ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ. " ΨΗΦΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ "

ΚΑΤΕΥΘΥΝΣΗ: ΔΙΚΤΥΟΚΕΝΤΡΙΚΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ

Ο ΜΕΤΑΦΟΡΙΚΟΣ ΛΟΓΟΣ ΣΤΟ TWITTER

Ονοματεπώνυμο: Καρανάσου Μαρία

Α.Μ.: ΜΕ12048

Επιβλέπων: Δουλκερίδης Χρήστος

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τους κο Χ. Δουλκερίδη και κα Μ. Χαλκίδη για την καθοδήγηση, τις ιδέες και το χρόνο που αφιέρωσαν και τον κο Μ. Θεμιστοκλέους ως εξεταστή και αξιολογητή του αποτελέσματος. Επίσης, θα ήθελα να ευχαριστήσω την οικογένεια μου. Χωρίς τη βοήθεια και την συμπαράσταση τους, ηθική και πρακτική, δεν θα ήταν δυνατόν να ολοκληρωθεί το εγχείρημα. Τέλος, ευχαριστώ όσους με στήριξαν καθ' όλη τη διάρκεια της έρευνας, υλοποίησης και συγγραφής του κειμένου αυτού.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΞΙΣΩΣΕΩΝ

ΠΡΟΛΟΓΟΣ

Σε μια εποχή που ο όγκος των δεδομένων που συλλέγονται καθημερινά είναι τεράστιος και ολοένα αυξανόμενος, δημιουργείται η ανάγκη κατανόησης και της εξαγωγής νοήματος από τα φαινομενικά ασύνδετα δεδομένα. Η παρούσα εργασία ασχολείται με τον τομέα της ανάλυσης συναισθήματος, ο οποίος έχει ως στόχο την εξαγωγή συμπερασμάτων σχετικά με τις απόψεις που επικοινωνούνται μέσω κειμένου. Πιο συγκεκριμένα, γίνεται έρευνα στον επιστημονικό χώρο της εξαγωγής συναισθήματος από κοινωνικά δίκτυα και ειδικότερα στον μεταφορικό λόγο στο Twitter.

1. ΕΙΣΑΓΩΓΗ

Στην εποχή των μεγάλων δεδομένων (Big Data), είναι επιτακτική η ανάγκη εξαγωγής πληροφορίας και γνώσης από φαινομενικά ασύνδετα δεδομένα. Πιο συγκεκριμένα, η ανάγκη γνώσης της γνώμης του χρήστη σχετικά με κάποιο γεγονός ή προϊόν έχει γίνει πολύ σημαντική για διάφορους τομείς, όπως για παράδειγμα το marketing.

Υπάρχει πλέον η ανάγκη για ανάλυση και εξαγωγή πληροφορίας από τα κοινωνικά δίκτυα, καθότι παρέχεται τεράστιος όγκος πληροφορίας που μπορεί να χρησιμοποιηθεί αναλόγως. Οι λόγοι πίσω από αυτή την ανάγκη κυμαίνονται από προώθηση προϊόντων και marketing, πρόβλεψη αποτελεσμάτων, έως αποτροπή εγκλημάτων. Σημαντική φαίνεται να είναι και η ανάγκη για πληροφόρηση σχετικά με τις πολιτικές απόψεις που κυριαρχούν. [1]

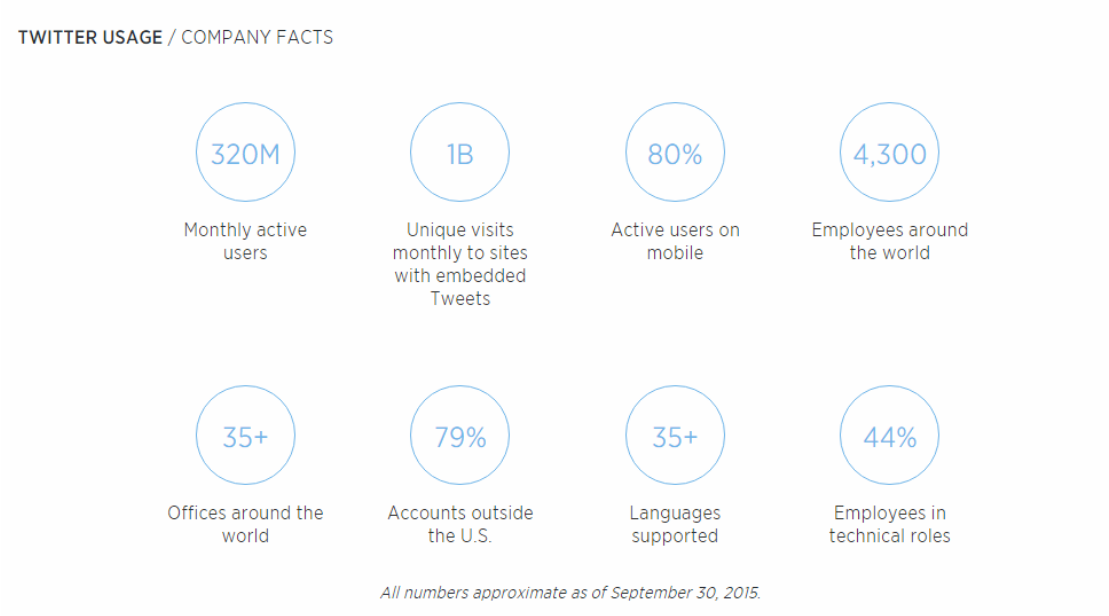
Τα τελευταία χρόνια, το κοινωνικό δίκτυο Twitter γίνεται όλο και πιο δημοφιλές, με περίπου 350 εκατομμύρια ενεργούς χρήστες μηνιαίως [2]. Κάθε δευτερόλεπτο δημιουργούνται 6000 tweets κατά μέσο όρο, κάτι το οποίο αντιστοιχεί σε πάνω από 350000 το λεπτό, 500 εκατομμύρια τη μέρα και 200 δισεκατομμύρια το χρόνο[1].

Καθώς βασικό στοιχείο της ανθρώπινης συμπεριφοράς είναι η συλλογή γνώσης, το τι σκέφτονται οι άλλοι άνθρωποι είναι σημαντικό μέρος αυτής της προσπάθειας. Έχοντας ως βοήθεια την ολοένα αυξανόμενη διαθεσιμότητα και δημοτικότητα πόρων που περιέχουν γνώμες, όπως αξιολογήσεις και κριτικές στο διαδίκτυο, προσωπικά ιστολόγια (blogs), κριτικές στα κοινωνικά δίκτυα, προκύπτουν νέες προκλήσεις και ευκαιρίες για εξόρυξη απόψεων ώστε να γίνει πιο κατανοητή η γνώμη των συνανθρώπων μας. Τεχνολογικά, τέτοιες προσπάθειες υποστηρίζονται από τους τομείς της εξόρυξης γνώμης (opinion mining) και της ανάλυσης συναισθήματος (sentiment analysis), οι οποίοι χειρίζονται την πολύπλοκη δουλειά της μοντελοποίησης τέτοιου είδους γνώσης ώστε να είναι υπολογιστικά κατανοητή.

Η σημασία φαίνεται στους ακόλουθους αριθμούς:

- Το 81% των χρηστών του διαδικτύου έχουν κάνει έρευνα αγοράς για ένα προϊόν τουλάχιστον μια φορά.
- Εκ των οποίων, το 20% κάνουν έρευνα αγοράς σε καθημερινή βάση.
- Το 73% - 87% των αναγνωστών κριτικών για ξενοδοχεία, ιατρούς, ταξιδιωτικά πρακτορεία κλπ. αναφέρουν πως οι κριτικές αυτές έπαιξαν μεγάλο ρόλο στην αγοραστική απόφασή τους.
- Καταναλωτές αναφέρουν ότι είναι πρόθυμοι να πληρώσουν από 20% έως και 99% περισσότερο για ένα προϊόν με την καλύτερη κριτική, σε σχέση με ένα προϊόν που έχει λάβει πολύ καλή αλλά λιγότερο καλή κριτική. [1]

[1] http://www.internetlivestats.com/twitter-statistics/



Σχήμα 1. Στατιστικά χρήσης του Twitter έως και τον Σεπτέμβριο του 2015 [2]

1.1 ΔΟΜΗ ΕΡΓΑΣΙΑΣ

Αρχικά, ορίζεται το πρόβλημα που καλείται η παρούσα εργασία να επιλύσει. Στη συνέχεια παρουσιάζεται μια βιβλιογραφική ανασκόπηση σχετικά με τις θεματικές ενότητες που εμπλέκονται στον ορισμό αλλά και την επίλυση του. Ακολουθεί η μοντελοποίηση του προβλήματος με όλες τις λεπτομέρειες που παίζουν ρόλο στον χειρισμό του, η περιγραφή του συστήματος που αναπτύχθηκε στα πλαίσια της συμμετοχής στο SemEval 2015 Task 11, η περιγραφή της βιβλιοθήκης που αναπτύχθηκε στα πλαίσια της εργασίας για την επεξεργασία των tweets και την εξαγωγή των σχετικών χαρακτηριστικών και η περιγραφή της δικτυακής εφαρμογής για την οπτικοποίηση των αποτελεσμάτων. Τέλος, αναλύονται τα πειράματα που έγιναν καθώς και τα αντίστοιχα αποτελέσματα, παρουσιάζονται τα αποτελέσματα του SemEval 2015 Task 11 και περιγράφονται τα συμπεράσματα και όποια μελλοντική εργασία και βελτιώσεις.

1.2 ΕΙΣΑΓΩΓΙΚΟΙ ΟΡΙΣΜΟΙ

BIG DATA

Με τον όρο Big Data περιγράφονται τα σύνολα δεδομένων που είναι τόσο μεγάλα και περίπλοκα που οι παραδοσιακοί τρόποι και εφαρμογές επεξεργασίας δεν επαρκούν για την ανάλυση και τον χειρισμό τους. Οι προκλήσεις συμπεριλαμβάνουν την ανάλυση, τη σύλληψη, επιμέλεια δεδομένων, αναζήτηση, κοινή χρήση, αποθήκευση, μεταφορά, απεικόνιση, και το ζήτημα της ιδιωτικότητας της πληροφορίας. Ο όρος αυτός αναφέρεται συνήθως στην χρήση predictive analytics ή άλλων εξελιγμένων δεδομένων εξαγωγής πληροφορίας και αξίας από δεδομένα και σπάνια στα δεδομένα καθ' αυτά [3]. Η ακρίβεια στα Big Data, μπορεί να

οδηγήσει στη λήψη πιο σωστών αποφάσεων και κατ' επέκταση σε καλύτερης ποιότητας αποφάσεις οδηγούν σε καλύτερη λειτουργική απόδοση, μείωση του κόστους και του ρίσκου [4] [5].

ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ - SOCIAL MEDIA

Ως Κοινωνικά Δίκτυα (Social Media) χαρακτηρίζονται σε γενικές γραμμές τα πολλά και σχετικά φθηνά και ευρέως προσβάσιμα ηλεκτρονικά εργαλεία που δίνουν την δυνατότητα σε οποιονδήποτε να δημοσιεύσει πληροφορίες, να έχει πρόσβαση σε πληροφορίες, να συνεργαστεί για ένα κοινό σκοπό ή να δημιουργήσει σχέσεις. Σε αυτή την κατηγορία βρίσκεται και το Twitter, το οποίο δίνει στους χρήστες του τη δυνατότητα να δημοσιεύσουν μικρού μήκους μηνύματα 140 χαρακτήρων [6].

SEMEVAL

Το SemEval (Σημασιολογική Αξιολόγηση - Semantic Evaluation) είναι μια σειρά από αξιολογήσεις υπολογιστικών συστημάτων σημασιολογικής ανάλυσης. Αρχικά είχε ξεκινήσει ως SenseEval, δηλαδή ως σειρά εννοιολογικής αξιολόγησης των λέξεων (Word sense evaluation series). Οι αξιολογήσεις έχουν ως στόχο την εξερεύνηση της φύσης της έννοιας στην γλώσσα. Ενώ έννοια και το νόημα είναι κάτι το ενστικτώδες για τους ανθρώπους, η μεταφορά της κατανόησης αυτής των εννοιών στην υπολογιστική ανάλυση έχει αποδειχτεί πολύ δύσκολο έως και απίθανο εγχείρημα. Το SemEval με τις αξιολογήσεις του προσφέρει ένα μηχανισμό για τον πιο ακριβή χαρακτηρισμό του τι ακριβώς είναι απαραίτητο να υπολογιστεί στα πλαίσια του νοήματος, συνεπώς θεσπίζει ασκήσεις που σκοπό έχουν την εξερεύνηση όσο το δυνατόν περισσότερων διαστάσεων και παραμέτρων που συσχετίζονται με τη γλώσσα και τη χρήση της. Αρχικά, στις τρεις πρώτες αξιολογήσεις (Senseval 1 έως 3), οι ασκήσεις αυτές περιορίζονταν στην προσπάθεια της αίσθησης των λέξεων (word senses) με υπολογιστικό τρόπο. Εξελίχθηκαν όμως, από το πρώτο SemEval το 2007, σε ασκήσεις διερεύνησης των συσχετίσεων μεταξύ των στοιχείων μιας πρότασης, των σχέσεων μεταξύ των προτάσεων και της φύσις του τι λέγεται (σημασιολογικές σχέσεις και συναισθηματική ανάλυση. [7] [8]

1.3 ΟΡΙΣΜΟΣ ΠΡΟΒΛΗΜΑΤΟΣ

Στην παρούσα εργασία, γίνεται έρευνα στον επιστημονικό χώρο της εξαγωγής συναισθήματος από κοινωνικά δίκτυα και ειδικότερα στον μεταφορικό λόγο στο Twitter, τις δυσκολίες που προκύπτουν και από την ανάλυση συναισθήματος αλλά και από τις ιδιαιτερότητες του μεταφορικού λόγου.

Η ανάλυση συναισθήματος σε μεταφορικό λόγο είναι δύσκολο εγχείρημα, το οποίο γίνεται ακόμα πιο δύσκολο όταν η αναγνώριση συναισθήματος αφορά κείμενο μικρού μήκους προερχόμενο από κοινωνικά δίκτυα, όπως το Tweet, το οποίο περιορίζεται σε 140 χαρακτήρες. Για παράδειγμα, ένα tweet μπορεί να είναι πλούσιο σε ειρωνεία, η οποία δηλώνεται είτε μέσω hashtags πχ. #irony, είτε εμμέσως, π.χ. "*I love working when I'm sick*". Ο προσδιορισμός του πραγματικού συναισθήματος του κειμένου αυτού αποτελεί πρόκληση, ιδιαιτέρως λόγω του

περιορισμένου μεγέθους του αλλά και χαρακτηριστικών όπως συντομογραφίες και αργκό (slang).

Συνεπώς, το να χαρακτηριστεί θετικό, αρνητικό ή ουδέτερο ένα κείμενο είναι αρκετά δύσκολο έργο. Το πραγματικό νόημα μπορεί να είναι πολύ διαφορετικό από αυτό που φαίνεται να δηλώνεται, καθώς, για παράδειγμα, στον ειρωνικό λόγο, αυτό που εκφράζεται μπορεί να είναι το εντελώς αντίθετο από αυτό που υπονοείται, π.χ. *“Oh, you don't like sarcasm? You must be so funny to hang around with.”*

Επιπλέον δυσκολία αποτελεί η αναγνώριση της έντασης του συναισθήματος, καθώς όσο περισσότερες κατηγορίες τόσο πιο στοχευμένα θα πρέπει να είναι τα χαρακτηριστικά που θα χρησιμοποιηθούν στην ανάλυση συναισθήματος ώστε να δοθεί η σωστή βαρύτητα στο αντίστοιχο tweet.

Για την αντιμετώπιση αυτής της πρόκλησης, προτείνεται ένα σύστημα για την ανάλυση συναισθήματος σε μεταφορικό λόγο, το οποίο βασίζεται στην επιλογή χαρακτηριστικών ενός tweet και εκπαιδεύει έναν αλγόριθμο ώστε να μπορεί να κάνει πρόβλεψη σχετικά με το συναίσθημα ενός tweet.

Το πρόβλημα μπορεί να συνοψιστεί ως εξής:

Δεδομένου ενός συνόλου από tweets τα οποία είναι πλούσια σε μεταφορά και ειρωνεία, στόχος είναι να καθοριστεί αν ο χρήστης έχει εκφράσει θετικό, αρνητικό ή ουδέτερο συναίσθημα, καθώς και ο βαθμός ο οποίος εκφράζει το πόσο έντονο είναι αυτό το συναίσθημα.

2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΗΣΗ

2.1 ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ – DATA MINING

ΔΕΔΟΜΕΝΑ

Τα δεδομένα είναι οποιαδήποτε γεγονότα, αριθμοί ή κείμενο που είναι δυνατόν να επεξεργαστεί ένας υπολογιστής. Στη σημερινή εποχή γίνεται συσσώρευση τεράστιων και αυξανόμενων ποσοτήτων δεδομένων σε διάφορες μορφές και διαφορετικές βάσεις δεδομένων, όπως για παράδειγμα διαδικαστικά δεδομένα ήτοι πωλήσεις, κόστος, μισθοδοσία και λογιστικές διαδικασίες, μη διαδικαστικά δεδομένα, δηλαδή οι πωλήσεις ενός τομέα, δεδομένα σχετικά με προβλέψεις και μακροοικονομικά δεδομένα και μετα-δεδομένα, δεδομένα που αφορούν και περιγράφουν δεδομένα, όπως η λογική σχεδίασης μιας βάσης ή δομές δεδομένων [3].

ΠΛΗΡΟΦΟΡΙΑ

Τα μοτίβα, οι συσχετίσεις και οι σχέσεις μεταξύ των δεδομένων οι οποίες μπορούν να φανερώσουν σημαντικά χρήσιμα στοιχεία, για παράδειγμα, η ανάλυση των δεδομένων των συναλλαγών ενός καταστήματος λιανικής πώλησης μπορεί να αποδώσει σημαντικά στοιχεία σχετικά με το ποια προϊόντα έχουν καλές πωλήσεις, ποιο το αγοραστικό κοινό και ποιο το χρονικό υπόβαθρο [3].

ΓΝΩΣΗ

Η πληροφορία μπορεί να μετατραπεί σε γνώση σχετικά με παρελθόντα μοτίβα και μελλοντικές τάσεις. Παραδείγματος χάριν, η συνοπτική πληροφορία σχετικά με πωλήσεις καταστημάτων μπορεί να αναλυθεί ώστε να βοηθήσει προσπάθειες προώθησης των πωλήσεων δίνοντας γνώση σχετικά με την αγοραστική συμπεριφορά των καταναλωτών. Η πληροφορία αυτή θα μπορούσε καταλήξει στην προώθηση συγκεκριμένων προϊόντων τα οποία φαίνεται να έχουν κάποια προοπτική [3].

ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ (DATA WAREHOUSING)

Τα τελευταία χρόνια έχουν γίνει δραματικές αλλαγές στην απόκτηση, μεταφορά και αποθήκευση δεδομένων, στις ικανότητες της επεξεργαστικής ισχύος, Το Data warehousing ορίζεται ως μια διαδικασία κεντρικής διαχείρισης και ανάκτησης δεδομένων και αντιπροσωπεύει την ιδέα της διατήρησης ενός κεντρικού repository όλων των δεδομένων ενός οργανισμού [3].

2.1.1 ΟΡΙΣΜΟΣ

Ως Εξόρυξη Δεδομένων ορίζεται η εξαγωγή χρήσιμων μοτίβων ή μοντέλων από μεγάλο όγκο δεδομένων ή η αυτοματοποιημένη ανάλυση μεγάλου όγκου δεδομένων [4]. Είναι η υπολογιστική διαδικασία ανακάλυψης μοτίβων σε μεγάλα σύνολα δεδομένων (Big Data) που χρησιμοποιεί μεθόδους τεχνητής νοημοσύνης, μηχανικής μάθησης, στατιστικής, και τα

συστήματα βάσεων δεδομένων και αποτελεί ένα διεπιστημονικό τομέα της επιστήμης των υπολογιστών [9] [10].

Σε γενικές γραμμές, η εξόρυξη δεδομένων, η οποία κάποιες φορές ονομάζεται ανακάλυψη γνώσης, είναι η διαδικασία ανάλυσης δεδομένων από διαφορετικές προοπτικές και η σύνοψή της σε χρήσιμη πληροφορία, σε πληροφορία δηλαδή που μπορεί να χρησιμοποιηθεί για να αυξήσει τις αποδόσεις, να μειώσει το κόστος ή και τα δύο. Ο τελικός στόχος της διαδικασίας εξόρυξης δεδομένων είναι η εξαγωγή πληροφοριών από ένα σύνολο δεδομένων και η μετατροπή της σε μια κατανοητή δομή για περαιτέρω χρήση [10]. Το λογισμικό που ασχολείται με το Data Mining είναι ένα από τα εργαλεία που χρησιμοποιούνται για την ανάλυση δεδομένων και δίνει τη δυνατότητα στο χρήστη του να εξάγει συμπεράσματα από πολλές διαφορετικές οπτικές, να τα κατηγοριοποιεί και να αναγνωρίζει τις σχέσεις που ανακαλύπτονται. Πρακτικά, ένας τρόπος να περιγραφεί η διαδικασία του Data Mining, είναι ως η διαδικασία ανακάλυψης συσχετισμών ή μοτίβων μεταξύ πολλών πεδίων σε μεγάλες σχεσιακές βάσεις δεδομένων.

Εκτός από την ανάλυση, στην εξόρυξη περιλαμβάνονται διαδικασίες διαχείρισης, προ-επεξεργασία και μοντελοποίηση δεδομένων, μετρικές ενδιαφέροντος, θέματα πολυπλοκότητας και συμπερασμάτων, μετα-επεξεργασία δεδομένων και δομών, οπτικοποίηση αποτελεσμάτων και διαδικασίας και ενημέρωση σε πραγματικό χρόνο.

Οι χρήσεις και οι εφαρμογές του Data Mining είναι σε εταιρίες και οργανισμούς που έχουν άμεσο στόχο τον καταναλωτή, την οικονομία, τις επικοινωνίες, την προώθηση αγαθών (marketing), το εμπόριο λιανικής κ.α. . Σε αυτούς τους οργανισμούς δίνεται η δυνατότητα, μέσω του data mining, να κατανοήσουν και να αξιοποιήσουν σχέσεις μεταξύ παραδείγματος χάριν τιμής και θέσης προϊόντος, δημογραφικών στοιχείων κλπ. και τι αντίκτυπο έχουν αυτές στην αγορά και τις πωλήσεις. [3] [11]

2.1.2 ΔΙΑΔΙΚΑΣΙΑ

Η διαδικασία εντοπισμού γνώσης (Knowledge Discovery in Databases (KDD)) αποτελείται σε γενικές γραμμές, αλλά και με πολλές παραλλαγές, από τα ακόλουθα στάδια:

1. Επιλογή δεδομένων (Selection)
   Πριν τη χρήση αλγόριθμων εξόρυξης γνώσης, πρέπει να συγκεντρωθεί ένα σύνολο δεδομένων. Η εξόρυξη γνώσης μπορεί να ανακαλύψει μοτίβα που προϋπάρχουν στα δεδομένα, οπότε το σύνολο των δεδομένων αυτών θα πρέπει να είναι αρκετά μεγάλο ώστε να περιέχει τα μοτίβα αυτά και ταυτόχρονα να είναι αρκετά ξεκάθαρο ώστε να είναι δυνατόν να επιτευχθεί εξόρυξη μέσα σε ένα αποδεκτό χρονικό διάστημα. Παράδειγμα πηγής δεδομένων προς εξόρυξη είναι η αποθήκη δεδομένων (data warehouse)

2. Προ-επεξεργασία (Pre-processing)
   Η προεπεξεργασία είναι απαραίτητη για την ανάλυση ποικίλων σετ δεδομένων πριν την εξόρυξη. Τα δεδομένα αυτά καθαρίζονται, αφαιρώντας τις παρατηρήσεις που περιέχουν θόρυβο και αυτές με ελλιπή στοιχεία.

3. Μετατροπή (Transformation)
   Μετατροπή των δεδομένων σε μορφή κατάλληλη ώστε να είναι δυνατόν να εφαρμοστούν αλγόριθμοι εξόρυξης γνώσης.

4. Εξόρυξη Γνώσης (Data Mining)
   Το Data mining συμπεριλαμβάνει τις ακόλουθες έξι κατηγορίες σχετικών διαδικασιών [3] [11]:

   Anomaly detection (Outlier/change/deviation detection)
   Η αναγνώριση μη συνηθισμένων δεδομένων, τα οποία θα μπορούσαν είτε να αποτελέσουν ενδιαφέρουσα πηγή γνώσης και ένδειξη πληροφορίας, είτε λάθη που επιβάλουν περαιτέρω ανάλυση.

   Association rule learning (Dependency modelling)
   Η αναζήτηση συσχετίσεων μεταξύ μεταβλητών, π.χ. των προϊόντων ενός εμπορικού καταστήματος και των τιμών τους, τα προϊόντα που αγοράζονται συχνά μαζί, οι αγορές σε σχέση με το χρονικό διάστημα στο οποίο πραγματοποιούνται κ.α.

   Clustering
   Η διαδικασία ανακάλυψης ομάδων και δομών σε δεδομένα που είναι κατά κάποιο τρόπο όμοια, χωρίς να υπάρχει κάποια γνώση για τις δομές τους και τις κατηγορίες στις οποίες εμπίπτουν.

   Classification
   Η διαδικασία γενίκευσης μιας γνωστής δομής και η εφαρμογή της σε νέα δεδομένα.

   Regression
   Η προσπάθεια ανακάλυψης μιας εξίσωσης που μοντελοποιεί τα δεδομένα με το ελάχιστο λάθος.

   Summarization
   Η προσπάθεια απόδοσης του νοήματος ή της παρουσίασης μιας πιο συμπυκνωμένης μορφής ενός συνόλου δεδομένων, συμπεριλαμβανομένου της οπτικοποίησης και της δημιουργίας αναφορών.

5. Αξιολόγηση/ Ερμηνεία (Evaluation/ Interpretation)
   Παρουσίαση των δεδομένων σε κατανοητή μορφή, όπως ένας πίνακας ή ένας γράφος.

   Η Εξόρυξη Δεδομένων από Data Warehouses αποτελείται από πέντε στοιχεία:
   - Εξαγωγή, μετατροπή, και φόρτωση όλων των δεδομένων που αφορούν διαδικασίες στο σύστημα αποθήκης δεδομένων.
   - Αποθήκευση και διαχείριση των δεδομένων σε πολυδιάστατο σύστημα βάσης δεδομένων (multidimensional database system)
   - Παροχή πρόσβασης σε επιχειρηματικούς αναλυτές και τους επαγγελματίες της τεχνολογίας των πληροφοριών.

- Ανάλυση των δεδομένων μέσω λογισμικού
- Παρουσίαση των δεδομένων σε κατανοητή μορφή, όπως ένας πίνακας ή ένας γράφος.

Ως απλοποίηση των παραπάνω βημάτων, μπορούμε να θεωρήσουμε τα ακόλουθα γενικά βήματα: προ-επεξεργασία, εξόρυξη γνώσης, επικύρωση αποτελεσμάτων.

2.2 ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΟΥ - TEXT ANALYTICS

Τα αδόμητα δεδομένα, μεγάλο μέρος των οποίων είναι τα δεδομένα κειμένου, αποτελούν μια από τις τρεις κύριες πηγές της έκρηξης δεδομένων που έχει συμβεί την τελευταία δεκαετία. Σχεδόν όλες οι επικοινωνίες είναι πλέον ψηφιακές, από email ως δεδομένα κοινωνικών δικτύων όπως τα tweets και ιστολόγια (blogs). Ακόμα και όταν η επικοινωνία γίνεται μέσω τηλεφώνου, μπορεί να μετατραπεί σε κείμενο ώστε να υποστεί περαιτέρω επεξεργασία. Τα δεδομένα κειμένου, ασχέτως από που προέρχονται, παρουσιάζουν προκλήσεις στην επεξεργασία και μετατροπή τους από ακατέργαστη (raw) μορφή σε μορφή κατάλληλη για μοντελοποίηση και εξαγωγή συμπερασμάτων.

2.2.1 ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ - INFORMATION RETRIEVAL (IR)

Η ανάκτηση πληροφορίας (Information Retrieval) είναι απαραίτητο βήμα σε κάθε διεργασία ανάλυσης κειμένου. Διάφορες διαδικασίες αποτελούν κοινά βήματα για την ανάκτηση πληροφορίας [12]:

Αναζήτηση σε αρχεία (File crawling): Είναι η διαδικασία επεξεργασίας αρχείων με σκοπό την εξαγωγή χρήσιμων πληροφοριών για μετέπειτα χρήση.

Αναζήτηση στον παγκόσμιο ιστό (Web crawling): Είναι παρόμοια διαδικασία με το file crawling, με τη διαφορά ότι λαμβάνει χώρα στο διαδίκτυο. Ένα παράδειγμα λογισμικού που εκτελεί τέτοιες διαδικασίες είναι τα "bots", τα οποία εκτελούν αυτοματοποιημένα επαναλαμβανόμενα βήματα για την ανάκτηση (scraping) κειμένου από ιστοσελίδες.

Εξαγωγή κειμένου (Text extraction): Η εξαγωγή κειμένου είναι συνήθως το βήμα που ακολουθεί μια διαδικασία web ή file crawling και συμπεριλαμβάνει τον διαχωρισμό του κειμένου από τη μορφοποίηση ενός αρχείου ή μιας ιστοσελίδας. Παραδείγματος χάριν, τέτοιου είδους διαδικασία μπορεί να ειναι η εξαγωγή του κειμένου και ο διαχωρισμός του από το format ενός Portable Document Format (PDF). Στο παράδειγμα αυτό, αφαιρείται η μορφοποίηση, το μέγεθος και το είδος των γραμμάτων, οπότε το τελικό προϊόν είναι το απλό κείμενο.

Ευρετήριο και αναζήτηση (Index and search): Οι υπηρεσίες ανάκτησης πληροφορίας καλούνται να χτίσουν ευρετήρια που μπορούν να χρησιμοποιηθούν σε αποτελεσματική αναζήτηση. Το 2008, η Google στο επίσημο blog της ανέφερε ότι έχει πιάσει το milestone της αναγνώρισης ενός τρισεκατομμυρίου μοναδικών urls σε μια δεδομένη στιγμή [13].Με καλής ποιότητας ευρετήρια, οι χρήστες έχουν τη δυνατότητα να κάνουν αναζήτηση όχι μόνο με μια έννοια, μια λέξη, αλλά και με πιο σύνθετους τρόπους, όπως για παράδειγμα αναζήτηση με Boolean λογική.

2.2.2 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΠΕΡΙΕΧΟΜΕΝΟΥ - CONTENT CATEGORIZATION

Η Κατηγοριοποίηση Περιεχομένου είναι η διαδικασία αξιολόγησης εγγράφων και της πρότασης ή της ανάθεσης τους σε κατηγορίες με βάση το περιεχόμενό τους. Αυτό μπορεί να επιτευχθεί με διάφορες μεθόδους, η ποιο διαδεδομένη εκ των οποίων είναι η bag-of-words μέθοδος. Οι λέξεις ενός κειμένου τοποθετούνται σε ένα σύνολο (bag) και ανακτώνται σε ομάδες. Η πιθανότητα των λέξεων να ανακτηθούν μαζί βοηθάει στην κατηγοριοποίηση του κειμένου σε κατηγορίες με βάση τη συνύπαρξη των λέξεων. Παραδείγματος χάριν οι λέξεις "άντρας", "γυναίκα", "άνθρωπος" είναι πιθανόν να βρεθούν μαζί σε κείμενα σχετικά με τον κλάδο της ψυχολογίας αλλά μπορεί να βρεθούν μαζί και σε κείμενα που αφορούν τη θρησκεία.

Υπάρχουν δύο κύριες μέθοδοι σχετικές με την κατηγοριοποίηση κειμένου, η αναγνώριση θέματος (text topic identification) και η συσταδοποίηση κειμένων (text clustering). Καθώς όλο και περισσότερα έγγραφα κατηγοριοποιούνται και αξιολογούνται ως παρόμοια, τα θέματα που περιέχονται σε αυτά προκύπτουν οργανικά, χωρίς προηγούμενους κανόνες ή κατευθύνσεις.

Σε κάθε έγγραφο δίνεται μια βαθμολογία για κάθε θέμα. Στη συνέχεια μια τιμή η οποία παίζει το ρόλο του ανώτατου ορίου, δηλαδή ένα κατώφλι (cutoff) εφαρμόζεται ώστε να αναγνωριστούν τα έγγραφα που ανήκουν σε μια συγκεκριμένη κατηγορία. Ένα κείμενο/ έγγραφο μπορεί να ανήκει σε παραπάνω από μια κατηγορίες.

2.2.3 ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ - TEXT MINING

Η Εξόρυξη κειμένου είναι ένας κλάδος που συνδυάζει την εξόρυξη δεδομένων και Text Analytics για τη χρήση μη δομημένων δεδομένων κειμένου ή μαζί με τα δομημένα δεδομένα για τους σκοπούς της εξερεύνησης, ανακάλυψης, και προγνωστικής μοντελοποίησης ή κατάταξης. Σε ένα μεγάλο σύνολο δεδομένων κειμένου, είναι επιθυμητό να γίνει εξαγωγή των ιδεών που το διέπουν, της κεντρικής ιδέας δηλαδή. Η δυνατότητα να φιλτράρονται και να αναλύονται μεγάλοι όγκοι δεδομένων, είναι κάτι που απαιτεί την ικανότητα επιτυχούς ανάλυσης και το κείμενο του φίλτρου προκειμένου να αποκαλύψει το πιο ουσιαστικό και σημαντικό περιεχόμενο που περιλαμβάνεται σε αυτό. Ένα από τα σημαντικότερα εμπόδια στη χρήση αδόμητων δεδομένων είναι η σύνοψή τους σε μια μορφή που μπορεί να χρησιμοποιηθεί για την ανακάλυψη του θέματος και για προγνωστική μοντελοποίηση. Όπως και οι περισσότερες διαδικασίες εξόρυξης δεδομένων, τα δεδομένα κειμένου απαιτούν πρόσθετη προετοιμασία για να είναι αποτελεσματική η ανάλυσή τους. [12]

2.3 ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ - NATURAL LANGUAGE PROCESSING

2.3.1 ΟΡΙΣΜΟΣ

Με τον όρο "φυσική γλώσσα", αναφερόμαστε σε μια γλώσσα που χρησιμοποιείται καθημερινά για την ανθρώπινη επικοινωνία, παραδείγματος χάριν, τα ελληνικά, τα αγγλικά, τα πορτογαλικά κ.π.α. Σε αντίθεση με τις τεχνητές γλώσσες, όπως προγραμματιστικές γλώσσες και μαθηματική σημασιολογία, οι φυσικές γλώσσες έχουν εξελιχθεί περνώντας από γενιά σε γενιά και είναι δύσκολο να εντοπιστούν οι ακριβείς κανόνες.

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing ή NLP) είναι μια ευρεία έννοια που καλύπτει οποιαδήποτε επεξεργασία ή χειρισμός μέσω η/υ της φυσικής γλώσσας. Από το πιο απλό, όπως το άθροισμα των συχνοτήτων των λέξεων για τον διαχωρισμό διαφορετικών στυλ συγγραφής, έως το πολύ περίπλοκο, όπως είναι το να γίνουν "κατανοητές" ανθρώπινες εκφράσεις, τουλάχιστον στο βαθμό που να είναι δυνατόν να δοθεί κάποια λογική απάντηση σε αυτές [14].

2.3.2 ΤΕΧΝΙΚΕΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

Κυρίαρχες τεχνικές στο NLP είναι οι sequence labeling, n-gram models, backoff, και evaluation, οι οποίες είναι χρήσιμες σε πολλούς τομείς [15]. Ακολούθως περιγράφονται οι πιο σχετικές με το πρόβλημα της εργασίας.

TEXT WRANGLING

Με τον όρο Text Wrangling εννοείται όλη η προεπεξεργασία που γίνεται στο κείμενο πριν καταλήξει να είναι κατάλληλη είσοδος για την εφαρμογή υπολογιστικών μεθόδων και αλγορίθμων, δηλαδή η προεπεξεργασία που απαιτείται ώστε το κείμενο να είναι κατανοητό από μηχανή (machine readable). Αυτό περιλαμβάνει καθαρισμό κειμένου, συγκεκριμένη προεπεξεργασία, tokenization, stemming, lemmatization και αφαίρεση των stop words [15].

ΚΑΤΑΚΕΡΜΑΤΙΣΗ – TOKENIZATION

Οποιοδήποτε κομμάτι κειμένου δεν μπορεί να επεξεργαστεί χωρίς να περάσει από τη διαδικασία της κατακερμάτισης (tokenization). Tokenization είναι η διαδικασία διαχωρισμού του κειμένου σε λογικά κομμάτια, όπως για παράδειγμα σε λέξεις. Η πολυπλοκότητα του tokenization ποικίλει αναλόγως τις ανάγκες της NLP εφαρμογής και την πολυπλοκότητα της γλώσσας. Παραδείγματος χάριν, ένα κομμάτι κειμένου στην αγγλική γλώσσα θα μπορούσε να κατακερματιστεί διαλέγοντας μόνο τις λέξεις και τους αριθμούς. Εάν το κείμενο περιέχει Κινέζικους ή Ιαπωνικούς χαρακτήρες, τότε το tokenization αποτελεί πολύ πολύπλοκη διαδικασία [15].

