που εξηγούνται στη διπλωματική εργασία του Κοσμόπουλου [21].[6]

Κάθε μήνυμα ή στην περίπτωσή μας κάθε πρόταση εκπαίδευσης ή αξιολόγησης παριστάνεται ως ένα διάνυσμα ιδιοτήτων. Κάθε ιδιότητα δείχνει χονδρικά πόσες φορές εμφανίζεται στο κείμενο μια συγκεκριμένη λέξη. Για την ακρίβεια, κάθε ιδιότητα έχει ως τιμή το tf·idf της αντίστοιχης λέξης. Το term frequency (tf) ισούται με τη συχνότητα εμφάνισης της λέξης στο εξεταζόμενο κείμενο, ενώ το inverse document frequency (idf) ισούται με $log_{10}\frac{N}{df_w}$, όπου df_w είναι

[6] Πιο συγκεκριμένα χρησιμοποιούμε τη μορφή του αφελούς ταξινομητή Bayes που το λογισμικό ονομάζει «`Multinomial with TF transformed attributes`».

το πλήθος των κειμένων που περιέχουν τη λέξη και N το συνολικό πλήθος των εγγράφων.

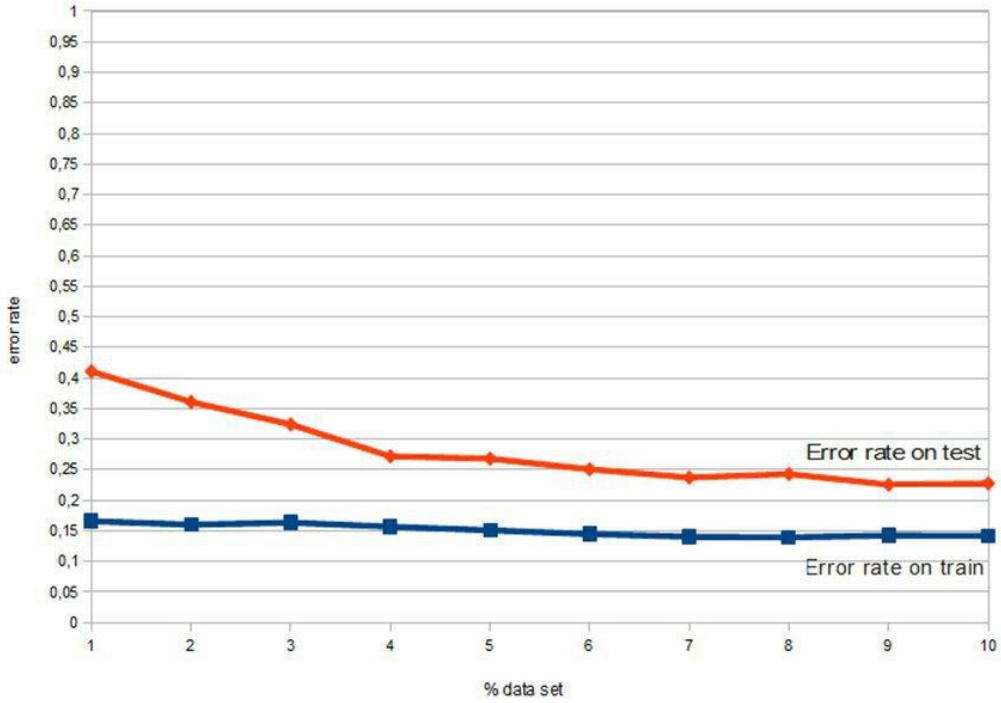
Οι λέξεις για τις οποίες θα υπάρχουν ιδιότητες στα διανύσματα επιλέγονται (feature selection) ως εξής:

1. Εντοπίζονται οι λέξεις οι οποίες υπάρχουν σε τουλάχιστον 5 κείμενα εκπαίδευσης.
2. Για κάθε μια από τις παραπάνω λέξεις, υπολογίζεται το κέρδος πληροφορίας IG(X,C):

   $IG(X,C) = \sum_{x\in\{0,1\},c\in\{c_+,c_-\}} P(X=x \cap C=c) \cdot log_2 \frac{P(X=x \cap C=c)}{P(X=x)\cdot P(C=c)}$

   όπου C τυχαία μεταβλητή που παριστάνει την κατηγορία (στην περίπτωσή μας θετική ή αρνητική γνώμη) του μηνύματος ή πρότασης και X δυαδική (Boolean) τυχαία μεταβλητή που δείχνει αν η λέξη εμφανίζεται ή όχι στο μήνυμα ή πρόταση. Οι ιδιότητες, δηλαδή, επιλέγονται χρησιμοποιώντας τις δυαδικές μορφές τους, ενώ μετά την επιλογή των ιδιοτήτων, οι επιλεγμένες ιδιότητες λαμβάνουν ως τιμές τα tf-idf των λέξεων. Χρησιμοποιούμε στα διανύσματα ιδιότητες για τις 1000 λέξεις με τα υψηλότερα κέρδη πληροφορίας.

