ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αναγνώριση αρθρογράφων ιστολογίων θετικής/αρνητικής επιρροής με εξόρυξη γνώμης

ΚΑΡΟΖΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

Επιβλέπων : Κατσαρός Δημήτριος, Λέκτορας

Μέλη Επιτροπής : Βάβαλης Εμμανουήλ, Καθηγητής

Μποζάνης Παναγιώτης, Αναπληρωτής Καθηγητής

Βόλος, Οκτώβριος 2013

UNIVERSITY OF THESSALY
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

MASTER THESIS

Identifying positive/negative influential bloggers with opinion mining

KAROZOS KONSTANTINOS

Supervisor : Katsaros Dimitrios, Lecturer

Committee Members : Vavalis Emmanouil, Professor

Bozanis Panayiotis, Associate Professor

Volos, October 2013

.....

ΚΑΡΟΖΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

Κάτοχος Μεταπτυχιακού Διπλώματος Ειδίκευσης στην «Επιστήμη και Τεχνολογία Υπολογιστών, Τηλεπικοινωνιών και Δικτύων» από το Πανεπιστήμιο Θεσσαλίας

Περίληψη

Το διαδίκτυο στην σύγχρονη μορφή του είναι μία αστείρευτη πηγή άντλησης πληροφοριών αλλά και απόψεων. Σε αυτό έχουν συμβάλει καθοριστικά τα ιστολόγια (blogs), τα άρθρα των οποίων καταναλώνονται καθημερινά από εκατομμύρια ανθρώπων ανά την υφήλιο.

Σκοπός μας είναι η εύρεση των αρθρογράφων με την μεγαλύτερη επιρροή στην κοινωνία. Προς επίτευξη αυτού, εξετάζουμε κάποιες χρονικά μεταβαλλόμενες μετρικές αξιολόγησης που έχουν προταθεί. Τις επεκτείνουμε και ορίζουμε την νέα βελτιωμένη έκδοση αυτών, στην οποία οι απόψεις που εκφράζονται στα σχόλια των άρθρων αξιοποιούνται στην διαδικασία αξιολόγησης των bloggers.

Εν συνεχεία, σχεδιάζουμε και υλοποιούμε Σύστημα το οποίο υπολογίζει τις μετρικές αυτές για δεδομένα ιστολογίων. Αναλυτικότερα, αποτελείται από Ι) το Υποσύστημα Εξόρυξης Γνώμης, το οποίο παίρνοντας ως είσοδο τα σχόλια των άρθρων, εντοπίζει την θετικότητα/αρνητικότητα που εκφέρουν τα σχόλια και τα ποσοτικοποιεί, ΙΙ) το Υποσύστημα Ταξινόμησης Bloggers, το οποίο παίρνει αυτά τα αποτελέσματα και συνυπολογίζοντάς τα μαζί με ορισμένες άλλες παραμέτρους ταξινομεί.

Εφαρμόζουμε το Σύστημα σε δεδομένα. Υλοποιούμε τις παλιές τεχνικές, τις οποίες επίσης εφαρμόζουμε στα ίδια δεδομένα. Κάνουμε συγκρίσεις και εξάγουμε συμπεράσματα.

Λέξεις Κλειδιά: Εξόρυξη Γνώμης, Επεξεργασία Φυσικής Γλώσσας, Ταξινόμηση Αρθρογράφων Ιστολογίων

Abstract

The web in its modern form is an endless source of information as well as opinions. Blogs have contributed thoroughly to this. Blog posts are being read every day by millions of people around the globe.

Our goal is to find the most influential to the society bloggers. In that direction, we consider some already cited time-aware metrics. We develop them further and define a new improved version where opinions in blog posts' comments are being evaluated during the ranking procedure.

Then, we design and implement a System running these techniques for web blogs' data. It is comprised of: I) the Opinion Mining Subsystem, which takes the posts' comments as input, tracks their positivity/negativity and evaluates them, II) the Ranking Bloggers Subsystem, where these results come in along with other parameters so as to generate the ranking.

We implement the System on our data. We also implement the old metrics on the same data. We make comparisons and draw conclusions.

Keywords: Opinion Mining, Sentiment Analysis, Natural Language Processing (NLP), Blogs Ranking, Bloggers Ranking

Πίνακας περιεχομένων

1 *Εισαγωγή*

1.1 Απόψεις στο διαδίκτυο

Στο διαδίκτυο, εκτός της πληθώρας πληροφοριών η οποία διατίθεται, συναντάμε πλέον και μία πληθώρα απόψεων, η οποία προέρχεται από τους πάντες και αναφέρεται στα πάντα. Ο οποιοσδήποτε πολίτης από κάθε μέρος της γης, έχει την ευχέρεια να τοποθετηθεί επί οποιουδήποτε θέματος εκείνος επιθυμεί. Πρόκειται για μία παγκόσμια πολυφωνία, η οποία στα πρώτα βήματά της βασίστηκε στα forums, αργότερα στα ιστολόγια (blogs) και εντέλει γιγαντώθηκε με την επικράτηση των κοινωνικών δικτύων (social networks, κυρίως twitter, facebook).

Τα blogs δε, εξακολουθούν να τυγχάνουν ευρείας αναγνώρισης και να χρησιμοποιούνται κατά κύριο λόγο από χρήστες οι οποίοι προτιμούν την ανταλλαγή απόψεων στηριζόμενων σε επιχειρήματα. Το εικονικό σύμπαν στο οποίο εμπεριέχονται όλα τα blogs, η blogόσφαιρα (blogosphere) προσφέρεται για καταγραφή προσωπικών εμπειριών, προβληματισμών, για κριτική προϊόντων, gadgets, βιβλίων, ταινιών, για έκφραση γνώμης πάνω στα κοινωνικοπολιτικά δρώμενα ή για οτιδήποτε άλλο φανταστεί κανείς. Εκτιμάται ότι σήμερα, υπάρχουν πάνω από 172 εκατομμύρια blog sites στα οποία δημοσιεύονται περισσότερα από 1 εκατομμύριο δημοσιεύσεις (posts) σε καθημερινή βάση [3]. Συνυπολογίζοντας το ότι οι απλοί χρήστες που απλώς διαβάζουν τα blogs είναι πολλαπλάσιοι των αρθρογράφων (bloggers), γίνονται αντιληπτά το χαοτικό μέγεθος της αντικειμενικής ή/και υποκειμενικής πληροφορίας που διατίθεται, καθώς επίσης και η τεράστια επιρροή που ασκείται μέσω αυτής.

Η πληροφορία για να αξιοποιηθεί πρέπει πρώτα να ανακτηθεί από το χάος. Για την ανάγκη αυτή επιστρατεύεται η επιστήμη της Ανάκτησης Πληροφορίας (Information Retrieval IR). Όσο αφορά την επιρροή, για να μπορέσουμε να την κάνουμε "εκμεταλλεύσιμη" (με την καλή έννοια), πρέπει να εντοπιστούν τα πιο δημοφιλή κέντρα άσκησης επιρροής (trend setters ή influentials [8]) που έχουν το μεγαλύτερο εκτόπισμα στην κοινή γνώμη και απολαμβάνουν την εμπιστοσύνη των περισσοτέρων. Σε ένα βαθμό υπερβολής, θέλοντας να υπογραμμίσουν ότι όντως υπάρχουν άτομα που επηρεάζουν σημαντικά τους υπολοίπους, οι Keller και Berry αναφέρουν χαρακτηριστικά ότι "ένας στους δέκα Αμερικανούς λέει στους άλλους εννέα πώς να ψηφίσουν, πού να φάνε και τί να αγοράσουν" [8], αναφερόμενοι στις influential προσωπικότητες, εντός αλλά και εκτός διαδικτύου. Κρίνεται σκόπιμο, λοιπόν, να κάνουμε κάποιου είδους αξιολόγηση και κατάταξη (ranking) των bloggers, λαμβάνοντας υπόψη όσα στοιχεία δύναται να μεταφράζονται (άμεσα και έμμεσα) σε δημοφιλία.

Η άντληση πληροφορίας από blogs, η αξιολόγηση των bloggers, ο συνδυασμός τους και πολλά άλλα συναφή, βρίσκονται στις πρώτες θέσεις προτίμησης ως αντικείμενο έρευνας των τελευταίων ετών στην ακαδημαϊκή/ερευνητική κοινότητα. Αν και η επιστήμη του IR έχει κάνει σημαντικά βήματα προόδου και έχει επιφέρει λύσεις που εφαρμόζονται επιτυχώς στην πράξη, με χαρακτηριστικά παραδείγματα τους PageRank [22], HITS [23], πολλές από αυτές τις μεθόδους αποδεικνύονται αναποτελεσματικές στην blogόσφαιρα λόγω των ιδιαιτεροτήτων της. Το χαρακτηριστικότερο αυτών είναι το περιορισμένο πλήθος συνδέσμων (links) μεταξύ των blog posts. Στις κλασικές ιστοσελίδες συνηθίζεται να υπάρχουν διάφοροι υπερσύνδεσμοι που οδηγούν σε άλλα σημεία του web, είτε για λόγους διαφήμισης, για επεξηγήσεις, παραπομπές, downloads είτε για πολλά άλλα. Το γεγονός αυτό εκμεταλλεύεται ο PageRank και σε αυτό βασίζει ολόκληρη την λογική του. Απεναντίας, η δομή των blogs είναι "φτωχή" σε links. Ο blogger το μόνο που κάνει είναι να παραθέτει τις απόψεις του μέσω μιας -συνήθως- λιτής διεπαφής (interface). Σπάνια παραπέμπει τον αναγνώστη αλλού. Ακόμα και στην περίπτωση που τον παραπέμψει, υπάρχει διαφορά εάν πρόκειται για link μέσα ή έξω από την blogόσφαιρα.

1.2 Αντικείμενο διπλωματικής

Είναι φανερό, επομένως, ότι η αναζήτηση και η κατάταξη των blogs αποτελούν ανοιχτό πρόβλημα βαρύνουσας σημασίας. Στην παρούσα διπλωματική προτείνουμε μεθόδους αξιολόγησης των bloggers. Βασιζόμαστε στα [1] και [2], όπου ο Λεωνίδας Ακριτίδης παρουσιάζει κάποιους τελεστές για ranking των bloggers βάση του πλήθους και της ηλικίας των posts, των σχολίων (comments) και των εισερχόμενων συνδέσμων (inlinks).

Σχεδιάσαμε και υλοποιήσαμε ένα Σύστημα Εξόρυξης Γνώμης (Opinion Mining), το οποίο εφαρμόζουμε στα comments. Επεκτείναμε αυτούς τους τελεστές κατάταξης αξιοποιώντας την γνώμη που εκφέρεται στα comments, η οποία εξάγεται από το Opinion Mining σύστημα.

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Μελετήσαμε τους [1], [2] τελεστές και τους επεκτείναμε.
2. Μελετήσαμε τις αρχές του Opinion Mining.
3. Σχεδιάσαμε και υλοποιήσαμε Opinion Mining Σύστημα.
4. Σχεδιάσαμε και υλοποιήσαμε τους νέους αλλά και τους βασικούς τελεστές.
5. Εφαρμόσαμε τους τελεστές σε βάση δεδομένων και κάναμε συγκρίσεις.

1.3 Οργάνωση κειμένου

Η οργάνωση της διπλωματικής έχει ως εξής:

Στο **δεύτερο κεφάλαιο** αναλύονται τα [1], [2]. Επιπλέον, παρουσιάζονται οι βασικές αρχές του Opinion Mining.

Στο **τρίτο κεφάλαιο** περιγράφονται κάποιες δημοσιεύσεις που συσχετίζονται με την προτεινόμενη προσέγγιση.

Στο **τέταρτο κεφάλαιο** γίνεται παρουσίαση των νέων τελεστών και επεξηγείται η λογική τους.

Το **πέμπτο κεφάλαιο** αναφέρεται στην ανάλυση και σχεδίαση του συστήματος που υλοποιεί τα περιγραφέντα στο τέταρτο.

Το **έκτο κεφάλαιο** έχει να κάνει με την υλοποίηση του συστήματος.

Στο **έβδομο κεφάλαιο** παρουσιάζονται τα αποτελέσματα των διαφορετικών μετρικών και γίνεται αποτίμηση αυτών.

Στο **όγδοο κεφάλαιο** γράφεται ο Επίλογος.

Στο **ένατο κεφάλαιο** δίνεται η βιβλιογραφία και γενικότερα οι πηγές από τις οποίες αντλήθηκαν οι απαραίτητες πληροφορίες για τη συγγραφή της διπλωματικής.

2 Υπόβαθρο

Παρουσιάζονται τα βασικά στοιχεία αξιολόγησης των bloggers (2.1) και εξόρυξης γνώμης (2.2) στα οποία θα βασιστούμε στα επόμενα κεφάλαια.

2.1 Ταξινόμηση των Bloggers

Οι συγγραφείς του [6] ασχολούνται με το περιγραφέν πρόβλημα της ταξινόμησης των bloggers. Το μοντέλο που προτείνουν βασίζεται σε τέσσερεις παραμέτρους: i) Την αναγνώριση που απολαμβάνει ο εκάστοτε blogger (ανάλογη των εισερχόμενων links), ii) την κινητικότητα, δραστηριότητα που προκαλούν τα posts του (ανάλογη του πλήθους των comments), iii) την καινοτομία, πρωτοτυπία των ιδεών του (αντιστρόφως ανάλογη των εξερχομένων links -outlinks-), iv) την ευγλωττία του (αντιστρόφως ανάλογη του μήκους των posts).

Με βάση αυτά τα στοιχεία υπολογίζεται μία βαθμολογία επιρροής I(p) για το κάθε ένα post p. Η συνάρτηση είναι η εξής:

$$I(p) = w(\lambda) \cdot (w_{com} \cdot \gamma_p + w_{in} \cdot \sum_{m=1}^{|I|} I_p(m) - w_{out} \cdot \sum_{n=1}^{|O|} I_p(n)) \qquad (2.1.1)$$

, όπου το w(λ) προκύπτει συναρτήσει του μήκους λ του post, γ_p είναι το πλήθος των comments, w_{com}, w_{in}, w_{out} τα βάρη που καθορίζουν το πόσο θα ληφθούν υπόψη στην διαμόρφωση του I(p) τα comments, τα inlinks και τα outlinks αντίστοιχα. Το I(p) υπολογίζεται με αναδρομή, με παρόμοια λογική με αυτή του PageRank.

Η επιρροή του ίδιου του blogger, σύμφωνα με το μοντέλο, συμπίπτει με την μεγαλύτερη τιμή I(p) όλων των posts του. Αυτός ο τρόπος αξιολόγησης αποτελεί μια υπεραπλούστευση και δεν είναι δύσκολο να αντιληφθεί κανείς ότι δεν είναι ο πιο

δίκαιος. Ο λόγος είναι ότι στην ουσία ο blogger χαρακτηρίζεται από ένα και μόνο post και όχι από το σύνολο των δημοσιεύσεών του. Ενδεχομένως κάποιος να έχει αναρτήσει ελάχιστα influential άρθρα, ενώ οι υπόλοιπες δημοσιεύσεις του στην πλειοψηφία τους να είναι αδιάφορες. Ο συγκεκριμένος, θα θεωρηθεί δημοφιλέστερος έναντι ενός άλλου με πολύ μεγαλύτερο πλήθος ενδιαφερόντων άρθρων, κανένα όμως εκ των οποίων δεν βαθμολογήθηκε παραπάνω από το κορυφαίο του πρώτου.

Ένα ακόμα μειονέκτημα αυτής της προσέγγισης είναι το γεγονός ότι το αποτέλεσμα έχει μεγάλη εξάρτηση από τα βάρη w_{com}, w_{in}, w_{out}. Μικρές αλλαγές στις τιμές τους επιφέρουν μεγάλες ανακατατάξεις, κάτι το οποίο δεν είναι επιθυμητό. Με τόσο μεγάλες αποκλίσεις που παρουσιάζονται, ο τελεστής εντέλει δεν είναι αντικειμενικός, δεν έχει σταθερή κρίση.

Συν τις άλλοις, δεν έχει ληφθεί μέριμνα για έναν καθοριστικό παράγοντα ο οποίος δεν είναι άλλος από τον χρόνο. Με την πάροδο του χρόνου, οι ισορροπίες αλλάζουν, ορισμένοι trend setters χάνουν την πάλε ποτέ αίγλη τους ή παύουν να είναι δραστήριοι και δίνουν την σκυτάλη σε άλλους των οποίων η αξία αναγνωρίζεται όλο και περισσότερο.

Λαμβάνοντας υπόψη τα παραπάνω, οι συγγραφείς του [1] προτείνουν νέες μετρικές κατάταξης των bloggers βασισμένες στον αντίκτυπο και τις αντιδράσεις που προκαλούν τα blog posts τους. Ένα post i του j blogger λαμβάνει το εξής score:

$$S_j^m(i) = c_1 \cdot (C_i^j + 1) \cdot R_i^j \cdot \left(\frac{\theta}{t - t_{i,p}^j + \theta} \right)^{\delta} \qquad (2.1.2)$$

, όπου c1 μία σταθερά η οποία χρησιμοποιείται απλώς για να μεγαλώσουμε το εύρος τιμών (επιλέγεται c1 = 4), C_i^j το πλήθος των comments του i post, R_i^j το σύνολο των inlinks, $t_{i,p}^j$ η χρονική στιγμή που καταχωρήθηκε το post, t το σήμερα. Η διαφορά $t - t_{i,p}^j$ επιστρέφει το χρονικό διάστημα που πέρασε από την δημοσίευση του post, μετρημένο σε δευτερόλεπτα. Οι θ, δ είναι άλλες δύο σταθερές. Συγκεκριμένα η θ ισούται με τα δευτερόλεπτα ενός εικοσιτετραώρου ($\theta = 60 \cdot 60 \cdot 24 \cdot \sec = 86400 \cdot \sec$). Η δ εκφράζει τον βαθμό απαξίωσης των άρθρων με την πάροδο του χρόνου. Προτιμάται $\delta = 1$.

Εν ολίγοις, η αποδιδόμενη στο post βαθμολογία εξαρτάται από τις δύο από τις τέσσερεις παραμέτρους του [6], δηλαδή τα εισερχόμενα links και το πλήθος των

comments, με την βασικότατη, όμως, διαφορά ότι συνδυάζονται με την μορφή γινομένου και όχι αθροίσματος με βάρη. Με την απλή αυτή διαφοροποίηση καταπολεμάται η έλλειψη αντικειμενικότητας. Τρίτος -καινοτόμος- παράγοντας είναι η παλαιότητα της δημοσίευσης. Όσο περνάει ο καιρός, αυτή χάνει την φρεσκάδα της και ο παρονομαστής του $\left(\frac{\theta}{t-t_{i,p}^{j}+\theta}\right)^{\delta}$ μεγαλώνει, με αποτέλεσμα να μικραίνει το $S_j^m(i)$. Τα εξερχόμενα links και το μήκος των posts συνειδητά δεν επιλέχθηκαν να συμμετέχουν στην διαμόρφωση της βαθμολογίας καθότι κρίνεται ότι δεν έχουν σχέση με το πόσο influential είναι ένα post. Η λογική ότι με τα outlinks "χάνεται" μέρος της αξίας προέρχεται από τον PageRank και τον HITS και αρμόζει περισσότερο στο web εκτός blogόσφαιρας. Ο ισχυρισμός ότι σε ένα blog, μεγάλο πλήθος outlinks μεταφράζεται σε έλλειψη καινοτομίας είναι μάλλον παρερμηνεία, λανθασμένη μετάφραση που έγινε κατά την "μεταβίβαση" από τα web sites στα blogs. Υπενθυμίζεται ότι τα ιστολόγια παρουσιάζουν ορισμένες ιδιαιτερότητες που καθιστούν ανέφικτη την πιστή αντιγραφή μεθοδολογιών του web. Τέλος, δεν υιοθετείται η σκέψη ότι το μήκος των posts έχει σχέση με το influence.

Η μετρική αξιολόγησης που προτείνεται, επομένως, καταφέρνει να βρει λύση στα προβλήματα της αντικειμενικότητας και του χρόνου. Εάν θεωρηθεί ότι ο κάθε blogger χαρακτηρίζεται από το μέγιστο $S_j^m(i)$, θα εξακολουθεί να υπάρχει το θέμα της μη αξιοποίησης του συνόλου των άρθρων του. Για αυτό το λόγο επιστρατεύεται η λογική του h-index [7], το οποίο χρησιμοποιείται ευρέως για την αξιολόγηση των επιστημόνων με βάση τις δημοσιεύσεις τους. Εν προκειμένω, για την βαθμολόγηση της επιρροής του j blogger, ορίζεται ο MEIBI (Metric for Evaluating and Identifying a Blogger's Influence) τελεστής:

Ένας blogger j έχει δείκτη MEIBI ίσο με m, εάν m posts του έχουν βαθμολογία $S_j^m(i) \geq m$, *ενώ τα υπόλοιπα έχουν μικρότερη.*

Κατ' αυτόν τον τρόπο, ο MEIBI επιβραβεύει τους αρθρογράφους με επιρροή οι οποίοι έχουν παράξει αξιόλογου μεγέθους έργο το οποίο συνεχίζουν ακόμα και σήμερα. Εάν κατά το τελευταίο χρονικό διάστημα έχουν παραμελήσει το καθήκον τους να δημοσιεύουν κείμενα για το κοινό τους, η MEIBI αξία τους θα σημειώσει πτώση. Στον αντίλογο αυτής της προσέγγισης, η επιρροή ενός blogger δεν μετριέται

από το πόσο πρόσφατα είναι τα posts του αλλά από το εάν αυτά έχουν σήμερα απήχηση. Τα posts βαθμολογούνται σύμφωνα με αυτό τον συλλογισμό ως εξής:

$$S_j^x(i) = c_1 \cdot (C_i^j + 1) \cdot \sum_{\forall x \in R_i^j} \left(\frac{\theta}{t - t_{x,l}^j + \theta} \right)^{\delta} \qquad (2.1.3)$$

Οι διαφορές με το $S_j^m(i)$ είναι ότι αντί του πλήθους των inlinks στο i post, στον τύπο υπολογίζεται η "ηλικιακή" τους επιρροή (δηλαδή το αντίστοιχο κλάσμα με πριν, με την διαφορά ότι εξετάζουμε το πόσο πρόσφατα δημοσιεύτηκε το εκάστοτε inlink και όχι το post), ενώ απουσιάζει η ηλικία του post.

Χρησιμοποιώντας το $S_j^x(i)$ αντί του $S_j^m(i)$, ορίζεται με τον ίδιο ακριβώς τρόπο ο νέος τελεστής MEIBIX (eXtended). Η διαφορά μεταξύ των MEIBI&MEIBIX έγκειται στο ότι ο μεν MEBI αναγνωρίζει ως influential αυτούς που δημοσιεύουν πρόσφατα, ο δε MEIBIX θεωρεί σημαντικότερους εκείνους που κερδίζουν inlinks πρόσφατα (θα δούμε αργότερα στις μετρήσεις ότι συνήθως πρόκειται για τους ίδιους). Οι συγγραφείς του [1], θέλοντας να διαχωρίσουν την παραγωγικότητα από την επιρροή, στο [2] εμπλουτίζουν την δυάδα των MEIBIX&MEIBI με τον δείκτη παραγωγικότητας και τον δείκτη επιρροής. Και οι δύο ακολουθούν εξίσου πιστά την h-index λογική.