ΑΦΑΙΡΕΣΗ ΛΕΞΕΩΝ ΧΩΡΙΣ ΑΞΙΑ – STOP WORD REMOVAL

Η αφαίρεση των λέξεων που δεν προσδίδουν κάποια αξία στο NLP, δηλαδή το Stop word removal είναι ένα από τα πιο συνηθισμένα βήματα προ-επεξεργασίας κειμένου. Η ιδέα είναι να αφαιρεθούν όλες οι κοινές λέξεις, οι οποίες απαντώνται συχνά σε όλα τα έγγραφα μιας συλλογής (corpus). Συνήθως τα άρθρα και οι αντωνυμίες κατατάσσονται ως stop words. Οι λέξεις αυτές δεν έχουν καμία αξία σε κάποιες NLP εφαρμογές, κάτι που σημαίνει ότι οι λέξεις αυτές δεν είναι ιδιαίτερα σημαντικές. Υπάρχουν και περιπτώσεις βέβαια στις οποίες οι stop words παίζουν πάρα πολύ μικρό ρόλο και η επιρροή τους σε μια NLP εφαρμογή είναι αμελητέα. Τις περισσότερες φορές, η λίστα με τις stop words έχει δημιουργηθεί με το χέρι και αποτελείται από λέξεις που εμφανίζονται συχνά σε μια συλλογή κειμένων. Εκτός από τις ήδη υπάρχουσες και έτοιμες λίστες από stop words, ένας απλός τρόπος για να παραχθεί μια τέτοια λίστα είναι με βάση της συχνότητα μιας λέξης στα κείμενα που εξετάζονται, όπου εάν η λέξη

είναι παρούσα σε όλα τα κείμενα τότε μπορεί να χαρακτηριστεί ως stop word.Έχει γίνει αρκετή έρευνα σχετικά με τη βέλτιστη λύση στο θέμα και κάποιες βιβλιοθήκες, όπως η NLTK, παρέχουν πρόσβαση σε λίστες με 2,400 stopwords για 11 γλώσσες [15] .

PART-OF-SPEECH TAGGING

Η διαδικασία κατά την οποία γίνεται κατηγοριοποίηση των λέξεων σε μέρη του λόγου και επισημαίνονται κατάλληλα ονομάζεται part-of-speech tagging, POS-tagging, ή απλά tagging, ή πιο απλά η επισήμανση ενός κειμένου με τα αντίστοιχα μέρη του λόγου που το απαρτίζουν. Τα μέρη του λόγου (Parts of speech) είναι γνωστά επίσης και ως κατηγορίες λέξεων (word classes) ή λεκτικές κατηγορίες (lexical categories). Το σύνολο των tags που χρησιμοποιείται για τη συγκεκριμένη διεργασία είναι γνωστό ώς tagset. Στόχος του NLP είναι η ξιοποίηση/ εκμετάλλευση των tags αλλά και το αυτόματο tagging [14].

Ο ακόλουθος πίνακας δείχνει τα Part-of-Speech tags όπως αυτά έχουν οριστεί από το Penn Treebank Project [16].

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

Πίνακας 1. Penn Tree Part-of-Speech tags [16] [17]

ΑΝΑΓΝΩΡΙΣΗ ΕΠΩΝΥΜΩΝ ΟΝΤΟΤΗΤΩΝ – NAMED ENTITY RECOGNITION (NER)

Εκτός από το Part-of-Speech tagging, ένα από τα ποιο κοινά προβλήματα κατηγοριοποίησης (labeling) είναι η αναγνώριση οντοτήτων σε κείμενο. Συνήθως, το NER αποτελούν όνόματα, τοποθεσίες και οργανισμοί. Υπάρχουν NER συστήματα που μπορούν να αναγνωρίσουν πολλές περισσότερες οντότητες από τις τρεις που αναφέρθηκαν. Υπάρχουν αρκετές έρευνες που εργάζονται σε αυτό το πεδίο του NLP, όπου γίνεται προσπάθεια να αναγνωριστούν βιοϊατρικές οντοτητες, προϊόντα κ.λ.π. [15].

ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΟΜΟΙΟΤΗΤΑ - SEMANTIC SIMILARITY

Η σημασιολογική ομοιότητα (Semantic Similarity) είναι ένα κεντρικό θέμα των κλάδων της τεχνητής νοημοσύνης, της ψυχολογίας και των γνωστικών επιστημών εδώ και αρκετά χρόνια. Έχει χρησιμοποιηθεί πολύ στην επεξεργασία φυσικής γλώσσας, στην ανάκτηση πληροφορίας, στην αποσαφήνιση της έννοιας των λέξεων (word sense disambiguation), στον κατακερματισμό του κειμένου, σε ερωταπαντήσεις και συστήματα συστάσεων, στην εξαγωγή πληροφορίας κ.α. [18] [19].

Η γνώση της σημασιολογικής ομοιότητας μεταξύ λέξεων είναι χρήσιμη και στην κατηγοριοποίηση κειμένων έτσι ώστε η αναζήτηση για έναν γενικό όρο, όπως για παράδειγμα «όχημα» θα δώσει ως αποτέλεσμα έγγραφα που περιέχουν συγκεκριμένους όρους, όπως αυτοκίνητο κλπ. [14]

WORDNET

Το WordNet είναι το προϊόν ενός ερευνητικού έργου στο Princeton University [20]. Είναι μια μεγάλη βάση δεδομένων της αγγλικής γλώσσας. Τα ουσιαστικά, ρήματα, επιρρήματα και τα επίθετα στο wordnet οργανώνονται μέσω διάφορων σημασιολογικών συσχετίσεων σε ομάδες συνωνύμων (synsets), τα οποία αντιπροσωπεύουν μια έννοια (concept). Παραδείγματα τέτοιου είδους σχέσεων είναι η συνωνυμία, ομαδοποίηση, ομοιότητα κλπ. Κάποιες από τις σχέσεις αφορούν τη μορφή των λέξεων και άλλες την σημασία. Χρησιμοποιώντας αυτές τις σχέσεις, δημιουργείται μια ιεραρχία λέξεων, η οποία αποτελεί πολύ σημαντικό εργαλείο στον τομέα της επεξεργασίας φυσικής γλώσσας και της υπολογιστικής γλωσσολογίας (computational linguistics) [18].

Το WordNet επιφανειακά μοιάζει με λεξικό, καθώς ομαδοποιεί τις λέξεις με βάση το νόημά τους. Παρόλα αυτά, υπάρχουν κάποιοι σημαντικοί διαχωρισμοί. Πρώτον, το WordNet συσχετίζει όχι μόνο μορφές λέξεων – π.χ. γράμματα – αλλά και έννοιες. Σαν αποτέλεσμα, οι λέξεις που βρίσκονται σε «κοντινή» περιοχή σε μορφή δικτύου είναι σημασιολογικά αποσαφηνισμένες (semantically disambiguated), το νόημά τους δηλαδή είναι ένα και είναι ξεκάθαρο. Δεύτερον, επισημαίνονται οι σημασιολογικές σχέσεις μεταξύ των λέξεων, ενώ η ομαδοποίηση των λέξεων σε λεξικό δεν ακολουθεί κάποιο μοτίβο παρά μόνο της ομοιότητας μεταξύ εννοιών (meaning similarity). Πολύ σημαντικό επίσης είναι ότι η πλειοψηφία των σχέσεων ενώνει λέξεις που αποτελούν το ίδιο μέρος του λόγου, κάτι που σημαίνει ότι στο WordNet υπάρχουν τέσσερις μεγάλες κατηγορίες, τέσσερα μεγάλα υποδίκτυα, των

ουσιαστικών (nouns), των ρημάτων (verbs), των επιθέτων (adjectives) και των επιρρημάτων (adverbs) [21].

Αποτελεί πολύ σημαντική πηγή μετρικών σημασιολογικής ομοιότητας και έχει δώσει πολύ καλά αποτελέσματα τα τελευταία χρόνια, με πολλούς και διαφορετικούς αλγόριθμους μέτρησης του semantic similarity. Η χρήση των συνόλων από συνώνυμα (synsets), όπως αυτά παρέχονται από το WordNet, μπορεί να βοηθήσει στη δημιουργία ευρετηρίων για κείμενα, όπως περιγράφηκε πιο πάνω. Σε γενικές γραμμές, το semantic similarity μπορεί να χωριστεί σε τέσσερις κατηγορίες: μετρικές βασιζόμενες στο μήκος διαδρομής (path length) σε σχέση με την ιεραρχία, μετρικές βασιζόμενες στο πληροφοριακό υπόβαθρο (information content), μετρικές βασιζόμενες σε χαρακτηριστικά (features) και υβριδικές μετρικές. Ο Πίνακας 2 συνοψίζει τις βασικές μετρικές που υπάρχουν στο WordNet και είναι προσβάσιμες από την βιβλιοθήκη NLTK (βλ. κεφάλαιο 2.3) [18].

category	Principle	measure	features	advantages	disadvantages
Path based	function of path length linking the concepts and the position of the concepts in the taxonomy	Shortest path	count of edges between concepts	simple	two pairs with equal lengths of shortest path will have the same similarity
		W&P	path length to subsumer, scaled by subsumer path to root	simple	two pairs with the same lso and equal lengths of shortest path will have the same similarity
		L&C	count of edges between and log smoothing	simple	two pairs with equal lengths of shortest path will have the same similarity
		Li	non-linear function of the shortest path and depth of lso	simple	two pairs with the same lso and equal lengths of shortest path will have the same similarity
IC based	The more common information two concepts share, the more similar the concepts are.	Resnik	IC of lso	simple	two pairs with the same lso will have the same similarity
		Lin	IC of lso and the compared concepts	take the IC of compared concepts into considerate	two pairs with the same summation of $IC(c_1)$ and $IC(c_2)$ will have the same similarity
		Jiang	IC of lso and the compared concepts	take the IC of compared concepts into considerate	two pairs with the same summation of $IC(c_1)$ and $IC(c_2)$ will have the same similarity
Feature based	Concepts with more common features and less non-common features are more similar	Tversky	compare concepts' feature, such as their definitions or glosses	take concept's feature into considerate	Computational complexity. It can't works well when there is not a complete features set.
Hybrid method	combine multiple information sources	Zhou	combines IC and shortest path	well distinguished different concepts pairs	parameter to be settled, turning is required. If the parameter can't be turned well it may bring deviation.

Πίνακας 2. Σύνοψη των μέτρων Semantic Similarity του WordNet [18]

ΑΝΑΠΑΡΑΣΤΑΣΗ BAG-OF-WORDS

Το μοντέλο Bag-of-Words (BoW ή Bag-of-features) είναι μια απλοποιημένη αναπαράσταση που χρησιμοποιείται στο NLP και στην εξόρυξη πληροφορίας. Σε αυτό το μοντέλο, ένα κείμενο, όπως για παράδειγμα μια πρόταση ή ένα έγγραφο, αναπαρίσταται ως ένα σύνολο (bag) από τις λέξεις που το αποτελούν, αγνοώντας τη γραμματική και ακόμα και τη σειρά των λέξεων, κρατώντας όμως την πολλαπλότητα, δηλαδή το πόσες φορές μια λέξη υπάρχει μέσα στο κείμενο [15]. Το μοντέλο αυτό χρησιμοποιείται συχνά σε μεθόδους κατηγοριοποίησης εγγράφων, όπου η συχνότητα εμφάνισης κάθε λέξης χρησιμοποιείται ως χαρακτηριστικό (feature) για την εκπαίδευση ενός classifier. Μια από τις πρώτες φορές που χρησιμοποιήθηκε ο όρος “Bag-of Words» είναι στο άρθρο του Zellig Harris “Distributional Structure” (1954) [22].

ΕΦΑΡΜΟΓΕΣ

Το NLP εμπλέκεται σε πολλών ειδών εφαρμογές που απαιτούν συντακτική και σημασιολογική ανάλυση σε διάφορα επίπεδα, όπως για παράδειγμα σε εφαρμογές εξόρυξης πληροφορίας, μηχανικής μετάφρασης, συναισθηματικής ανάλυσης και απάντησης ερωτήσεων. Ακολούθως αναλύονται οι πιο σχετικές με το θέμα της παρούσας εργασίας.

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΕΙΜΕΝΟΥ - TEXT CLASSIFICATION

Η κατηγοριοποίηση κειμένου είναι μια πολύ ενδιαφέρουσα εφαρμογή των NLP κανόνων και μεθόδων. Είναι κάτι που συμβαίνει καθημερινά χωρίς να το συνειδητοποιούμε, για παράδειγμα, η χρήση των spam filters, η κατηγοριοποίηση των email σε σημαντικά, η συλλογή πληροφοριών που αποτελούν ειδήσεις κλπ. Όλες αυτές οι εφαρμογές χρησιμοποιούν text classification, το οποίο είναι ένα καλά ορισμένο και σχετικά λυμένο πρόβλημα το οποίο έχει εφαρμογή σε πολλούς τομείς. Συνήθως, οποιαδήποτε διαδικασία κατηγοριοποίησης κειμένου χρησιμοποιεί τις λέξεις του κειμένου ή κάποιον συνδυασμό τους. Παρόλο που είναι ένα τυπικό πρόβλημα μηχανικής μάθησης, πολλά από τα βήματα προεπεξεργασίας προέρχονται από το domain του NLP [15].

ΕΞΟΡΥΞΗ ΠΛΗΡΟΦΟΡΙΑΣ - INFORMATION EXTRATION (IE)

Εξόρυξη πληροφορίας είναι η διαδικασία εξαγωγής ουσιαστικής πληροφορίας από μη δομημένο κείμενο. Σε γενικές γραμμές, γίνεται συλλογή από μεγάλο αριθμό αδόμητων εγγράφων και παράγεται μια δομημένη ή ημι-δομημένη γνωστική βάση (Knowledge Base) η οποία μπορεί να χρησιμοποιηθεί ως βάση για τη δημιουργία κάποιας εφαρμογής. Ένα παράδειγμα είναι η δημιουργία μιας πολύ καλής οντολογίας χρησιμοποιώντας έναν τεράστιο όγκο κειμένων. Ένα έργο που κινείται στα πλαίσια του IE είναι η DBpedia, όπου όλα τα άρθρα της Wikipedia έχουν χρησιμοποιηθεί για να παράξουν μια οντολογία από artifacts τα οποία είναι άμεσα συσχετιζόμενα ή έχουν κάποιας μορφής σχέση μεταξύ τους.

Υπάρχουν δύο κύριοι τρόποι εξόρυξης πληροφορίας:

- Εξόρυξη με κανόνες: Η μέθοδος αυτή μπορεί να περιγραφεί ως η διαδικασία συμπλήρωσης προτύπων, έχοντας ως ιδέα το ακόλουθο σκεπτικό: η θέσπιση των περιπτώσεων με

προκαθορισμένη χρήση για τα αναμενόμενα αποτελέσματα και η προσπάθεια εξόρυξης του αδόμητου κειμένου ώστε να εφαρμόσει στο συγκεκριμένο πρότυπο.

- Εξόρυξη μέσω μηχανικής μάθησης: Η προσέγγιση αυτή συμπεριλαμβάνει μεθόδους που κατά κύριο λόγο βασίζονται σε NLP, όπως η δόμηση ενός parser ο οποίος είναι συγκεκριμένος ως προς το πρόβλημα προς επίλυση και κατάλληλος για την συγκεκριμένη Knowledge Base [14].

ΑΝΑΓΝΩΡΙΣΗ ΣΥΝΕΠΑΓΩΓΗΣ ΚΕΙΜΕΝΩΝ - RECOGNIZING TEXTUAL ENTAILMENT

Recognizing textual entailment (RTE) είναι η διαδικασία προσδιορισμού/εκτίμησης για το αν ένα συγκεκριμένο κομμάτι του κειμένου περιλαμβάνει ένα άλλο κείμενο που ονομάζεται "υπόθεση"[14].

2.3.3 ΠΕΡΙΟΡΙΣΜΟΙ

Παρόλο που έχει γίνει μεγάλη πρόοδος, τα συστήματα επεξεργασίας φυσικής γλώσσας που έχουν αναπτυχθεί για να καλύψουν πραγματικά προβλήματα ακόμα δεν είναι σε θέση να κάνουν common-sense reasoning ή να εξάγουν γνώση με έναν ομοιόμορφο τρόπο. Από την αρχή, ένας σημαντικός στόχος της έρευνας στον τομέα του NLP ήταν να γίνει πρόοδος στο δύσκολο εγχείρημα της δημιουργίας τεχνολογίας που «κατανοεί τη γλώσσα», χρησιμοποιώντας απλές αλλά ισχυρές τεχνικές αντί για απεριόριστη γνώση και δυνατότητες εκλογίκευσης (reasoning) [14].

2.4 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ - MACHINE LEARNING (ML)

2.4.1 ΟΡΙΣΜΟΣ

Το πεδίο της μηχανικής μάθησης (ML) ασχολείται με το ζήτημα της δημιουργίας προγραμμάτων ηλεκτρονικών υπολογιστών που βελτιώνονται αυτόματα, αποκτώντας εμπειρία. Η Μηχανική Μάθηση βρίσκεται στη “διασταύρωση” της επιστήμης των υπολογιστών, της μηχανικής και της στατιστικής και αρκετές φορές σε άλλους τομείς [23].

Machine learning είναι ο επιστημονικός κλάδος που διερευνά την δημιουργία και μελέτη των αλγοριθμων που μπορούν να εκπαιδευτούν από δεδομένα. Τέτοιοι αλγόριθμοι λειτουργούν μέσω της δημιουργίας ενός μοντέλου από τυχαιες εισόδους (example inputs) χρησιμοποιώντας το για προβλέψεις ή αποφάσεις, αντί να ακολουθήσουν αυστηρές στατικές οδηγίες. Η μηχανική μάθηση είναι στενά συνδεδεμένη και συχνά συγχέεται με την υπολογιστική στατιστική, μια επιστημονική αρχή η οποία επίσης ειδικεύεται στην πρόβλεψη αποφάσεων (prediction-making) [24].

Το ML αποτελεί μέρος του πεδίου της επιστήμης των υπολογιστών και προέρχεται από την έρευνα στον τομέα της τεχνητής νοημοσύνης. Ασχολείται με το ερώτημα του πως να κατασκευαστούν προγράμματα υπολογιστών που αυτομάτως βελτιώνονται αποκτώντας εμπειρία [25]. Κάποιες φορές συγχέεται με την εξόρυξη δεδομένων, παρόλο που αυτή εστιάζει περισσότερο στην διερευνητική ανάλυση δεδομένων. Χρησιμοποιεί στατιστική και

μαθηματική βελτιστοποίηση (mathematical optimization) και χρησιμοποιείται σε ένα ευρύ πεδίο υπολογιστικών διαδικασιών, όπου η σχεδίαση ενός προγράμματος με κανόνες είναι ανέφικτο. Στις χρήσεις του ML κατατάσσονται και το φιλτράρισμα της ανεπιθύμητης αλληλογραφίας και η οπτική αναγνώριση χαρακτήρων (OCR), οι μηχανές αναζήτησης και το computer vision [26].