Οι καμπύλες μάθησης του αφελούς ταξινομητή Bayes, που προέκυψαν όπως στην περίπτωση των γλωσσικών μοντέλων, είναι οι παρακάτω:



Διάγραμμα 4: Καμπύλες μάθησης για εξόρυξη γνώμης με χρήση του αφελούς ταξινομητή Bayes

Παρατηρούμε πως τα αποτελέσματα αυτής της μεθόδου είναι καλύτερα από εκείνα των γλωσσικών μοντέλων.

4.2.3 Εξόρυξη γνώμης με μηχανή διανυσμάτων υποστήριξης

Στη συνέχεια προσπαθήσαμε να βελτιώσουμε την ακρίβεια του συστήματος χρησιμοποιώντας μια μηχανή διανυσμάτων υποστήριξης (Support Vector Machine, SVM) που έχει χρησιμοποιηθεί ευρύτατα για κατηγοριοποίηση κειμένων [22] με πολύ καλά αποτελέσματα.

Για τα πειράματά μας χρησιμοποιήσαμε την υλοποίηση libsvm με πυρήνα RBF.[7]

Εκτός από την αλλαγή του αλγορίθμου μάθησης, αφαιρούμε πλέον επίσης και τις πολύ συχνές λέξεις (stopwords), με βάση μια χειρωνακτικά κατασκευασμένη λίστα, τα σημεία στίξης και τις καταλήξεις από τα ρήματα.[8] Η επιλογή ιδιοτήτων καθώς και οι τιμές τους, υπολογίζονται όπως και στην περίπτωση του αφελούς ταξινομητή Bayes, αλλά κρατάμε πλέον τις 700 καλύτερες ιδιότητες.[9] Οι καμπύλες μάθησης, που προέκυψαν όπως και στην περίπτωση των γλωσσικών μοντέλων, φαίνονται παρακάτω:

Διάγραμμα 5: Καμπύλες μάθησης για εξόρυξη γνώμης με χρήση SVMs

[7] Βλ. http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Κάνοντας επιλογή παραμέτρων (tuning) στα δεδομένα εκπαίδευσης, παρατηρήσαμε πως οι τιμές των παραμέτρων του SVM που δίνουν το καλύτερο το αποτέλεσμα είναι C=2 και g=0.125. Αυτές οι τιμές χρησιμοποιήθηκαν για την κατάταξη των δεδομένων αξιολόγησης.

[8] Βλ. http://tartarus.org/~martin/PorterStemmer/

[9] Χρησιμοποιήσαμε λιγότερες ιδιότητες (αντί των 1.000), επειδή αρχικά πειράματα έδειξαν μεγάλη απόσταση μεταξύ των καμπυλών error rate on test και error rate on train, ενδεικτική προβλήματος high variance, για την αντιμετώπιση του οποίου συνιστάται, μεταξύ άλλων, η μείωση του πλήθους των ιδιοτήτων.

Παρατηρούμε πως με ολόκληρο το σύνολο των δεδομένων εκπαίδευσης επιτυγχάνουμε χαμηλότερο ποσοστό λάθους (21%) στα δεδομένα αξιολόγησης, ενώ με τον ταξινομητή Naïve Bayes το αντίστοιχο αποτέλεσμα ήταν ελαφρά χειρότερο (22,7%).

Η υλοποίηση της μηχανής διανυσμάτων υποστήριξης που χρησιμοποιούμε επιστρέφει και ένα βαθμό βεβαιότητας (στο [0,1]) για κάθε απόφαση κατάταξης. Θεωρούμε πως αν το ποσοστό βεβαιότητας για την ταξινόμηση μιας πρότασης (στη θετική ή αρνητική κατηγορία) είναι μεγαλύτερο από 60%, είμαστε αρκετά σίγουροι ότι η πρόταση ταξινομήθηκε σωστά, αλλιώς δε λαμβάνουμε την πρόταση υπόψη μας. Η τελική πολικότητα του χαρακτηριστικού c ενός προϊόντος υπολογίζεται ως εξής:

$$\mathrm{pol(c)} = \frac{\textit{αριθμός θετικών προτάσεων που αναφέρουν το c}}{\textit{αριθμός θετικών και αρνητικών προτάσεων που αναφέρουν το c}} .$$

Τέλος, για λόγους παρουσίασης, κρατάμε για κάθε χαρακτηριστικό το σύνολο των προτάσεων που το αναφέρουν (και κατετάγησαν με βεβαιότητα πάνω από 80%) ως ενδεικτικές φράσεις της γνώμης του κοινού για το χαρακτηριστικό αυτό.