Για τον BP (Blogger's Productivity), τα άρθρα βαθμολογούνται ως εξής:

$$U_{i,p}^j(t) = \gamma \cdot \frac{L_i^j}{\overline{L}} \cdot \left(\frac{\theta}{\mathrm{t} - \mathrm{t}_{\mathrm{i,p}}^{\mathrm{j}} + \theta} \right)^{\delta} \qquad (2.1.4)$$

, όπου γ μία σταθερά αντίστοιχη της c1 στην (2.1.2) (θέτουμε γ=100), L_i^j το μήκος του i post του j blogger σε λέξεις, $\overline{L}$ το μέσο μήκος ενός post. Είναι φανερό ότι σε αυτή την μετρική το ενδιαφέρον εστιάζεται αποκλειστικά στην ποσότητα που δημοσιεύεται.

Στην τελευταία μετρική που παρουσιάζεται, η οποία είναι αυτή του BI (Blogger's Influence), τα άρθρα παίρνουν τιμές:

$$V_{i,p}^j(t) = w_l \cdot \sum_{\forall x \in R_i^j} \left(\frac{\theta}{\mathrm{t} - \mathrm{t}_{\mathrm{x,l}}^{\mathrm{j}} + \theta} \right)^{\delta} + w_c \cdot \sum_{\forall x \in C_i^j} \left(\frac{\theta}{t - t_{x,c}^j + \theta} \right)^{\delta} \qquad (2.1.5)$$

, όπου οι σταθερές w_l, w_c αναλαμβάνουν διπλό ρόλο: φροντίζουν ώστε ο $V_{i,p}^j(t)$ να λαμβάνει τα επιθυμητά επίπεδα τιμών (αντίστοιχα με c1 (2.1.2) και γ (2.1.4)) και

παράλληλα δίνουν βάρος στα $\sum_{\forall x \in R_i^j} \left(\frac{\theta}{\mathrm{t} - \mathrm{t}_{\mathrm{x,1}}^{\mathrm{j}} + \theta} \right)^{\delta}$ και $\sum_{\forall x \in C_i^j} \left(\frac{\theta}{t - t_{x,c}^{j} + \theta} \right)^{\delta}$. Ανατίθενται τιμές $w_l = 100$, $w_c = 10$. Αυτό σημαίνει πως ένας εισερχόμενος σύνδεσμος θεωρείται δέκα φορές πιο σημαντικός από ένα σχόλιο.

Από τις κατατάξεις που παρουσιάζονται στα [1], [2] προκύπτει ότι οι περιγραφέντες τελεστές συμπλέουν, ιδιαίτερα ο MEIBI με τον MEIBIX.

2.2 Εξόρυξη Γνώμης

2.2.1 Γενικά

Οι απόψεις αποτελούν πρωταρχικό στοιχείο στην πλειονότητα των ανθρώπινων δραστηριοτήτων και επηρεάζουν από λίγο έως πολύ όλους μας. Τα πιστεύω και οι αξίες μας και η συνολική αντίληψη του καθενός μας για τον κόσμο έχουν καθοριστεί, έως ένα βαθμό, από τις γνώμες του κοινωνικού τους περίγυρου.

Η Εξόρυξη Γνώμης (Opinion Mining), ή αλλιώς Ανάλυση Συναισθήματος (Sentiment Analysis) είναι ο τομέας που ασχολείται με την ανάλυση των απόψεων, συναισθημάτων, εκτιμήσεων και συμπεριφορών ως προς οντότητες όπως είναι προϊόντα, υπηρεσίες, δραστηριότητες, απόψεις άλλων αλλά και κοινωνικά, πολιτικά, κοινωνικά ή άλλης φύσης θέματα [9]. Υπάγεται στην γενικότερη επιστήμη της Γλωσσολογίας και της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing -NLP-). Την τελευταία δεκαετία, λόγω της υπερπληθώρας ψηφιοποιημένων εγγράφων με άποψη (opinionated), στον συγκεκριμένο τομέα του Opinion Mining έχει σημειωθεί ραγδαία αύξηση ενδιαφέροντος όσο αφορά την ερευνητική δραστηριότητα.

Δοθέντος ενός εγγράφου κειμένου, ο σκοπός είναι ο προσδιορισμός-κατηγοριοποίησή του (classification), βάση της γνώμης που εκφέρεται και του συναισθήματος που αποπνέει, σε θετικά/αρνητικά προσανατολισμένο. Η απλή κατηγοριοποίηση είναι το πρώτο βήμα. Δεύτερο βήμα είναι η βαθμολόγηση της έντασης του συναισθήματος, η οποία είθισται να γίνεται με τιμές [-1,0) για τα αρνητικά και (0,1] για τα θετικά έγγραφα.

Προκειμένου η ανάλυση συναισθήματος και η εξόρυξη γνώμης να γίνονται με αυτόματο τρόπο, εφαρμόζονται μέθοδοι μηχανικής μάθησης (machine learning), κατά τις οποίες το σύστημα "μαθαίνει" να ταξινομεί τα έγγραφα ανάλογα με το ύφος του περιεχομένου τους. Συγκεκριμένα, οι κανόνες, το κριτήριο σύμφωνα με το οποίο γίνεται ο διαχωρισμός των κειμένων προκύπτει κατόπιν επεξεργασίας "εκπαιδευτικών δεδομένων" (training data) που είναι επιλεγμένα για να εξυπηρετούν αυτόν τον σκοπό.

Στην μηχανική μάθηση υπάρχουν δύο γενικές κατηγορίες μεθόδων[1][10]:

- **Οι επιβλεπόμενες (supervised)** μέθοδοι, στις οποίες τα training data τα οποία χρησιμοποιούνται έχουν "ετικέτες" που χαρακτηρίζουν κάθε έγγραφο ως θετικό ή αρνητικό. Οι μέθοδοι λέγονται επιβλεπόμενες, διότι κάποιος άνθρωπος επιβλέπει την διαδικασία και βάζει τις ετικέτες καθορίζοντας τί θεωρείται θετικό, τί αρνητικό μεταξύ των εκπαιδευτικών δεδομένων.

  Κατά την "εκπαίδευση" του συστήματος, τα σημαντικότερα στοιχεία στα οποία πρέπει να δίνεται βάση είναι τα ακόλουθα:

  - ✓ *Οι όροι (terms) και η συχνότητά τους (term frequency tf)* - Σημαντικό ρόλο παίζουν οι όροι που συσχετίζονται με το θέμα, δηλαδή το αντικείμενο συζήτησης, όπως π.χ. τα πολιτικά, οι επιστήμες ή τα αθλητικά. Στο Sentiment Analysis, όμως, τον καθοριστικότερο ρόλο διαδραματίζουν οι λέξεις εκείνες που υποδεικνύουν υποκειμενικότητα, όπως π.χ. "good", "amazing", "bad", "awful". Για την εύρεση της συχνότητας εμφάνισής τους δανειζόμαστε εργαλεία από την IR, όπως το tf-idf weighting.
  - ✓ *Τα μέρη του λόγου (Part Of Speech POS)* - Λέξεις διαφορετικών κατηγοριών μερών του λόγου τυγχάνουν άλλης αντιμετώπισης. Έχει διαπιστωθεί ότι τα μέρη του λόγου που έχουν την μεγαλύτερη υποκειμενικότητα, δηλαδή εκφέρουν γνώμη (sentiment words), είναι τα επίθετα, τα επιρρήματα και σε μικρότερο βαθμό τα ρήματα [4][11][12].
  - ✓ *Αναστροφείς συναισθήματος (Sentiment shifters)* - Ορισμένες λέξεις/φράσεις αναστρέφουν την πολικότητα των υπόλοιπων όρων, από θετικά σε αρνητικά και αντίθετα. Είναι οι επονομαζόμενες

[1] Οι επιβλεπόμενες-μη επιβλεπόμενες μέθοδοι προέρχονται από την ευρύτερη θεωρία της αναγνώρισης προτύπων.

"Negation words", με την πιο χαρακτηριστική από αυτές το επίρρημα "not" το οποίο δηλώνει ξεκάθαρη άρνηση, που μεταφράζεται σε αναστροφή πολικότητας στους κοντινούς σε αυτό όρους.

- **Οι μη επιβλεπόμενες (unsupervised)** μέθοδοι, στις οποίες δεν υπάρχει συμβολή ανθρώπου κατά την κατηγοριοποίηση των εγγράφων κειμένου. Βασίζονται κυρίως στις sentiment words και σε καταγεγραμμένους συνδυασμούς μερών του λόγου στους οποίους έχει αποδειχθεί ότι είναι πιθανό να εκφέρεται άποψη.

  Μη επιβλεπόμενες είναι οι προσεγγίσεις που χρησιμοποιούν λεξικά (lexicon-based) τα οποία έχουν καταχωρημένες και βαθμολογημένες εκτιμήσεις συναισθημάτων για όλες τις λέξεις. Χάρις στις έρευνες και μελέτες πολλών ετών, έχουν συνταχθεί αρκετά λεξικά τέτοιου είδους και διατίθενται δωρεάν για ερευνητικούς κυρίως σκοπούς:

  - ✓ Sentiment lexicon (Hu and Liu, 2004): (http://www.cs.uic.edu/~liub/FBS/ sentiment-analysis.html)
  - ✓ MPQA subjectivity lexicon (Wilson, Wiebe and Hoffmann, 2005): (http://www.cs.pitt.edu/mpqa/subj _lexicon .html)
  - ✓ SentiWordNet (Esuli and Sebastiani, 2006): (http://sentiwordnet.isti.cnr.it/)
  - ✓ Emotion lexicon (Mohammad and Turney, 2010): (http://www.purl.org/net/emolex)

2.2.2 SentiWordNet Lexicon

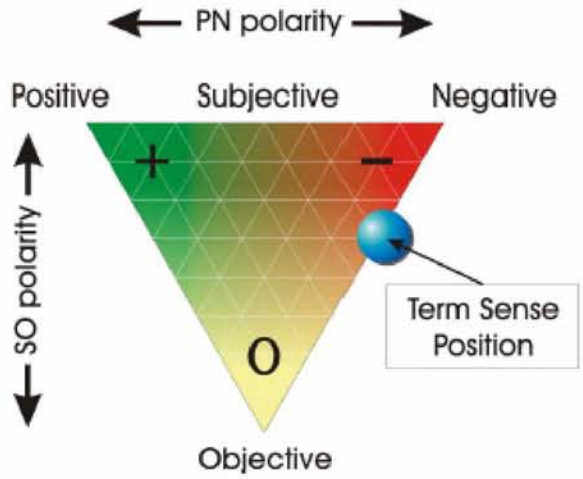
Εστιάζουμε την μελέτη μας στο SentiWordNet[4]. Το SentiWordNet έχει στηριχθεί στο λεξικό WordNet[15], όπου οι λέξεις είναι οργανωμένες σε ομάδες συνωνύμων, τα λεγόμενα synsets (sets of synonyms). Κάθε ένας όρος (term) ανήκει σε ένα ή περισσότερα synsets, ανάλογα με το αν έχει μία μοναδική, ξεκάθαρη έννοια ή ενδεχομένως μπορεί να χρησιμοποιηθεί με παραπάνω από έναν τρόπο. Για παράδειγμα ο όρος "beat" ανήκει σε 34 synsets: στα 10 ως ουσιαστικό, στα 23 ως ρήμα και στο 1 ως επίθετο (με την έννοια του κουρασμένου).

Το πρόβλημα εξόρυξης γνώμης - ανάλυσης συναισθήματος σε ένα κείμενο έχει διττή φύση: Πρέπει να εντοπιστεί το πόσο θετικό/αρνητικό είναι (πολικότητα), αλλά και το πόσο υποκειμενικό είναι, δηλαδή κατά πόσο το περιεχόμενό του είναι η άποψη του συγγραφέα ή διατυπώνονται αντικειμενικές αλήθειες (π.χ."Η γη γυρίζει"). Το SentiWordNet δίνει την λύση προτείνοντας μία τριάδα βαθμολογιών. Κάθε synset s

χαρακτηρίζεται από τρία scores: το Obj(s), το Pos(s) και το Neg(s), τα οποία προσδιορίζουν την αντικειμενικότητα (Objectivity), την θετικότητα (Positivity) και την αρνητικότητα (Negativity) των λέξεων του synset. Κάθε ένα score παίρνει τιμές στο [0,1] και το άθροισμα όλων είναι πάντοτε 1.

$$Obj(s) + Pos(s) + Neg(s) = 1 \qquad (2.2.2.1)$$

Κατά αυτόν τον τρόπο, δίνεται η δυνατότητα στα synsets να εκφράζουν ταυτόχρονα στοιχεία θετικότητας, αρνητικότητας αλλά και αντικειμενικότητας με τρόπο ελεγχόμενο. Η τριπλέτα τιμών ενός synset μπορεί να αναπαρασταθεί και γραφικά.



Εικόνα 2.2.2.1: Γραφική αναπαράσταση (Obj, Pos, Neg) ενός synset

Το SentiWordNet είναι ένα εξαιρετικά χρήσιμο εργαλείο για Opinion Mining εφαρμογές, το οποίο εάν χρησιμοποιηθεί σωστά μπορεί να αποφέρει πολύ υψηλή ακρίβεια αποτελεσμάτων. Έχουν προταθεί διάφοροι τρόποι αξιοποίησής του.

Στο [11] προτείνεται η αξιολόγηση της θετικότητας/αρνητικότητας κειμένων με βασικό γνώμονα το πλήθος των θετικών, αρνητικών όρων κατά το SentiWordNet (term counting). Επιπλέον, λαμβάνεται υπόψη το μέρος του εγγράφου, καθώς θεωρείται ότι σε τμήματα όπως η αρχή και το τέλος ενός κειμένου, συνηθίζεται να εκφέρονται ξεκάθαρα απόψεις σε αντίθεση με το ενδιάμεσο όπου παρατίθενται επιχειρήματα. Το μοντέλο, τέλος, ενισχύεται με έναν αλγόριθμο εντοπισμού Negation, τον NegEx[16].

Σε αντίθεση με το [11], στα [13], [17] υποστηρίζεται ότι είναι προτιμότερο πρώτα να γίνεται εκτίμηση της πολικότητας στο επίπεδο των προτάσεων και ύστερα σε όλο το κείμενο (και όχι από τις λέξεις να εξάγονται συμπεράσματα απευθείας για το κείμενο στο σύνολό του). Σύμφωνα, λοιπόν, με αυτή την λογική (η οποία ακολουθήθηκε και στην σχεδίαση του δικού μας Συστήματος), υπάρχουν τρία επίπεδα προσδιορισμού

πόλωσης του συναισθήματος: i) το αρχικό επίπεδο της λέξης (word-level polarity), ii) αυτό της πρότασης (sentence-level polarity) και iii) του κειμένου (post-level polarity).

3 Σχετικές Εργασίες

Σε αυτό το κεφάλαιο παρουσιάζονται μερικές δημοσιεύσεις οι οποίες έχουν να κάνουν με Αξιολόγηση Bloggers ή με τεχνικές Εξόρυξης Γνώμης από κείμενα ή συνδυάζουν και τα δύο. Πρόκειται για εργασίες με υψηλό ερευνητικό ενδιαφέρον στις οποίες προτείνονται και εφαρμόζονται μερικές από τις πιο καινοτόμες ιδέες και μεθόδους. Πολλά στοιχεία από αυτές επηρέασαν, βοήθησαν στην σύλληψη και τον σχεδιασμό των δικών μας προτάσεων.

3.1 Opinion Mining from Blogs [5]

Σε αυτή την δημοσίευση οι συγγραφείς παρουσιάζουν το αναπτυχθέν σύστημα AMOD (Automatic Mining of Opinion Dictionaries). Πρόκειται για ένα σύστημα που κάνει κατηγοριοποίηση εγγράφων βασιζόμενο σε εξειδικευμένα (ανάλογα με το θέμα -subject-) σύνολα θετικών/αρνητικών λέξεων. Η μέθοδος που ακολουθείται είναι μη επιβλεπόμενη. Αναλυτικότερα, χωρίζεται σε τρεις φάσεις:

- **Φάση 1: Άντληση εγγράφων για εκπαίδευση**

  Δεδομένων ενός θέματος και δύο σύντομων λιστών P, Q οι οποίες περιέχουν μερικές θετικές, αρνητικές λέξεις (επίθετα), αναζητούνται έγγραφα κατάλληλα για εκπαίδευση. Η αναζήτηση γίνεται στο διαδίκτυο με την χρήση κάποιας μηχανής αναζήτησης. Η όλη διαδικασία γίνεται με στοχευμένο τρόπο, έτσι ώστε να ανακτηθούν δύο ξεχωριστά σύνολα εγγράφων: Στα έγγραφα του ενός εξ' αυτών συναντώνται λέξεις από την P αλλά όχι από την Q, ενώ στα έγγραφα του

δεύτερου το αντίστροφο. Με αυτόν τον τρόπο δημιουργείται ένα σύνολο από θετικής διάθεσης έγγραφα και ένα από αρνητικής.

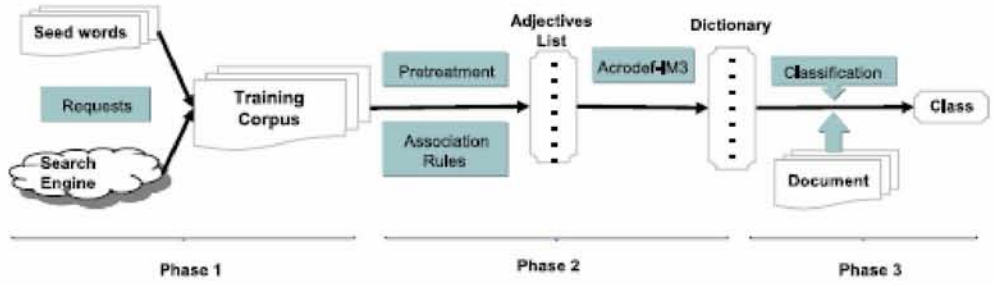
- **Φάση 2: Εξαγωγή λίστας με θετικά/αρνητικά επίθετα**

  Από τα δύο αυτά σύνολα, σκοπός είναι η επέκταση, ο εμπλουτισμός των αρχικών σύντομων, βασικών λιστών P και Q με περισσότερα επίθετα. Για να γίνει αυτό, μελετώνται οι συσχετίσεις των βασικών επιθέτων με άλλα επίθετα στο σύνολο των ανακτηθέντων εγγράφων και εξάγονται συσχετισμοί και κανόνες σχέσεων. Όσα εκ των νέων επιθέτων έχουν "ικανά" δυνατούς συσχετισμούς με τα βασικά επίθετα, γίνονται δεκτά. Τα υπόλοιπα απορρίπτονται.

  Με αυτόν τον τρόπο, οι λίστες P, Q έχουν, πλέον, μέσα τους έναν σεβαστό αριθμό θετικών, αρνητικών λέξεων σχετικών με το επιλεχθέν θέμα.

- **Φάση 3: Κατηγοριοποίηση εγγράφων**

  Η κατηγοριοποίηση των εγγράφων γίνεται με την απλή μέθοδο της καταμέτρησης των θετικών, αρνητικών λέξεων και της εύρεσης της διαφοράς.



Εικόνα 3.1.1: Σύστημα AMOD

Πέραν της περιγραφής της διαδικασίας, οι συγγραφείς παραθέτουν μετρήσεις, οι οποίες αναδεικνύουν την καταλληλότητα του AMOD.

3.2 Reviews classification using SentiWordNet [12]

Στο [12] παρουσιάζεται μία πιο απλή προσέγγιση για εξόρυξη γνώμης, στην οποία αξιοποιείται το SentiWordNet λεξικό, ενώ δεν γίνεται καθόλου εκπαίδευση δεδομένων. Η αξιολόγηση των κειμένων στηρίζεται στις βαθμολογίες των λέξεων με υποκειμενικότητα (sentiment words) υπό την κρίση του SentiWordNet.

Η όλη διαδικασία χωρίζεται σε δύο φάσεις:

- **Φάση 1: Υπολογισμός τελικής βαθμολογίας λέξεων**

Όπως έχει ειπωθεί, κάθε λέξη-όρος στο SentiWordNet ενδεχομένως να ανήκει σε παραπάνω από ένα synset, επομένως χαρακτηρίζεται από πολλές τριάδες (Obj, Pos, Neg). Στην παρούσα εργασία, για να εξαχθεί το τελικό score της εκάστοτε λέξης, εντοπίζεται η κατηγορία του Μέρους του Λόγου (Part-Of-speech-P.O.S.) στην οποία ανήκει και υπολογίζεται ο Μέσος Όρος των βαθμολογιών των synsets στα οποία ανήκει ο όρος και είναι του ίδιου P.O.S.

- **Φάση 2: Κατηγοριοποίηση εγγράφων**

  Σκοπός των συγγραφέων είναι να εφαρμόσουν κάτι αποτελεσματικότερο από το υπεραπλουστευμένο term counting. Δοκιμάζουν δύο μεθόδους. Κατά την πρώτη, υπολογίζονται ξεχωριστά τα αθροίσματα όλων των θετικών και όλων των αρνητικών όρων κάθε εγγράφου, ενώ στην δεύτερη προσέγγιση, το σύστημα υπολογίζει τους αντίστοιχους Μ.Ο. Κατά τους υπολογισμούς αυτούς, τίθενται ορισμένα thresholds ώστε να μην λαμβάνονται υπόψη οι λέξεις με πολύ μεγάλη αντικειμενικότητα. Η πολικότητα του εγγράφου καθορίζεται από το/τον μεγαλύτερο άθροισμα/Μ.Ο.

Κατόπιν μετρήσεων και συγκρίσεων, προκύπτει ότι η δεύτερη μέθοδος είναι αποτελεσματικότερη της πρώτης, ενώ και οι δύο υπερέχουν του term counting. Επιπλέον, οι συγγραφείς οδηγούνται στην διαπίστωση ότι τα υψηλά thresholds υποκειμενικότητας κατά την επιλογή των λέξεων (δηλαδή το να αγνοούνται οι όροι με χαμηλό Objectivity) έχουν δυσμενείς επιπτώσεις στην ακρίβεια του Συστήματος.

3.3 Identifying Bloggers with Marketing Influence [20]

Οι συγγραφείς του [20] επιχειρούν να βρουν λύση στο πρόβλημα της εύρεσης των bloggers με την μεγαλύτερη επιρροή, δίνοντας έμφαση σε ό,τι έχει να κάνει με προώθηση προϊόντων και μάρκετινγκ.

Η ιδιαιτερότητα του μοντέλου έγκειται στο ότι διαχωρίζουν την συνολική "αξία" του blogger (ως προς το επίπεδο επιρροής του) σε δύο κατηγορίες, οι οποίες είναι:

- **Η "δικτυακή" αξία (network-based value -NV-)**

  Η NV αξιολογεί όλα τα στοιχεία που έχουν σχέση με την δικτύωση του blogger, δηλαδή με την σχέση του ιστολογίου του με το υπόλοιπο web και τους χρήστες του. Συγκεκριμένα, αυτά τα στοιχεία είναι οι εισερχόμενοι, εξερχόμενοι

σύνδεσμοι (in-links, out-links), το πλήθος των σχολίων *(c)* στα posts του, οι παραπομπές -citations- *(η),* οι οποίες στην ουσία είναι σαν in-links, ο αριθμός των επισκεπτών *(α)* του ιστολογίου και η φήμη του blogger *(γ)* η οποία προκύπτει από το συνολικό έργο του. Η NV του blogger b ορίζεται από την σχέση

$$NV(b) = a_b \cdot PR(b) + \gamma_b \cdot f(c_b, \eta_b) \quad (3.3.1)$$

, όπου *PR(b)* η τιμή του PageRank του blog.