2.4.2 ΚΑΤΗΓΟΡΙΕΣ

Υπάρχει πλέον ένας μεγάλος αριθμός αλγορίθμων μηχανικής μάθησης, ο οποίοι είναι δυνατόν να κατηγοριοποιηθούν με διάφορους τρόπους, λαμβάνοντας υπόψιν διαφορετικές οπτικές και παραμέτρους, όπως για παράδειγμα, με βάση το μέγεθος της προσπάθειας ή της επαγωγής που χρειάζεται να χρησιμοποιηθεί από τον αλγόριθμο μάθησης, με βάση την ανάδραση (feedback) κατά τη διάρκεια της μάθησης, με βάση τα υποκείμενα πρότυπα (underlying paradigms) και με βάση τον τρόπο αναπαράστασης των δεδομένων εισόδων [26].

Περιγράφονται ακολούθως οι πιο κύριες και επικρατούσες.

2.4.3 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΜΕ ΒΑΣΗ ΤΟΝ ΤΡΟΠΟ ΜΑΘΗΣΗΣ

Μια σημαντική διάκριση των μεθόδων μηχανικής μάθησης είναι από τον τρόπο με τον οποίο γίνεται η διαδικασία της μάθησης, δηλαδή, με επίβλεψη, χωρίς επίβλεψη κλπ..

SUPERVISED LEARNING

Επιβλεπόμενη μάθηση ή μάθηση με επίβλεψη (supervised learning), όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους σε γνωστές, επιθυμητές εξόδους (σύνολο εκπαίδευσης), με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο (σύνολο ελέγχου). Στον υπολογιστή δηλαδή παρουσιάζονται παραδείγματα εισόδων με τα αντίστοιχα επιθυμητά αποτελέσματα (κατηγορίες) και ο στόχος είναι να μάθει έναν γενικό κανόνα που αντιστοιχίζει τις εισόδους στις επιθυμητές εξόδους, καταγράφοντας δομικά στοιχεία των εισόδων αυτών [27].

REINFORCEMENT LEARNING

Το πρόγραμμα αλληλεπιδρά με ένα δυναμικό περιβάλλον μέσα στο οποίο πρέπει να επιτευχθεί συγκεκριμένος στόχος χωρίς να δοθούν συγκεκριμένες οδηγίες σχετικά με το εάν πλησιάζεται ο στόχος ή όχι. (Feldman) Ένα παράδειγμα είναι η εκμάθηση ενός παιχνιδιού έχοντας έναν αντίπαλο. Αυτός είναι ένας τρόπος/ μια μορφή της μηχανικής μάθησης όπου το 'μέλος' μπορεί να προγραμματιστεί από μια επιβράβευση/ανταμοιβή αλλά και τιμωρία, χωρίς να διευκρινίζεται πως θα επιτευχθεί ο στόχος (task) [15].

SEMI-SUPERVISED

Σε αυτή την κατηγορία, το training set είναι ελλιπές, δηλαδή κάποιες από τις εισόδους λείπουν. Η μεταγωγή (transduction) είναι μια ειδική περίπτωση της αρχής αυτής, όπου το σύνολο των περιπτώσεων (instances) του προβλήματος είναι γνωστό σε χρόνο εκμάθησης, αλλά λείπει μέρος των κατηγοριών (targets). Αυτή η κατηγορία μάθησης και των αντίστοιχων τεχνικών που χρησιμοποιούνται, χρησιμοποιεί μη κατηγοριοποιημένα (unlabeled) δεδομένα. Από το όνομα

είναι κατανοητό ότι βρίσκεται κάπου στο ενδιάμεσο από την επιβλεπόμενη και τη μη επιβλεπόμενη μάθηση, όπου χρησιμοποιείται μικρός όγκος κατηγοριοποιημένων (labeled) δεδομένων και μεγάλος όγκος μη κατηγοριοποιημένων δεδομένων για να χτιστεί ένα μοντέλο μηχανικής μάθησης [15].

UNSUPERVISED

Ανεπίβλεπτη μάθηση ή μάθηση χωρίς επίβλεψη (unsupervised learning), όπου ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων χωρίς να γνωρίζει επιθυμητές εξόδους για το σύνολο εκπαίδευσης [27].

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΜΕ ΒΑΣΗ ΤΟ ΑΠΟΤΕΛΕΣΜΑ

Μια ακόμη κατηγοριοποίηση των τεχνικών μηχανικής μάθησης προκύπτει όταν εξετάζεται το επιθυμητό αποτέλεσμα του συστήματος [15]. Οι κύριες κατηγορίες με βάση αυτό το κριτήριο είναι οι ακόλουθες:

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ – CLASSIFICATION

Στην κατηγοριοποίηση οι είσοδοι του συστήματος χωρίζονται σε δύο ή περισσότερες κλάσεις και ο αλγόριθμος εκμάθησης πρέπει να παράξει ένα μοντέλο το οποίο να είναι ικανό να κατηγοριοποιήσει άγνωστες (unseen) εισόδους σε μία ή περισσότερες κλάσεις (multi-label classification). Αυτό συνήθως αποτελεί πρόβλημα μάθησης με επίβλεψη (supervised learning). Παράδειγμα τέτοιας περίπτωσης είναι το spam filtering, όπου οι είσοδοι είναι η ηλεκτρονική αλληλογραφία ή άλλα μηνύματα και οι κλάσεις είναι «ανεπιθύμητο» και «μη ανεπιθύμητο». Ως γενική περιγραφή, μπορούμε να πούμε ότι η κατάταξη χρησιμοποιείται όταν χρειάζεται να προβλέψουμε εάν ένα δείγμα ανήκει σε μια (ή περισσότερες) από τις γνωστές κατηγορίες. Εάν το δείγμα μπορεί να ανήκει σε μια μόνο κατηγορία τότε έχουμε Binary Classification, εάν μπορεί να ανήκει σε περισσότερες από μια κατηγορίες έχουμε Multiclass Classification. Το classification που εφαρμόζεται σε κείμενο αφορά τον τομέα της κατηγοριοποίησης κειμένου, όπου η αποστολή του αλγορίθμου είναι η αυτόματη ταξινόμηση του κειμένου σε δεδομένες κατηγορίες. Αυτή η διαδικασία χρησιμοποιείται στην κατηγοριοποίηση ιστοσελίδων και στην ανάλυση συναισθήματος κλπ. [15]. Οι πιο δημοφιλείς classifiers στην κατηγορία αυτή είναι ο Naïve Bayes, ο K-nearest neighbor και ο Support Vector Machines [27].

ΠΑΛΙΝΔΡΟΜΗΣΗ – REGRESSION

Η παλινδρόμηση είναι ένα supervised πρόβλημα, μόνο που σε αυτή την περίπτωση οι είσοδοι δεν είναι διακριτές αλλά συνεχείς. Ένα παράδειγμα διακριτών τιμών είναι ο δείκτης των μετοχών ή η τιμή των ακινήτων. Παράδειγμα classifier που ανήκει σε αυτή την κατηγορία αποτελεί ο Stochastic Gradient Descent (SGD) [15].

ΣΥΣΤΑΔΟΠΟΙΗΣΗ – CLUSTERING

Στο clustering, ένα σύνολο εισόδων πρόκειται να χωριστεί σε ομάδες. Αντιθέτως με την περίπτωση της κατηγοριοποίησης, οι ομάδες δεν είναι γνωστές εκ των προτέρων, κάτι που τοποθετεί το clustering στους αλγόριθμους χωρίς επίβλεψη. Σε γενικές γραμμές, το clustering αντιμετωπίζει την ανακάλυψη μοντέλων μέσω της εύρεσης ομαδοποιημένων δεδομένων τα οποία ικανοποιούν ένα κριτήριο, ενώ ελαχιστοποιεί την ομοιότητα των δεδομένων τα οποία

δεν το ικανοποιούν και έτσι τοποθετούνται σε διαφορετικές ομάδες. Παράδειγμα classifier που ανήκει σε αυτή την κατηγορία αποτελεί ο K-means [27].

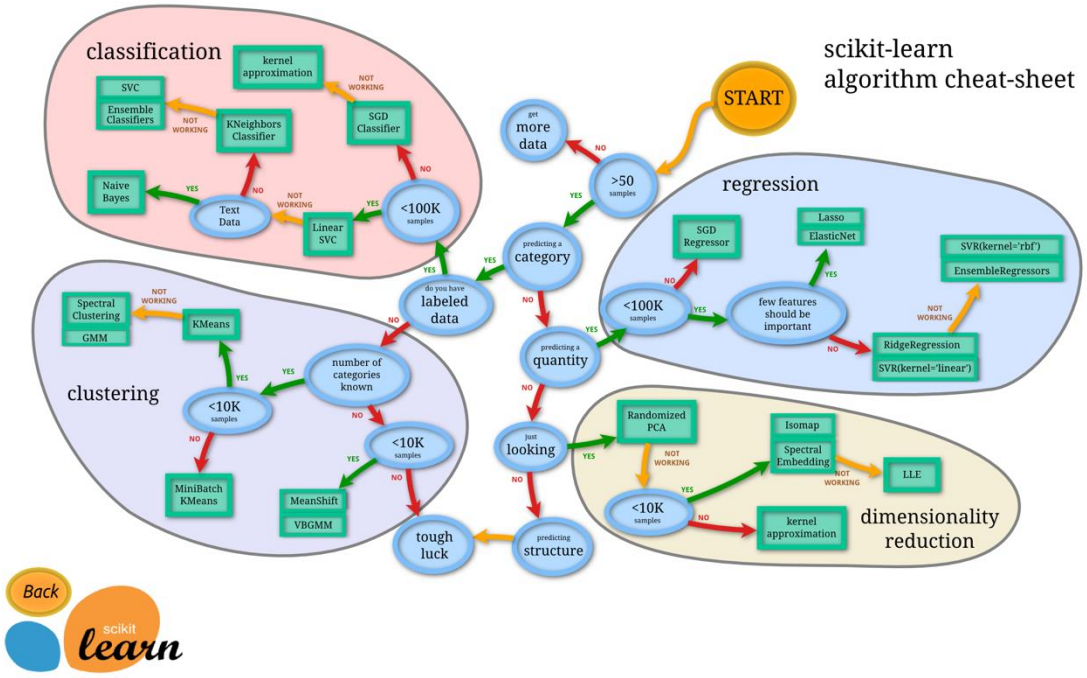
ΕΚΤΙΜΗΣΗ ΠΥΚΝΟΤΗΤΑΣ - DENSITY ESTIMATION

Η εκτίμηση πυκνότητας επιχειρεί να μάθει την κατανομή των εισόδων σε κάποιο χώρο και χωρίζεται σε παραμετρική και μη παραμετρική. Η παραμετρική λαμβάνει ως είσοδο κάποιο σύνολο από παραμέτρους ενώ η μη παραμετρική λαμβάνει υπόψιν της το σύνολο των εισόδων [23].

ΜΕΙΩΣΗ ΔΙΑΣΤΑΣΗΣ – DIMENSIONALITY REDUCTION

Το Dimensionality Reduction απλοποιεί τις εισόδους με την μετατροπή τους σε χώρο λιγότερων διαστάσεων. Η μοντελοποίηση θεμάτων (Topic modeling) είναι ένα σχετικό πρόβλημα, όπου σε ένα πρόγραμμα δίνεται ως είσοδος μια λίστα από έγγραφα που περιέχουν ανθρώπινη (φυσική) γλώσσα και στόχος του είναι να βρει ποια έγγραφα καλύπτουν παρόμοια θέματα.

Το ακόλουθο σχήμα περιγράφει έναν απλό τρόπο εύρεσης της κατάλληλης μεθόδου και του κατάλληλου αλγόριθμου για την επίλυση ενός προβλήματος μηχανικής μάθησης, αναλόγως των χαρακτηριστικών των διαθέσιμων δεδομένων.



Σχήμα 2. Συλλογιστική εύρεσης κατάλληλης μεθόδου, όπως προτείνεται από τη βιβλιοθήκη scikit-learn [28]

2.4.4 ΒΑΣΙΚΑ ΒΗΜΑΤΑ

Στη συνέχεια περιγράφονται τα γενικά βασικά βήματα που συνήθως ακολουθούνται σε μια διαδικασία μηχανικής μάθησης [23].

ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ

Η διαδικασία ξεκινά με την απόκτηση των δεδομένων που είναι σχετικά με τον στόχο της ανάλυσης και που θεωρείται ότι θα βοηθήσουν στο έργο. Η ανάκτηση μπορεί να γίνει είτε από τον παγκόσμιο ιστό είτε από μια βάση δεδομένων, είτε από όποια πηγή μπορεί να δώσει χρήσιμα και μετρήσιμα στοιχεία, π.χ. από αισθητήρες.

ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Τα δεδομένα αυτά σπανίως είναι σε μορφή αξιοποιήσιμη, οπότε το επόμενο βήμα είναι η μετατροπή τους σε ένα πρότυπο που μπορεί να χρησιμοποιηθεί στη συνέχεια. Η διαδικασία της μετατροπής μπορεί να περιλαμβάνει διάφορους αλγόριθμους και εξωτερικά δεδομένα. Επίσης, μπορεί να χρειαστεί να μετατραπούν τα δεδομένα σε πολύ συγκεκριμένη μορφή η οποία είναι κατάλληλη για χρήση ως είσοδος σε κάποιον αλγόριθμο. Παραδείγματος χάριν, κάποιοι αλγόριθμοι δεν μπορούν να χρησιμοποιήσουν αρνητικές αριθμητικές τιμές ως είσοδο, άρα θα πρέπει να γίνει μια μετατροπή με βάση αυτόν τον κανόνα.

ΑΝΑΛΥΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΕΙΣΟΔΟΥ

Αυτό το βήμα βοηθάει στην κατανόηση του τι πρέπει να αφαιρεθεί ως μη σημαντικό ή θόρυβος. Σε αυτό το βήμα γίνεται η αναγνώριση μοτίβων (patterns) ή εμφανών περίεργων σημείων όπως κάποια σημεία στα δεδομένα (data points) τα οποία είναι εντελώς διαφορετικά από τα υπόλοιπο σύνολο. Σε αυτό το σημείο συνηθίζεται η γραφική απεικόνιση των δεδομένων σε μία, δύο ή τρεις διαστάσεις.

ΕΚΠΑΙΔΕΥΣΗ ΑΛΓΟΡΙΘΜΟΥ

Σε αυτό το σημείο γίνεται η τροφοδότηση του αλγορίθμου με καλά «καθαρισμένα» δεδομένα, από τα δύο πρώτα βήματα και γίνεται η εξαγωγή της γνώσης και της πληροφορίας. Αυτή η γνώση, συχνά αποθηκεύεται σε ένα format είναι προς χρήση από ένα μηχάνημα για τα επόμενα δύο βήματα. Στην περίπτωση του unsupervised learning, δεν θα υπάρχει κάποιο training βήμα, επειδή δεν έχετε κάποια αναμενόμενη (target) τιμή.

ΕΚΤΙΜΗΣΗ ΑΠΟΔΟΣΗΣ ΑΛΓΟΡΙΘΜΟΥ

Εδώ είναι το σημείο όπου η πληροφορίες που αποκτήθηκαν στο προηγούμενο βήμα χρησιμοποιούνται. Στην περίπτωση του supervised learning, έχουμε μερικές γνωστές τιμές που μπορούμε να τις χρησιμοποιήσουμε για να αξιολογήσουμε τον αλγόριθμο. Στο unsupervised learning, ίσως χρειαστεί να χρησιμοποιηθούν κάποιου άλλου είδους μετρικές για να αξιολογηθεί η επιτυχία. Σε κάθε περίπτωση, είναι δυνατόν να πάμε πίσω στο βήμα 4, να κάνουμε αλλαγές και να ξανακάνουμε το test. Συχνά η συλλογή και προετοιμασία των δεδομένων μπορεί να αποτελέσει πρόβλημα οπότε συνηθίζεται η επιστροφή στο πρώτο βήμα.

ΧΡΗΣΗ

Το τελευταίο βήμα είναι η χρήση του συστήματος που περιγράφηκε στα πιο πάνω βήματα στην πράξη, για την επίλυση του προβλήματος για το οποίο δημιουργήθηκε. Έχοντας κάνει τις δοκιμές που χρειάζονται ώστε να είναι κατανοητό το πως λειτουργεί ο αλγόριθμος με τα δεδομένα που υπάρχουν και με την προεπεξεργασία που έχει γίνει, γίνεται πλέον η εφαρμογή του σε διαφορετικά δεδομένα.

2.4.5 ΒΑΣΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ

ΝΑΪVE BAYES

Οι Bayesian μεθόδοι βασίζονται στο θεώρημα του Bayes σχετικά με τις υπό συνθήκη πιθανότητες. Παρόλο που προέρχεται από τον τομέα της στατιστικής, αυτές οι μεθόδοι μελετώνται σε σχέση με τη μηχανική μάθηση και την εξόρυξη γνώσης.

Ο Naïve Bayes classifier βασίζεται στην υπόθεση ότι τα στοιχεία, οι είσοδοι είναι υπό συνθήκη ανεξάρτητα για κάθε δεδομένη υπόθεση (class). Τα Bayesian networks συνήθως προτείνουν τις πιο σωστές λύσεις, καθώς δεν υποθέτουν καμία ανεξαρτησία εκ των προτέρων.

Το θεώρημα Bayes ορίζεται μαθηματικά ως:

$$\mathrm{P}(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Εξίσωση 1. Θεώρημα Bayes

όπου A και B είναι γεγονότα, P(A) και P(B) είναι οι πιθανότητες των A και B που είναι ανεξάρτητα μεταξύ τους, P(A|B), η υπό συνθήκη πιθανότητα, είναι η πιθανότητα του A δεδομένου του B να είναι αληθής και P(B|A), είναι η πιθανότητα του B δεδομένου του A να είναι αληθής.

SUPPORT VECTOR MACHINES (SVM)

Ως Support Vector Machine ορίζεται ο classifier που διαιρεί τον χώρο εισόδου σε δύο περιοχές, οι οποίες χωρίζονται από ένα γραμμικό όριο.

Αποτελεί διακριτή προσέγγιση αλλά όχι πιθανοτικό μοντέλο. Η βασική ιδέα είναι ότι, εάν ο στόχος είναι να προβλεφθεί με ακρίβεια το αναμενόμενο αποτέλεσμα σύμφωνα με μια συνάρτηση κόστους, τότε αυτός θα πρέπει να είναι ο πρωταρχικός στόχος αντί να γίνεται εκτίμηση της κατανομής πιθανοτήτων, το οποίο είναι αρκετά πιο δύσκολο πρόβλημα.