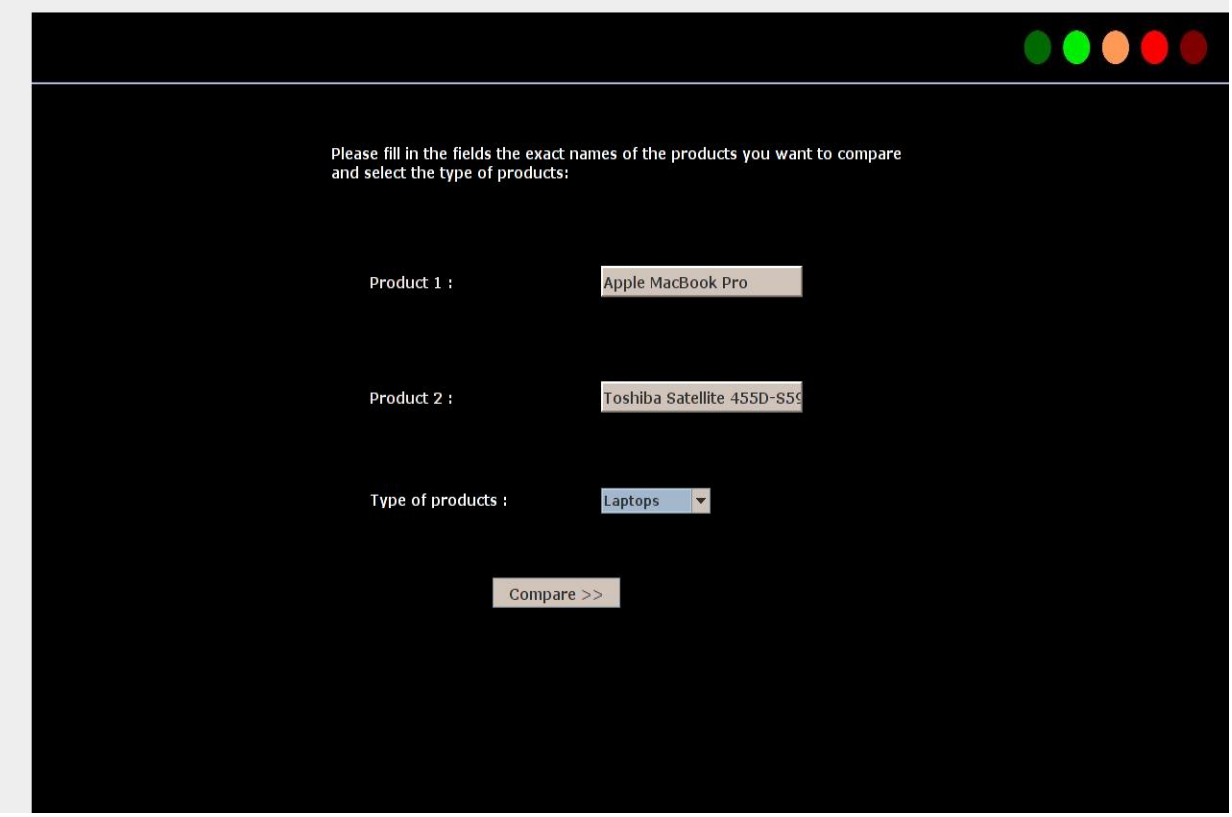
5. Τελική παρουσίαση των προϊόντων

Στα προηγούμενα κεφάλαια περιγράψαμε (α) πώς βρίσκουμε τα κυριότερα χαρακτηριστικά των προϊόντων που θέλουμε να συγκρίνουμε και (β) πώς βρίσκουμε τη γνώμη (πολικότητα) των χρηστών για κάθε χαρακτηριστικό των δύο συγκρινόμενων προϊόντων. Για την τελική παρουσίαση των προϊόντων, δίνουμε έμφαση στα χαρακτηριστικά που οι χρήστες αναφέρουν συχνά και για τα δύο προϊόντα, αλλά με μεγάλη διαφορά πολικότητας, δηλαδή μας ενδιαφέρουν ιδιαιτέρως κοινά χαρακτηριστικά (π.χ. οθόνη, μπαταρία) ως προς τα οποία τα δύο προϊόντα αξιολογούνται πολύ διαφορετικά (π.χ. πολύ κακή και πολύ καλή, αντίστοιχα, οθόνη ή μπαταρία). Ακόμη συμπεριλαμβάνουμε χαρακτηριστικά τα οποία μπορεί να αναφέρονται μόνο σε ένα από τα δύο προϊόντα, αλλά με έντονα θετική ή αρνητική πολικότητα.