- **Η αξία ως προς το περιεχόμενο των άρθρων (content-based value -CV-)**
  Η CV εξετάζει το περιεχόμενο των blog posts αυτό κάθε αυτό. Τα μεγέθη που το χαρακτηρίζουν είναι η υποκειμενικότητα *(s)* (η οποία καθορίζεται από το σύνολο των υποκειμενικών θετικών και αρνητικών λέξεων που περιέχονται σε όλα τα posts ενός blogger), το μέγεθος *(λ)* των posts, η ηλικία *(τ)* του ιστολογίου. Η CV του blogger b:

  $$CV(b) = \tau_b \cdot (s_b + \lambda_b) \quad (3.3.2)$$

Ο υπολογισμός των NV, CV γίνεται με την βοήθεια νευρωνικών δικτύων. Η συνολική αξία του blogger προκύπτει από το άθροισμα των NV, CV με βάρη.

3.4 Trackback-Rank [19]

Εξετάζεται το ίδιο θέμα, δηλαδή το πρόβλημα της αξιολόγησης της αξίας των bloggers, στο ευρύτερο, όμως, πλαίσιο της αναζήτησης πληροφορίας στην blogόσφαιρα. Αυτό που ενδιαφέρει, δηλαδή, τους συγγραφείς είναι η εύρεση των πιο αξιόλογων blog posts σχετικά με ένα συγκεκριμένο θέμα, αντικείμενο, το οποίο καθορίζεται από ένα tag *t*.

Επομένως, κάθε blog post (entry, *e*) λαμβάνει διαφορετικό score αναλόγως του tag *t* που μας ενδιαφέρει. Τα χαρακτηριστικά τα οποία παίζουν ρόλο στην διαμόρφωση του score *ES(e, t)* (Entry Score) είναι τα εξής:

- **Η αξία, φήμη του blogger b στο θέμα *t***
  Η φήμη του blogger *b* σε ένα θέμα *t* εκφράζεται με το μέγεθος *BS(t, b)* (Blogger Score), το οποίο είναι το κανονικοποιημένο άθροισμα των βαθμολογιών:
  - Blogger Entry Score (BES): Πρόκειται για τον Μ.Ο. των ES του παρελθόντος (Past Entry Score, PES) ως προς το tag *t:*

$$BES(t,b) = \frac{\sum_{e_i \in E} PES(e_i,t,b)}{\left|E_t^b\right|} \quad (3.4.1)$$

, όπου e το εκάστοτε blog entry, E_t^b το σύνολο των entries του blogger b που είναι σχετικά με το tag t και

$$PES(e_i,t,b) = 1 + \sum_{j=1}^{K} TBS_j(t,e_i,b_j) + (a_t^b + \beta) \cdot CS(e_i) \quad (3.4.2)$$

, όπου K το σύνολο των trackback links[2] στο entry e_i , b_j ο blogger που ανάρτησε το j-trackback link στο e_i entry, $TBS_j(t,e_i,b_j)$ το score του i trackback link (αναλύεται παρακάτω στην 3.4.4), a_t^b ο λόγος του συνόλου των trackback links προς αυτό των comments, β μία σταθερά που φροντίζει να μην μηδενιστεί ο παράγοντας σε περίπτωση ανυπαρξίας trackbacks, $CS(e_i)$ το σύνολο των comments στο e_i blog entry.

– Blogger Activity Score (BAS): Το μέγεθος αυτό εκφράζει την δραστηριότητα, παραγωγικότητα του blogger στο θέμα t. Συγκεκριμένα:

$$BAS(t,b) = N_{en}^t(b) \cdot \frac{N_{en}^{tr}(b)}{N_{en}(b)} \quad (3.4.3)$$

, όπου $N_{en}^t(b)$ το σύνολο των blog entries του blogger b που είναι σχετικά με το tag t, $N_{en}^{tr}(b)$ το σύνολο των entries του blogger b που έχουν trackback links, $N_{en}(b)$ το σύνολο όλων των entries του blogger b.

- **Τα trackback links**

  Εισάγεται το μέγεθος $TBS_i(t, e_i, b_j)$ το οποίο αναδεικνύει την αναγνωρισιμότητα που έχει ο b_j blogger ως προς το t ζήτημα, υπό μορφή trackback links:

$$TBS_i(t,e_i,b_j) = \begin{cases} 0 & \text{, εάν δεν υπάρχουν καθόλου trackback links} \\ 1 + BS(t, b_j) & \text{, εάν υπάρχουν.} \end{cases} \quad (3.4.4)$$

  , όπου $BS(t, b_j)$ το Blogger Score του blogger bj που ανάρτησε το j-trackback link.

- **Τα comments**

  Το πλήθος των comments $CS(e)$ του entry e.

[2] Στα περισσότερα blogspots υπάρχει η δυνατότητα όταν σε ένα blog post αναρτάται link προς ένα δεύτερο blog post, να στέλνεται αυτόματα ειδοποίηση στο δεύτερο και στο τέλος αυτού να προστίθεται ένα (trackback) link προς το πρώτο. Στην ουσία, σε ό,τι αφορά το ranking των bloggers, τα trackback links είναι τα in-links.

Όλα τα παραπάνω συνδυάζονται για να προκύψει εντέλει η γενική αξιολόγηση του άρθρου e ως προς το tag t:

$$ES(e,t) = \begin{cases} 0 & \text{, εάν το } e \text{ δεν έχει το tag } t \\ BS(t,b) + \sum_{i=1}^{K} TBS_i(t,e,b_i) + (a_t^b + \beta) \cdot CS(e) & \text{, εάν το έχει.} \end{cases} \quad (3.4.5)$$

Προς απόδειξη της αποτελεσματικότητας της αξιολόγησης κατά την περιγραφείσα μέθοδο, οι συγγραφείς παραθέτουν ενδεικτικές μετρήσεις. Το TrackBack-Rank είναι ένα αρκετά επιτυχημένο μοντέλο καθώς καταφέρνει έμπρακτα να εφαρμόζει παρεμφερή λογική με τα PageRank, HITS συνδυάζοντάς την με τις ιδιαιτερότητες των blogs.

4 *Ταξινόμηση Bloggers με Συναξιολόγηση Σχολίων*

4.1 Επιρροή στα blogs

Απώτερος σκοπός της διπλωματικής είναι να βελτιώσουμε, επεκτείνουμε τους MEIBI&MEIBIX διατηρώντας τον χαρακτήρα και την λογική τους. Αν και θεωρούμε ότι υπερτερούν έναντι των προαναφερθέντων μεθόδων που θέτουν παραπλήσιους στόχους, κρίνουμε ότι υπάρχει περιθώριο βελτίωσης, εξέλιξης.

Το περιθώριο αυτό εντοπίζεται στο γεγονός ότι οι MEIBI&MEIBIX, προκειμένου να αξιολογήσουν την επιρροή ενός i blog post, λαμβάνουν υπόψη, μεταξύ άλλων, το πλήθος C_i^j των σχολίων σε αυτό, τα οποία έχουν δημοσιευτεί από αναγνώστες. Πρόκειται για ένα μέγεθος του οποίου η τιμή όσο μεγαλύτερη είναι, τόσο περισσότερη επιρροή θεωρείται ότι ασκείται από το i post και κατ' επέκταση από τον j blogger. Είναι αλήθεια ότι η πληθώρα αντιδράσεων (σχολίων) που προκαλεί ένα κείμενο, θα μπορούσε να μεταφραστεί σε υψηλό βαθμό επιρροής, ενώ η αδιαφορία του κοινού ως προς ένα άλλο συνεπάγεται, συνήθως, τον μη επηρεασμό των περισσοτέρων (η έλλειψη σχολιασμού μερικές φορές οφείλεται απλώς στο ότι τα γραφόμενα δεν προσφέρονται για κάτι τέτοιο, για παράδειγμα η διήγηση ενός περιστατικού που συνέβη στον blogger μπορεί να μην αφήνει πολλά περιθώρια για comments όμως υποσυνείδητα να επηρεάζει τους αναγνώστες). Το τελικό ζητούμενο, όμως, της αξιολόγησης και της βαθμολόγησης, είναι να εντοπιστεί τι είδους είναι η επιρροή που ασκεί ο blogger μέσα από τα άρθρα του.

Σε αυτό το σημείο, κρίνεται σκόπιμο να διευκρινίσουμε τί ακριβώς εννοούμε με τον όρο "επιρροή" και ποιούς θεωρούμε influentials. Επιρροή ασκείται σε ένα άτομο όταν κάτι επιδρά στον πνευματικό, ψυχικό του κόσμο με αποτέλεσμα να παρατηρούνται αλλαγές στις εκδηλώσεις ή στις αποφάσεις του. Στην περίπτωση των χρηστών που διαβάζουν blogs, επηρεασμένοι θεωρούνται όσοι εξ' αυτών διαφοροποιούν την -εντός και εκτός διαδικτύου- δραστηριότητά τους χάρις στα κελεύσματα ενός ή παραπάνω blogger. Από την στιγμή που η εκτός διαδικτύου δραστηριότητα δεν παρακολουθείται, εστιάζουμε στην εντός, δηλαδή στα σχόλια και στους εισερχόμενους συνδέσμους. Υπάρχουν και άλλα στοιχεία που μαρτυρούν σημάδια επιρροής όπως η επισκεψιμότητα, τα "likes" και τα "shares" (στην ορολογία του facebook) και άλλα πολλά. Αρκούμαστε, όμως, προς το παρόν, σε αυτά τα δύο τα οποία είναι σχεδόν πάντοτε διαθέσιμα. Το πλήθος των comments και των inlinks μαρτυρά την ύπαρξη (ή μη) επιρροής, χωρίς όμως να γίνεται σαφές εάν οι χρήστες συμφωνούν ή διαφωνούν, εάν ταυτίζονται με τις απόψεις του blogger ή αντιτίθενται σε αυτές, εάν διατίθενται να ακολουθήσουν τις προτροπές του ή κάνουν ακριβώς τα αντίθετα. Από το πλήθος των comments και των inlinks προκύπτει το μέτρο της επιρροής. Από τον τρόπο με τον οποίο αυτά χρησιμοποιούνται καθορίζεται το πρόσημο αυτής. Έχουμε θετική επιρροή σε περιπτώσεις ταύτισης με τα γραφόμενα και αρνητική όταν υπάρχει διαφωνία. Influentials, επομένως, θεωρούνται οι bloggers στα άρθρα των οποίων γράφεται ικανοποιητικός αριθμός comments, τα οποία έχουν θετικό χαρακτήρα.

Δυστυχώς για εμάς, η συμφωνία/διαφωνία των χρηστών δεν διαφαίνεται άμεσα. Ο πιο απλός και βολικός τρόπος να ξεχωρίζουν οι ευρέως αποδεχθείσες απόψεις από τις υπόλοιπες είναι το thumbs up-thumbs down (vote up-vote down) σύστημα, το οποίο στις περιπτώσεις που εφαρμόζεται διευκολύνει σε πολύ μεγάλο βαθμό τον σκοπό μας. Για όλες τις άλλες περιπτώσεις, χρειάζεται να γίνει κάποιου είδους επεξεργασία του κειμένου από την οποία να προκύπτει κάποια θετικότητα/αρνητικότητα (ή πιθανόν ουδετερότητα). Σε αυτό το σημείο υπεισέρχεται η επιστήμη της Επεξεργασίας Φυσικής Γλώσσας (NLP - Natural Language Processing) και συγκεκριμένα το Opinion Mining and Sentiment Analysis. Σκοπός μας είναι ο σχεδιασμός ενός Opinion Mining συστήματος το οποίο θα επιστρέφει την πολικότητα των comments ενός blog ώστε να προσδιορίζεται το πρόσημο της επιρροής που αυτό ασκεί. Όσο

αφορά τον τρόπο με τον οποίο χρησιμοποιούνται τα inlinks, δεν θα ασχοληθούμε με αυτόν, αλλά θα αρκεστούμε μόνο στο πλήθος τους.

Στον παρόν κεφάλαιο δεν εξετάζουμε τον τρόπο ανάλυσης-σχεδίασης του Συστήματος Εξόρυξης Γνώμης. Αυτό το κάνουμε στο κεφάλαιο 5. Στο παρόν, αναλύουμε τον τρόπο και την φιλοσοφία σύμφωνα με την οποία επεκτείναμε τους MEIBI&MEIBIX ώστε να λαμβάνεται υπόψη το μέτρο (πλήθος comments) και το πρόσημο (προϊόν Opinion Mining) του δείκτη επιρροής comments.

4.2 Υπολογισμός Συνολικής Γνώμης κάθε blog post

Από το Opinion Mining Σύστημα αναμένουμε να επιστραφεί το $score \in [-1,1]$ για κάθε σχόλιο. Οι βαθμολογίες αυτές των επιμέρους comments δεν είναι άμεσα αξιοποιήσιμες από τα $S_j^m(i)$, $S_j^x(i)$ των MEIBI, MEIBIX. Προκειμένου να καθιστούν χρήσιμες, είναι απαραίτητο να συμψηφιστούν στο επίπεδο των post.

Για να βρεθεί η επικρατούσα άποψη των comments ενός post, μια απλοϊκή προσέγγιση θα ήταν να βρούμε τον Μέσο Όρο των scores τους. Κατ' αυτόν τον τρόπο, όμως, ορισμένοι οι οποίοι έχουν γράψει δύο ή τρία ή πέντε ή και παραπάνω σχόλια θα καταφέρνουν να επηρεάζουν την συνολική opinion περισσότερο απ' όσο τους αναλογεί. Θεωρούμε ότι όλοι όσοι σχολιάζουν έχουν την ίδια βαρύτητα, ανεξάρτητα από το πόσες φορές επιλέξουν να στηρίξουν την άποψή τους.

Ως εκ τούτου, προτιμάμε να συγχωνεύσουμε τις γνώμες των διαφορετικών, ξεχωριστών σχολιαστών (**distinct commentators**) από το να συγχωνεύσουμε "ακέφαλα", απρόσωπα comments. Ο "θόρυβος" **(Buzz)** που προκαλείται από ένα δημοσίευμα μετράται πιο δίκαια με το πλήθος των σχολιαστών παρά με των σχολίων.

Το πιο σύνηθες φαινόμενο είναι ένας σχολιαστής να αναρτά μόνο ένα σχόλιο. Σε αυτή την περίπτωση η άποψή του περί του blog αναδεικνύεται από το score του σχολίου του. Συναντάμε, όμως, ορισμένες περιπτώσεις στις οποίες είτε λόγω κάποιου ερεθίσματος από τυχόν λεγόμενα αλλουνού, είτε επειδή ξέχασε να αναπτύξει την σκέψη του ολοκληρωμένα, επιστρέφει για να σχολιάσει για δεύτερη, για τρίτη, για πέμπτη φορά. Εάν πρόκειται για απαντήσεις σε σχόλια άλλων σχολιαστών, απλώς αγνοούνται διότι ενδιαφερόμαστε μόνο για ό,τι αναφέρεται στο blog post αυτό κάθε

αυτό. Εάν όμως κάποιος εξακολουθεί να σχολιάζει τα γραφόμενα του blogger, το σύστημα θα είναι υποχρεωμένο να τον λάβει υπόψη.

Στην συντριπτική πλειοψηφία τέτοιου είδους καταστάσεων με αλλεπάλληλα comments από το ίδιο άτομο, η βασική άποψη εκφέρεται στο πρώτο σχόλιο και ίσως στο δεύτερο. Όσο προχωράμε στα επόμενα, τόσο πιο ανούσια πράγματα γράφονται. Τα γεγονότα, λοιπόν, συνηγορούν στην εύρεση ενός τρόπου συμψηφισμού των comments ενός distinct commentator όπου η βαρύτητα αυτών θα φθίνει (σε αντίθεση με τον απλό μέσο όρο) όσο προστίθενται νέα. Για αυτή την ανάγκη επινοήσαμε μία αναδρομική (recursive) συνάρτηση μέσου όρου με δυναμικά βάρη:

```
double weighted_avg (int count, int num_of_c)
{
        if (count < num_of_c – 1)
                output = 0.8*(C.get(count)) + 0.2*weighted_avg(++count, num_of_c);
        else if (count == num_of_c – 1)
                output = 1*(C.get(count));
        return output;
}
```

Για κάθε blog post και για κάθε commentator αυτού , λαμβάνουν μέρος στον υπολογισμό τα πρώτα δέκα (εάν υπάρχουν τόσα) comments του commentator (οι βαθμολογίες από το sentiment analysis των οποίων αντλούνται από το *array list C*). Το πρώτο χρονικά comment συμβάλλει στη διαμόρφωση του score του commentator κατά 80%, ενώ όλα τα υπόλοιπα μαζί συμβάλλουν κατά 20% με φθίνουσα βαρύτητα. Η διαδικασία λαμβάνει χώρα για τα score_a, score_b του commentator. Εάν για παράδειγμα κάποιος πόσταρε τρία σχόλια με βαθμολογίες 0,6 / 0,7 / -0,3

η σύνοψη αυτών θα είναι: 0,8*0,6+0,2*[0,8*0,7+0,2*(-0,3)] = 0,8*0,6+0,16*0,7+0,04*(-0,3) = 0,58

Παρατηρούμε ότι το αρνητικό -0,3 έχει ανεπαίσθητη επιρροή. Τα βάρη 0,8 και 0,2 μπορούν να αλλάξουν ανάλογα με τις προτιμήσεις μας.

Με αυτή την λογική θα εξάγεται το score του κάθε σχολιαστή ενός άρθρου.

Για να διαμορφωθεί εντέλει η συνολική άποψη **(Opinion)** όλων όσων έχουν σχολιάσει ένα post, υπολογίζεται ο Μέσος Όρος των score τους ($score_i^j$). Η συνολική αυτή Opinion χαρακτηρίζει την θετικότητα/αρνητικότητα του post. Η ισχύς της, η εγκυρότητά της εξαρτώνται από το πλήθος των distinct commentators ($\#d\_commentators_i^j$).

4.3 MEIBI-MEIBIX με Γνώμη

Υπενθυμίζεται ότι οι MEIBI&MEIBIX του j blogger προκύπτουν από τα $S_j^m(i)$ & $S_j^x(i)$ (2.1.2 & 2.1.3) αντίστοιχα, με την λογική του h-index.

Σκοπός μας είναι η αντικατάσταση του πλήθους C_i^j των comments από ένα μέγεθος που θα συνδυάζει την Opinion με το Buzz του i post. Το νέο C_i^j θα είναι το γινόμενο των $score_i^j, \#d\_commentators_i^j$. Καθότι είναι της ίδιας τάξης μεγέθους με το παλιό C_i^j, δεν χαλάει τις ισορροπίες. Εκ φύσεώς του, ευνοεί τα θετικά υψηλά score που συνδυάζονται με μεγάλο πλήθος σχολιαστών. Απεναντίας, παρουσιάζει μικρές τιμές σε περιπτώσεις χαμηλών score ή/και λίγων σχολιαστών.

Υπάρχει ένας προβληματισμός στο κατά πόσο θα πρέπει η αρνητικότητα (ή μη) του $score_i^j$ να αποδίδεται σε ολόκληρο τον $S_j^m(i) ή S_j^x(i)$. Κατά μία άποψη, εφόσον παράγεται βαθμολογία από τα σχόλια η οποία κατηγοριοποιεί τα posts σε θετικά και αρνητικά, αυτή η κατηγοριοποίηση θα πρέπει να διαφαίνεται άμεσα από τους $S_j^m(i)$ & $S_j^x(i)$ και άρα τους MEIBI&MEIBIX, εκχωρώντας τους αρνητικές τιμές. Κατ' αυτόν τον τρόπο, τα χαρακτηρισμένα ως "αρνητικά" blog posts αποκτούν ένα μείον και αποκλείονται άμεσα από το ενδιαφέρον του -έχοντος τυφλή εμπιστοσύνη στο σύστημά μας- αναγνώστη.

Στον αντίλογο, το σύστημά μας δεν θα μπορεί να θεωρηθεί τόσο έμπιστο και ακριβές ώστε να διαχωρίζει αποτελεσματικά και με ασφάλεια ένα "ελαφρώς" αρνητικό κείμενο από ένα "ελαφρώς" θετικό ή ένα ουδέτερο (άνευ υποκειμενικότητας). Γι' αυτό και δεν παίρνουμε την βαριά ευθύνη της ρητής κατηγοριοποίησης όλων των λεπτεπίλεπτων, ιδιαίτερων τέτοιων περιπτώσεων στις πλάτες μας. Για την ακρίβεια, δεν έχει επινοηθεί ως σήμερα κανένα NLP (Natural Language Processing) σύστημα που να μπορεί να αναλάβει με σιγουριά τέτοια ευθύνη.

Επιπλέον, η τιμή του τελεστή δεν διαμορφώνεται μόνο από τα comments. Στον MEIBI παίζουν ακόμη ρόλο το πλήθος των inlinks και η ηλικία του post, ενώ στον MEIBIX η ηλικία και το πλήθος των inlinks. Δεν είναι είναι πρόθεσή μας να παραγκωνιστεί η σημασία αυτών για χάρη της συνολικής άποψης των comments,

αλλά όλα αυτά τα στοιχεία να συνδιαμορφώνουν τους MEIBI&MEIBIX διατηρώντας τον χαρακτήρα και την λογική τους.

Για αυτούς τους λόγους επιλέγεται να μην διαφοροποιείται το πρόσημό τους. Για να επιτευχθεί αυτό, απλά προσθέτουμε στο $score_i^j$ την μονάδα. Κατ' αυτόν τον τρόπο αποφεύγεται ο υποβιβασμός υπό του μηδενός, ενώ παράλληλα ευνοείται το μέτρο των θετικά βαθμολογημένων posts σε αντίθεση με τα αρνητικά. Έχουμε:

$$\begin{aligned} C_i^j &= (score_i^j + 1) \cdot \#d\_commentators_i^j = \\ &= \frac{sum(d\_commentators\_scores)_i^j}{\#d\_commentators_i^j} \cdot \#d\_commentators_i^j + \#d\_commentators_i^j = \\ &= sum(d\_commentators\_scores)_i^j + \#d\_commentators_i^j \end{aligned} \tag{4.3.1}$$

Παρατηρούμε ότι το νέο C_i^j είναι στην ουσία το άθροισμα των $score_i^j$ των σχολιαστών προστιθέμενο στο πλήθος αυτών. Τα νέα $S_j^m(i) \,\&\, S_j^x(i)$ λέγονται $S_j^{mo}(i) \,\&\, S_j^{xo}(i)$ και (εκ πρώτης όψεως) είναι ίδια με τα παλιά:

$$S_j^{mo}(i) = c_1 \cdot (C_i^j + 1) \cdot R_i^j \cdot \left(\frac{\theta}{t - t_{i,p}^j + \theta} \right)^{\delta} \tag{4.3.2}$$

$$S_j^{xo}(i) = c_1 \cdot (C_i^j + 1) \cdot \sum\nolimits_{\forall x \in R_i^j} \left(\frac{\theta}{t - t_{x,l}^j + \theta} \right)^{\delta} \tag{4.3.3}$$

Οι καινούργιοι τελεστές, τους οποίους θα αποκαλούμε **MEIBIO** και **MEIBIXO** (**O**pinionated), προκύπτουν με την ίδια λογική του h-index.