DECISION TREES

Τα Decision Trees ανήκουν στα πιο δημοφιλή μοντέλα για την επίλυση προβλημάτων κατηγοριοποίησης. Όταν χτίζεται ένα Decision Tree, χωρίζεται ο χώρος εισόδου με έναν ιεραρχικό τρόπο αναδρομικά. Αυτή η μέθοδος είναι γνωστή και ως «διαίρει και βασίλευε» και έχει εφαρμοστεί σε διάφορους αλγόριθμους.

Υπάρχουν δύο βασικοί αλγόριθμοι που συσχετίζονται με τα Decision Trees. Ο κύριος αλγόριθμος ονομάζεται ID3 [29] και τα βασικά βήματα αναφέρονται ακολούθως. Η επέκταση αυτού ονομάζεται C4.5 αλγόριθμος [30], μπορεί να διαχειριστεί και αριθμητικά δεδομένα και η συνθήκη τερματισμού είναι πιο "χαλαρή", δίνοντας τη δυνατότητα στον αλγόριθμο να διαχειριστεί δεδομένα με "θόρυβο" τα οποία μπορεί να περιέχουν συνδυασμό κατηγορικών και αριθμητικών χαρακτηριστικών.

Τα βήματα του ID3:

1. Επιλέγεται ένα χαρακτηριστικό εισόδου ως η «ρίζα» ενός συγκεκριμένου (υπο)δένδρου
2. Διαιρούνται τα δεδομένα αυτής της ρίζας σε υποσύνολα, σε σχέση με την τιμή του κάθε επιλεγμένου χαρακτηριστικού και προστίθεται ένας νέος κόμβος για κάθε υποσύνολο.
3. Εάν κάποιος κόμβος περιέχει παραδείγματα από διαφορετικές κατηγορίες, πήγαινε στο βήμα

Τα χαρακτηριστικά πρέπει να είναι κατηγορικά (διακριτά), εάν δεν είναι πρέπει να γίνουν με προεπεξεργασία. Ο αλγόριθμος τερματίζει όταν, είτε όλα τα παραδείγματα ενός κόμβου βρίσκονται σε μια κλάση, είτε όταν δεν υπάρχουν άλλα χαρακτηριστικά για περαιτέρω τμηματοποίηση, είτε δεν υπάρχουν άλλα δείγματα για κατηγοριοποίηση.

Η ουσία του αλγορίθμου βρίσκεται στον τρόπο που επιλέγεται η διαίρεση του χαρακτηριστικού καθώς ένα κριτήριο που αντιστοιχεί στην ικανότητα να διαχωρίζει σωστά τα instances διαφορετικών κλάσεων υπολογίζεται για κάθε διαθέσιμο χαρακτηριστικό (input attribute) στον κόμβο διαχωρισμού [26].

3. ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ - SENTIMENT ANALYSIS

3.1 ΟΡΙΣΜΟΣ

Ως Ανάλυση Συναισθήματος (Sentiment Analysis) ή Εξόρυξη Γνώμης (Opinion Mining) ενός κειμένου ορίζεται το εγχείρημα της ανάκτησης της γνώμης του συγγραφέως σχετικά με συγκεκριμένες οντότητες. Η διαδικασία λήψης αποφάσεων των ανθρώπων επηρεάζεται από τις γνώμες που σχηματίζουν μέσω ηγετών ("thought-leaders") και συνανθρώπων. Όταν κάποιος σκέφτεται να αγοράσει ένα προϊόν διαδικτυακά, συνήθως πρώτα κάνει έρευνα για απόψεις και reviews άλλων σχετικά με το προϊόν αυτό [31] .

Η ανάλυση συναισθήματος αναφέρεται στην γενική μέθοδο εξαγωγής πολικότητας και υποκειμενικότητας από το κείμενο (δυνητικά και από την ομιλία). Ο όρος Σημασιολογικός Προσανατολισμός (Semantic Orientation ή SO) αναφέρεται στην πολικότητα και την ένταση των λέξεων, των φράσεων ή των κειμένων [32].

Σε γενικές γραμμές η Ανάλυση Συναισθήματος, η οποία είναι γνωστή και ως Εξόρυξη Γνώμης (Opinion Mining), αναφέρεται στη χρήση NLP, ανάλυσης κειμένου και υπολογιστικής γλωσσολογίας (computational linguistics) για τον εντοπισμό και την εξαγωγή υποκειμενικής πληροφορίας. Είναι δηλαδή η υπολογιστική μελέτη που ασχολείται με τον σωστό εντοπισμό των απόψεων, αισθημάτων, της αξιολόγησης και των συναισθημάτων που εκφράζονται σε κειμενική μορφή, όπως για παράδειγμα στις ειδήσεις, στα ιστολόγια (blogs), στις συζητήσεις, στα microblogs και στα κοινωνικά δίκτυα. Ο τομέας αυτός μπορεί να εντοπιστεί με διάφορες ονομασίες όπως για παράδειγμα sentiment analysis [33], subjectivity (Lyons 1981; Langacker 1985), opinion mining [33], analysis of stance (Biber and Finegan 1988; Conrad and Biber 2000), appraisal (Martin and White 2005), point of view (Wiebe 1994; Scheibman 2002), evidentiality (Chafe and Nichols 1986) κ.π.α. [34].

Μεταξύ των δύο κύριων τύπων κειμενικής πληροφορίας (textual information) – των γεγονότων και των απόψεων, ένα μεγάλο μέρος των σημερινών μεθόδων επεξεργασίας πληροφοριών, όπως η αναζήτηση στο διαδίκτυο και την εξόρυξη κειμένου (text mining), χρησιμοποιούν τον πρώτο. Το Opinion Mining αναφέρεται στην ευρύτερη έννοια/περιοχή της επεξεργασίας φυσικής γλώσσας, στα computational linguistics και εξόρυξη κειμένου που αφορούν την μελέτη των απόψεων, συναισθημάτων και αισθημάτων που εκφράζονται στο κείμενο. Μια σκέψη, άποψη ή συμπεριφορά βασισμένα σε ένα συναίσθημα και όχι στην λογική, αναφέρεται συχνά ως συναίσθημα. Ο εναλλακτικός ορισμός του Opinion Mining, δηλαδή η Ανάλυση Συναισθήματος έχει πολύ μεγάλη σημασία σε τομείς στους οποίους οργανισμοί ή ιδιώτες επιθυμούν να μάθουν το γενικό συναίσθημα, την κοινή γνώμη σχετικά με μια οντότητα, όπως ένα προϊόν, μια ταινία, ένα άτομο κλπ., και έχει εφαρμογή σε πολλούς τομείς, συμπεριλαμβανομένων της επιστήμης και της τεχνολογίας, στην εκπαίδευση, στην διασκέδαση, στην πολιτική, στο marketing, στη λογιστική, στη νομοθεσία και στην έρευνα και ανάπτυξη (R&D). Η προσέγγιση για την κατηγοριοποίηση του κειμένου, εμπλέκει την κατασκευή classifiers από κατηγοριοποιημένα (labeled) δείγματα κειμένου ή προτάσεων, κάτι που σημαίνει ότι είναι διαδικασία επιβλεπόμενης μάθησης [32]. Ένας ακόμα όρος που σχετίζεται με την ανάλυση συναισθήματος είναι η πολικότητα της γνώμης (opinion orientation, sentiment orientation, polarity of opinion, semantic orientation), ο οποίος αναφέρεται στον συναισθηματικό προσανατολισμό ενός κειμένου, μιας πρότασης ή μιας λέξης. Για παράδειγμα

η πολικότητα ενός κειμένου μπορεί να είναι «θετική», «αρνητική» ή «ουδέτερη», που σημαίνει ότι επικρατεί ενός είδους συναίσθημα [35].

3.2 ΚΑΤΗΓΟΡΙΕΣ

3.2.1 DOCUMENT-LEVEL

Είναι η πιο απλή μορφή sentiment analysis και θεωρείται ότι το έγγραφο περιέχει την άποψη του συγγραφέως πάνω σε ένα κύριο αντικείμενο. Αρκετές έρευνες έχουν ασχοληθεί με το συγκεκριμένο αντικείμενο. Υπάρχουν δύο κύριες προσεγγίσεις του θέματος: Προσέγγιση με Επιβλεπόμενη Μάθηση (Supervised Learning) και προσέγγιση με Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning).

Η επιβλεπόμενη προσέγγιση προϋποθέτει ότι υπάρχουν δεδομένα για εκπαίδευση (training data) και υποθέτει ότι υπάρχει ένα πεπερασμένο σύνολο κατηγοριών στις οποίες ανήκει το έγγραφο. Η πιο απλή περίπτωση είναι οι δύο κατηγορίες: θετικό και αρνητικό, ενώ η πολυπλοκότητα μπορεί να αυξηθεί απλά προσθέτοντας την ουδέτερη κλάση ή και δίνοντας βαθμίδες στο πόσο θετική ή αρνητική μπορεί να είναι μια άποψη (5-star-class Amazon). Δεδομένων των training data, το σύστημα χτίζει ένα μοντέλο κατάταξης χρησιμοποιώντας κάποιον από τους συνηθισμένους αλγόριθμους κατάταξης όπως SVM, Naïve Bayes, Logistic Regression, ή KNN. Στη συνέχεια, αυτό το μοντέλο χρησιμοποιείται για την πρόβλεψη κατηγορίας σε νέα έγγραφα. Στην περίπτωση που χρειάζεται να αποδοθεί αριθμός ως κλάση του εγγράφου τότε είναι δυνατόν να χρησιμοποιηθεί regression. Έχει αποδειχθεί μέσω ερευνών ότι μέσω της αναπαράστασης ενός εγγράφου ως Bag Of Words είναι δυνατόν να επιτευχθεί καλή ακρίβεια στα αποτελέσματα. Πιο πολύπλοκες και προηγμένες τεχνικές αναπαράστασης εγγράφου χρησιμοποιούν TFIDF, POS (Part-Of-Speech) πληροφορία, sentiment lexicons και ανάλυση της δομής και της μορφολογίας του εγγράφου (parse structures).

Η προσέγγιση της Μη Επιβλεπόμενης Μάθησης, βασίζεται σε SO (Semantic Orientation polarity) συγκεκριμένων φράσεων μέσα στο έγγραφο. Εάν ο μέσος όρος του SO των φράσεων αυτών είναι πάνω από κάποιο όριο (threshold), τότε το έγγραφο κατηγοριοποιείται ως θετικό, αλλιώς ως αρνητικό. Οι κύριες τεχνικές που ακολουθούνται για την ανίχνευση των φράσεων που θα χρησιμοποιηθούν για τον προσδιορισμό του SO polarity είναι δύο: χρήση συγκεκριμένων POS μοτίβων και χρήση λεξικών αποτελούμενων από λέξεις κατηγοριοποιημένες ως προς το συναίσθημά τους (sentiment lexicons). Κλασσική μέθοδος υπολογισμού SO μιας λέξης ή φράσης αποτελεί ο υπολογισμός της διαφοράς του PMI (Pointwise Mutual Information) μεταξύ δύο λέξεων που εκφράζουν συναίσθημα. Το PMI(P,W) μετράει την στατιστική εξάρτηση μεταξύ μιας φράσης P και της λέξης W, βασιζόμενο στην συνύπαρξή τους μέσα σε μια συλλογή λέξεων/ φράσεων/ εγγράφων ή στο διαδίκτυο με τη χρήση Web search queries [31].

Για την ανάλυση κειμένων που περιέχουν γλώσσες όπως Κινέζικα και Ισπανικά, για τις οποίες δεν υπάρχουν αρκετοί γλωσσικοί πόροι, είναι συνήθης η χρήση της μηχανικής μετάφρασης για την μετατροπή του κειμένου πρώτα σε Αγγλικά, για τα οποία υπάρχουν πολλοί πόροι για αυτή τη διαδικασία και μετά της εφαρμογής του όποιου αλγόριθμου για την ανάλυση συναισθήματος.

3.2.2 SENTENCE-LEVEL SENTIMENT ANALYSIS

Η ανάλυση συναισθήματος μπορεί να γίνει σε επίπεδο πρότασης κάτι που γίνεται περίπλοκο από το γεγονός ότι η σημασιολογική ερμηνεία των λέξεων εξαρτάται πάρα πολύ από το πλαίσιο στο οποίο αναφέρονται. Η ανάλυση συναισθήματος σε επίπεδο πρότασης δίνει μια πιο αναλυτική οπτική των διαφορετικών απόψεων που μπορεί να εκφράζονται σε ένα κείμενο σχετικά με τις οντότητες που αναφέρονται. Σε αυτή την περίπτωση γίνεται η υπόθεση ότι η ταυτότητα της οντότητας για την οποία εκφράζεται κάποια άποψη σε μια πρόταση είναι γνωστή. Επίσης, γίνεται η υπόθεση ότι η πρόταση εκφράζει ένα κύριο συναίσθημα μόνο και συνήθως μόνο οι υποκειμενικές προτάσεις αναλύονται, καθώς θεωρείται ότι οι αντικειμενικές προτάσεις δεν εμπεριέχουν κάποιο συναίσθημα. Κάποιες προσεγγίσεις κάνουν χρήση των αντικειμενικών προτάσεων, η ανάλυση των οποίων όμως αποτελεί αρκετά δυσκολότερο εγχείρημα [31] [36].

ASPECT-BASED SENTIMENT ANALYSIS

Η ανάλυση συναισθήματος σε σχέση με ένα αντικείμενο ή στόχο ονομάζεται Aspect Based Sentiment Analysis (ABSA). Τα συστήματα που έχουν ως στόχο την ABSA λαμβάνουν ως είσοδο ένα σύνολο κειμένων, όπως για παράδειγμα, κριτικές προϊόντων ή μηνύματα από κοινωνικά δίκτυα, τα οποία καταπιάνονται με μια συγκεκριμένη οντότητα, π.χ. ένα προϊόν, ένα κινητό τηλέφωνο κ.α. Στη συνέχεια, προσπαθούν να εντοπίσουν την περιοχή ενδιαφέροντος, για παράδειγμα το πιο συχνό θέμα με τα χαρακτηριστικά του, δηλαδή την οθόνη ενός φορητού υπολογιστή και να υπολογίσουν το συνολικό συναίσθημα προς αυτό. Παρόλο που πολλά συστήματα ABSA έχουν προταθεί, και πρόκειται ως επί το πλείστον για ερευνητικά πρωτότυπα (Liu, 2012), δεν υπάρχει καθιερωμένη διαδικασία για ABSA, ούτε υπάρχουν θεσπισμένα μέτρα αξιολόγησης για τις δευτερεύουσες εργασίες που τα ABSA συστήματα καλούνται να εκτελέσουν [31].

3.2.3 COMPARATIVE SENTIMENT ANALYSIS

Μια συγκριτική πρόταση συνήθως εκφράζει μια σειριακή σχέση μεταξύ δύο συνόλων οντοτήτων σε σχέση με κάποια χαρακτηριστικά ή θέματα.

Οι συγκρίσεις συσχετίζονται με τις άμεσες απόψεις, αλλά ταυτόχρονα είναι και αρκετά διαφορετικές. Ένα παράδειγμα μια τυπική πρόταση που εκφράζει άμεση άποψη είναι «Η ποιότητα του Χ προϊόντος είναι άψογη!». Σε μια συγκριτική πρόταση θα είχαμε το ακόλουθο: «Η ποιότητα του Χ προϊόντος είναι καλύτερη από την ποιότητα του Ψ προϊόντος». Είναι προφανές ότι οι συγκρίσεις χρησιμοποιούν διαφορετικού τύπου εκφράσεις από την έκφραση άποψης. Συνήθως οι συγκρίσεις εκφράζουν μια συγκριτική άποψη για δύο ή περισσότερες οντότητες σχετικά με κοινά χαρακτηριστικά, π.χ. «κατασκευαστική ποιότητα», «τιμή» κλπ.

Παραδείγματος χάριν, η συγκριτική πρόταση *"Canon's optics are better than those of Sony and Nikon"* εκφράζει την συγκριτική σχέση (better, {optics}, {Canon}, {Sony, Nikon}). Οι συγκριτικές προτάσεις χρησιμοποιούν διαφορετικού τύπου γλωσσικές ιδιότητες και γλωσσικές “κατασκευές” από τις τυπικές προτάσεις που εκφράζουν απόψεις, για παράδειγμα "Cannon's optic is great" [31] [37] [38].

3.2.4 SENTIMENT LEXICON ACQUISITION

Η προσέγγιση που βασίζεται στα λεξικά, περιλαμβάνει τον υπολογισμό του συναισθηματικού προσανατολισμού ενός κειμένου από τον σημασιολογικό προσανατολισμό των λέξεων ή των φράσεων που το αποτελούν. [31] [39].

Τα λεξικά για τη λεξιλογική αυτή προσέγγιση είναι δυνατόν να κατασκευαστούν είτε με μη αυτόματο τρόπο (manually) είτε αυτόματα, χρησιμοποιώντας αρχικές λέξεις, ονομαζόμενες ως seed words, ώστε να επεκταθεί η λίστα από λέξεις. Πολλές από τις μεθόδους που βασίζονται σε λεξικά έχουν εστιάσει στη χρήση επιθέτων (adjectives) ως ενδείξεις για τον σημασιολογικό προσανατολισμό ενός κειμένου. Αρχικά μια λίστα επιθέτων και οι αντίστοιχες SO τιμές μαζεύονται σε ένα λεξικό (dictionary) και στη συνέχεια όλα τα επίθετα ενός κειμένου μαρκάρονται με τα σκορ που υπάρχουν στο λεξικό. Στη συνέχεια, τα σκορ αυτά μετατρέπονται σε μέσο όρο, ο οποίος τελικά θα καθορίσει την πολικότητα (polarity) του κειμένου [32].

Τα Sentiment Lexicons είναι λίστες από λέξεις και εκφράσεις συναισθημάτων ή γνώμης. Εκτός από λέξεις, μπορεί να αποτελούνται και από φράσεις ή και ιδιωματισμούς. Πολλά από τα λεξικά αυτού του τύπου είναι διαθέσιμα στο διαδίκτυο. Συνήθως περιέχουν χιλιάδες όρους και είναι αρκετά χρήσιμα ως συναισθηματικές λέξεις και φράσεις (sentiment words, polar words, opinion bearing words κλπ.). Παραδείγματος χάριν, οι λέξεις όμορφος, υπέροχος, καλός, φανταστικός κατηγοριοποιούνται ως θετικές λέξεις, ενώ οι λέξεις κακός, ελλιπής, απαίσιος ως αρνητικές.

Υπάρχουν τρεις κύριοι τρόποι δημιουργίας Sentiment Lexicons:

Μη αυτόματη (χειροκίνητη) προσέγγιση

Ο τρόπος αυτός είναι αρκετά καλή ιδέα όταν η διαδικασία πρόκειται να γίνει μια φορά μόνο. Περιλαμβάνει την χειροκίνητη κατηγοριοποίηση λέξεων και φράσεων σε κάποια συναισθηματική κατηγορία, π.χ. “θετική”, “αρνητική”.