Έχοντας υπολογίσει, όπως στα προηγούμενα κεφάλαια, την πολικότητα των κυριότερων χαρακτηριστικών των δύο προϊόντων, ακολουθούμε τα παρακάτω βήματα για την τελική παρουσίασή τους:

- Υπολογίζουμε την απόλυτη διαφορά πολικότητας (στα δύο προϊόντα) των κοινών τους χαρακτηριστικών και ταξινομούμε τα κοινά τους χαρακτηριστικά κατά φθίνουσα σειρά αυτής της απόλυτης διαφοράς.
- Ταξινομούμε κατά φθίνουσα πολικότητα τα χαρακτηριστικά που αναφέρουν οι κριτικές μόνο για ένα από τα δύο προϊόντα.
- Στην τελική παρουσίαση των προϊόντων περιλαμβάνουμε τα χαρακτηριστικά του πρώτου βήματος και αν αυτά δεν ξεπερνούν τα 10, συμπληρώνουμε με τα καλύτερα χαρακτηριστικά του δεύτερου βήματος.

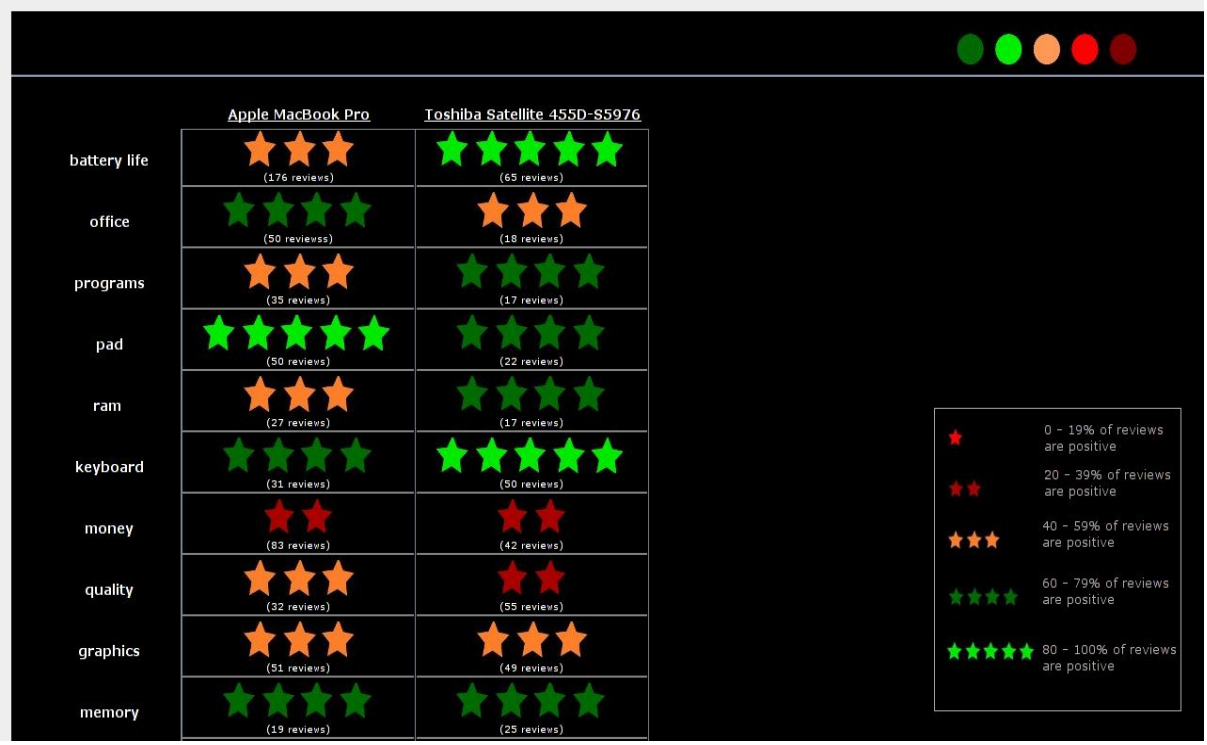
Παρακάτω φαίνεται ένα παράδειγμα χρήσης του συστήματος της εργασίας.[10] Αρχικά ο χρήστης συμπληρώνει τα ονόματα των προϊόντων που θέλει να συγκρίνει και επιλέγει το είδος τους:



Έστω πως ο χρήστης επιλέγει να συγκρίνει τα laptop Apple MacBook Pro και Toshiba Satellite 455D-S5976. Συμπληρώνει τα ονόματα των προϊόντων και επιλέγει τον τύπο τους (π.χ. laptop και όχι φωτογραφική μηχανή) και στη συνέχεια επιλέγει Compare.

Τα αποτελέσματα που προκύπτουν φαίνονται παρακάτω:

[10] Η υλοποίηση της διεπαφής παρουσίασης των αποτελεσμάτων έγινε σε Java χρησιμοποιώντας τις βιβλιοθήκες Swing και AWT.



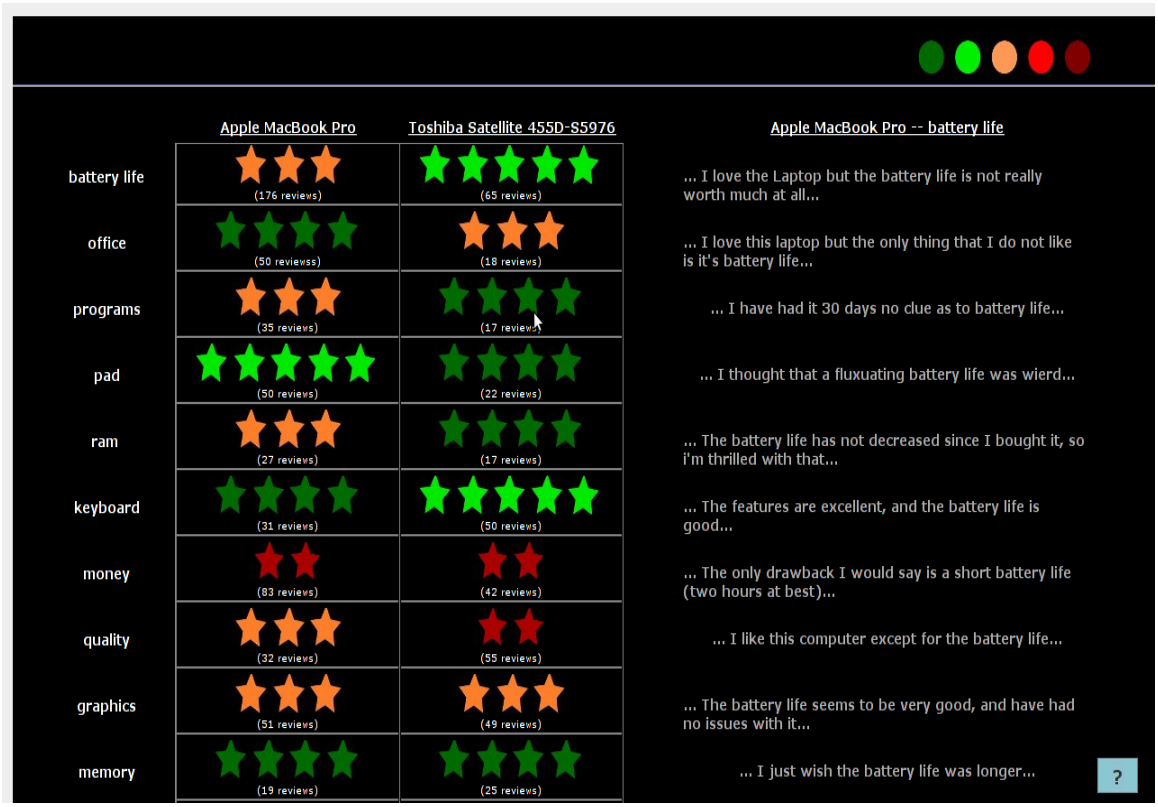
Στην πρώτη στήλη υπάρχουν τα ονόματα των χαρακτηριστικών που επιλέγονται και στις δύο επόμενες στήλες φαίνεται η γνώμη των χρηστών για καθένα από τα προϊόντα. Επίσης, κάτω από τη βαθμολογία του κάθε χαρακτηριστικού φαίνεται ο αριθμός των κριτικών από τις οποίες προέκυψε η βαθμολογία αυτή.