5 Ανάλυση-Σχεδίαση Συστήματος

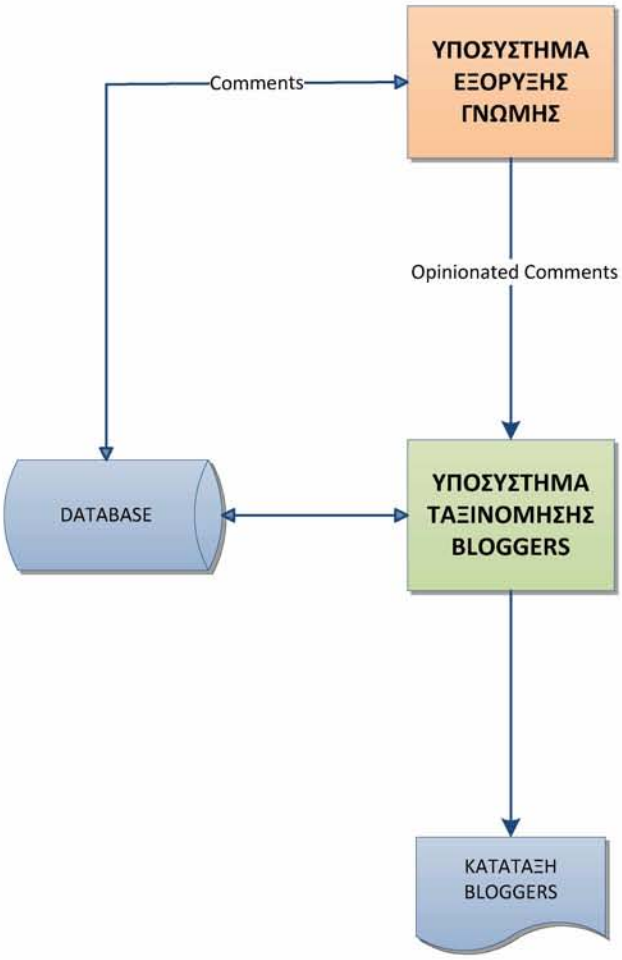
5.1 Γενική Αρχιτεκτονική

Όπως φαίνεται από το Διάγραμμα της Εικόνας 5.1.1, το Σύστημα αποτελείται από δύο Υποσυστήματα.

Στο πρώτο επίπεδο βρίσκεται το **Υποσύστημα Εξόρυξης Γνώμης**, το οποίο επεξεργάζεται τα comments των blog posts και με βάση το SentiWordNet lexicon, αξιολογεί, βαθμολογεί τις λέξεις των comments που εμπεριέχουν υποκειμενικότητα. Στη συνέχεια αποδίδει βαθμό γνώμης στα ίδια τα comments.

Στο δεύτερο επίπεδο βρίσκεται το **Υποσύστημα Ταξινόμησης Bloggers**, το οποίο με βάση τις βαθμολογίες των comments υπολογίζει και αποδίδει βαθμό γνώμης στους συγγραφείς των comments (commentators). Στη συνέχεια υπολογίζει για το κάθε post την συνισταμένη άποψη καθώς και τις νέες μετρικές αξιολόγησης. Τέλος ενημερώνει τις τιμές των ΜΕΙΒΙΟ&ΜΕΙΒΙΧΟ τελεστών κατάταξης των Bloggers με βάση την αξιολόγηση των blog posts του καθενός.

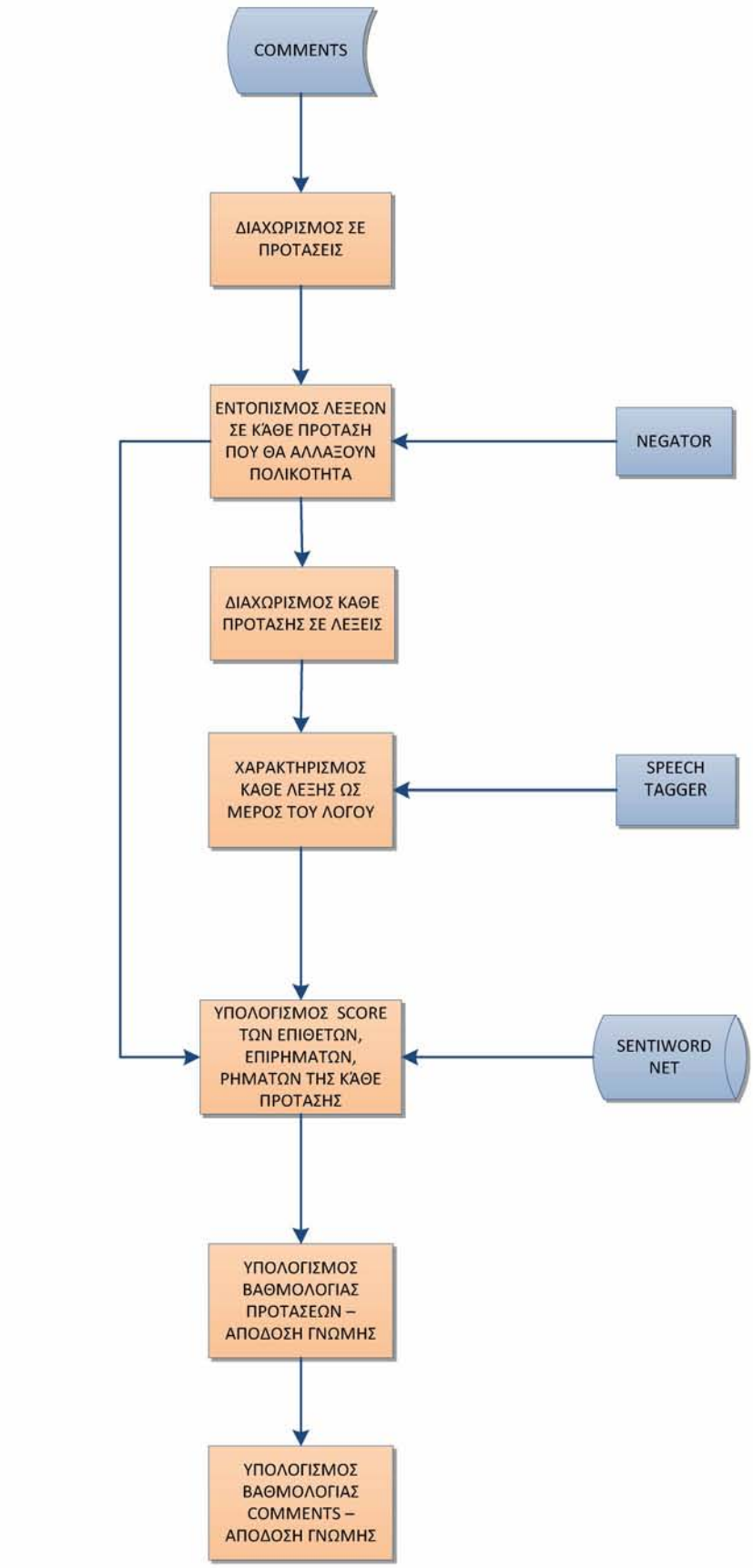
ΓΕΝΙΚΗ ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΣΥΣΤΗΜΑΤΟΣ



Εικόνα 5.1.1: Γενική Αρχιτεκτονική Συστήματος

5.2 Υποσύστημα Εξόρυξης Γνώμης

Το Υποσύστημα Εξόρυξης Γνώμης (Opinion Mining) λαμβάνει σαν input τα comments που αναφέρονται στα διάφορα blog posts ενός data set και επιχειρεί να εκτελέσει μια NLP διαδικασία με στόχο την κατηγοριοποίηση των γνωμών των αναγνωστών σε θετικές/αρνητικές καθώς επίσης και την απόδοση βαθμολογίας στα comments.



Εικόνα 5.2.1: Υποσύστημα Εξόρυξης Γνώμης

Το Υποσύστημα χρησιμοποιεί το SentiWordNet. Το SentiWordNet διατίθεται ελεύθερα ως text αρχείο[3]. Για τον αποδοτικότερο χειρισμό αυτού και για την δική μας διευκόλυνση το SentiWordNet φορτώθηκε (με τη μορφή δύο πινάκων με σχέση ένα-προς-πολλά) στην βάση δεδομένων.

Οι διαδικασίες του Υποσυστήματος φαίνονται συνοπτικά στην Εικόνα 5.2.1.

Το Υποσύστημα διαβάζει από την βάση δεδομένων τα comments που αναφέρονται σε ένα blog post. Καθότι σχεδιάζουμε ένα Opinion Mining Σύστημα, επιθυμούμε τα δεδομένα μας να είναι όσο το δυνατόν πιο προσανατολισμένα στο θέμα, δηλαδή τα comments να αναφέρονται στο blog post αυτό κάθε αυτό. Με αυτή τη λογική αποκλείονται από τις διεργασίες όλα τα comments της βάσης τα οποία ξεκινούν με '@' και άρα απαντάνε στα λεγόμενα άλλων σχολιαστών. Τέτοια στη βάση engadget είναι 192.601 στο σύνολο των 3.672.819 comments.

Για κάθε ένα Comment που διαβάζεται εκτελεί τα ακόλουθα:

- Το comment διαχωρίζεται σε προτάσεις.
- **Για κάθε μια Πρόταση του comment** εκτελούνται τα ακόλουθα:
  - Καλείται το πρόγραμμα Negator GeneralNegEx.2.0[4] το οποίο δέχεται στην είσοδο την πρόταση, εντοπίζει τις αρνήσεις –εάν υπάρχουν- μέσα σε αυτή και επιστρέφει το τμήμα της πρότασης στο οποίο επιδρά η άρνηση (negation scope sentence). Σε μια πρόταση μπορεί να εμφανίζονται περισσότερες από μια αρνήσεις.
  - **Για κάθε Άρνηση που εντοπίζεται στην πρόταση** γίνονται τα εξής:
    - Γίνεται η παραδοχή ότι η άρνηση συνήθως επιδρά στην πρώτη ή στην δεύτερη λέξη της negation scope sentence.
    - Καλείται το πρόγραμμα Speech Tagger POS Tagger 3.2.0 – left3words[5] το οποίο δέχεται στην είσοδο μια λέξη της Αγγλικής γλώσσας και επιστρέφει την λέξη χαρακτηρισμένη ως προς το μέρος του λόγου (Part Of Speech) στο οποίο ανήκει. Εάν διαπιστωθεί ότι η πρώτη λέξη της negation scope sentence είναι επίθετο ή επίρρημα, θα αντιστραφεί η πολικότητα της πρώτης λέξης. Εάν δεν είναι η πρώτη λέξη επίθετο ή επίρρημα, εξετάζεται η δεύτερη λέξη της negation

[3] http://sentiwordnet.isti.cnr.it
[4] http://code.google.com/p/negex/downloads/detail?name=JAVA_GENERALNEGEX_v1.01.zip&can=2&q=
[5] http://nlp.stanford.edu/software/tagger.shtml

scope sentence και εάν αυτή είναι επίθετο ή επίρρημα ή ρήμα, θα αντιστραφεί η πολικότητα της δεύτερης λέξης.

- Εάν η πρώτη ή η δεύτερη λέξη της negation scope sentence αλλάξει πολικότητα, η λέξη επισημαίνεται και η negation scope sentence καταγράφεται σε δομή στη μνήμη, αλλά καταχωρείται και στη βάση δεδομένων προκειμένου να είναι εφικτός ο έλεγχος των αποτελεσμάτων της διαδικασίας αργότερα.

- Η Πρόταση διαχωρίζεται σε Λέξεις.
- **Για κάθε Λέξη της πρότασης** εκτελούνται τα ακόλουθα:
  - ✓ Εξετάζεται αν υπάρχουν καταχωρημένες αρνήσεις για την πρόταση. Αν ναι, ερευνάται αν η εξεταζόμενη λέξη έχει επισημανθεί για να αντιστρέψει την πολικότητά της.
  - ✓ Καλείται ο Speech Tagger για την λέξη για να γνωστοποιηθεί τί μέρος του λόγου είναι.
  - ✓ Εάν η λέξη είναι επίθετο ή επίρρημα ή ρήμα, γίνεται ανάκτηση από το SentiWordNet του Μ.Ο. των θετικών scores και του Μ.Ο. των αρνητικών scores της λέξης από τα synsets του SentiWordNet τα οποία περιέχουν την λέξη και ταυτόχρονα έχουν και την ίδια κατηγορία μέρους του λόγου με την λέξη.
  - ✓ Από τους Μ.Ο. των θετικών και αρνητικών scores της λέξης, προκύπτει το τελικό score της λέξης:

    Εάν ο ένας από τους δύο Μ.Ο. έχει τιμή και ο άλλος έχει μηδέν, σημαίνει ότι υπάρχουν μόνο θετικά ή μόνο αρνητικά scores για την λέξη και επομένως το τελικό score είναι ο μη μηδενικός θετικός ή αρνητικός Μ.Ο.

    Εάν όμως υπάρχουν θετικά και αρνητικά score για την λέξη (και για την ίδια κατηγορία μέρους του λόγου), τότε εφαρμόζεται ένα μικρό threshold (0.05) στην διαφορά της απόλυτης τιμής των δύο Μ.Ο., προκειμένου η λέξη να πάρει ως τελικό score τον θετικό ή τον αρνητικό Μ.Ο. ή να χαρακτηριστεί ουδέτερη (τελικό score=0).
  - ✓ Εάν έχει επισημανθεί από τον Negator αντιστρέφεται το πρόσημο του τελικού score της λέξης.
  - ✓ Η λέξη καταγράφεται σε δομή στη μνήμη, αλλά καταχωρείται και στη βάση με όλα τα χαρακτηριστικά της (comment που ανήκει,

πρόταση που ανήκει, μέρος του λόγου, αντιστροφή πολικότητας, μέσοι όροι αρνητικών και θετικών score, τελικό score, θέση της λέξης στην πρόταση).

- Αφού ολοκληρωθεί η επεξεργασία όλων των λέξεων της πρότασης που εκφράζουν υποκειμενικότητα, ακολουθεί ο υπολογισμός της βαθμολογίας της πρότασης από τα τελικά score των βαθμολογημένων λέξεων.

  Η βαθμολογία της πρότασης γίνεται με δύο προσεγγίσεις:

  Στην πρώτη (a) προσέγγιση υπολογίζεται απλά ο Μ.Ο. των score των λέξεων της πρότασης με μη μηδενική τιμή.

  Στην δεύτερη (b) προσέγγιση υπολογίζεται ο Μ.Ο. των θετικών score των λέξεων (δεν λαμβάνουν μέρος λέξεις με μηδενικά score) και ο Μ.Ο. των αρνητικών score των λέξεων. Εάν ο ένας από τους δυο έχει μηδενική τιμή, το score της πρότασης παίρνει την τιμή του άλλου. Εάν όμως και οι δυο Μ.Ο. έχουν μη μηδενικές τιμές, τότε εφαρμόζεται η ίδια λογική που εφαρμόστηκε στον υπολογισμό των λέξεων: Εφαρμόζεται ένα threshold (0.2) στην διαφορά των απολύτων τιμών των δυο Μ.Ο. και η πρόταση παίρνει το score του "ικανά" μεγαλύτερου θετικού ή αρνητικού Μ.Ο., αλλιώς χαρακτηρίζεται ουδέτερη (score=0).

  Για κάθε βαθμολογήσιμη λέξη, το σύστημα καταχωρεί και τις δύο βαθμολογίες **(score_a, score_b)**. Την διαφορετικότητα των δύο αυτών προσεγγίσεων θα την χρειαστούμε αργότερα στην διεξαγωγή συγκρίσεων.

- Η πρόταση καταγράφεται σε δομή στη μνήμη, αλλά και καταχωρείται στη βάση με όλα τα χαρακτηριστικά της (comment που ανήκει, score α΄τρόπου, score β΄τρόπου, Μ.Ο. θετικών score λέξεων, Μ.Ο. αρνητικών score λέξεων.

- Αφού ολοκληρωθεί η επεξεργασία των προτάσεων του comment, ακολουθεί ο υπολογισμός της βαθμολογίας του ίδιου του comment από τα score των προτάσεων που το αποτελούν:

  Εφόσον στις προτάσεις του comment έχουν αποδοθεί δυο scores και στο comment υποχρεωτικά αποδίδονται δυο score που το καθένα προκύπτει από τους μέσους όρους των score α΄τρόπου και score β΄τρόπου των προτάσεων του comment. Στον υπολογισμό δεν λαμβάνουν μέρος προτάσεις ουδέτερες, δηλαδή με μηδενικό score.

- Ενημερώνεται το comment στη βάση δεδομένων με τις δυο βαθμολογίες που υπολογίστηκαν.

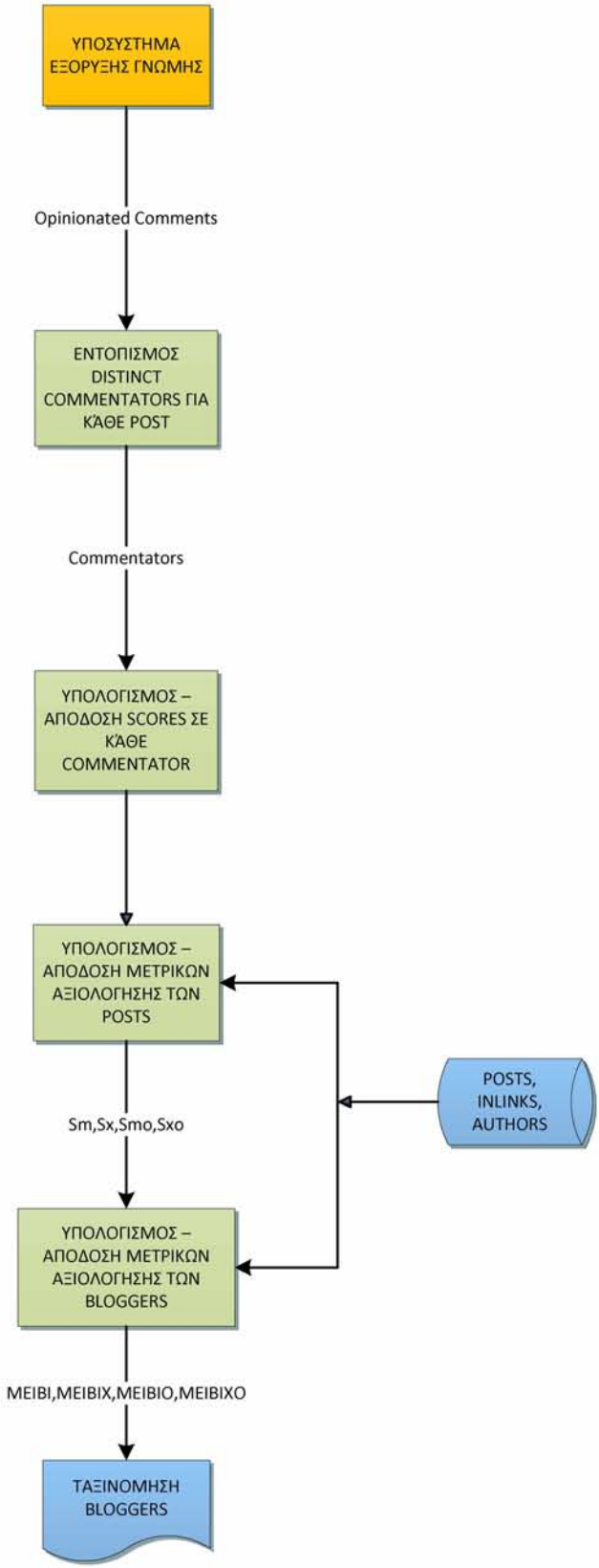
5.3 Υποσύστημα Ταξινόμησης Bloggers

Το Υποσύστημα Ταξινόμησης Bloggers (Bloggers Ranking) λαμβάνει ως είσοδο τα βαθμολογημένα ως προς την θετική ή αρνητική γνώμη που εκφέρουν comments για τα blog posts στα οποία αναφέρονται, μεταφέρει την συνολική γνώμη στα ίδια τα posts και μαζί με άλλους παράγοντες (ηλικία post, inlinks,….) τα αξιολογεί με απώτερο σκοπό την βαθμολογική κατάταξη των bloggers ως προς την επιρροή που ασκούν.

Οι διαδικασίες του Υποσυστήματος, οι οποίες φαίνονται συνοπτικά στην Εικόνα 5.3.1, είναι οι παρακάτω:

- Κάθε ένα blog post ενημερώνεται με τον αριθμό των comments τα οποία έγραψαν καταχωρημένοι (με username) χρήστες και αναφέρονται στο post αυτό κάθε αυτό και δεν απαντούν σε άλλα comments (comments_opinionated_verified). Αξιοσημείωτο είναι ότι με το φιλτράρισμα αυτό μειώνεται κατά πολύ το πλήθος των comments. Στο engadget dataset τα comments_opinionated_verified είναι 1.434.508 στο σύνολο των 3.672.819 comments.
- Στη συνέχεια, ενημερώνεται κάθε ένα post με τον αριθμό των καταχωρημένων χρηστών οι οποίοι έγραψαν τουλάχιστον ένα comment που αναφέρεται απευθείας στο post (distinct commentators).
- Εντοπίζονται για κάθε blog post οι distinct commentators οι οποίοι καταχωρούνται στην βάση -ως ξεχωριστή οντότητα- μαζί με τον αριθμό των comments του καθενός.
- Η διαδικασία για να αποδοθεί ένα score (γνώμη) σε κάθε distinct commentator είναι αυτή που έχει περιγραφεί στο 4.2.

ΥΠΟΣΥΣΤΗΜΑ ΤΑΞΙΝΟΜΗΣΗΣ BLOGGERS



Εικόνα 5.3.1: Υποσύστημα Ταξινόμησης Bloggers

- Μετά την απόδοση score στους commentators με βάση τα score των comments του καθενός, ενημερώνεται κάθε ένα blog post με τον Μ.Ο. των score των commentators (avg-commentators-score-a, avg-commentators-score-b) οι οποίοι έχουν σχολιάσει το συγκεκριμένο blog post.
- **Για κάθε ένα blog post** υπολογίζονται τα παρακάτω μεγέθη τα οποία επιχειρούν να εκτιμήσουν το μέγεθος της επιρροής των posts. Από αυτά θα προκύψουν οι παραλλαγές των MEIBI, MEIBIX.

➢ $$S_j^m(i) = 4 \cdot (\#comments_i^j + 1) \cdot \#inlinks_i^j \cdot \frac{86400}{post's\_age_i \cdot 86400 + 86400}$$

$$S_j^x(i) = 4 \cdot (\#comments_i^j + 1) \cdot \sum_{x\_inlink_i^j} \left(\frac{86400}{x\_inlink's\_age_i^j \cdot 86400 + 86400} \right) \quad (5.3.1)$$

Τα S_j^m & S_j^x είναι η πιστή εφαρμογή αυτών που περιγράφονται στο 2.1.