Corpus - based προσέγγιση

Σε αυτόν τον τρόπο συνήθως χρησιμοποιείται το double propagation μεταξύ των λέξεων που περιέχουν άποψη και των αντικειμένων στα οποία αναφέρονται. Απαιτεί μεγάλο όγκο κειμένων για να υπάρξει καλό ποσοστό κάλυψης.

Dictionary-based προσέγγιση

Συνήθως χρησιμοποιούνται ομάδες συνώνυμων (synsets) και ιεραρχίες ώστε να γίνει εξαγωγή των λέξεων που εκφράζουν κάποιο συναίσθημα και το τελικό αποτέλεσμα είναι λέξεις των οποίων η πολικότητά τους δεν εξαρτάται από το γενικό πλαίσιο μέσα στο οποίο βρίσκονται. Βασίζονται σε συντακτικά μοτίβα και χρειάζονται μεγάλο όγκο κειμένου [39].

ΠΑΡΑΔΕΙΓΜΑΤΑ ΛΕΞΙΚΩΝ

General Inquirer Lexicon [40]:

- Θετικές (1915) και αρνητικές λέξεις (2291)
- Δυνατές (Strong) και Αδύναμες (Weak), Δυναμικές (Active) και Παθητικές), Υπερεκτιμημένες (Overstated) και Υποτιμημένες (Understated)
- Ευχαρίστηση (Pleasure), Πόνος (Pain), Αρετή (Virtue), Ελάττωμα (Vice), Παρακίνηση (Motivation), Γνωστικός Προσανατολισμός (Cognitive Orientation)
- Πηγή: http://www.wjh.harvard.edu/~inquirer
- Άδεια χρήσης: Ελεύθερη χρήση στον τομέα της έρευνας.

MPQA Subjectivity Cues Lexicon [41] [42]:

- 6885 λέξεις από 8221 λήμματα (lemmas)
- 2718 θετικές λέξεις και 4912 αρνητικές
- Κάθε λέξη έχει χαρακτηριστεί ως προς την ένταση
- Πηγή: http://www.cs.pitt.edu/mpqa/subj_lexicon.html
- Άδεια χρήσης: GNU GLP

Opinion Lexicon [43]:

- 6786 λέξεις
- 2006 θετικές και 4783 αρνητικές
- Πηγή: http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

SentiWordNet [44]:

- Όλα τα σύνολα συνωνύμων (synsets) του WordNet χαρακτηρισμένα αυτομάτως με βαθμούς θετικότητας, αρνητικότητας και με ουδετερότητα.
- Πηγή: http://sentiwordnet.isti.cnr.it/
- Άδεια χρήσης: Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0).

	Opinion Lexicon	General Inquirer	SentiWordNet
MPQA	33/5402 **(0.6%)**	49/2867 **(2%)**	1127/4214 **(27%)**
Opinion Lexicon		32/2411 **(1%)**	1004/3994 **(25%)**
General Inquirer			520/2306 **(23%)**
SentiWordNet			

Πίνακας 3. Διαφωνίες μεταξύ των λεξικών [45]

3.3 ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΤΟ TWITTER

Τα tweets έχουν μοναδικά χαρακτηριστικά σε σχέση με άλλες συλλογές δεδομένων (corpora). Στα χαρακτηριστικά αυτά συμπεριλαμβάνονται emoticons, συντομεύσεις, hashtags, δημιουργικές συντομεύσεις κ.α. Σχετικά με την ανάλυση συναισθήματος σε ένα τέτοιο corpus/ domain, έχει αποδειχθεί ότι η χρήση αυτών των χαρακτηριστικών μπορεί να συνεισφέρει στην καλή απόδοση πρόβλεψης συναισθήματος. Συγκεκριμένα, η χρήση emoticons ως features θεωρείται ένας αρκετά αποτελεσματικός τρόπος έκφρασης θετικού ή αρνητικού συναισθήματος [46] [47]. Οι Go, Bhayani και Huang [48] στην έρευνά τους έδειξαν ότι οι αλγόριθμοι μηχανικής μάθησης μπορούν να έχουν ακρίβεια πάνω από 80% όταν εκπαιδεύονται με δεδομένα που περιέχουν emoticons. Επίσης, υπάρχουν ενδείξεις ότι η χρήση hashtags και η παρουσία intensifiers, όπως για παράδειγμα λέξεις με κεφαλαία και τα σημεία στίξης, μπορεί να παίξει ρόλο στην αναγνώριση συναισθήματος [49]. Οι Agarwal et al. [50] στην έρευνά τους αναφέρουν ότι τέτοιου είδους χαρακτηριστικά (features) μπορεί να προσθέσουν αξία στην αναγνώριση συναισθήματος, αλλά μόνο οριακά, δηλαδή θεωρούν ότι τα χαρακτηριστικά αυτά παίζουν μικρό ρόλο στην διαδικασία σωστής κατηγοριοποίησης. Χαρακτηριστικά που εμπίπτουν στην κατηγορία natural language, παραδείγματος χάριν Part-Of-Speech tags, και η χρήση λεξικών συναισθήματος συμβάλουν σημαντικά στην ανίχνευση της διάθεσης του συγγραφέα ενός tweet. Οι Agarwal, Xie, Vovsha, Rambow και Passonneau [50] καταλήγουν στο ότι τα πιο σημαντικά χαρακτηριστικά είναι αυτά που συνδυάζουν prior polarity των λέξεων και τα αντίστοιχα part-of-speech tags.

3.4 ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΕ ΚΕΙΜΕΝΟ ΠΟΥ ΠΕΡΙΕΧΕΙ ΜΕΤΑΦΟΡΙΚΟ ΛΟΓΟ

Η ανάλυση συναισθήματος σε μεταφορικό λόγο (Sentiment Analysis on Figurative Language) έχει αντιμετωπισθεί με διάφορους τρόπους. Έχει διερευνηθεί η χρήση λεξικολογικών και συντακτικών χαρακτηριστικών για την σωστή αναγνώριση του μεταφορικού λόγου και του συναισθήματος που εκφράζεται μέσα από αυτόν. Η πολυπλοκότητα ενός τέτοιου εγχειρήματος είναι μεγάλη, ιδίως αν ληφθεί υπόψη το γεγονός ότι η ειρωνεία και σαρκασμό συχνά αναμειγνύονται [51]. Ο σαρκασμός χρησιμοποιείται συνήθως για να μειώσει τον στόχο του σχολίου και είναι σχετικά πιο εύκολος στην αναγνώρισή του, σε σχέση με τις υπόλοιπες κατηγορίες. Η ειρωνεία λειτουργεί ως άρνηση (negation), μετατοπίζοντας το συναίσθημα, αλλά σχεδόν πάντα προς το αρνητικό. Μπορεί να εκφράζεται μέσω ενός θετικού πλαισίου βέβαια, γεγονός που καθιστά τη διάκριση του πραγματικού νοήματος δύσκολη [52] [53]. Οι [53] έδειξαν ότι μοτίβα όπως “As * As *”, “about as * as *” είναι χρήσιμα για την ανίχνευση ειρωνικών παρομοιώσεων (ironic similes).

Οι [54] εξέτασαν τον σαρκασμό που εκφράζεται από τον χρήστη μέσω hashtag για να εντοπίσουν εάν τέτοιου είδους tweets είναι αξιόπιστη πηγή σαρκασμού. Συμπέραναν πως τα tweets που κατηγοριοποιούνται από τους χρήστες ως σαρκαστικά μπορεί να περιέχουν “θόρυβο” και ότι αποτελούν την πιο δύσκολη μορφή για τη σωστή κατηγοριοποίηση του σαρκασμού. Οι [55] εντόπισαν ότι ο σαρκασμός προκύπτει από την αντίθεση μεταξύ ενός θετικού συναισθήματος και μιας αρνητικής κατάστασης. Παραδείγματος χάριν “Don’t you just love it when it rains on your wedding day…”. Οι Reyes, Rosso και Buscaldi, χρησιμοποίησαν στην έρευνά τους χαρακτηριστικά που εκφράζουν αυτή την ανισορροπία εκ των

συμφραζόμενων (contextual imbalance), χαρακτηριστικά που προκύπτουν από τον τομέα της φυσικής γλώσσας (natural language), όπως επίσης και διάφορα συντακτικά και μορφολογικά χαρακτηριστικά ενός tweet. [52]. Σε αρκετές έρευνες, γίνεται χρήση αυτής της ανισορροπίας των συμφραζόμενων (contextual imbalance), η οποία υπολογίζεται ως η σημασιολογική ομοιότητα μεταξύ των λέξεων, αλλά και διάφορων λεξιλογικών πόρων (WordNet, Whisel's dictionary) για τον εντοπισμό χαρακτηριστικών όπως "συναισθηματικό περιεχόμενο" (emotional content), πολικότητα λέξεων (polarity of words), τερπνότητα (pleasantness) και επιρρήματα που υπονοούν άρνηση (negation) ή εκφράζουν χρονικό παράθυρό και συγχρονισμό (implying negation or expressing timing). Οι [56] έχουν αναπτύξει μια μη επιβλεπόμενη μέθοδο (un-supervised method), η οποία έχει υψηλή ακρίβεια στην αναγνώριση μεταφορών με τη χρήση συνώνυμων μέσω του WordNet. Τέλος, χαρακτηριστικά που φαίνεται να είναι χρήσιμα στη διαδικασία αναγνώρισης συναισθήματος είναι τα σημεία στίξης, τα οποία συνήθως δείχνουν ένταση συναισθήματος, τα emoticons, οι κεφαλαίες λέξεις, οι οποίες και αυτές φαίνεται να δείχνουν ένταση, τα n-grams και τα skip-grams [51].

4. ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΠΡΟΒΛΗΜΑΤΟΣ

Το πρόβλημα ασχολείται με την ταξινόμηση των tweets που περιέχουν ειρωνεία και μεταφορικό λόγο. Εμπίπτει στην κατηγορία της ανάλυσης κειμένου (Text Analysis), και πιο συγκεκριμένα της ανάλυσης συναισθήματος (Sentiment Analysis), εμπλέκοντας άμεσα στοιχεία επεξεργασίας φυσικής γλώσσας, καθώς το κείμενο του tweet περιέχει ανεπίσημη φυσική γλώσσα. Από τον ορισμό του επίσης προκύπτει ότι είναι ένα πρόβλημα που εμπεριέχει επιβλεπόμενη μάθηση, καθώς τα διαθέσιμα δεδομένα, τα οποία χωρίζονται σε δεδομένα εκπαίδευσης (train set) και δεδομένα ελέγχου (test set), αποτελούνται από το κείμενο του tweet ως είσοδο και την αντίστοιχη συναισθηματική του αξιολόγηση (label).

Τα δεδομένα αυτά περιλαμβάνουν μια συμπυκνωμένη ποσότητα αυτών των φαινομένων. Απώτερος στόχος είναι να εξεταστεί κατά πόσο το συμβατικό sentiment analysis μπορεί να χειριστεί αποτελεσματικά τις ιδιαιτερότητες της δημιουργικής γλώσσας και του μεταφορικού γραπτού λόγου, ιδιαιτέρως στον περιορισμό των 140 χαρακτήρων και την γενικώς πολύπλοκη φύση ενός tweet και εάν συστήματα τα οποία είναι μοντελοποιημένα με τέτοιο τρόπο ώστε να καλύπτουν όσο πιο καλά γίνεται τους ιδιαιτερότητες του μεταφορικού λόγου αποδίδουν καλύτερα από τους κλασικές τεχνικές.

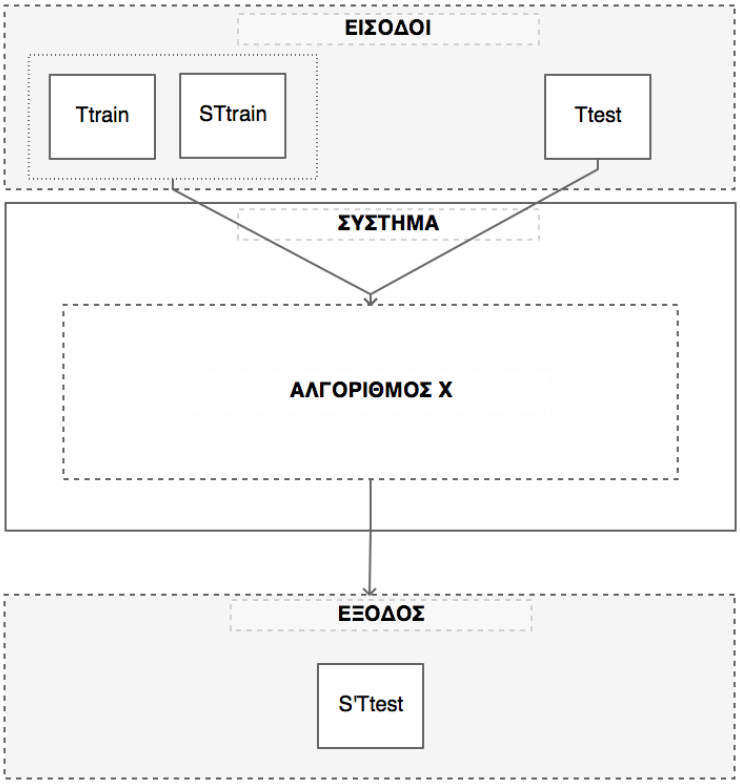
Το πρόβλημα μπορεί να περιγραφεί ως εξής:

Δεδομένων:

- ενός συνόλου δοκιμής $\boldsymbol{T_{train}} = \{\boldsymbol{t_1}, \boldsymbol{t_2}, \ldots \boldsymbol{t_m}\}$, όπου $\boldsymbol{t_i}$ το κείμενο ενός tweet, με $\boldsymbol{i} \in [1, m]$ και m το πλήθος των tweets του συνόλου,
- του αντίστοιχου συνόλου $\boldsymbol{S_{Ttrain}} = \{\boldsymbol{s_{t1}}, \boldsymbol{s_{t2}}, \ldots \boldsymbol{s_{tm}}\}$, όπου $\boldsymbol{s_{ti}}$ η βαθμολόγηση του αντίστοιχου $\boldsymbol{t_i}$, σχετικά με το συναίσθημα που περιέχεται.
- το $\boldsymbol{s_{ti}}$ ανήκει σε ένα εύρος τιμών R= [a, b], το οποίο χρησιμοποιείται για να εκφραστεί το συναίσθημα ενός tweet.

Να υπολογιστούν σε ένα σύνολο δοκιμής $\boldsymbol{T_{test}} = \{\boldsymbol{q_1}, \boldsymbol{q_2}, \ldots \boldsymbol{q_n}\}$, όπου $\boldsymbol{q_j}$ το κείμενο ενός tweet, με $\boldsymbol{j} \in [1, n]$ και n το πλήθος των tweets του συνόλου, το αντίστοιχο σύνολο $\boldsymbol{S'_{Ttest}} = \{\boldsymbol{s'_{t1}}, \boldsymbol{s'_{t2}}, \ldots \boldsymbol{s'_{tn}}\}$, όπου $\boldsymbol{s'_{tj}}$ η προσέγγιση της βαθμολόγησης του αντίστοιχου $\boldsymbol{t_j}$, σχετικά με το συναίσθημα που περιέχεται και για το οποίο ισχύει $\boldsymbol{s'_{tj}} \in R$ Στόχος είναι το $\boldsymbol{S'_{Ttest}}$ να προσεγγίσει όσο το δυνατόν το $\boldsymbol{S_{gold}} = \{\boldsymbol{s_{t1}}, \boldsymbol{s_{t2}}, \ldots \boldsymbol{s_{tn}}\}$, το οποίο περιέχει τη ζητούμενη αντίστοιχη βαθμολογία.

Ο υπολογισμός θα γίνει μέσω ενός αλγόριθμου X, ο οποίος λαμβάνει ως εισόδους το $\boldsymbol{T_{train}}$ με το αντίστοιχο $\boldsymbol{S_{Ttrain}}$ και το $\boldsymbol{T_{test}}$, και η αναμενόμενη έξοδός του θα είναι ένα $\boldsymbol{S'_{Ttest}}$, το οποίο θα ονομάζεται στη συνέχεια ως $\boldsymbol{S_{system}}$.



Σχήμα 3. Επισκόπηση προβλήματος

Για τις βαθμολογίες έχει χρησιμοποιηθεί το συνεχόμενο διάστημα R= [-5, 5], ώστε να συλληφθεί η επίδραση της ειρωνείας και της μεταφοράς σε σχέση με το αντιληπτό το συναίσθημα ενός tweet. Το σύστημα θα πρέπει να αναθέσει σε κάθε tweet βαθμολογία που να αντιστοιχεί στην δεδομένη κλίμακα, η οποία θα απεικονίζει το πόσο θετικό, αρνητικό ή ουδέτερο είναι το συναίσθημα που εκφράζεται από τον συγγραφέα του.

Η αξιολόγηση των αποτελεσμάτων του συστήματος γίνεται με τη χρήση του αλγόριθμου Cosine Similarity. Τα αποτελέσματα συγκρίνονται με το σταθμισμένο μέσο όρο των βαθμολογιών που παρέχονται από τους ανθρώπινους βαθμολογητές και παρέχονται μέσω της πλατφόρμας CrowdFlower [57].