Παρακάτω φαίνεται ο τρόπος με τον οποίο προέκυψε η βαθμολογία για κάθε χαρακτηριστικό:

5 αστεράκια	Το χαρακτηριστικό βαθμολογήθηκε θετικά σε περισσότερο από το 80% των προτάσεων των κριτικών.
4 αστεράκια	Το χαρακτηριστικό βαθμολογήθηκε θετικά σε ποσοστό μεταξύ 60% και 79% των προτάσεων των κριτικών.
3 αστεράκια	Το χαρακτηριστικό βαθμολογήθηκε θετικά σε

	ποσοστό μεταξύ 40% και 59% των προτάσεων των κριτικών.
2 αστεράκια	Το χαρακτηριστικό βαθμολογήθηκε θετικά σε ποσοστό μεταξύ 20% και 39% των προτάσεων των κριτικών.
1 αστεράκι	Το χαρακτηριστικό βαθμολογήθηκε θετικά σε λιγότερο από το 20% των προτάσεων των κριτικών.

Επίσης, δίνεται η δυνατότητα στο χρήστη επιλέγοντας οποιοδήποτε χαρακτηριστικό ενός προϊόντος να δει ενδεικτικά ένα σύνολο φράσεων που έχουν χρησιμοποιηθεί από διάφορους χρήστες για το σχολιασμό του χαρακτηριστικού αυτού.



Στο παραπάνω παράδειγμα, φαίνεται η φράση που εμφανίζεται στο χρήστη, όταν αυτός επιλέγει το χαρακτηριστικό battery life του laptop Toshiba Satellite 455D-S5976.

Στον παρακάτω πίνακα φαίνονται ενδεικτικά τα αποτελέσματα πολικότητας (ως αστεράκια) της μεθόδου μας και τα αντίστοιχα αποτελέσματα που παράγει το Google shopping για μερικά προϊόντα.[11] Παρατηρούμε πως στις περισσότερες περιπτώσεις συμφωνούμε με τα αποτελέσματα του Google shopping ή τα προσεγγίζουμε αρκετά ικανοποιητικά.

	Χαρακτηριστικό	**Το σύστημά μας**	**Google shopping**
Apple MacBook Pro - Core 2 Duo 2.4 GHz - 13.3”	Battery life		
	Value		
	Size		
	Style		
	Performance		
	Picture		
Apple MacBook Pro Core i5 2.4 GHz 15.4	Customer support		
	Battery life		
	Ease of use		
	Design		
Dell Inspiron 15R Core i3 2.53GHz 15.6	Ease of use		
	Battery life		

[11] Βλ. http://www.google.com/prdhp.

	Performance		
	Design		
Apple MacBook Air - Core 2 Duo 1.86 GHz	Battery		
	Performance		
Canon PowerShot SD1400 IS Digital ELPH 14.1 MP Digital Camera (Pink)	Size		
	Zoom		
	Color		
Sony Cyber-shot DSC-HX5V 10.2 MP Digital Camera (Black)	Video		
	Zoom		
	Size		
	Battery life		

Kodak EASYSHARE M580 14 MP Digital Camera (Blue)	Zoom/lens		
	Value		
Nikon Coolpix 3700 3.2 MP Digital Camera	Zoom/lens		
	Battery life		
Nikon Coolpix S630 12 MP Digital Camera	Zoom		
	Picture		

Για να βρούμε την απόκλιση μας με το σύστημα του Google shopping χρησιμοποιούμε τον παρακάτω τύπο σε ένα σύνολο 30 προϊόντων:

$$απόκλιση = \frac{\sum_1^i |\ \#\ Google\ stars - \ \#\ System\ stars|}{i}$$

Όπου #Google stars είναι η βαθμολογία του Google shopping σε αστεράκια για το χαρακτηριστικό ενός προϊόντος και #System stars είναι η βαθμολογία του συστήματός μας σε αστεράκια για το ίδιο χαρακτηριστικό του προϊόντος.

Ο βαθμός διαφωνίας μας με το Google shopping είναι 0.25, συνεπώς συμφωνούμε αρκετά στα αποτελέσματά μας.