➢ $$S_j^{m1}(i) = 4 \cdot (\#comments\_opinionated\_verified_i^j + 1) \cdot \#inlinks_i^j \cdot \cdot \frac{86400}{post's\_age_i \cdot 86400 + 86400}$$

$$S_j^{x1}(i) = 4 \cdot (\#comments\_opinionated\_verified_i^j + 1) \cdot \cdot \sum_{x\_inlink_i^j} \left(\frac{86400}{x\_inlink's\_age_i^j \cdot 86400 + 86400} \right) \quad (5.3.2)$$

Τα S_j^{m1} & S_j^{x1} είναι τα S_j^m & S_j^x εφαρμοσμένα σε υποσύνολο του αρχικού dataset το οποίο περιέχει τα σχόλια μόνο όσων έχουν verified ταυτότητα.

➢ $$S_j^{m2}(i) = 4 \cdot (\#d\_commentators_i^j + 1) \cdot \#inlinks_i^j \cdot \cdot \frac{86400}{post's\_age_i \cdot 86400 + 86400}$$

$$S_j^{x2}(i) = 4 \cdot (\#d\_commentators_i^j + 1) \cdot \cdot \sum_{x\_inlink_i^j} \left(\frac{86400}{x\_inlink's\_age_i^j \cdot 86400 + 86400} \right) \quad (5.3.3)$$

Τα S_j^{m2} & S_j^{x2} είναι οι ίδιοι τελεστές με την βασική διαφορά ότι μελετούν τις γνώμες των distinct commentators αντί των ίδιων των comments.

➢ $$S_j^{mo\_aa}(i) = 4 \cdot (C\_a_i^j + 1) \cdot \#inlinks_i^j \cdot \frac{86400}{post's\_age_i \cdot 86400 + 86400}$$

$$S_j^{xo\_aa}(i) = 4 \cdot (C\_a_i^j + 1) \cdot \sum_{x\_inlink_i^j} \left(\frac{86400}{x\_inlink's\_age_i^j \cdot 86400 + 86400} \right)$$

$$C\_a_i^j = (avg(d\_commentator\_score\_a)_i^j + 1) \cdot \#d\_commentators_i^j \quad (5.3.4)$$

Τα $S_j^{mo\_aa}$ & $S_j^{xo\_aa}$ είναι οι opinionated τελεστές από τους οποίους προκύπτουν οι ΜΕΙΒΙΟ&ΜΕΙΒΙΧΟ. Οι βαθμολογήσεις των λέξεων έχουν γίνει με την πρώτη προσέγγιση (score_a), όπως περιγράφεται στο 5.2. Εφαρμόζονται στο ίδιο dataset με τα S_j^{m2} & S_j^{x2}.

- $$S_j^{mo\_bb}(i) = 4 \cdot (C\_b_i^j + 1) \cdot \#inlinks_i^j \cdot \frac{86400}{post's\_age_i \cdot 86400 + 86400}$$

$$S_j^{xo\_bb}(i) = 4 \cdot (C\_b_i^j + 1) \cdot \sum_{x\_inlink_i^j} \left(\frac{86400}{x\_inlink's\_age_i^j \cdot 86400 + 86400} \right)$$

$$C\_b_i^j = (avg(d\_commentator\_score\_b)_i^j + 1) \cdot \#d\_commentators_i^j \quad (5.3.5)$$

Τα $S_j^{mo\_bb}$ & $S_j^{xo\_bb}$ είναι οι opinionated τελεστές που προέκυψαν με την δεύτερη προσέγγιση (score_b). Ομοίως, εφαρμόζονται στα ίδια δεδομένα.

Η ηλικία του i post $post's\_age_i$ και η ηλικία του x inlink $x\_inlink's\_age_i^j$ μετρώνται σε ημέρες. Πολλαπλασιάζοντάς τες επί 86400, τις ανάγουμε σε δευτερόλεπτα.

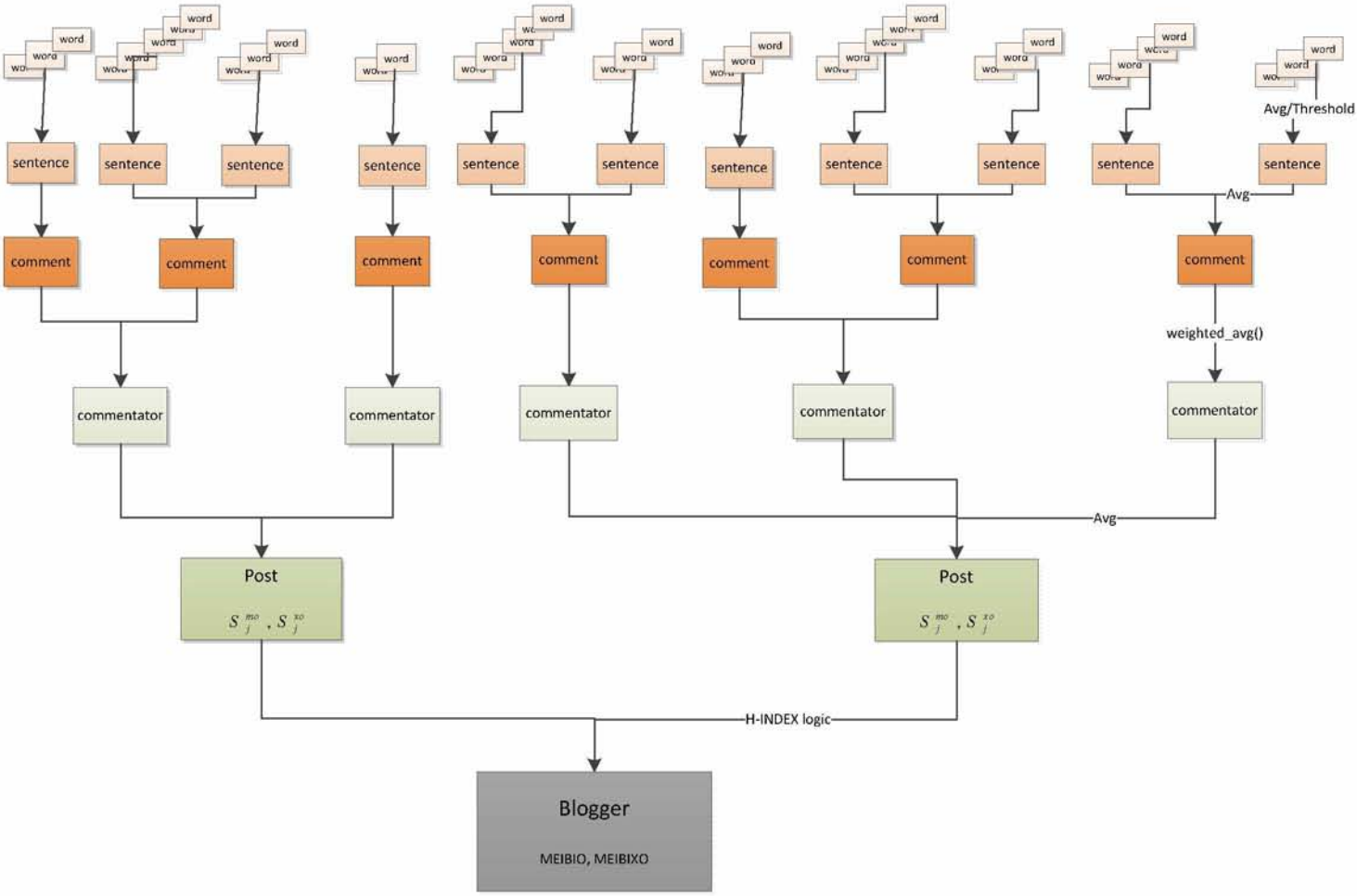
- **Για κάθε blogger** υπολογίζονται οι παρακάτω μετρικές οι οποίες επιχειρούν να συνεκτιμήσουν την παραγωγικότητα και την επιρροή που ασκεί ο καθένας ώστε να είναι δυνατή η ταξινόμησή τους. Η λογική που εφαρμόζεται σε όλες είναι η λογική του **h_index** ο οποίος προσπαθεί να εκτιμήσει την παραγωγικότητα και την επιρροή ενός επιστήμονα με βάση τον αριθμό των δημοσιεύσεών του και τις αναφορές που υπάρχουν σε άλλα δημοσιεύματα για κάθε μία από αυτές.

Ένας επιστήμονας έχει h_index = h, εάν έχει κάνει h δημοσιεύσεις με τουλάχιστον h αναφορές η κάθε μια, ενώ οι υπόλοιπες δημοσιεύσεις του έχουν λιγότερες από h αναφορές η κάθε μια.

- H_INDEX : Επιστήμονας είναι ο blogger, δημοσιεύσεις είναι τα blog post του blogger και αναφορές είναι τα inlinks του post.

- ΜΕΙΒΙ : Ο υπολογισμός γίνεται με βάση τα posts του blogger και τα S^m τους.
- ΜΕΙΒΙΧ : Ο υπολογισμός γίνεται με βάση τα posts του blogger και τα S^x τους.
- ΜΕΙΒΙ1 : Ο υπολογισμός γίνεται με βάση τα posts του blogger και τα S^{m1} τους. Ο υπολογισμός γίνεται μόνο για συγκρίσεις.
- ΜΕΙΒΙΧ1 : Ο υπολογισμός γίνεται με βάση τα posts του blogger και τα S^{x1} τους. Ο υπολογισμός γίνεται μόνο για συγκρίσεις.
- ΜΕΙΒΙ2 : Ο υπολογισμός γίνεται με βάση τα posts του blogger και τα S^{m2} τους. Ο υπολογισμός γίνεται μόνο για συγκρίσεις.
- ΜΕΙΒΙΧ2 : Ο υπολογισμός γίνεται με βάση τα posts του blogger και τα S^{x2} τους. Ο υπολογισμός γίνεται μόνο για συγκρίσεις.
- ΜΕΙΒΙΟ_aa : Ο υπολογισμός γίνεται με βάση τα posts του blogger και τα S^{mo_aa} τους.
- ΜΕΙΒΙΧΟ_aa : Ο υπολογισμός γίνεται με βάση τα posts του blogger και τα S^{xo_aa} τους.
- ΜΕΙΒΙΟ_bb : Ο υπολογισμός γίνεται με βάση τα posts του blogger και τα S^{mo_bb} τους.
- ΜΕΙΒΙΧΟ_bb : Ο υπολογισμός γίνεται με βάση τα posts του blogger και τα S^{xo_bb} τους.

Στην Εικόνα 5.1.2 αναπαρίστανται συνοπτικά όλες οι διαδικασίες του Συστήματος, ξεκινώντας από το επίπεδο της βαθμολόγησης των λέξεων και φτάνοντας σε αυτό της αξιολόγησης των bloggers.



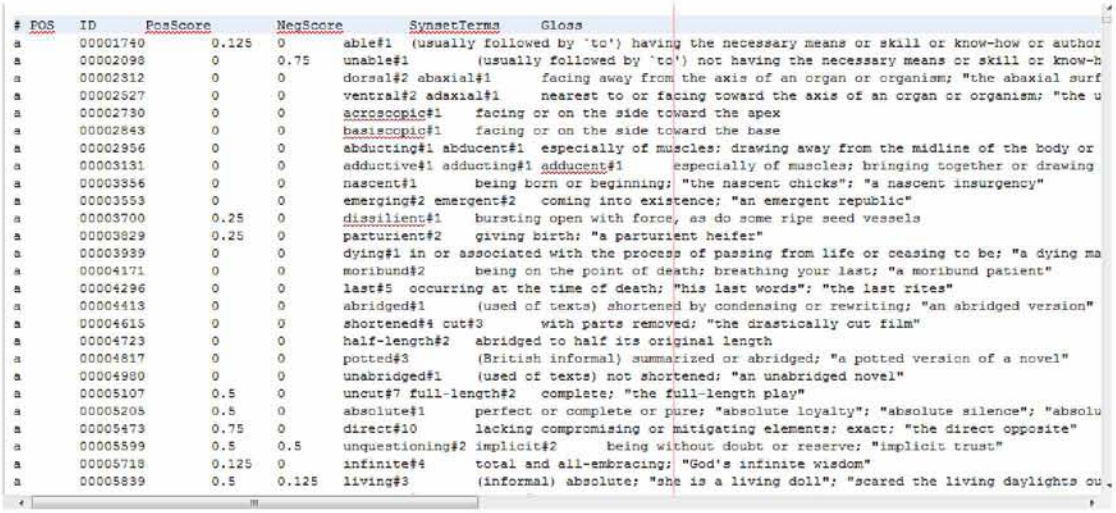
Εικόνα 5.1.2.: Από τις λέξεις στους bloggers

6 Υλοποίηση Συστήματος

6.1 Βάση Δεδομένων

Εφαρμόζουμε το περιγραφέν Σύστημα σε μία mySQL βάση δεδομένων της οποίας οι κύριες οντότητες είναι ο author, το post, το comment και το inlink. Στο παρών περιγράφουμε την δομή της βάσης δεδομένων. Αργότερα, στην ενότητα 7.1 γίνεται αναφορά στα χαρακτηριστικά των δεδομένων αυτών κάθε αυτών που συλλέχθησαν.

Προκειμένου να γίνει πλήρη αξιολόγηση των bloggers με βάση όλα τα αναφερθέντα κριτήρια, η βάση δεδομένων εμπλουτίστηκε με πολλά νέα στοιχεία (νέα attributes σε ήδη υπάρχοντα tables αλλά και νέα tables), το πρώτο εκ των οποίων είναι το SentiWordNet. Το SentiWordNet διατίθεται στο διαδίκτυο ως text αρχείο. Συγκεκριμένα η έκδοση που κατεβάσαμε είναι η SentiWordNet_3.0.0_20130122.txt :



# POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	00001740	0.125	0	able#1	(usually followed by `to') having the necessary means or skill or know-how or author
a	00002098	0	0.75	unable#1	(usually followed by `to') not having the necessary means or skill or know-h
a	00002312	0	0	dorsal#2 abaxial#1	facing away from the axis of an organ or organism; "the abaxial surf
a	00002527	0	0	ventral#2 adaxial#1	nearest to or facing toward the axis of an organ or organism; "the u
a	00002730	0	0	acrascopic#1	facing or on the side toward the apex
a	00002843	0	0	basiscopic#1	facing or on the side toward the base
a	00002956	0	0	abducting#1 abducent#1	especially of muscles; drawing away from the midline of the body or
a	00003131	0	0	adductive#1 adducting#1 adducent#1	especially of muscles; bringing together or drawing
a	00003356	0	0	nascent#1	being born or beginning; "the nascent chicks"; "a nascent insurgency"
a	00003553	0	0	emerging#2 emergent#2	coming into existence; "an emergent republic"
a	00003700	0.25	0	dissilient#1	bursting open with force, as do some ripe seed vessels
a	00003829	0.25	0	parturient#2	giving birth; "a parturient heifer"
a	00003939	0	0	dying#1	in or associated with the process of passing from life or ceasing to be; "a dying ma
a	00004171	0	0	moribund#2	being on the point of death; breathing your last; "a moribund patient"
a	00004296	0	0	last#5	occurring at the time of death; "his last words"; "the last rites"
a	00004413	0	0	abridged#1	(used of texts) shortened by condensing or rewriting; "an abridged version"
a	00004615	0	0	shortened#4 cut#3	with parts removed; "the drastically cut film"
a	00004723	0	0	half-length#2	abridged to half its original length
a	00004817	0	0	potted#3	(British informal) summarized or abridged; "a potted version of a novel"
a	00004980	0	0	unabridged#1	(used of texts) not shortened; "an unabridged novel"
a	00005107	0.5	0	uncut#7 full-length#2	complete; "the full-length play"
a	00005205	0.5	0	absolute#1	perfect or complete or pure; "absolute loyalty"; "absolute silence"; "absolu
a	00005473	0.75	0	direct#10	lacking compromising or mitigating elements; exact; "the direct opposite"
a	00005599	0.5	0.5	unquestioning#2 implicit#2	being without doubt or reserve; "implicit trust"
a	00005718	0.125	0	infinite#4	total and all-embracing; "God's infinite wisdom"
a	00005839	0.5	0.125	living#3	(informal) absolute; "she is a living doll"; "scared the living daylights ou

Εικόνα 6.1.1: Το SentiWordNet ως txt αρχείο

Το αρχείο αυτό μεταφέρθηκε στην βάση στους πίνακες:

sentiwordnet_synset

(synset_id integer (PK), pos varchar(1), offset integer, pos_score real, neg_score real, gloss longvarchar)

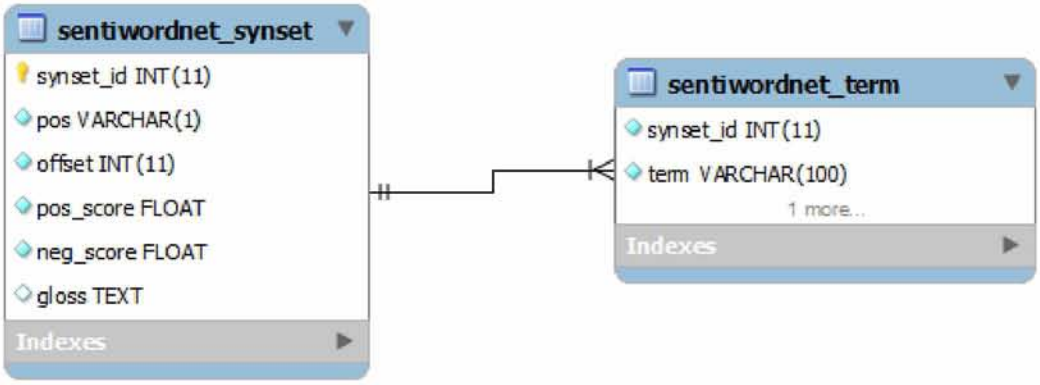
sentiwordnet_term

(synset_id integer (FK), term varchar(100))

Ο πίνακας **sentiwordnet_synset** περιέχει τα synsets με όλα τα στοιχεία του καθενός (pos: part-of-speech, offset: pos+offset→ID, positive score, negative score, gloss: sense description) εκτός από τα terms. Το Primary Key: *synset_id* πήρε τιμές κατά την δημιουργία του πίνακα και έχει τιμές από 1 έως 117.659, όσες είναι και οι γραμμές του SentiWordNet txt αρχείου.

Τα offset και gloss δεν χρησιμοποιούνται από το σύστημα.

Ο πίνακας **sentiwordnet_term** περιέχει τους όρους των synsets. Ο συνδυασμός *synset_id* + *term* είναι μοναδικός στον πίνακα, θα μπορούσε να οριστεί ως Primary Key, δεν κρίθηκε, όμως, σκόπιμο καθότι δεν γίνεται καμία αναζήτηση με δεδομένα: *synset_id* + *term*. Η αναζήτηση που γίνεται έχει δεδομένα: *term* + *pos*.



Εικόνα 6.1.2: Το SentiWordNet καταχωρημένο στην βάση δεδομένων

Το δοθέν dataset αποτελείται από τους παρακάτω βασικούς πίνακες:

authors

(author_id integer (PK), *author_name* varchar(250))

Ο πίνακας με τα id's και τα ονόματα όλων των συγγραφέων (bloggers).

posts

(post_id integer (PK), *post_title* varchar(250), *post_author* varchar(250), *post_author_id* integer (FK), *post_comments* integer, *post_url* longvarchar, *post_date* date, *post_inlinks_retrieved* integer, *post_comments_retrieved* integer, *post_length* integer, *post_outlinks* integer)

Ο πίνακας με όλα τα blog posts και τα στοιχεία αυτών.

comments

(comment_id integer (PK), *comment_post_id* integer (FK), *comment_body* longvarchar, *comment_author* varchar(250), *comment_date* date, *comment_vote* integer)

Ο πίνακας με τα comments που έχουν γραφτεί στα blog posts.

inlinks

(inlink_id integer (PK), *inlink_post_id* integer (FK), *inlink_title* varchar(250), *inlink_author* varchar(250), *inlink_date* date, *inlink_url* varchar(250))

Ο πίνακας με τα inlinks σε blog posts.