Ένα vector space χρησιμοποιείται για να αξιολογήσει την ομοιότητα των προβλέψεων και τους πραγματικής βαθμολογίας που προέρχεται από τη λίστα αναμενόμενων αποτελεσμάτων (gold standard) που έχει προκύψει από την ανθρώπινη βαθμολόγηση. Η λίστα των αναμενόμενων βαθμολογιών θα χρησιμοποιηθεί για να κατασκευαστεί ένα κανονικοποιημένο διάνυσμα (gold standard vector) και από τη λίστα των προβλέψεων θα προκύψει ένα αντίστοιχο διάνυσμα. Η σύγκρισή τους με τη βοήθεια του cosine similarity, το συνημίτονο της μεταξύ τους γωνίας δηλαδή, θα χρησιμοποιηθεί για να μετρηθεί το πόσο καλά οι εκτιμήσεις του συστήματος προσεγγίζουν τα gold standards στο σύνολο δοκιμών. Τους ο τρόπος αξιολόγησης σημαίνει ότι το σύστημα δεν αναμένεται να προβλέψει ακριβώς την αναμενόμενη βαθμολογία. Η αξιολόγηση είναι συνεχόμενη και όχι διακριτή και δεν διαφέρει για αμελητέες διαφορές. Εάν δύο συστήματα προβλέπουν συνεχώς λάθος βαθμολογία αλλά το ένα είναι συνεχώς πιο κοντά στο gold standard από το άλλο, για παράδειγμα προβλέποντας -3.1 για ένα tweet που η ανθρώπινη βαθμολόγηση είναι -4.2, τότε το σύστημα αυτό που είναι συνεπώς πιο κοντά θα έχει μεγαλύτερη τελική βαθμολογία. [58]

Η τελική βαθμολογία όπως ορίστηκε, μπορεί να περιγραφεί από την ακόλουθη εξίσωση, όπου $\boldsymbol{final\_score_{system_r}}$ είναι το τελικό σκορ για το σύστημα, $\boldsymbol{S_{gold}}$ είναι η λίστα των σκορ του gold standard, $\boldsymbol{S_{system}}$ είναι η λίστα των σκορ που έχει προβλέψει το σύστημα και **N** ο αριθμός των δειγμάτων (tweets) που περιέχονται στο $\boldsymbol{T_{test}}$.

$$final\_score_{system_r} = \cos(S_{gold}, S_{system}) = \frac{\sum_{i=1}^{N} S_{gold\_standard\_i} * S_{system\_i}}{\sqrt{\sum_{i=1}^{N} S_{gold\_i}{}^2} * \sqrt{\sum_{i=1}^{N} S_{system\_i}{}^2}}$$

Εξίσωση 2. Ο υπολογισμός της τελικής βαθμολογίας για κάθε σύστημα

Ένα ακόμη μέτρο που χρησιμοποιείται στα πλαίσια της αξιολόγησης των συστημάτων είναι και το MSE (βλ. κεφάλαιο 8.1)

id	text	initial_score
5074054276	Change British spelling to American spelling or risk being hung as a spy for the Queen.	-2
5987998129	Vegetarians, environmentalists, and animal rights activists may be collectively referred to as 'Communists.'	-2.2
8634493242	It's better to plagiarize from Encarta than from Wikipedia, because people actually read Wikipedia.	-2.4
8923309095	If you feel like your technology column is lacking something, it's probably condescension.	-2.6
12178656008	When summer comes and California starts burning, try to act surprised.	-2.4
16847979647	Presidential missteps are always the fault of the previous administration. See also: Presidential accomplishments.	-2.6
3134587228131328	No, i haven't gained weight.. Your eyes just got fat #sarcastweet	-3.4
10481341816639488	@collinslateshow jeremy hunt cuture media sports sect shows his disgust at BBC over panorama's nail in coffin by giving them 4.5 billion ?	-2.8
15063904711348224	Just how many planets *do* I have to blow up before I'm named TIME's Person of the Year?	-2.8
65448069851906048	'Don't assume you can walk into the New York Times and get a job cleaning toilets. You have to work as a features editor first.	0
94909968855203841	I think church and state are secretly fucking.	-2.6
111900485040078848	I hate those unrealistic movies where women are friends.	-2.8
122147560646393857	'There's a guy at your wife's office who possesses many of the qualities you lack.'	-3
143731085979811841	I wanted to write 'stop fucking texting me' but instead wrote 'I love you too'.	-2.4
190248272944836608	Oh, you took a picture in the bathroom? You must be an upcoming model.	-2.4
190880032166641666	Oh, you cheated on your beautiful girlfriend for a hoe that looks like Hellboy? Good choice!	-3.4
191202468112244740	Oh, 10 hashtags in that tweet? You must be such a trend-setter.	-2.4

Σχήμα 4. Δείγμα των δεδομένων δοκιμής με συνεχόμενα σκορ

id	text	initial_score
537239855777185792	@SeifertESPN states the Vikings aren't being moral cause AP was going to play. ..simultaneously uses the situation to push an agenda #irony	-2
537239886957666304	Truer words... RT @ianollis: So the ANC is bussing cadres into Parliament to prove that the speaker is not biased? #Irony #BalekaMustGo	-3
537239914824630273	Gonna finish this cup of coffee then take a nap. #irony	-1
537239969577050112	I don't speak Spanish.. So idk how to reply to you	-1
537239996932300800	Nash can you help with my homework plz. I'm literally asking someone with 3.42 followers. That's how desperate I am plz help! @Nashgrier	0
537240024476319744	I wish I had money so I could go to conventions and do panels. I am totally willing to speak openly about these things #GamerGate	-1
537240051751854080	@AnheuserBusch odds are those players were drinking that poison you sell when they were doing their #baddeeds LOL #Irony @nfl	-2
537240079056769024	@jdbftcdallas AT FIRST HE LOOKS SO FREAKING CUTE AND THEN LITERALLY A SEX GOD	4
537240106948890624	I remember my brother got me tickets to the b2k concert I literally cried the entire time	2
537240134794903552	@ProfessorFlynn not surprised a teacher who teaches out of outdated textbooks has no clue about cutting-edge science. #gobacktoschool #iro	-3
537240162724741120	he'll literally just come up to him and talk like a normal person and my dad goes "yeah yeah ok whatever " i want to punch him	-4
537240191090823168	When someone hurts you so bad and you just wonder what gives you the right to still speak to me? Felt violated all over again.	-3
537240247202234370	RT @Jazzbmxo: Want to have someone to speak to I'm so bored	-1
537240275228577792	Why is there shit all over my shirt? I literally lint rolled four times.	-3
537240303091347456	And now they run ads for ppl to join the CIA on TV and radio #Irony	-2
537240331134455808	@chaandbeti Oh wow you are such an expert on the matter. How old are you again? #Sarcasm @SheikhLarki	-3
537240359106273280	#irony India Is A Country Where Some Contribute To Poor By Facebook Like And Share.	-1
537240386419580928	RT @JeaneDidThat: Literally at this computer hollerin while I'm workin. I can't.	-2

Σχήμα 5. Δείγμα των δεδομένων ελέγχου με διακριτά σκορ

Επίσης, πρέπει να σημειωθεί πως οι βαθμολογίες των δεδομένων ελέγχου $\boldsymbol{S_{Ttest}}$ έγιναν γνωστές μετά την αξιολόγηση και αποτελούν διακριτά σκορ, όχι συνεχόμενα. Τέλος, στο $\boldsymbol{T_{test}}$ υπάρχει και μικρό ποσοστό tweets που δεν περιέχουν μεταφορικό λόγο.

5. ΠΕΡΙΓΡΑΦΗ ΣΥΣΤΗΜΑΤΟΣ

5.1ΣΚΟΠΟΣ

Σκοπός του συστήματος που αναπτύχθηκε και ονομάζεται *Figurative Text Analysis* είναι να υποστηρίξει την επίλυση του προβλήματος, όπως αυτό περιγράφηκε στα προηγούμενα κεφάλαια και όπως αυτό τέθηκε από τους διοργανωτές του SemEval 2015 Task 11 [58], αλλά να είναι και δομημένο με τέτοιο τρόπο ώστε να είναι εύκολο, με μικρές αλλαγές, να χρησιμοποιηθεί σε οποιαδήποτε παρόμοια διαδικασία ανάλυσης συναισθήματος σε tweets ή κείμενο[2].

5.2ΣΧΕΤΙΚΟΙ ΟΡΙΣΜΟΙ

Στη συνέχεια παρουσιάζονται και αναλύονται κάποιοι όροι που θα χρησιμοποιηθούν κατά τη διάρκεια της περιγραφής του συστήματος.

VECTOR SPACE MODEL

Η έννοια του vector space model (VSM) προέρχεται από την γεωμετρία. Αποτελεί τον τρόπο για να απεικονιστούν έγγραφα σε έναν πολυδιάστατο χώρο, ως διάνυσμα (vector)[15]. Μπορούμε να αναπαραστήσουμε το διάνυσμα με διάφορους τρόπους, αλλά ο πιο χρήσιμος και αποδοτικός είναι η χρήση του TF-IDF.

Το Term Frequency (TF) είναι η συχνότητα στο έγγραφο, ενώ το IDF είναι το αντίστροφο της συχνότητας του εγγράφου, το οποίο είναι ο αριθμός των εγγράφων στο σύνολο (corpus) όπου εμφανίζεται ο όρος.

$$\mathrm{tf}(t,d) = 0.5 + \frac{0.5 \times \mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

Εξίσωση 3. Term Frequency (TF)

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Εξίσωση 4. Inverse Document Frequency (IDF)

Στο TF-IDF, σημειώνουμε έναν όρο για το πόσο συχνά εμφανίζεται στο παρόν έγγραφο και πόσο συχνά εμφανίζεται σε όλο το σύνολο των εγγράφων (corpus). Αυτό μας δίνει μια ιδέα

[2] Ο κώδικας που αναπτύχθηκε στα πλαίσια της εργασίας βρίσκεται στο https://bitbucket.org/mkaranasou/figurative-text-analysis

για τα χαρακτηριστικά, που δεν είναι κοινά σε ολόκληρο το σύνολο και για αυτά που έχουν υψηλή συχνότητα.

$$\mathrm{tfidf}\left(t,d,D\right)=\mathrm{tf}\left(t,d\right)\times\mathrm{idf}\left(t,D\right)$$

Εξίσωση 5. TF-IDF

COUNTVECTORIZER

Κλάση του scikit-learn[3] η οποία μετατρέπει ένα σύνολο κειμένων σε έναν πίνακα από token counts. Η υλοποίηση αυτή παράγει έναν "αραιό" (sparse) πίνακα, ο οποίος περιέχει την καταμέτρηση των λέξεων που περιέχονται σε αυτά τα κείμενα. Εάν δεν δοθεί ένα λεξικό εκ των προτέρων και δεν χρησιμοποιηθεί ένας αναλυτής που να κάνει κάποιο feature selection τότε ο αριθμός των χαρακτηριστικών θα είναι ίδιος με το μέγεθος του λεξικού των αρχικών δεδομένων.

DICTVECTORIZER

Αποτελεί κλάση του scikit-learn, η οποία μετατρέπει λίστες από ζευγάρια χαρακτηριστικών με τις αντίστοιχες τιμές τους σε πίνακες διανυσμάτων σε μορφή Numpy arrays. Στην πράξη δηλαδή, μετατραπεί ένα python dictionary σε ένα πίνακα διανυσμάτων (vector array). Όταν οι τιμές των features είναι κειμενικές, ο transformer αυτός θα κάνει κωδικοποίηση των τιμών όπως περιγράφεται ακολούθως: ένα χαρακτηριστικό με δυαδική τιμή κατασκευάζεται για κάθε μία από τις πιθανές τιμές που μπορεί να πάρει αυτό το χαρακτηριστικό. Παραδείγματος χάριν, ένα χαρακτηριστικό "f" που μπορεί να πάρει τιμή "a" και "b" θα μετατραπεί σε δύο χαρακτηριστικά, "f=a" και "f=b". Τα χαρακτηριστικά που δεν υπάρχουν σε ένα δείγμα, θα έχουν μηδενική τιμή στο παραγόμενο πίνακα διανυσμάτων.

DISCRETIZATION

Ορισμός χαρακτηριστικών διακριτών τιμών που τμηματοποιούν τη συνεχή τιμή ενός χαρακτηριστικού σε ένα διακριτό σύνολο διαστημάτων[4], το οποίο, συνήθως πάντα επιφέρει ένα ποσοστό λάθους (discretization ή truncation error).

[3] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[4] https://en.wikipedia.org/wiki/Discretization_of_continuous_features

TFIDFTRANSFORMER

Αποτελεί κλάση του scikit–learn [5] η οποία μετατρέπει έναν πίνακα με αποτελέσματα καταμέτρησης (χαρακτηριστικών, λέξεων κλπ.) σε έναν κανονικοποιημένο πίνακα tf ή tf-idf (use_idf=False ή use_idf=True αντιστοίχως). Αυτό είναι μια συνηθισμένη μέθοδος στάθμισης όρων (term weighting scheme) στην ανάκτηση πληροφορίας, το οποίο έχει αποδειχθεί ότι έχει καλή εφαρμογή στην κατηγοριοποίηση κειμένων. Ο στόχος είναι να χρησιμοποιηθεί tf ή tf-idf αντί της απλής χρήσης της συχνότητας εμφάνισης μιας ένδειξης (token), π.χ. μιας λέξης, σε ένα σύνολο κειμένων, ώστε να μειωθεί το αντίκτυπο των ενδείξεων που εμφανίζονται πολύ συχνά και συνεπώς παρέχουν λιγότερη πληροφορία από αυτές που εμφανίζονται πολύ αραιά. Το tf-df υπολογίζεται με βάση την ακόλουθη συνάρτηση, αντί για tf * idf, όπως δηλαδή ορίζεται στην Εξίσωση 6, ώστε οι ενδείξεις που εμφανίζονται σε όλα τα κείμενα να μην αγνοηθούν εντελώς.

$$tfidf = tf * (idf + 1) = tf + tf * idf$$

Εξίσωση 6. Ο υπολογισμός tf-idf από τον TfidfVectorizer

PAIRWISE COSINE SIMILARITY

Ως Pairwise Cosine Similarity ορίζεται ο υπολογισμός εσωτερικού γινόμενου των κανονικοποιημένων (L2) διανυσμάτων[6]. Πιο συγκεκριμένα, εάν x και y είναι δυο σειρές από διανύσματα, το cosine similarity τους k ορίζεται ως:

$$k(x, y) = \frac{xy^\top}{||x||||y||}$$

Ονομάζεται cosine similarity, διότι η Ευκλείδεια (L2) κανονικοποίηση προβάλλει τα διανύσματα πάνω στο χώρο και το εσωτερικό γινόμενό τους είναι το συνημίτονο της μεταξύ των σημείων που υποδηλώνονται από τα διανύσματα.

CLASSIFIER FEATURES

Με τον όρο αυτό αναφέρονται τα χαρακτηριστικά που «μαθαίνει» ένας classifier με την εκπαίδευσή του. Παράδειγμα τέτοιων χαρακτηριστικών, όπως αυτά έχουν μετατραπεί για να είναι κατάλληλη είσοδος για τον αλγόριθμο, είναι τα ακόλουθα:

[5] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

[6] http://scikit-learn.org/stable/modules/metrics.html

```
['ABC=NN', 'AFC=NN', 'AFDC=NN', 'AL=NN', 'AP=NN', 'APB=NN', 'AR=NN', 'ASPCA=NN',
 'AS_GROUND_AS_VEHICLE=False', 'AS_GROUND_AS_VEHICLE=True', 'ATM=NN', 'AWOL=NN',
 'Aarhus=NN', 'Aaron=NN', 'Abdel=NN', 'Achilles=NN', 'Adah=NN', 'Adam=NN',
 'Adams=NN', 'Adele=NN', 'Adhara=NN', 'Adidas=NN', 'Adrea=NN', 'Adrian=NN',
'Advil=NN', 'Afghanistan=NN', 'Africa=NN', 'African=ADJ', 'Agna=NN',
'Aguilar=NN', 'Ahmadabad=NN', 'Aksel=NN', 'Alabama=NN', 'Alberich=NN',
'Alberta=NN', 'Albertina=NN', 'Albertine=NN', 'Albuquerque=NN', 'Alejandra=NN',
'Alex=NN', 'Alexander=NN', 'Alexia=NN', 'Alexis=NN', 'Algeria=NN',
'Alhambra=NN', 'Allah=NN', 'Allen=NN', 'Allhallows=VB', 'Allie=NN',
'Allstate=NN', 'Alonso=NN', 'Altair=NN', 'Amanda=NN', 'Ame=NN', 'Amelia=NN',
'America=NN', 'American=NN', 'Americanization=NN', 'Americans=NN', 'Ames=NN',
'Amish=NN', 'Amtrak=NN', 'Anatole=NN', 'Andersen=NN', 'Anderson=NN',
'Andra=NN', 'Andrea=NN', 'Andrew=NN', 'Andy=NN', 'Angelina=NN',
'Angelle=NN', 'Angelou=NN', 'Anjela=NN', 'Anna=NN', 'Annabelle=NN',
'Annetta=NN', 'Anny=NN', 'Antares=NN', 'Antigua=NN', 'Antonio=NN',
'Appleton=NN', 'April=NN', 'Arabic=NN', 'Arabs=NN', 'Archy=NN',
'Argentinian=ADJ', 'Argus=NN', 'Arius=NN', 'Arizona=NN', 'Arjuna=NN',
'Arkansas=NN', 'Arleyne=NN', 'Arlington=NN', 'Armour=NN', 'Arnaldo=NN',
'Arpanet=NN', 'Artemis=NN', 'Ashil=NN', 'Ashley=NN', 'Ashton=NN',
'Ashurbanipal=NN', 'Asia=NN', 'Asian=ADJ', 'Asians=NN', 'Assad=NN',
'Assam=NN', 'Aston=NN', 'Athens=NN', 'Atkins=NN', 'Atlanta=NN',
'Atlantic=NN', 'Au=NN', 'Audrey=NN', 'Augustus=NN', 'Aussie=NN',
'Austen=NN', 'Austin=NN', 'Australia=NN', 'Australian=ADJ',
'Australopithecus=NN', 'Austria=NN', 'Austrian=ADJ', 'Ava=NN', 'Avicenna=NN',
'Avictor=NN', 'Avila=NN', 'Axe=NN', 'Azania=NN', 'Aztec=NN', 'BB=NN', 'BBC=NN',
'BBQ=NN', 'BC=NN', 'BMW=NN', 'Babs=NN', 'Babylonian=NN', 'Bakersfield=NN',
'Baltimore=NN', 'Baluchistan=NN', 'Bamby=NN', 'Bangalore=NN', 'Barbara=NN',
'Barbra=NN', 'Barlow=NN', 'Bartlett=NN', 'Barton=NN', 'Batista=NN', 'Bauer=NN',
'Bea=NN', 'Becca=NN', 'Bechtel=NN', 'Becky=NN', 'Beelzebub=NN', 'Behan=NN',
'Beirut=NN',  'Belgium=NN', 'Bella=NN', 'Bellamy=NN', 'Belmont=NN', 'Ben=NN',
'Bengali=NN', 'Benghazi=NN', 'Benoit=NN', 'Benson=NN', 'Benton=NN',
'Berglund=NN', 'Bernard=NN', 'Bethany=NN', 'Bevon=NN', 'Bhutan=NN',
'Bierce=NN', 'Blaine=NN', 'Blake=NN', 'Boris=NN', 'Boston=NN', 'Boyle=NN',
'Bradley=NN', 'Brahma=NN', 'Brampton=NN', 'Brandi=NN', 'Brandon=NN',
'Brantley=NN', 'Brazil=NN', 'Breanne=NN', 'Breathalyzer=NN', 'Breckenridge=NN',
 'Brendan=NN', 'Brett=NN', 'Bridgewater=NN', 'Brie=NN', 'Brighton=NN',
 'Brillo=NN', 'Bristol=NN', 'Britain=NN', 'British=ADJ', 'Britney=NN',
 'Briton=NN', 'Brittany=NN', 'Broadway=NN', 'Brock=NN', 'Brooklyn=NN',
 'Brose=NN', 'Brownian=NN', 'Bruce=NN', 'Bruno=NN', 'Bryan=NN', 'Bryna=NN',
 'Buckingham=NN', 'Bulgaria=NN', 'Bundy=NN', 'Burbank=NN', 'Burlington=NN',
 'CBC=NN', 'CBS=NN', 'CD=NN', 'CDC=NN', 'CEO=NN', 'CFC=NN', 'CIA=NN', … ]
```