6.Συμπεράσματα και μελλοντικές κατευθύνσεις

Στόχος της εργασίας ήταν η δημιουργία ενός συστήματος συγκριτικής παρουσίασης δύο προϊόντων, που να βασίζεται σε κριτικές χρηστών δημοσιευμένες σε ιστολόγια και ειδικούς ιστοτόπους κριτικών. Τα δύο βασικά στάδια της εργασίας ήταν (α) η εξαγωγή λέξεων-κλειδιών, οι οποίες περιγράφουν τα χαρακτηριστικά των δύο προϊόντων που σχολιάζονται πιο συχνά και (β) ο προσδιορισμός του συναισθήματος (θετική ή αρνητική γνώμη) των χρηστών για τα χαρακτηριστικά αυτά.

Για το πρώτο στάδιο, βασιστήκαμε στο μοντέλο LDA, ακολουθώντας την προσέγγιση των Brody και Elhadad [3], όσον αφορά την εφαρμογή του μοντέλου LDA σε επίπεδο προτάσεων, αλλά με χειρωνακτική επιλογή του επιθυμητού αριθμού των χαρακτηριστικών και πρόσθετα κριτήρια επιλογής των λέξεων-κλειδιών που αντιπροσωπεύουν κάθε χαρακτηριστικό. Καταφέραμε έτσι να εξάγουμε λέξεις-κλειδιά που περιγράφουν τα πιο συχνά χαρακτηριστικά που αναφέρουν οι χρήστες στις κριτικές τους.

Για το δεύτερο στάδιο, βασιστήκαμε στα αποτελέσματα του προηγούμενου βήματος. Θεωρήσαμε ότι οι προτάσεις των κριτικών που περιλαμβάνουν λέξεις-κλειδιά αναφέρουν γνώμες των χρηστών για τα χαρακτηριστικά που ονοματίζονται από τις λέξεις-κλειδιά. Εκπαιδεύσαμε ταξινομητές που επιχειρούν να κατατάξουν κάθε τέτοια πρόταση ως θετική (θετική γνώμη) ή αρνητική (αρνητική γνώμη) και όταν μια πρόταση κατατάσσεται ως θετική (αντίστοιχα αρνητική), θεωρούμε ότι εκφράζει θετική (αντίστοιχα αρνητική) γνώμη για τα χαρακτηριστικά που αναφέρει. Πειραματιστήκαμε με

ταξινομητές βασισμένους σε γλωσσικά μοντέλα, έναν αφελή ταξινομητή Bayes και μια μηχανή διανυσμάτων υποστήριξης, που οδήγησε και στα καλύτερα αποτελέσματα. Παρ' όλο που οι μέθοδοι αυτές απαιτούν παραδείγματα εκπαίδευσης επισημειωμένα με τις σωστές τους κατηγορίες, καταφέραμε να δημιουργήσουμε αυτόματα παραδείγματα εκπαίδευσης, θεωρώντας ως θετικές όλες τις προτάσεις κριτικών που συνοδεύονταν από πολύ υψηλές συνολικές βαθμολογίες (5/5 αστεράκια) και αρνητικές όλες τις προτάσεις κριτικών που συνοδεύονταν από πολύ χαμηλές συνολικές βαθμολογίες (0/5 αστεράκια).

Το σύστημα της παρούσας εργασίας μπορεί να επεκταθεί και να χρησιμοποιηθεί σε μια πληθώρα εφαρμογών. Κατ' αρχάς, μπορούμε να το χρησιμοποιήσουμε και σε άλλα πεδία ενδιαφέροντος, όπως σύγκριση παροχής υπηρεσιών (π.χ. ξενοδοχεία, εστιατόρια) ή εξόρυξη γνώμης για πρόσωπα (π.χ. πολιτικοί, αθλητές). Ακόμη, επεκτείνοντάς το σε περισσότερες από δύο εξεταζόμενες οντότητες, μπορεί να χρησιμοποιηθεί σε περιπτώσεις συζητήσεων ή debates, για να παρουσιάσει συνοπτικά τις απόψεις που εκφράζει κάθε συμμετέχων ή τις απόψεις που εκφράστηκαν ανά θέμα. Τέλος, μπορεί να χρησιμοποιηθεί ως βάση για δημιουργία άλλων εφαρμογών, όπως ένα σύστημα που θα παρουσιάζει την άποψη του κοινού για διάφορα ζητήματα (π.χ. πολιτικούς ή πολιτικά κόμματα) στην πάροδο του χρόνου. Ακόμη, παραλλαγές του συστήματος, μπορεί να περιλαμβάνουν την παρουσίαση των καλύτερων προϊόντων ανά κατηγορία (π.χ. laptop, φωτογραφική μηχανή) βάσει π.χ. των βαθμολογιών τους ανά χαρακτηριστικό.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] David Blei, Andrew Ng, Michael Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research.*, 3(5): 93-1022.