Για τις απαιτήσεις της διαδικασίας Sentiment Analysis και προκειμένου να γίνει καταγραφή όλων των βημάτων, δημιουργήθηκαν οι εξής πίνακες:

- *comment_sentence*
- *comment_sentence_word*
- *comment_sentence_negation_scope*
- *commentators*

Επιπλέον, προστέθηκαν νέες columns στους πίνακες authors, posts, comments. Ακολουθεί η τελική εικόνα της Βάσης Δεδομένων:

authors (author_id : PK)

(author_id integer, *author_name* varchar(250), *H_INDEX* integer, *MEIBI* integer, *MEIBIX* integer, *MEIBI1* integer, *MEIBI2* integer, *MEIBIX1* integer, *MEIBIX2* integer, *MEIBIO_aa* integer, *MEIBIO_bb* integer, *MEIBIXO_aa* integer, *MEIBIXO_bb* integer*)*

Ο πίνακας authors εμπλουτισμένος με όλες τις παραλλαγές MEIBI&MEIBIX οι οποίες επεξηγούνται στην 5.3.

posts (post_id : PK)

(post_id integer, *post_title* varchar(250), *post_author* varchar(250), *post_author_id* integer (FK), *post_comments* integer, *post_content* longvarchar, *post_url* longvarchar, *post_date* date, *post_inlinks_retrieved* integer, *post_comments_retrieved* integer, *post_length* integer, *post_outlinks* integer, *post_comments_opinion_no_unverified* integer, *post_distinct_commentators_opinionated* integer, *commentators_avg_score_a* double, *commentators_avg_score_b* double, *Sm* double, *Sx* double, *Sm1* double, *Sm2* double, *Sx1* double, *Sx2* double, *Smo_aa* double, *Smo_bb* double, *Sxo_aa* double, *Sxo_bb* double, *sum_inlinks_age* double , *post_age* double*)*

Ο πίνακας posts εμπλουτισμένος με όλους τους Sm&Sx δείκτες.

comments (comment_id : PK)

(comment_id integer , *comment_post_id* integer (FK), *comment_body longvarchar, comment_author* varchar(250), *comment_date* date, *comment_vote* integer, *comment_score_a* real, *comment_score_b* real)

Ο πίνακας comments με νέες columns τα *comment_score_a, comment_score_b.*

comment_sentence (comment_id + sentence_id : PK)

(*comment_id* integer (FK), *sentence_id* integer, *sentence_body* longvarchar, *sentence_score_a* real, *sentence_score_b* real, *avg_pos_score* real, *avg_neg_score* real*)*

Νέος πίνακας comment_sentence ο οποίος περιέχει όλες τις προτάσεις των comments. Κάθε εγγραφή είναι μία ξεχωριστή πρόταση. Ένα comment έχει τουλάχιστον μία πρόταση.

comment_sentence_negation_scope

(*comment_id* integer (FK), *sentence_id* integer (FK),

negation_scope_sentence longvarchar, *word* varchar(50), *index_in_sentence* integer, *tag_label* varchar(10))

Νέος πίνακας comment_sentence_negation_scope στον οποίο καταχωρούνται όλες οι negation_scope προτάσεις (σε περιπτώσεις που έχουμε negation).

comment_sentence_word (comment_id + sentence_id + word_id : PK)

(*comment_id* integer (FK), *sentence_id* integer (FK), *word_id* integer, *word* longvarchar, *tag_pos* varchar(4), *negation* varchar(5), *avg_positive* real, *avg_negative* real, *score* real, *index_in_sentence* integer)

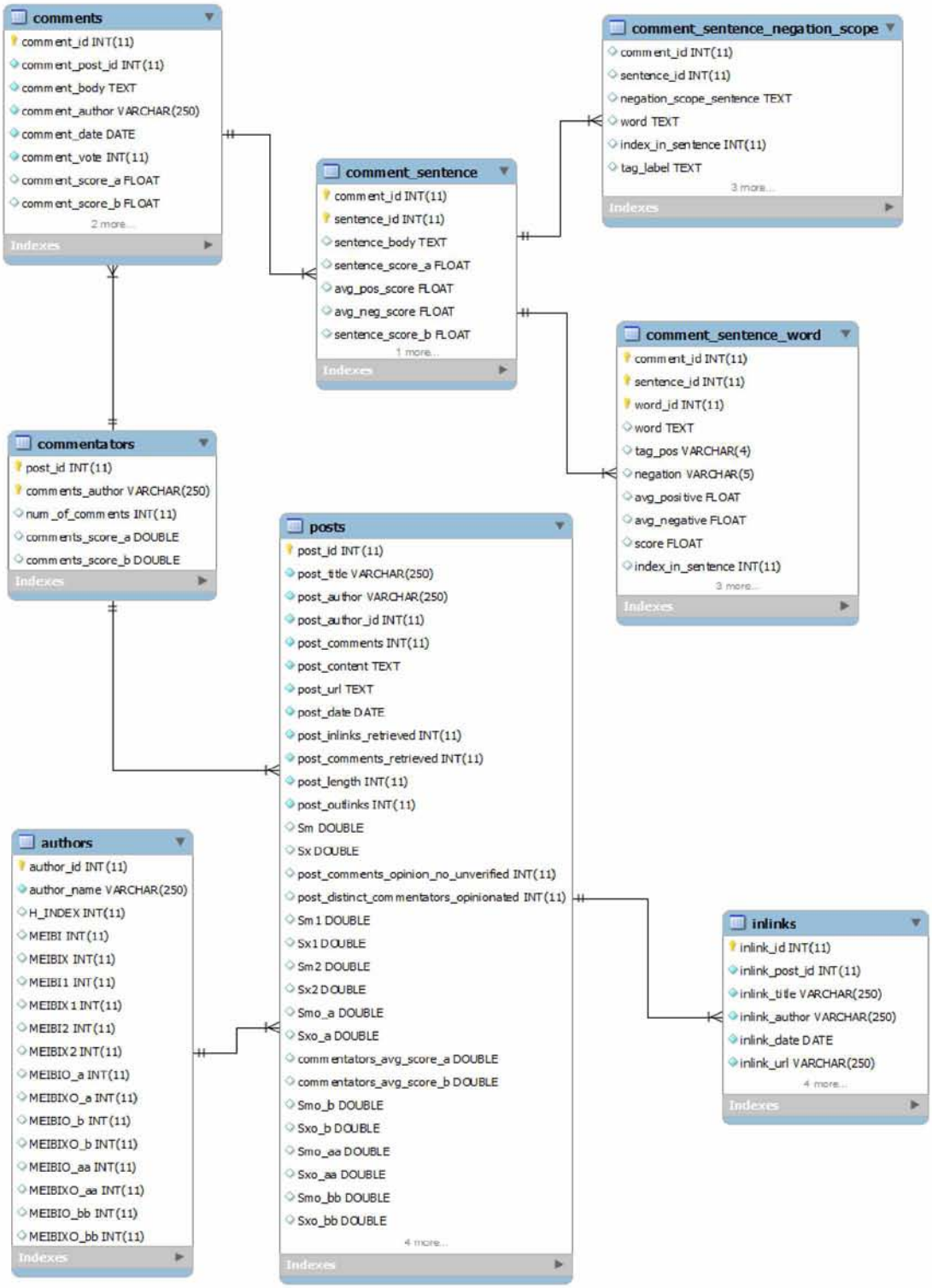
Νέος πίνακας comment_sentence_word. Σε αυτόν καταχωρούνται όλες οι sentiment words.

commentators (post_id + comments_author : PK)

(*post_id* integer (FK), *comments_author* varchar(250),

num_of_comments integer, *comments_score_a* double, *comments_score_b* double)

Νέος πίνακας commentators. Περιέχει όλους τους distinct commentators για κάθε blog post.



Εικόνα 6.1.3: Οι βασικοί πίνακες της βάσης δεδομένων

Εκτός από τους παραπάνω βασικούς πίνακες, υπάρχουν και οι βοηθητικοί πίνακες *for_h_index, for_meibi, for_meibix, for_meibi1, for_meibi2, for_meibix1,*

for_meibix2, for_meibio_aa, for_meibio_bb, for_meibixo_aa, for_meibixo_bb, οι οποίοι δημιουργήθηκαν για να καταγράφουν την διαδικασία υπολογισμού των H_INDEX, MEIBI, MEIBIX, MEIBI1, MEIBI2, MEIBIX1, MEIBIX2, MEIBIO_aa, MEIBIO_bb, MEIBIXO_aa, MEIBIXO_bb.

Για παράδειγμα, ο υπολογισμός του H-INDEX για τον blogger με id=13, είναι ο εξής:

author_id	total_posts	post_inlinks	H_INDEX
23	1	26	<NULL>
23	2	25	<NULL>
23	3	24	<NULL>
23	4	23	<NULL>
23	5	19	<NULL>
23	6	18	<NULL>
23	7	17	<NULL>
23	8	15	<NULL>
23	9	14	<NULL>
23	10	14	<NULL>
23	11	14	<NULL>
23	12	13	<NULL>
23	13	13	13

Εικόνα 6.1.4: Υπολογισμός H-INDEX

,ενώ ο υπολογισμός του MEIBI για τον ίδιο blogger:

author_id	total_posts	Sm	MEIBI
23	1	132.0	<NULL>
23	2	131.0	<NULL>
23	3	100.0	<NULL>
23	4	88.0	<NULL>
23	5	69.0	<NULL>
23	6	68.0	<NULL>
23	7	67.0	<NULL>
23	8	26.0	<NULL>
23	9	26.0	<NULL>
23	10	24.0	<NULL>
23	11	22.0	<NULL>
23	12	21.0	<NULL>
23	13	19.0	<NULL>
23	14	19.0	<NULL>
23	15	14.0	14

Εικόνα 6.1.5: Υπολογισμός MEIBI

6.2 Υποσύστημα Εξόρυξης Γνώμης

6.2.1 Sentiment_Analysis: Java Class

Μεταβλητές κλάσης: Connection con

Μέθοδος: main()

- Δημιουργεί το αντικείμενο *MaxentTagger tagger* (English - left3words) με το οποίο θα γίνει η επισήμανση των λέξεων με το Μέρος του Λόγου της κάθε μίας.
- Δημιουργεί το αντικείμενο *GenNegEx g* με το οποίο θα εντοπιστούν οι αρνήσεις στις προτάσεις.
- Δημιουργεί το αντικείμενο *db_functions db* το οποίο διαχειρίζεται την σύνδεση – επικοινωνία με την βάση δεδομένων.
- Διαβάζει τα αρχεία: *hitex_negexrules.negex* και *terms_to_check_if_negated.rules* τα οποία διατίθενται με τον tagger και τα εισάγει στα array lists: *rules* και *phrases* αντίστοιχα.
- Δημιουργεί το αντικείμενο *get_comment_score gcm* το οποίο κάνει την επεξεργασία/αξιολόγηση ενός Comment και αποδίδει δύο βαθμούς γνώμης σε αυτό. Στον constructor της κλάσης περνούν τα αντικείμενα *MaxentTagger tagger, GenNegEx g* καθώς και τα array lists *rules* και *phrases.*
- Προκειμένου να αποφευχθούν τα μεγάλα result set και δεδομένου ότι στο *engadge*t dataset περιλαμβάνονται 3.672.819 comments, η εκτέλεση και η ενημέρωση της βάσης γίνεται τμηματικά ανά 1000 comments. Με τον τρόπο αυτόν αποφεύγεται η επεξεργασία από την αρχή όταν η εκτέλεση για οποιοδήποτε λόγο κοπεί.
- Γίνεται ανάκτηση από την βάση δεδομένων των στοιχείων των Comments τα οποία απαντούν κατευθείαν στο ίδιο το blog post, δηλαδή δεν αρχίζουν με "@".
- Για κάθε ένα Comment εκτελείται η μέθοδος *comment_process()* του *gcm* αντικειμένου με παραμέτρους τα comment_id και comment_body του Comment.

6.2.2 get_comment_score: Java Class

Μεταβλητές κλάσης: MaxentTagger tagger

GenNegEx g

ArrayList rules

ArrayList phrases

Constructor: get_comment_score(MaxentTagger new_tagger, GenNegEx new_g, ArrayList new_rules, ArrayList new_phrases)

Μεταφέρει τα αντικείμενα *MaxentTagger*, *GenNegEx* και τα array lists *rules, phrases* στις μεταβλητές της κλάσης.

Μέθοδος: comment_process(String comment_body, Integer comment_id)

Παίρνει παραμετρικά τα comment_id και comment_body ενός Comment.

- Αφαιρεί από το comment body παραπλανητικά σημεία στίξης που μπορεί να σηματοδοτούν τέλος πρότασης, αντικαθιστώντας τα με "."
- Δημιουργεί το αντικείμενο *BreakIterator: iterator_s* με το οποίο θα γίνει διάσπαση του comment body σε προτάσεις.
- Δημιουργεί το array list *aListSentence* από objects comment_sentence_features για να καταχωρηθούν τα χαρακτηριστικά των προτάσεων του comment.
- Διαγράφονται από τους πίνακες της βάσης comment_sentence_word, comment_sentence_negation_scope, comment_sentence οι λέξεις, οι υπο-προτάσεις άρνησης και οι προτάσεις αντίστοιχα του συγκεκριμένου comment, που τυχόν έχουν αποθηκευτεί στην βάση από προηγούμενη εκτέλεση.
- Χωρίζεται με το iterator_s αντικείμενο το comment body σε προτάσεις και **για κάθε πρόταση του comment:**
  - ✓ Προκειμένου να χωριστεί η πρόταση σε λέξεις και να γίνει αποτίμηση για την κάθε μια, αντικαθιστώνται όλα τα μη αλφαβητικά σύμβολα με κενό.

- ✓ Δημιουργείται το array list *negList* από objects *sentence_negation_features* για να καταχωρηθούν τα χαρακτηριστικά των υπο-προτάσεων άρνησης της πρότασης.
- ✓ Δημιουργείται το array list *sentences* στο οποίο τοποθετείται η εξεταζόμενη πρόταση για τις απαιτήσεις του negator και μόνο.
- ✓ Δημιουργείται το array list *afterNegCheck* προκειμένου ο negator (object *g*) να τοποθετήσει σε αυτό τις υπο-προτάσεις άρνησης που τυχόν περιέχονται στην εξεταζόμενη πρόταση.
- ✓ Εκτελείται η μέθοδος *negCheck(...)* του negator με αποδέκτη της (των) επιστρεφόμενης (επιστρεφόμενων) υπο-πρότασης (προτάσεων) το array list *afterNegCheck.*
- ✓ Με τη βοήθεια iterator γίνεται περιδιάβαση του array list *afterNegCheck* για να μελετηθούν οι υπο-προτάσεις άρνησης. **Για κάθε υπο-πρόταση άρνησης** (*negation_scope_sentence*)**:**
  - Εντοπίζεται σε ποιό index (θέση) της εξεταζόμενης πρότασης, αρχίζει το *negation_scope_sentence* → *index_negation_scope_sentence*.
  - Τμηματοποιείται το negation_scope_sentence.
  - Αποσπάται η 1η λέξη και υπολογίζεται η θέση της στην πρόταση με τη βοήθεια του *index_negation_scope_sentence.*
  - Εκτελείται η μέθοδος *tagString* του *tagger* για την $1^{η}$ λέξη και αποκαλύπτεται το μέρος του λόγου της $1^{ης}$ λέξης. Εάν είναι επίθετο ή επίρρημα θα αντιστραφεί η πολικότητα της $1^{ης}$ λέξης. Εάν δεν είναι, αποσπάται η $2^{η}$ λέξη, και όπως έγινε με την $1^{η}$ λέξη, υπολογίζεται η θέση της στην πρόταση, εκτελείται ο tagger και γίνεται γνωστό το μέρος του λόγου της $2^{ης}$ λέξης. Εάν είναι επίθετο ή επίρρημα ή ρήμα, θα αντιστραφεί η πολικότητα της $2^{ης}$ λέξης.
  - Για όποια λέξη αποφασίστηκε ότι θα αντιστραφεί η πολικότητα, δημιουργείται το αντικείμενο *sentence_negation_features sentence_neg* το οποίο αποτελείται από τα χαρακτηριστικά της λέξης που θα αντιστρέψει την πολικότητά της (πρόταση στην οποία ανήκει, υπο-πρόταση άρνησης στην οποία ανήκει, η ίδια η λέξη, θέση της λέξης στην πρόταση, μέρος του λόγου της λέξης).
  - Το αντικείμενο *sentence_neg* προστίθεται στην negList.

- Εφόσον έχει προκύψει ότι κάποια λέξη θα αντιστρέψει την πολικότητά της, η *negation_scope_sentence* καταχωρείται στη βάση δεδομένων στον πίνακα *comment_sentence_negation_scope* με όλα τα χαρακτηριστικά της (comment και πρόταση, λέξη $1^{η}$ ή $2^{η}$, θέση της λέξης στην πρόταση, μέρος του λόγου της λέξης).

✓ Δημιουργείται το array list *aListWord_features* από objects *word_features* για να καταχωρηθούν τα χαρακτηριστικά των λέξεων της πρότασης.

✓ Η πρόταση με τη βοήθεια StringTokenizer χωρίζεται στις λέξεις που την απαρτίζουν. **Για κάθε μια λέξη της πρότασης:**

- Σημειώνεται η θέση του occurence της εξεταζόμενης λέξης στην πρόταση (μετά την θέση της προηγούμενης λέξης)
- Εάν στο array list negList υπάρχει η ίδια λέξη με την εξεταζόμενη λέξη και με την ίδια θέση στην πρόταση με την εξεταζόμενη λέξη, τότε *change_word_polarity = true* που σημαίνει ότι αφού η εξεταζόμενη λέξη βαθμολογηθεί, θα αλλάξει την πολικότητά της.
- Εκτελείται η μέθοδος *tagString()* του *tagger* για την εξεταζόμενη λέξη και αποκαλύπτεται το μέρος του λόγου της.
- Εάν η λέξη είναι επίθετο ή επίρρημα ή ρήμα, από τους πίνακες της βάσης *sentiwordnet_synset* και *sentiwordnet_term* συλλέγεται ο Μ.Ο. των θετικών score *(avg_word_pos)* και ο Μ.Ο. των αρνητικών score *(avg_word_neg)* του term ο οποίος ισούται με την εξεταζόμενη λέξη και ανήκει στο ίδιο μέρος του λόγου με αυτήν.
- Αν δεν υπάρχει conflict στους 2 Μ.Ο. (δηλαδή δεν υπάρχουν θετικές και αρνητικές βαθμολογίες για το ίδιο term στο ίδιο μέρος του λόγου):

  Εάν *avg_word_pos* !=0 ΚΑΙ *avg_word_neg*=0 ➔
  final_word_score = avg_word_pos
  Αλλιώς εάν *avg_word_pos*=0 ΚΑΙ *avg_word_neg*!=0 ➔
  *final_word_score = avg_word_neg * (-1)*
  Αλλιώς εάν *avg_word_pos*=0 ΚΑΙ *avg_word_neg*=0 ➔
  word : neutral

final_word_score = 0

Διαφορετικά, εάν υπάρχει conflict, εφαρμόζεται thredhold= 0.05. Εάν

avg_word_pos > *avg_word_neg* ΚΑΙ η διαφορά τους είναι >= thredhold ➔ *final_word_score = avg_word_pos*

Αλλιώς εάν *avg_word_neg* > *avg_word_pos* ΚΑΙ η διαφορά τους είναι >= threshold➔*final_word_score=avg_word_neg*(-1)*

Αλλιώς word : neutral

final_word_score= 0

- Εάν *change_word_polarity* είναι true, το πρόσημο του *final_word_score* αντιστρέφεται.
- Δημιουργεί το αντικείμενο *word_features : word* το οποίο το οποίο αποτελείται από τα χαρακτηριστικά της εξεταζόμενης λέξης (η ίδια η λέξη, αντιστροφή πρόσημου ή όχι, μέρος του λόγου, *final_word_score).* Το αντικείμενο προστίθεται στο *aListWord_features* array list.
- Η λέξη με όλα τα παραπάνω χαρακτηριστικά εισάγεται στη βάση στον πίνακα *comment_sentence_word.*

✓ Γίνεται περιδιάβαση του array list *aListWord_features* με τις λέξεις της πρότασης και υπολογίζεται το score της πρότασης από τις βαθμολογημένες λέξεις με δυο τρόπους:

✓ Ο πρώτος τρόπος αποδίδει στην πρόταση τον Μ.Ο. των *final_word_score* όλων των λέξεων χωρίς να συνυπολογίζονται οι ουδέτερες λέξεις. ➔ *sentence_score_a*

✓ Ο δεύτερος τρόπος υπολογίζει τον Μ.Ο. των θετικών λέξεων και τον Μ.Ο. των αρνητικών λέξεων. ➔ *avg_sentence_pos*, *avg_sentence_neg* (και οι δύο θετικοί αριθμοί).

Οι λέξεις με μηδενικό *final_word_score* δεν λαμβάνονται υπόψη.

Εάν *avg_sentence_pos* > *avg_sentence_neg* τουλάχιστον κατά threshold=0.2 ➔ *sentence_score_b = avg_sentence_pos*

Αλλιώς εάν *avg_sentence_neg* > *avg_sentence_pos* τουλάχιστον κατά threshold=0.2 ➔ *sentence_score_b = avg_sentence_neg* * (-1)

Αλλιώς ➔ *sentence_score_b* = 0

- ✓ Δημιουργείται το αντικείμενο *comment_sentence_features sentence_f* το οποίο αποτελείται από την πρόταση και τα χαρακτηριστικά της (*sentence_id, sentence, sentence_score_a, avg_sentence_pos, avg_sentence_neg, sentence_score_b).*
- ✓ Το αντικείμενο *comment_sentence_features* προστίθεται στην array list *aListSentence.*
- ✓ Η πρόταση με όλα τα ανωτέρω χαρακτηριστικά εισάγεται στην βάση δεδομένων στον πίνακα *comment_sentence.*

- ➢ Αφού ολοκληρωθεί η επεξεργασία όλων των προτάσεων πραγματοποιείται ο υπολογισμός της βαθμολογίας του comment από τα *sentence_score_a* και *sentence_score_b* των προτάσεών του. Με την βοήθεια του *aListSentence* array list, υπολογίζεται ο Μ.Ο. των *sentence_score_a* → *comment_score_a* και ο Μ.Ο. των *sentence_score_b* → *comment_score_b.*
- ➢ Ενημερώνεται στη βάση ο πίνακας *comments* με τα *comment_score_a* και *comment_score_b.*

6.2.3 comment_sentence_features: Java Class

Μεταβλητές κλάσης: int sentence_id

String sentence_body

Float sentence_score_a

Float avg_pos_score

Float avg_neg_score

Float sentence_score_b

Constructor: comment_sentence_features(int sentence_id, String sentence_body, float sentence_score_a, float avg_pos_score, float avg_neg_score, float sentence_score_b)

Μεταφέρει τα int sentence_id, String sentence_body, float sentence_score_a,

float avg_pos_score,float avg_neg_score, float sentence_score_b στις μεταβλητές της κλάσης.

6.2.4 sentence_negation_features: Java Class

Μεταβλητές κλάσης: int sentence_id

String negation_scope_sentence

String word

int index_in_sentence

String tag_label

Constructor: sentence_negation_features(int sentence_id, String negation_scope_sentence, String word, int index_in_sentence, String tag_label)

Μεταφέρει τα int sentence_id, String negation_scope_sentence,String word,int index_in_sentence,String tag_labelστις μεταβλητές της κλάσης.

6.2.5 word_features: Java Class

Μεταβλητές κλάσης: String word

String negation

String tag

float average_pos

float average_neg

float final_score

Constructor: word_features(String w, String neg, String t, float average_p, float average_n, float final_sc)

Μεταφέρει τα String word, String negation, String tag, float average_pos, float average_neg, float final_score στις μεταβλητές της κλάσης.

6.2.6 db_functions: Java Class

Μεταβλητή κλάσης: *Connection myConnection*

Constructor: db_functions()

Μέθοδος: init()

Αποκαθιστά επικοινωνία με την *localhost:3306/engadget* database μέσω του *jdbc:mysql* driver.

Μέθοδος: getMyConnection()

Getter για την μεταβλητή κλάσης *myConnection.*

Μέθοδος : close

Κλείνει ένα ενεργό result set.

Μέθοδος: destroy()

Κλείνει την επικοινωνία με την βάση.

6.2.7 create_sentiwordnet_term: MySQL Stored Procedure

- Με χρήση cursor διαβάζεται μια-μια εγγραφή από τον πίνακα *sentiwordnet_synset* η οποία περιέχει το synset_id ενός synset του SentiWordNet, σε συνδυασμό (joined) με τα terms αυτού του synset τα οποία είναι καταγεγραμμένα σε προσωρινό πίνακα.
- Τα terms του synset διαχωρίζονται (παρεμβάλλεται μεταξύ τους ο χαρακτήρας : #) και κάθε ένα μαζί με το αντίστοιχο synset_id εισάγεται στον πίνακα *sentiwordnet_term.*

6.3 Υποσύστημα Ταξινόμησης Bloggers

6.3.1 update_posts: MySQL Script

- Ενημέρωση πίνακα posts με το post_comments_opinion_no_unverified.
- Ενημέρωση πίνακα posts με το post_distinct_commentators_opinionated.

6.3.2 insert_post_commentators: MySQL Stored Procedure

- Με χρήση cursor διατρέχονται οι εγγραφές του πίνακα posts.

- Για κάθε εγγραφή του posts συλλέγονται από τον πίνακα comments οι επιβεβαιωμένοι commentators που έγραψαν comments για το συγκεκριμένο post, με το πλήθος των comments του καθενός και τον ΜΟ των comment_score_a, τον ΜΟ των comment_score_b των comments του καθενός.
- Γίνεται insert στον πίνακα commentators κάθε σχολιαστής του post μαζί με τα ανακτηθέντα στοιχεία (*comments_author, comments_post_id, num_of_comments, comments_avg_score_a, comments_avg_score_b*).

6.3.3 update_commentators_with_score: Java Class

Μεταβλητές κλάσης: Connection con

ArrayList C

Μέθοδος: main()

- Ανακτώνται από την βάση δεδομένων από τον πίνακα *posts* το *post_id* όλων των posts.
- **Για κάθε ένα *post_id*:**
  - ✓ Επιλέγονται από τον πίνακα *commentators* οι commentators (με το όνομά τους και το πλήθος των comments) οι οποίοι απευθύνονται στο συγκεκριμένο post *(post_id)*.