[2] Ivan Titov, Ryan McDonald. (2008b). Modeling online reviews with multi-grain topic models*. Proceedings of the 17th International Conference on the World Wide Web, Νέα Υόρκη, ΗΠΑ, σσ. 111-120.*

[3] Samuel Brody, Noemie Elhadad. (2010). An unsupervised aspect-sentiment model for online reviews. *Proceedings of the 2010 Annual Conference of the North American Chapter*, σσ. 804-812. Λος Άντζελες, Καλιφόρνια.

[4] Minqing Hu, Bing Liu. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Νέα Υόρκη, ΗΠΑ, σσ. 168-177.

[5] Bing Liu, Wynne Hsu, Yiming Ma. (1998). Integrating classification and association rule mining. *Knowledge discovery and data mining*, Νέα Υόρκη, ΗΠΑ, σσ. 80-86.

[6] Ana-Maria Popescu, Oren Etzioni. (2005). Extracting product features and opinions from reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Morristown, ΗΠΑ, σσ.339-346.

[7] Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis, Jeff Reynar. (2008). Building a sentiment summarizer for local service reviews. *WWW Workshop on NLP Challenges in the Information Explosion Era,* Πεκίνο, Κίνα.

[8] Wayne Xin Zhao, Jing Jiang, Hongfei Yan, Xiaoming Li. (2010). Jointly modelling aspects and opinions with a MaxEnt-LDA hybrid. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Stroudsburg, ΗΠΑ, σσ. 56-65.

[9] Soo-Min Kim, Eduard Hovy (2004). Determining the sentiment of opinions. *Proceedings of the 20th international conference on Computational Linguistics*, Stroudsburg, ΗΠΑ.

[10] Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena (2007). Large-scale sentiment analysis for news and blogs. *Proceedings of the International Conference on Weblogs and Social Media, Κολοράντο, ΗΠΑ*.

[11] George A. Miller (1995). WordNet: a lexical database for English. *Communications of the ACM*, Νέα Υόρκη, ΗΠΑ, σσ. 39-41.

[12] Theresa Wilson, Janyce Wiebe,Paul Hoffmann (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)* , Stroudsburg, ΗΠΑ, σσ. 347-354.

[13] Ivan Titov and Ryan McDonald (2008a). A joint model of text and aspect ratings for sentiment summarization. *In Proceedings of the 17th International Conference on World Wide Web*, Οχάιο, ΗΠΑ, σσ. 308–316.

[14] Aris Kosmopoulos, Georgios Paliouras, Ion Androutsopoulos (2008). Adaptive spam filtering using only Naive Bayes text classifiers. *Fifth Conference on Email and AntiSpam*, Mountain View, Καλιφόρνια ΗΠΑ.

[15] Andrea Esuli, Fabrizio Sebastiani (2006). Determining term subjectivity and term orientation for opinion mining. *Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics*, Τρέντο, Ιταλία, σσ. 193-200.

[16] Alina Andreevskaia, Sabine Bergler (2006). Mining WordNet for fuzzy sentiment: Sentiment tag extraction fromWordNet glosses. *Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics*, Τρέντο, Ιταλία, σσ. 209-216.

[17] Tilman Lange, Volker Roth, Mikio L. Braun (2004). Stability-based validation of clustering solutions. *Neural Computation*. 16(6): 1299-1323.

[18] Manning, Christopher D., Schutze, Hinrich (1999). Foundations of statistical natural language processing. Cambridge Massachusetts: MIT Press.

[19] Ellen Riloff, Janyce Wiebe (2003). Learning extraction patterns for subjective expressions. Proceedings of the 2003 conference on Empirical methods in natural language processing, Stroudsburg, ΗΠΑ, σσ. 105-112.

[20] V. Metsis, I. Androutsopoulos , G. Paliouras, "Spam filtering with Naive Bayes -- Which Naive Bayes?". Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS 2006), Mountain View, CA, USA, 2006.

[21] Α. Κοσμόπουλος, «Διήθηση ανεπιθύμητης ηλεκτρονικής αλληλογραφίας με διάφορες μορφές του απλοϊκού ταξινομητή Bayes και διαμοιρασμό φίλτρων μεταξύ χρηστών», μεταπτυχιακή διπλωματική εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2007.

[22] T. Joachims, Learning to classify text using support vector machines: Methods, theory, and algorithms. Kluwer, 2002.