    Για κάθε commentator που επιλέγεται:
    - Επιλέγονται από τον πίνακα *comments* τα: *comment_id, comment_score_a, comment_score_b* των comments που έχει γράψει ο συγκεκριμένος commentator για το post *post_id.*
    - Δημιουργούνται τα array list *C1, C2* στα οποία καταχωρούνται τα *comment_score_a, comment_score_b* αντίστοιχα των 10 πρώτων χρονικά comments του commentator που επιλέχθηκαν.
    - Εκτελείται η weighted_avg() μέθοδος δυο φορές και εξάγει τα commentator score: *final_s_a, final_s_b.*
    - Ενημερώνεται στον πίνακα *commentators* ο επεξεργαζόμενος commentator του post *post_id*, με τα score *final_s_a, final_s_b* στις columns *comments_score_b, comments_score_b.*

Μέθοδος: weighted_avg() → Recursive function

- Παίρνει παραμετρικά το πλήθος των comments που έχουν ήδη συμμετάσχει στον υπολογισμό του score του commentator *(count)* και το συνολικό πλήθος των comments που συμμετέχουν *(num_of_c)*.
- Παίρνει από το array list C1 ή C2 την επόμενη τιμή του αποθηκευμένου comment score και παράγει ως output το 80% αυτού, συν το 20% του output της ίδιας function. Έτσι, το πρώτο χρονικά comment συμβάλλει στη διαμόρφωση του score του commentator κατά 80%, ενώ όλα τα υπόλοιπα μαζί συμβάλλουν κατά 20% με φθίνουσα βαρύτητα.

6.3.4 update_posts_1: MySQL Script

- Ενημερώνονται σε κάθε post (πίνακας *posts*) οι columns *commentators_avg_score_a, commentators_avg_score_b* με τους ΜΟ των *comments_avg_score_a, comments_avg_score_b* των commentators αυτού του post.

6.3.5 update_posts_with_sm_sx: MySQL Stored Procedure

- Με χρήση cursor διατρέχονται οι εγγραφές του πίνακα *posts*.
- **Για κάθε εγγραφή *posts***:
  - Υπολογίζεται η μεταβλητή *vpost_age* = $\frac{86400}{(to\_days('2010-03-29') - to\_days(post\_date)) \cdot 86400 + 86400}$
  - Υπολογίζεται η μεταβλητή *vSm* = *4*(post_comments_retrieved+1)*(post_inlinks_retrieved)*vpost_age*
  - Από τον πίνακα inlinks για τις εγγραφές που αναφέρονται στο εξεταζόμενο post, υπολογίζεται η μεταβλητή *vsum_inlinks_age* = $sum\left(\frac{86400}{(to\_days('2010-03-29') - to\_days(post\_date)) \cdot 86400 + 86400}\right)$
  - Υπολογίζεται η παράσταση *vSx* = *4*(post_comments_retrieved+1)*vsum_inlinks_age*
  - Ενημερώνεται το εξεταζόμενο post με *Sm = vSm, Sx = vSx*

6.3.6 update_posts_with_sm1_sx1: MySQL Procedure

- Με χρήση cursor διατρέχονται οι εγγραφές του πίνακα *posts*.
- **Για κάθε εγγραφή *posts*:**
  - Υπολογίζεται η μεταβλητή *vpost_age* όπως στο 6.3.5
  - Υπολογίζεται η παράσταση *vSm1 =*
    *4 * (post_comments_opinion_no_unverified +1) **
    *(post_inlinks_retrieved) * vpost_age*
  - Υπολογίζεται η μεταβλητή *vsum_inlinks_age* όπως στο 6.3.5
  - Υπολογίζεται η παράσταση *vSx1 =*
    *4 * (post_comments_opinion_no_unverified+1) * vsum_inlinks_age*
  - Ενημερώνεται το εξεταζόμενο post με *Sm1 = vSm1, Sx1 = vSx1*

6.3.7 update_posts_with_sm2_sx2: MySQL Stored Procedure

- Με χρήση cursor διατρέχονται οι εγγραφές του πίνακα *posts*.
- **Για κάθε εγγραφή *posts*:**
  - Υπολογίζεται η μεταβλητή *vpost_age* όπως στο 6.3.5
  - Υπολογίζεται η παράσταση *vSm2 =*
    *4 * (post_distinct_commentators_opinionated+1) **
    *(post_inlinks_retrieved) * vpost_age*
  - Υπολογίζεται η μεταβλητή *vsum_inlinks_age* όπως στο 6.3.5
  - Υπολογίζεται η παράσταση *vSx2 =*
    *4 * (post_distinct_commentators_opinionated+1) * vsum_inlinks_age*
- Ενημερώνεται το εξεταζόμενο post με *Sm2 = vSm2, Sx2 = vSx2*

6.3.8 update_posts_with_smo_sxo: MySQL Stored Procedure

- Με χρήση cursor διατρέχονται οι εγγραφές του πίνακα *posts*.
- **Για κάθε εγγραφή *posts*:**
  - Υπολογίζεται η μεταβλητή *vpost_age* όπως στο 6.3.5
  - Υπολογίζεται η παράσταση *vSmo_a =*
    *4 * (post_distinct_commentators_opinionated **
    *(commentators_avg_score_a+1)+1)**
    *(post_inlinks_retrieved)*vpost_age*
  - Υπολογίζεται η παράσταση *vSmo_b =*

*4 * (post_distinct_commentators_opinionated * (commentators_avg_score_b+1)+1)* (post_inlinks_retrieved)*vpost_age*

- Υπολογίζεται η μεταβλητή *vsum_inlinks_age* όπως στο 6.3.5
- Υπολογίζεται η παράσταση *vSxo_a = 4 * (post_distinct_commentators_opinionated * (commentators_avg_score_a+1) + 1) * vsum_inlinks_age*
- Υπολογίζεται η παράσταση *vSxo_b = 4 * (post_distinct_commentators_opinionated * (commentators_avg_score_b+1) + 1) * vsum_inlinks_age*
- Ενημερώνεται το εξεταζόμενο post με : *Smo_aa = vSmo_a, Smo_bb = vSmo_b, Sxo_aa=vSxo_a, Sxo_bb=vSxo_b*

6.3.9 update_authors_with_H_INDEX: MySQL Stored Procedure

- Με χρήση cursor διατρέχονται οι εγγραφές του πίνακα *authors* (bloggers)
- **Για κάθε *author (author_id)*:**
  - Με χρήση cursor διατρέχονται τα posts του author, με φθίνουσα σειρά κατά *post_inlinks_retrieved.*
  - Οι μεταβλητές *total_posts, vcitations, previous_vcitations* αρχικοποιούνται στο 0.
  - **Για κάθε post του author (post_id):**
    - Η μεταβλητή *total_posts* (πλήθος posts μέχρι τώρα) αυξάνεται κατά 1.
    - Ανακτάται η τιμή της column *post_inlinks_retrieved* (αριθμός αναφορών στο post) και εισάγεται στην μεταβλητή *vcitations.*
    - Εάν *vcitations* > *total_posts* σημαίνει ότι δεν έχει ακόμα προσδιοριστεί ο h_index του author και απλά γίνεται μια εγγραφή στον πίνακα *for_h_index (author_id, total_posts, post_inlinks_retrieved)* προορισμός του οποίου είναι η καταγραφή των βημάτων και ο μετέπειτα έλεγχος του προσδιορισμού του h_index.
    - Εάν *vcitations = total_posts* , το h_index βρέθηκε. Γίνεται εγγραφή στον πίνακα *for_h_index (author_id, total_posts,*

post_inlinks_retrieved, H_INDEX = total_posts) και ενημερώνεται ο author (πίνακας *authors*) με τον h_index = *total_posts.*

- Εάν *vcitations < total_posts* και *previous_vcitations > total_posts-1* σημαίνει ότι το h_index βρέθηκε και είναι το πλήθος των posts στον προηγούμενο κύκλο δηλαδή *total_posts-1.* Γίνεται εγγραφή στον πίνακα *for_h_index (author_id, total_posts, post_inlinks_retrieved, H_INDEX = total_posts-1)* και ενημερώνεται ο author (πίνακας *authors*) με τον h_index = *total_posts-1.*
- Σε όλες τις περιπτώσεις η μεταβλητή *previous_vcitations* παίρνει την τιμή : *vcitations.*

- Εάν διαβαστούν όλα τα posts του author και δεν έχει βρεθεί ο h_index, τότε, εάν στην τελευταία εγγραφή (post) που διαβάστηκε ισχύει : *total_posts > 0* και *vcitations > total_posts* δηλαδή υπάρχουν κάποια posts του author και οι αναφορές του τελευταίου (ας μην ξεχνάμε ότι τα posts έχουν ανακτηθεί order by *post_inlinks_retrieved=vcitations* desc) υπερέχουν του αριθμού των post, το h_index είναι η μεταβλητή *total_posts.* Γίνεται εγγραφή στον πίνακα *for_h_index (author_id, total_posts, post_inlinks_retrieved, H_INDEX = total_posts)* και ενημερώνεται ο author (πίνακας *authors*) με τον h_index = *total_posts.*

- Εάν δεν υπάρχει h_index για τον author, αυτός ενημερώνεται με h_index=0.

6.3.10 update_authors_with_MEIBI: MySQL Stored Procedure

- Με χρήση cursor διατρέχονται οι εγγραφές του πίνακα *authors* (bloggers).
- **Για κάθε *author (author_id)*:**
  - Με χρήση cursor διατρέχονται τα posts του author (πίνακας *posts*), order by *Sm* descending.
  - Οι μεταβλητές *total_posts, vSm, previous_vSm* αρχικοποιούνται στο 0.
  - **Για κάθε post του author (post_id):**

- Η μεταβλητή *total_posts* (πλήθος posts μέχρι τώρα) αυξάνεται κατά 1.
- Ανακτάται η τιμή της column Sm (score S_j^m επιρροής του post) και εισάγεται στην μεταβλητή *vSm.*
- Εάν round(*vSm)* > *total_posts* σημαίνει ότι δεν έχει ακόμα προσδιοριστεί ο MEIBI του author (ο οποίος έχει ακριβώς την φιλοσοφία του h_index) και απλά γίνεται μια εγγραφή στον πίνακα *for_meibi (author_id, total_posts, Sm)* προορισμός του οποίου είναι η καταγραφή των βημάτων και ο μετέπειτα έλεγχος του προσδιορισμού του MEIBI.
- Εάν round(*vSm)* = *total_posts*, ο MEIBI βρέθηκε. Γίνεται εγγραφή στον πίνακα *for_meibi (author_id, total_posts, Sm, MEIBI = total_posts)* και ενημερώνεται ο author (πίνακας *authors*) με τον *MEIBI = total_posts.*
- Εάν round(*vSm)* < *total_posts* και *previous_Sm* > *total_posts-1* σημαίνει ότι ο *MEIBI* βρέθηκε και είναι το πλήθος των posts στον προηγούμενο κύκλο δηλαδή *total_posts-1.* Γίνεται εγγραφή στον πίνακα *for_meibi (author_id, total_posts, Sm, MEIBI = total_posts-1)* και ενημερώνεται ο author (πίνακας *authors*) με *MEIBI = total_posts-1.*
- Σε όλες τις περιπτώσεις η μεταβλητή *previous_Sm* παίρνει την τιμή *Sm.*

- Εάν διαβαστούν όλα τα posts του author και δεν έχει βρεθεί ο *MEIBI*, τότε, εάν στην τελευταία εγγραφή (post) που διαβάστηκε ισχύει : *total_posts > 0* και *Sm > total_posts* δηλαδή υπάρχουν κάποια posts του author και το Sm score του τελευταίου post (ας μην ξεχνάμε ότι τα post έχουν ανακτηθεί order by *Sm* desc) υπερέχουν του αριθμού των post, το *MEIBI* είναι η μεταβλητή *total_posts.* Γίνεται εγγραφή στον πίνακα *for_meibi (author_id, total_posts, Sm, MEIBI = total_posts)* και ενημερώνεται ο author (πίνακας *authors*) με τον *MEIBI = total_posts.*

- Εάν δεν υπάρχει *MEIBI* για τον author, αυτός ενημερώνεται με *MEIBI* =0.

6.3.11 update_authors_with_MEIBIX: MySQL Stored Procedure

- Η διαδικασία είναι ακριβώς η ίδια με το 6.3.10 με τη μόνη διαφορά ότι αντί για την column *Sm* του πίνακα *posts*, χρησιμοποιείται για τον υπολογισμό η column *Sx* (score S_j^x επιρροής του post).

6.3.12 update_authors_with_MEIBI1: MySQL Stored Procedure

- Η διαδικασία είναι ακριβώς η ίδια με το 6.3.10 με τη μόνη διαφορά ότι αντί για την column *Sm* του πίνακα *posts*, χρησιμοποιείται για τον υπολογισμό η column *Sm1*.

6.3.13 update_authors_with_MEIBI2: MySQL Procedure

- Η διαδικασία είναι ακριβώς η ίδια με το 6.3.10 με τη μόνη διαφορά ότι αντί για την column *Sm* του πίνακα *posts*, χρησιμοποιείται για τον υπολογισμό η column *Sm2*.

6.3.14 update_authors_with_MEIBIX1: MySQL Procedure

- Η διαδικασία είναι ακριβώς η ίδια με το 6.3.10 με τη μόνη διαφορά ότι αντί για την column *Sm* του πίνακα *posts*, χρησιμοποιείται για τον υπολογισμό η column *Sx1*.

6.3.15 update_authors_with_MEIBIX2: MySQL Procedure

- Η διαδικασία είναι ακριβώς η ίδια με το 6.3.10 με τη μόνη διαφορά ότι αντί για την column *Sm* του πίνακα *posts*, χρησιμοποιείται για τον υπολογισμό η column *Sx2*

6.3.16 update_authors_with_MEIBIO_aa: MySQL Procedure

- Η διαδικασία είναι ακριβώς η ίδια με το 6.3.10 με τη μόνη διαφορά ότι αντί για την column *Sm* του πίνακα *posts*, χρησιμοποιείται για τον υπολογισμό η column *Smo_aa*.

6.3.17 update_authors_with_MEIBIO_bb: MySQL Procedure

- Η διαδικασία είναι ακριβώς η ίδια με το 6.3.10 με τη μόνη διαφορά ότι αντί για την column *Sm* του πίνακα *posts*, χρησιμοποιείται για τον υπολογισμό η column *Smo_bb*.

6.3.18 update_authors_with_MEIBIXO_aa: MySQL Procedure

- Η διαδικασία είναι ακριβώς η ίδια με το 6.3.10 με τη μόνη διαφορά ότι αντί για την column *Sm* του πίνακα *posts*, χρησιμοποιείται για τον υπολογισμό η column *Sxo_aa*.

6.3.19 update_authors_with_MEIBIXO_bb: MySQL Procedure

- Η διαδικασία είναι ακριβώς η ίδια με το 6.3.10 με τη μόνη διαφορά ότι αντί για την column *Sm* του πίνακα *posts*, χρησιμοποιείται για τον υπολογισμό η column *Sxo_bb*.

7 Αξιολόγηση Αποτελεσμάτων

Παρουσιάζονται τα αποτελέσματα και οι κατατάξεις που προκύπτουν από τις προϋπάρχουσες και τις νέες μετρικές.

7.1 Χαρακτηριστικά δεδομένων

Τα δεδομένα τα οποία χρησιμοποιούμε προέρχονται από την γνωστή ιστοσελίδα www.engadget.com και συγκεκριμένα από την ενότητα reviews. Οι bloggers σε κάθε άρθρο τους παρουσιάζουν κάποιο gadget, εκφέροντας παράλληλα την δική τους άποψη σχετικά με το πόσο τους αρέσει ή όχι. Οι αναγνώστες σχολιάζουν ελεύθερα.

Όπως φαίνεται στην υποενότητα 6.1, στην διάθεσή μας βρίσκονται όλες οι πληροφορίες σχετικά με τα posts, τα comments, τους authors (bloggers) και τους commentators.

Τα δεδομένα τα οποία ανακτήθηκαν αναφέρονται στην περίοδο από 2004-03-02 έως 2010-03-26. Αυτές είναι η μικρότερη και η μεγαλύτερη ημερομηνία δημοσίευσης άρθρων. Δυστυχώς για εμάς, η αντίστοιχη περίοδος ανάρτησης σχολίων (2005-12-19 με 2010-03-28) δεν ταυτίζεται, καθότι για τα δύο πρώτα έτη (2004 & 2005) δεν υπάρχουν καταχωρημένα σχόλια. Επιπροσθέτως, την ημέρα "έναρξης" των σχολιασμών παρατηρείται ένας άνευ προηγουμένου υπερβάλλον ζήλος από πλευράς των commentators, όπως προλαβαίνουν μέσα σε 24 ώρες να αναρτήσουν 223690 comments (ενώ η μέση τιμή είναι 2354 comments την ημέρα). Είναι προφανές ότι έχει γίνει κάποιο σφάλμα κατά την άντληση των δεδομένων από το web (crawling), όσο αφορά τις ημερομηνίες των σχολίων, γεγονός το οποίο λαμβάνουμε υπόψη στις μετρήσεις μας.

Ένα ακόμη στοιχείο που αποτελεί εμπόδιο στην σύγκριση των τελεστών κατάταξης είναι η ύπαρξη 'Unverified' χρηστών-σχολιαστών. Από το σύνολο των 3.672.819 comments, τα 2.072.668 έχουν γραφτεί από τέτοιους. Αν και η ταυτότητα των commentators δεν μας απασχολεί κατά την εφαρμογή των MEIBI&MEIBIX, στους MEIBIO&MEIBIXO είναι απαραίτητο να ξεχωρίζουμε τους distinct commentators.

Τέλος, 192.601 σχόλια ξεκινούν με τον '@' χαρακτήρα, που σημαίνει ότι απαντούν στα λεγόμενα ενός άλλου σχολιαστή.

Κατά το στάδιο της αξιολόγησης, λαμβάνουμε υπόψη αυτές τις παρατηρήσεις και εφαρμόζουμε τους τελεστές σε φιλτραρισμένα υποσύνολα των αρχικών δεδομένων, ώστε οι συγκρίσεις να γίνουν επί ίσοις όροις.

7.2 H-INDEX, MEIBI, MEIBIX

Αρχικά, εφαρμόζουμε το Σύστημα που αναπτύχθηκε στο σύνολο των δεδομένων D. Η οικογένεια μετρικών MEIBI&MEIBIX, χρονικά ευαίσθητη ούσα (time-aware), υποβαθμίζει την βαθμολογία των posts όσο περνάει ο καιρός. Επομένως, εάν η αξιολόγηση των δεδομένων D γίνονταν με σημερινή ημερομηνία, τα score θα ήταν πολύ μικρού εύρους και δεν θα μας διευκόλυναν στην εξαγωγή συμπερασμάτων. Για αυτό τον λόγο επιλέγουμε σαν "σημερινή" ημερομηνία την 2010-3-29 (τρεις ημέρες μετά την τελευταία ανάρτηση post και μία ημέρα μετά την τελευταία ανάρτηση σχολίου).

Ακολουθεί η αξιολόγηση των bloggers σύμφωνα με τον ευρέως αποδεκτό h-index, τον MEIBI και τον MEIBIX. Παρατίθενται τα ονόματα των δεκαπέντε πρώτων αρθρογράφων κατά φθίνουσα σειρά του αθροίσματος των τριών αυτών μετρικών.

	author_name	H_INDEX	MEIBI	MEIBIX
▶	Joshua Topolsky	52	111	125
	Darren Murph	34	112	134
	Chris Ziegler	33	113	129
	Nilay Patel	42	104	122
	Vladislav Savov	26	100	122
	Thomas Ricker	42	91	107
	Paul Miller	40	89	105
	Ross Miller	27	77	85
	Ryan Block	53	58	66
	Laura June	20	76	81
	Donald Melanson	24	67	80
	Tim Stevens	20	59	68
	Joseph L. Flatley	19	55	67
	Joanna Stern	19	45	50
	Richard Lawler	19	40	43

Εικόνα 7.2.1: Ταξινόμηση Bloggers στο D (H-INDEX, MEIBI, MEIBIX)

Όπως αναμένεται, οι MEIBI&MEIBIX συμπλέουν. Εν γένει, αυτό συμβαίνει και με τον h-index, με μερικές, όμως, εξαιρέσεις. Ο Ryan Block, αν και έχει τον υψηλότερο h-index, τα MEIBI&MEIBIX του είναι πολύ μέτρια. Ο λόγος είναι ότι μετά το 2008 η δραστηριότητά του είναι μηδαμινή, σε αντίθεση για παράδειγμα με τον Joshua Topolsky ο οποίος δημοσιεύει ασύστολα μέχρι το τέλος της περιόδου μέτρησης. Ως εκ τούτου, οι MEIBI-MEIBIX θεωρούν ότι η επιρροή του Ryan Block έχει παρέλθει και του προσδίδουν σχεδόν την μισή αξία από τον "φρέσκο" Joshua Topolsky, παρά το γεγονός ότι ο πρώτος έχει 5.643 δημοσιεύσεις ενώ ο δεύτερος 2.057.

Τα συγκεκριμένα αποτελέσματα επαληθεύονται από τα [1], [2].

7.3 MEIBIO&MEIBIXO vs MEIBI&MEIBIX

Εξετάζουμε κατά πόσο διαφοροποιείται η ταξινόμηση των Opinionated τελεστών σε σχέση με των προκατόχων τους. Για να το κάνουμε αυτό, ορίζουμε το υποσύνολο $D1 \in D$ από το οποίο έχουν αποκλειστεί όσα comments ξεκινούν με τον χαρακτήρα '@' ή έχουν γραφεί από 'Unverified' χρήστη. Σε αυτά τα δεδομένα, οι 15 επικρατέστεροι bloggers φαίνονται στην Εικόνα 7.3.1.

	author_name	MEIBI1	MEIBIX1	MEIBIO_aa	MEIBIXO_aa	MEIBIO_bb	MEIBIXO_bb
▶	Chris Ziegler	84	94	82	91	81	90
	Darren Murph	81	95	79	93	78	93
	Joshua Topolsky	83	93	75	86	75	86
	Nilay Patel	78	94	73	86	72	86
	Vladislav Savov	75	90	74	85	73	84
	Paul Miller	62	79	59	71	58	71
	Thomas Ricker	63	75	60	70	59	69
	Ross Miller	58	64	53	60	53	59
	Laura June	53	58	54	57	54	56
	Tim Stevens	46	55	44	52	44	51
	Donald Melanson	44	54	43	50	42	50
	Joseph L. Flatley	39	51	38	49	38	49
	Sean Hollister	35	35	35	35	35	35
	Ryan Block	36	39	30	34	30	34
	Joanna Stern	32	36	32	35	33	35

Εικόνα 7.3.1: Ταξινόμηση Bloggers στο $D1$
(MEIBI, MEIBIX, MEIBIO, MEIBIXO)

Τα MEIBI1 & MEIBIX1 είναι οι κλασικοί τελεστές. Οι MEIBIO_aa & MEIBIXO_aa, MEIBIO_bb & MEIBIXO_bb είναι οι Opinionated τελεστές που προκύπτουν από τα score_a και score_b αντίστοιχα.

Αρχικά να επισημάνουμε ότι η γενική μείωση των MEIBI & MEIBIXO σε σχέση με τα αποτελέσματα της Εικόνας 7.2.1 είναι αναμενόμενη και οφείλεται στη μείωση του πλήθους των comments.

Παρατηρούμε ότι οι περισσότεροι bloggers έχουν βαθμολογίες MEIBIO & MEIBIXO ελάχιστα πιο κάτω από αυτές των MEIBI & MEIBIX. Καθότι, όπως έχουμε επισημάνει στην 4.3, τα νέα S^{mo}, S^{xo} διαφέρουν από τα αντίστοιχα παλιά μόνο ως προς τον $C_i^j + 1$ παράγοντα, είμαστε σε θέση να γνωρίζουμε τις αιτίες για αυτή την συμπεριφορά των νέων τελεστών. Οι ελάχιστα μειωμένες τιμές τους οφείλονται στο ότι το πλήθος των σχολιαστών είναι μικρότερο από αυτό των σχολίων καθώς και στο ότι ορισμένοι αρθρογράφοι έχουν αρνητική μέση βαθμολογία. Για να γίνει κατανοητό, υπενθυμίζουμε το ποιόν του C_i^j στους παλιούς και στους Opinionated τελεστές:

MEIBI&MEIBIX: $C_i^j = \#comments_i^j$ (7.3.1)

MEIBIO&MEIBIXO: $C_i^j = (avg(d\_commentator\_score)_i^j + 1) \cdot \#d\_commentators_i^j$

Το πλήθος των distinct commentators ($\#d\_commentators_i^j$) του i post είναι μικρότερο ή ίσο του αριθμού των comments στο post αυτό ($\#comments_i^j$). Ίσο είναι στην περίπτωση που κάθε σχολιαστής έχει γράψει μόνο από ένα σχόλιο. Ο παράγοντας $avg(d\_commentator\_score)_i^j + 1$ από την μεριά του, έχει εύρος τιμών το διάστημα [0, 2]: Μικρότερο της μονάδας για αρνητικές μέσες βαθμολογίες και μεγαλύτερο για θετικές. Συνοψίζουμε:

$$\left.\begin{array}{l} 0 \le \#d\_commentators_i^j \le \#comments_i^j \\ -1 \le avg(d\_commentator\_score)_i^j \le 1 \Leftrightarrow 0 \le avg(d\_commentator\_score)_i^j + 1 \le 2 \end{array}\right\}$$

$$\Rightarrow 0 \le (avg(d\_commentator\_score)_i^j + 1) \cdot \#d\_commentators_i^j \le 2 \cdot \#comments_i^j$$

δηλαδή: $0 \le C_{i-NEW}^j \le 2 \cdot C_{i-OLD}^j$ (7.3.2)

Δείξαμε ότι το νέο C_i^j θεωρητικά έχει το διπλάσιο εύρος τιμών από το προηγούμενο. Στην πράξη, όμως, για να πιάσει πολύ υψηλές ή χαμηλές τιμές, χρειάζεται σχόλια με σημαντικά μεγάλες πολικότητες. Συγκεκριμένα χρειάζεται τα σχόλια που έχουν αναρτηθεί σε όλα τα άρθρα ενός blogger να είναι ή πολύ θετικά ή πολύ αρνητικά στο σύνολό τους, κάτι το οποίο γενικά είναι σπάνιο.

Αντιθέτως, τα δεδομένα από την τεχνολογικού ενδιαφέροντος ιστοσελίδα engadget, παρουσιάζουν μια σχετική ουδετερότητα. Ακολουθούν αναλυτικότερα οι Μέσοι Όροι του πλήθους των comments, των distinct commentators, καθώς επίσης οι Μέσοι Όροι των score_a, score_b και των διαφόρων S^m, S^x για τα posts των 15 αρθρογράφων της Εικόνας 7.3.1.

	author_name	avg_comments	avg_d_commentators	avg_score_a	avg_score_b	avg_Sm1	avg_Sx1	avg_Smo_aa	avg_Sxo_aa	avg_Smo_bb	avg_Sxo_bb
▶	Chris Ziegler	32.3901	27.8077	0.0149	-0.0049	16.9230	21.0831	15.5991	19.3126	15.2979	18.9474
	Darren Murph	24.9020	22.1405	0.0197	0.0025	3.4984	4.9351	3.2125	4.4364	3.1701	4.3732
	Joshua Topolsky	53.2032	46.0676	0.0210	0.0031	29.0576	40.7647	25.1736	35.2415	24.8499	34.7188
	Nilay Patel	28.6978	23.1278	0.0146	-0.0052	19.6450	27.9085	17.8489	25.1776	17.5027	24.6745
	Vladislav Savov	27.3008	24.1070	0.0244	0.0100	27.2813	36.4539	26.0555	34.2832	25.6893	33.7840
	Paul Miller	18.5680	15.4632	0.0191	-0.0009	9.2590	11.4470	8.6579	10.5696	8.5759	10.4600
	Thomas Ricker	15.3687	12.5821	0.0191	-0.0012	5.3517	7.5356	4.7974	6.7048	4.7060	6.5684
	Ross Miller	29.0751	23.6679	0.0187	0.0013	11.5933	14.6296	10.3171	12.9710	10.1266	12.7119
	Laura June	268.3783	262.1877	0.0190	0.0032	24.0990	32.8242	24.3959	30.3649	24.2496	30.0826
	Tim Stevens	26.2653	22.7935	0.0227	0.0054	8.4712	11.1517	8.0110	10.4554	7.9191	10.3271
	Donald Melanson	13.8297	11.4582	0.0196	0.0019	2.5305	3.5869	2.3416	3.2356	2.3079	3.1914
	Joseph L. Flatley	26.0666	21.8232	0.0205	0.0047	8.0033	10.6059	7.5150	9.8094	7.4092	9.6817
	Sean Hollister	20.9057	20.0943	0.0306	0.0214	75.0888	77.4341	73.9876	76.2062	73.2563	75.4196
	Joanna Stern	23.6090	22.6992	0.0389	0.0275	32.7120	41.8241	32.5026	41.5025	32.0345	40.8827
	Ryan Block	19.8706	18.2867	0.0252	0.0051	0.9886	1.3937	0.8837	1.2034	0.8739	1.1864

Εικόνα 7.3.2: Μέσοι Όροι στοιχείων των 15 bloggers της Εικόνας 7.3.1[6]

Τα λεγόμενά μας επαληθεύονται από τα απεικονιζόμενα δεδομένα. Σε κάθε post, οι σχολιαστές είναι λίγο λιγότεροι από τα σχόλια, οι απόψεις συμψηφίζονται για να προκύψει μία άχρωμη ουδετερότητα και οι τιμές των S^m, S^x μειώνονται ελαφρώς. Βέβαια, η έλλειψη πόλωσης που εξάγεται από το Υποσύστημα Εξόρυξης Γνώμης μπορεί να θεωρηθεί αναμενόμενη λόγω της φύσης των δεδομένων. Οι "bloggers" του engadget απλώς παρουσιάζουν προϊόντα τεχνολογίας. Οι σχολιαστές άλλες φορές αρέσκονται σε αυτά, άλλες όχι. Όταν στο τέλος συμψηφίζονται οι γνώμες πάνω σε όλα τα άρθρα ενός blogger, για να υπερισχύσει η θετική ή η αρνητική πλευρά θα έπρεπε είτε να τύχαινε να παρουσίαζε συνέχεια μόνο κακά -ή μόνο καλά- προϊόντα είτε απλώς να μην ήξερε την δουλειά του οπότε όλες οι κριτικές του να τύγχαναν κάκιστων σχολίων -ή το ανάποδο- (το γεγονός ότι δεν υπάρχει subjectivity detection [14] είναι ένα άλλο πρόβλημα). Είναι εύλογο ότι με τις παρούσες συνθήκες δεν θα μπορούσε να προκύψει συνολική πόλωση απόψεων.

Υπάρχουν, ωστόσο, άρθρα τα οποία λαμβάνουν σχόλια με μια συνέπεια στην πολικότητά τους. Παρατίθενται μερικά από αυτά προς επίδειξην των αλλαγών στα S^m, S^x, οι διαφορές των οποίων υπολογίζονται στις αντίστοιχες στήλες.

[6] Η εναλλαγή των δύο τελευταίων θέσεων μεταξύ Joanna Stern και Ryan Block δεν οφείλεται σε λάθος. Απλώς τυχαίνει να έχουν ίσο άθροισμα τελεστών.

author_name	post_id	comments	d_commentators	score_b	Sm1	Smo_bb	Smo_bb-Sm1	Sx1	Sxo_bb	Sxo_bb-Sx1
Ross Miller	9568	18	18	-0.2149	26.7407	21.2968	-5.4440	38.1816	30.4084	-7.7732
Ross Miller	10223	30	29	-0.1742	124.0000	99.7934	-24.2066	126.2143	101.5755	-24.6388
Thomas Ricker	10363	67	61	-0.2057	11.4526	8.3287	-3.1239	11.4862	8.3531	-3.1331
Ryan Block	11970	121	113	0.3079	15.0519	18.3580	3.3061	15.1922	18.5291	3.3369
Chris Ziegler	12247	26	24	-0.1886	30.0984	22.8243	-7.2741	30.3701	23.0304	-7.3398
Ryan Block	16096	1465	1456	0.2579	111.5434	139.4312	27.8878	111.9457	139.9341	27.9884
Ryan Block	16123	1958	1944	0.2979	180.7490	232.8974	52.1485	181.5177	233.8879	52.3703
Paul Miller	26511	30	28	0.2628	21.6279	25.3658	3.7379	21.7127	25.4652	3.7525

Εικόνα 7.3.3: Παραδείγματα posts με υπολογίσιμη πολικότητα σχολίων

Επιλέχθηκαν να απεικονιστούν μόνο τα score_b διότι με αυτά (δεύτερη προσέγγιση στην βαθμολόγηση λέξεων, §5.2) δημιουργείται πάντα μεγαλύτερη πόλωση από τα score_a.

7.4 Αξιολόγηση Υποσυστήματος Εξόρυξης Γνώμης

Οι παραπάνω αξιολογήσεις έγιναν με την βασική προϋπόθεση ότι το Opinion Mining Υποσύστημα αποφέρει αξιοπρεπή αποτελέσματα. Προς επιβεβαίωση αυτής, εφαρμόζουμε το αναπτυχθέν Υποσύστημα σε κάποια δεδομένα ελέγχου (test data) [25]. Στα ίδια δεδομένα, εφαρμόζουμε επίσης τρεις -απλές- μεθόδους ταξινόμησης εγγράφων (γίνεται ταξινόμηση -classification- σε θετικά και αρνητικά χωρίς υπολογισμό score) οι οποίες περιγράφονται στα [11], [12]. Σύμφωνα με αυτές, το κριτήριο με βάση το οποίο γίνεται το classification είναι το πλήθος των θετικών/αρνητικών λέξεων, τα αθροίσματά τους και οι Μέσοι Όροι τους.

Τα test data που χρησιμοποιήθηκαν αποτελούνται από 25000 έγγραφα κειμένου, τα μισά εκ των οποίων έχουν θετική χροιά και τα άλλα μισά αρνητική. Συγκεκριμένα, είναι κριτικές για ταινίες (η πιο δημοφιλής επιλογή στα δεδομένα ελέγχου για εξόρυξη γνώμης) και διατίθενται από το Πανεπιστήμιο του Stanford[7].

Ακολουθούν τα αποτελέσματα των μεθόδων ταξινόμησης καθώς και του δικού μας Υποσυστήματος Εξόρυξης Γνώμης.

[7] http://ai.stanford.edu/~amaas/data/sentiment

Term Counting	Positive	Negative
Predicted Positive	9560	6682
Predicted Negative	1382	4005
Predicted Zero	1558	1813
Total	12500	12500
Recall	76,48%	32,04%

Εικόνα 7.4.1: Ταξινόμηση με Term Counting

Sum	Positive	Negative
Predicted Positive	8359	4062
Predicted Negative	3082	7518
Predicted Zero	1059	920
Total	12500	12500
Recall	66,87%	60,14%

Εικόνα 7.4.2: Ταξινόμηση με βάση το Άθροισμα (Sum) των βαθμολογιών των λέξεων

Avg	Positive	Negative
Predicted Positive	3648	1356
Predicted Negative	7792	10227
Predicted Zero	1060	917
Total	12500	12500
Recall	29,18%	81,82%

Εικόνα 7.4.3: Ταξινόμηση με βάση τον Μ.Ο. (Avg) των βαθμολογιών των λέξεων

Ours (score_a)	Positive	Negative
Predicted Positive	9264	4616
Predicted Negative	3200	7828
Predicted Zero	36	56
Total	12500	12500
Recall	74,11%	62,62%

Εικόνα 7.4.4: Ταξινόμηση με το δικό μας Σύστημα με score_a

Ours (score_b)	Positive	Negative
Predicted Positive	7702	3219
Predicted Negative	4677	9158
Predicted Zero	121	123
Total	12500	12500
Recall	61,61%	73,26%

Εικόνα 7.4.5: Ταξινόμηση με το δικό μας Σύστημα με score_b

	Accuracy
Term Counting	54,26%
Sum	63,51%
Avg	55,50%
Ours (score_a)	68,37%
Ours (score_b)	67,44%

Εικόνα 7.4.6: Σύγκριση της Accuracy των μεθόδων

Σε όλες τις παραπάνω μεθόδους η βαθμολογία των λέξεων προκύπτει όπως έχει περιγραφεί στην ενότητα 5.2. Η αξιολόγησή τους γίνεται με τα θεμελιώδη μεγέθη του IR Recall και Accuracy. Σε κάθε πείραμα αξιολογούνται τα θετικά ή τα αρνητικά δεδομένα με την εκάστοτε μέθοδο. Λόγω αυτής της χειραγώγησης των δεδομένων, το Precision είναι πάντοτε 1.

Παρατηρούμε ότι το μεγαλύτερο Recall στην ανάκτηση των θετικών εγγράφων επιτυγχάνεται με την μέθοδο του Term Counting (76,48%). Η συγκεκριμένη - απλούστατη- μέθοδος, όμως, ταυτόχρονα παρουσιάζει το χαμηλότερο ποσοστό ανάκτησης αρνητικών εγγράφων, μόλις 32,04%, καθώς προκύπτει υπερβολικά μεγάλος αριθμός false positives. Αυτή η αδυναμία εντοπισμού αρνητικών εγγράφων και η θεώρηση πολλών εξ αυτών ως θετικά οφείλεται στο ότι οι λέξεις με θετική χροιά είναι περισσότερες από αυτές με αρνητική ακόμα και στα αρνητικά δεδομένα. Προς απόδειξη αυτού, παραθέτονται τα ποσοστά των θετικών, ουδέτερων και αρνητικών όρων στα δύο σύνολα δεδομένων.

POSITIVE DATA	
pos terms	0,2334
neutral terms	0,6252
neg terms	0,1414

NEGATIVE DATA	
pos terms	0,2032
neutral terms	0,6228
neg terms	0,1740

Εικόνα 7.4.7

Είναι, επομένως, εύλογο το συμπέρασμα ότι το Term Counting δεν είναι άξιο εμπιστοσύνης.

Ανάλογες αδυναμίες παρουσιάζουν και οι μέθοδοι του Αθροίσματος αλλά και του Μέσου Όρου των score των λέξεων. Η πρώτη επιτυγχάνει μέτρια επίπεδα ανάκτησης θετικών (66,82%) - αρνητικών (60,14%) εγγράφων κατατασσόμενη στην τρίτη θέση και στα δύο, ενώ η δεύτερη έχει κάκιστη απόδοση στα θετικά (τελευταία με 29,18%) και την καλύτερη όλων στα αρνητικά δεδομένα (81,82%).

Αντιθέτως, οι δύο μέθοδοι του Υποσυστήματος που αναπτύξαμε παρουσιάζουν ικανοποιητικά αποτελέσματα και -το κυριότερο- συνέπεια. Δεν υπάρχει αναντιστοιχία στα ποσοστά ανάκτησης θετικών - αρνητικών κειμένων, γεγονός που αποδεικνύεται και από τις τιμές Accuracy (68,37% και 67,44%) οι οποίες είναι οι υψηλότερες.

Ως εκ τούτου, αποδείχθηκε εμπράκτως η ικανοποιητική επίδοση του Opinion Mining Υποσυστήματος στα παραχθέντα αποτελέσματα του οποίου στηρίζεται το Υποσύστημα Ταξινόμησης Bloggers.

8 Επίλογος

8.1 Σύνοψη

Στην παρούσα διπλωματική επεκτείναμε τους MEIBI & MEIBIX τελεστές με opinion mining των comments, ώστε στον υπολογισμό του rank να μην παίζει ρόλο το πλήθος των comments αλλά η συνολική opinion που έχει διαμορφωθεί από όλα τα comments.

Για να υλοποιηθεί αυτή η ιδέα, φτιάξαμε ένα Opinion Mining Σύστημα το οποίο βαθμολογεί την διάθεση (θετική/αρνητική) των σχολίων. Χρησιμοποιήσαμε το λεξικό SentiWordNet το οποίο αποδίδει πολικότητα στις λέξεις.

8.2 Μελλοντικές επεκτάσεις

Το θέμα της συγκεκριμένης διπλωματικής μπορεί να επεκταθεί ως προς πολλούς τομείς. Αρχικά, είναι ενδιαφέρον να ασχοληθούμε με την εκμετάλλευσή των opinionated τελεστών MEIBIO & MEIBIXO πέρα από την απλή ταξινόμηση των bloggers, στον ευρύτερο τομέα του Information Retrieval, όπως γίνεται στο [3].

Επιπλέον, κρίνεται σκόπιμη η εφαρμογή του αναπτυχθέντος Συστήματος σε κάποιο άλλο dataset που να περιέχει κείμενα όπου εκφράζονται γνώμες με πιο ευθύ τρόπο, όπως για παράδειγμα blogs με κοινωνικοπολιτικές ή ακόμα και αθλητικές συζητήσεις. Η λογική με την οποία αναπτύξαμε τους MEIBIO & MEIBIXO ταιριάζει περισσότερο σε ιστολόγια αυτής της φύσης και είναι σίγουρο ότι σε αυτά θα αναδειχθεί η αξία και η χρησιμότητά τους έναντι των απλών MEIBI & MEIBIX. Αξίζει να σημειωθεί ότι σε τέτοιου είδους blogs η γνώμη των σχολιαστών έχει μεγαλύτερη βαρύτητα. Ως εκ τούτου, ενδέχεται να χρειαστεί να κάνουμε μία μικρή

παρέμβαση στην σχέση 4.3.1 επιτρέποντας στο C_j^i να παίρνει αρνητικές τιμές (άρα και στα $S_j^{mo}(i)$ & $S_j^{xo}(i)$, οπότε και στους ΜΕΙΒΙΟ & ΜΕΙΒΙΧΟ).

Όσο αφορά το Opinion Mining Υποσύστημα, θα ήταν χρήσιμη η σύγκρισή του με κάποια άλλα περισσότερο εξελιγμένα μοντέλα από αυτά στην ενότητα 7.4 και εάν κριθεί απαραίτητο να βελτιωθεί.

Τέλος, ένας πιο μακρινός στόχος είναι η ενασχόληση με ranking σε κοινωνικά δίκτυα και κυρίως στο twitter, το οποίο είναι το πλέον επικρατέστερο μέσο έκφρασης απόψεων με άμεσο, σύντομο τρόπο. Η σωστή αξιολόγηση των influentials στον κόσμο του twitter είναι μια πρόκληση.

9 Βιβλιογραφία

[1]	Akritidis, L., Katsaros, D., Bozanis, P.: Identifying influential bloggers: time does matter. In: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conferences onWeb Intelligence and Intelligent Agent Technologies (WI-IAT'09), vol. 1, pp. 76–83 (2009)
[2]	Akritidis, L., Katsaros, D., Bozanis, P.: Identifying the productive and influential bloggers in a community. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **41**(5), 759–764 (2011)
[3]	Akritidis L., Bozanis P.. Improving opinionated blog retrieval effectiveness with quality measures and temporal features. Springer Science + Business Media New York 2013
[4]	Esuli A, Sebastiani F.. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Proceedings from International Conference on Language Resources and Evaluation (LREC), Genoa, 2006.
[5]	Dray G., Plantie M., Harb A., Poncelet P., Roche M., Trousset F.: Opinion Mining From Blogs. International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM) ISSN: 2150-7988 Vol.1 (2009), pp.205-213
[6]	N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In Proceedings of ACM WSDM Conf., pages 207–218, 2008
[7]	Wikipedia. The Hirsch h-index, Jan. 2009. Available from http://en.wikipedia.org/wiki/H-index
[8]	E. Keller and J. Berry. One American in ten tells the other nine how to vote, where to eat and, what to buy. They are The Influentials. The Free Press, 2003
[9]	Liu B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, May 2012
[10]	Manning C. D., Raghavan P., Schutze H.: An Introduction to Information Retrieval, Cambridge University Press, April 2009
[11]	Ohana B., Tierney B.: Sentiment Classification of Reviews Using SentiWordNet. IT&T Conference, October 2009
[12]	Hamouda A., Rohaim M.: Reviews Classification Using SentiWordNet Lexicon. The Online Journal on Computer Science and Information

	Technology (OJCSIT) Vol. (2) - No. (1) (2010)
[13]	G. Gezici, B. Yanikoglu, D. Tapucu and Yucel Saygın. New Features for Sentiment Analysis: Do Sentences Matter? In proceedings of Sentiment Discovery from Affective Data (SDAD 2012)
[14]	B. Pang, L. Lee: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1-135
[15]	WordNet: A Lexical Database for English. Princeton University. Available at http://wordnet.princeton.edu
[16]	Chapman W, Bridewell W, Hanbury P, Cooper G, Buchanan B. (2001). Evaluation of Negation Phrases in Narrative Clinical Report. Proceedings of 2001 AMIA Symposium, 105-109
[17]	Chalothorn T., Ellman J.: Using SentiWordNet and Sentiment Analysis for Detecting Content on Web Forums. Sixth International Conference on Software Knowledge Information Management and Applications (SKIMA 2012), Chengdu University, China, September, 2012
[18]	Asmi A., Ishaya T.: Negation Identification and Calculation in Sentiment Analysis. IMMM 2012: The Second International Conference on Advances in Information Mining and Management, 2012
[19]	Kim J.-H., Yoon T.-B., Kim K.-S., Lee J.-H.: Trackback-Rank: An effective Ranking Algorithm for the Blog Search. Proccedings of the 2008 Second International Symposium on Intelligent Information Technology Application - Volume 3 -P.503-507, 2008
[20]	Li Y.M., Lai C.-Y., Chen C.-W.: Identifying Bloggers with Marketing Influence in the Blogosphere. Proceedings of the 11th International Conference on Electronic Commerce, P.335-340, 2009
[21]	Kritikopoulos, A., Sideri, M. and Varlamis, I. 2006. BlogRank: ranking weblogs based on connectivity and similarity features. Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications
[22]	Wikipedia PageRank [http://en.wikipedia.org/wiki/PageRank]
[23]	Wikipedia HITS [http://en.wikipedia.org/wiki/HITS_algorithm]
[24]	Bross J., Richly K., Kohnen M., Meinel C.: Identifying the top-dogs of the blogosphere. Social Network Analysis and Mining, Springer-Verlag 2011
[25]	Maas A.L., Daly R.E., Pham P.T., Huang D., Ng Y.A., Potts C.Q Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).