Σχήμα 6. Παράδειγμα classifier features

CLASSIFIER VOCABULARY

Ο όρος αυτός αναφέρεται στο «λεξικό» των χαρακτηριστικών που αποκτά ένα classifier μετά την εκπαίδευσή του. Παράδειγμα vocabulary είναι το ακόλουθο:

```
{
  'ran=VB': 10039,  'individual=ADJ': 6802,  'intellectual activities=NN': 6908,
  'shows=VB': 11231,  'spiders=NN': 11667,  'cashier=NN': 3018,
  'land=NN': 7341,  'carver=NN': 3012,  'lead=VB': 7425,  'band camp=ADJ': 2143,
  'mudtztbsqt=NN': 8343,  'woodwork=NN': 13627,  'locknut=NN': 7646,
  'hailing=VB': 6122,  'lookalike=VB': 7679,  'sickening=ADJ': 11250,
  'emotion=NN': 4799,  'abjure=NN': 1471,  'relieved=VB': 10270,
  'dragging=VB': 4558,  'sportscast=NN': 11711,  'apple=NN': 1846,
  'knight=NN': 7274,  'wetlands=NN': 13453,  'Suarez=NN': 1263,
  'supernatant=ADJ': 12116,  'Santa=NN': 1173,  'Taft=NN': 1287,
  'abolitionist=VB': 1473,  'ear buds=NN': 4681,  'burning=VB': 2817,
  'laurel=NN': 7392,  'teammates=NN': 12333,  'keen=ADJ': 7195,
  'shiner=NN': 11161,  'danger=NN': 4040,  'edition=NN': 4725,
  'Pete=NN': 1080,  'express=VB': 5050,  'fully=RB': 5639,
  'districts=NN': 4436,  'fake=ADJ': 5129,  'adoration=NN': 1583,
  'handicap=NN': 6148,  'revolving=VB': 10444,  'cleaning=VB': 3320,
  'determined=VB': 4255,  'footballers=NN': 5475,  'downhill=RB': 4540,
  'someone\\sll=NN': 11543, 'serration=NN': 11056,  'mental=ADJ': 8038,
  'unleash=VB': 13047,  'so funny=ADJ': 11485,  'aqbwmw=NN': 1865,
  'bannock=NN': 2165,  'club=NN': 3377,  'teen mom=ADJ': 12349,
  'scouts=NN': 10915,  'Waikiki=NN': 1395,  'looks fine=ADJ': 7682,
  'apocalypse=NN': 1822,  'Lufthansa=NN': 822,  'Madonna=NN': 842,
  'the blaze=NN': 12444,  's_word-4=somewhat_positive': 10718,
  'graduate=VB': 5942,  'braveness=NN': 2672,  'Franck=NN': 471,
  'nascent=ADJ': 8420,  'dicking=VB': 4286,  'amazed=VB': 1728,
  'tailgated=VB': 12249,  'hyper dunks=NN': 6619,  'fighting=VB': 5283,
  'gladyouguysareassholes=NN': 5811,  'cereal=NN': 3081,
  'ass=NN': 1932,  'dirt=NN': 4349,  'powering=VB': 9610,
  'purchase=VB': 9900,  'thriving=NN': 12540,  'fudge=NN': 5630,
  'total joke=ADJ': 12693,  'lays=VB': 7415,  'very classy=ADJ': 13181,
  'believing=VB': 2315,  'Enif=NN': 414,  'Duncan=NN': 385,
  'Boston=NN': 174,  'western journalism=NN': 13449,  'dog facts=NN': 4479,
  'schoolwork=NN': 10888,  'carefully=RB': 2981,  'tampon=NN': 12281,
  'marginalized=VB': 7876,  'triangles=NN': 12806,  'taxes=NN': 12315,
  'location=NN': 7640,  'souped=VB': 11602,  'kennel=NN': 7203,
  'simulation=NN': 11285,  'commercials=NN': 3493,  'glen=VB': 5818,
  'anyone=NN': 1811,  'thongs=NN': 12518,  'fighters=NN': 5282,
  'contact=NN': 3647,  'balling=VB': 2134,  'beach body=ADJ': 2246,
  'Broadway=NN': 198,  'signing=NN': 11270,  'shotgunned=VB': 11207,
  'hanger=NN': 6158,  'ins=VB': 6854,  'pleasedontkickmeinthenuts=NN': 9453,
  'upper=ADJ': 13096,  'starting=VB': 11806,  'wastefulness=NN': 13353,
  'dealer=NN': 4084,  'friends forever=ADJ': 5592,  'freethinking=VB': 5574,
  'city=NN': 3280,
  …
}
```


Σχήμα 7. Παράδειγμα classifier vocabulary

5.3 ΔΕΔΟΜΕΝΑ

5.3.1 ΕΞΩΤΕΡΙΚΑ ΔΕΔΟΜΕΝΑ

TWITTER EMOTICONS

Για τον υπολογισμό των χαρακτηριστικών POS_EMOTICON και NEG_EMOTICON, χρησιμοποιήθηκαν τα πρώτα δέκα όσον αφορά τη συχνότητα χρήσης τους emoticons από [59], μαζί με όποιες παραλλαγές είχαν νόημα για να καλυφθούν τυχόν λάθος πληκτρολογήσεις και παραλλαγές. Ο χωρισμός σε θετικά (Positive) και αρνητικά (Negative) έγινε χειροκίνητα.

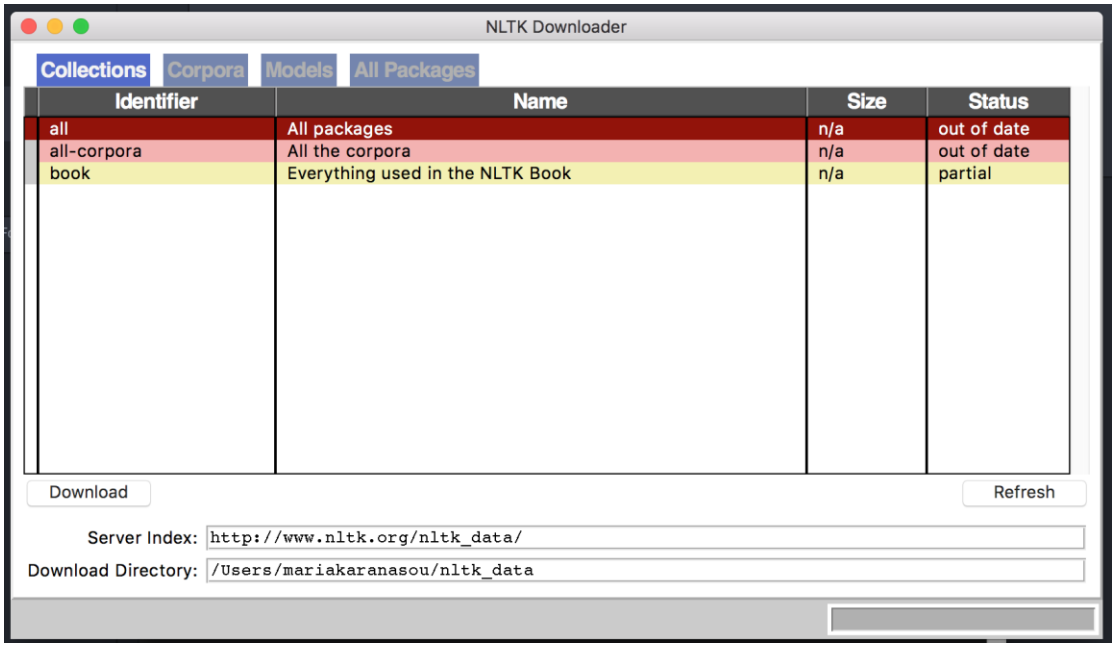
POSITIVE	NEGATIVE
:)	:-\
:-)	:\
:-D	:-(
:))	:-((
:-))	:(
(:	:o
:D	:O
:)	D:
;)	=/
;-)	=(
XD	:'-(
=]	:\\
;D	:/
:]	:S
:o)	

Πίνακας 4. Τα emoticons που χρησιμοποιήθηκαν, κατηγοριοποιημένα ως προς το συναίσθημα.

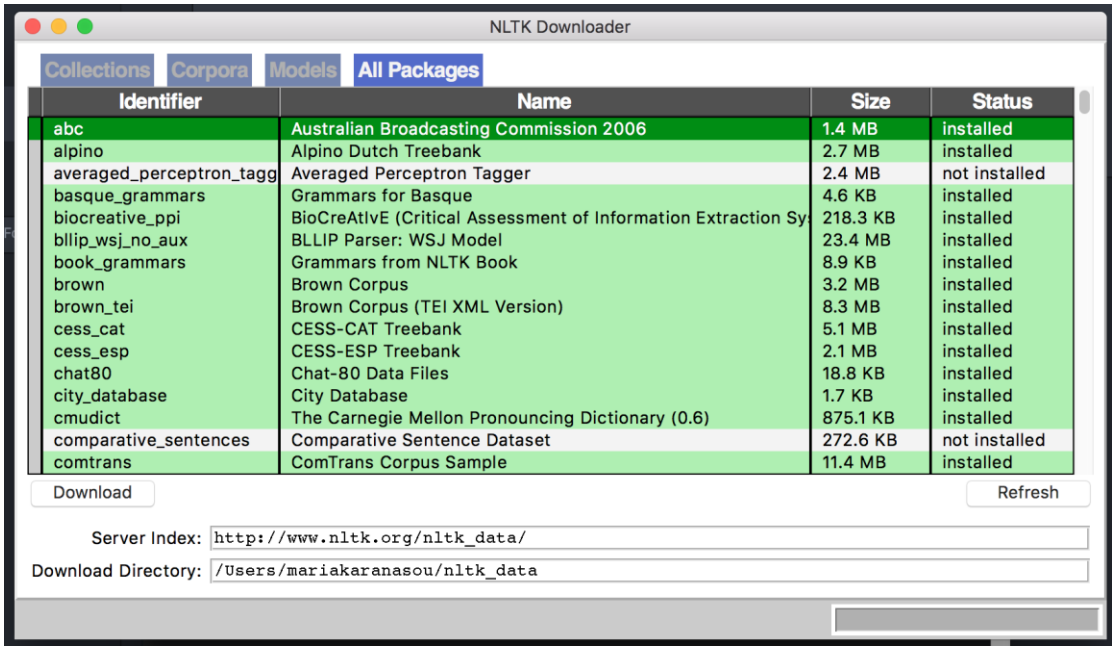
WORDNET 3.0

Από το WordNet χρησιμοποιήθηκαν δεδομένα για την εύρεση συνώνυμων και για το semantic similarity (Resnik, Wup, Path και Lin) [60], τα οποία περιγράφονται στη συνέχεια.

Για να γίνει χρήση των δεδομένων του WordNet, εκτελέστηκε η εντολή *nltk.download(),* η οποία ξεκινάει ένα python πρόγραμμα το οποίο κατεβάζει τα επιλεγμένα δεδομένα σε συγκεκριμένο φάκελο (nltk_data), όπως φαίνεται στο Σχήμα 8 και στο Σχήμα 9.



Σχήμα 8. NLTK downloader



Σχήμα 9. NLTK Packages

PATH-BASED

PATH SIMILARITY MEASURE

Το Path Similarity ή Shortest Path Similarity αναθέτει μια βαθμολογία στο εύρος [0,1] που βασίζεται στο πιο σύντομο μονοπάτι που συνδέει τις έννοιες στην ιεραρχία τύπων (hypernymy)[7]. Στη περίπτωση που δεν υπάρχει κάποιο κοινό μονοπάτι, η βαθμολογία παίρνει τιμή -1 [14] [18] [19].

Για μία συγκεκριμένη έκδοση του WordNet, το deep_max, δηλαδή το μέγιστο βάθος στην ταξονομία, είναι συγκεκριμένη τιμή. Η ομοιότητα μεταξύ δύο εννοιών (c1, c2) είναι η συνάρτηση του μικρότερου μήκους μονοπατιού μεταξύ τους, len(c1,c2). [14] [18]

$$sim_{path}(c_1, c_2) = 2 * deep\_\max - len(c_1, c_2)$$

Εξίσωση 7. Μέτρο ομοιότητας Shortest Path

WU-PALMER SIMILARITY MEASURE

Οι Wu και Palmer εισήγαγαν ένα κλιμακωτό μέτρο ομοιότητας. Το μέτρο αυτό παίρνει τη θέση των εννοιών (concepts) c1, c2 στην ταξονομία σε σχέση με τη θέση της πιο συγκεκριμένης έννοιας lso(c1, c2). Υποθέτει πως η ομοιότητα μεταξύ των δύο εννοιών είναι συναρτήσει του μήκους του μονοπατιού και του βάθους.

$$sim_{WP}(c_1, c_2) = \frac{2 * depth(lso(c_1, c_2))}{len(c_1, c_2) + 2 * depth(lso((c_1, c_2))}$$

Εξίσωση 8. Μέτρο ομοιότητας Wu-Palmer

όπου η ομοιότητα μεταξύ δύο εννοιών (c1, c2) είναι η συνάρτηση της απόστασής τους και της πιο κοντινής κοινής έννοιας που είναι πρόγονος των c1, c2 (least common subsumer – LCS [19], όπου δηλαδή υπάρχει σχέση is-a μεταξύ των c1 - LCS και c2 - LCS. Εάν το lso(c1,c2) είναι η "ρίζα" του δέντρου, τότε το βάθος depth(lso(c1,c2)) είναι ίσο με 1 και συνεπώς simWP(c1,c2) >0. Εάν οι δύο έννοιες έχουν κοινό νόημα, τότε το c1, το c2 και το lso(c1, c2) είναι ο ίδιος κόμβος και το μήκος του μονοπατιού len(c1,c2) ισούται με μηδέν. Σε αυτή την περίπτωση, η συνάρτηση simWP(c1,c2) παίρνει τη μέγιστη τιμή της, ίση με 1. Σε όποια άλλη περίπτωση ισχύουν 0<depth(lso(c1,c2))< deep_max, 0<len(c1,c2)< 2*deep_max, όπου deep_max το

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

μέγιστο βάθος στην ταξονομία, κάτι που σημαίνει ότι οι τιμές της simWP(c1,c2) βρίσκονται στο διάστημα (0, 1].) [14] [18] [19].

INFORMATION CONTENT-BASED

Γίνεται η υπόθεση ότι κάθε έννοια περικλείει πολλή πληροφορία στο WordNet. Οι μετρικές ομοιότητας αυτής της κατηγορίας βασίζονται στο περιεχόμενο της πληροφορίας (Information Content - IC) κάθε έννοιας. Όσο ποιο κοινή πληροφορία υπάρχει μεταξύ των δύο εννοιών τόσο πιο όμοιες είναι.

RESNIK SIMILARITY MEASURE

Το 1995, ο Resnik πρότεινε ένα μέτρο ομοιότητας που βασίζεται στο Information Content. Υποθέτει ότι για δύο έννοιες c1, c2, η ομοιότητα εξαρτάται από το IC που τις ορίζει στην ταξονομία. Όπως φαίνεται από την Εξίσωση 9, οι τιμές εξαρτώνται μόνο από την χαμηλότερη στην ταξονομία έννοια που αποτελεί πρόγονος των c1, c2.

$$sim_{Resnik}(c_1, c_2) = -\log p(lso(c_1, c_2)) = IC(lso(c_1, c_2))$$

Εξίσωση 9. Μέτρο ομοιότητας Resnik

LIN SIMILARITY MEASURE

Το μέτρο που έχει προταθεί από τον Lin χρησιμοποιεί και την ποσότητα πληροφορίας που χρειάζεται για να οριστούν τα κοινά χαρακτηριστικά των δύο εννοιών c1, c2, και την πληροφορία για την πλήρη περιγραφή τους. Καθώς ισχύει ότι IC(lso(c1,c2)) <=IC(c1) και IC(lso(c1,c2)) <=IC(c2), οι τιμές που μπορεί να πάρει το μέτρο αυτό είναι στο διάστημα (0,1] [14] [18] [19].

$$sim_{Lin}(c_1, c_2) = \frac{2 * IC(lso(c_1, c_2)))}{IC\ (c_1) + IC\ (c_2)}$$

Εξίσωση 10. Μέτρο ομοιότητας Lin

SENTIWORDNET

Για την εξαγωγή του χαρακτηριστικού της πολικότητας των λέξεων, χρησιμοποιήθηκε το λεξικό SentiWordNet. Το αρχείο του SentiWordNet[8] επεξεργάστηκε ώστε να καταλήξει στη μορφή μια γραμμή ανά λέξη και αποθηκεύτηκε στον ομώνυμο πίνακα της βάσης SentiFeed στη μορφή:

[8] http://sentiwordnet.isti.cnr.it/

Id: Auto-increment id.

SentiWordNetId: Το id που δίνεται από το SentiWordNet. Λέξεις που ανήκουν στο ίδιο synset έχουν το ίδιο SentiWordNetId.

Category: το μέρος του λόγου στο οποίο ανήκει η λέξη, “n”, “v”, “a”, “r”.

Pos: η θετική βαθμολογία της λέξης[9].

Neg: η αρνητική βαθμολογία της λέξης που ανήκει σε ένα synset.

Systerms: η λέξη που ανήκει σε ένα synset.

Description: η περιγραφή της λέξης[8].

SentimentAssessment: δηλαδή ο υπολογισμός (1+Pos-Neg), πιο συγκεκριμένα

$$\text{SentimentAssessment}_w = 1 + wScore_p - wScore_n$$

Εξίσωση 11. Ο υπολογισμός της βαθμολογίας για κάθε λέξη στο SentiWordNet

με $wScore_p$, η θετική βαθμολογία της λέξης w και $wScore_n$ η αρνητική βαθμολογία της λέξης w.

Η βαθμολογία αυτή ονομάζεται και prior polarity, διότι εκ των προτέρων ορίζεται η πολικότητα των λέξεων, η οποία μπορεί να χρησιμοποιηθεί για να προβλεφθεί η πολικότητα ενός κειμένου, όπως π.χ. ενός tweet.

GATE POS-TAG MODEL

Μετά από χρήση των κλασσικών pos-taggers του nltk παρατηρήθηκε ότι λόγω της ιδιαιτερότητας των tweets οι προβλέψεις δεν είναι σωστές σε μεγάλο ποσοστό. Για παράδειγμα, η λέξη «apple» αναγνωριζόταν κατά κόρον ως ρήμα. Καθώς το Part-of-Speech tagging σε tweets είναι ακόμα πιο δύσκολο από ότι σε κανονικά κείμενα, επιλέχθηκε να χρησιμοποιηθεί ένα μοντέλο για pos-tagging, το οποίο είναι ειδικευμένο στην περίπτωση των tweets. Το μοντέλο αυτό έχει 91% accuracy στα tokens του συνόλου δεδομένων στο οποίο αξιολογήθηκε. Αναγνωρίζει εκτός από τα tags που αναφέρθηκαν στο κεφάλαιο σχετικά με το NLP και επιπλέον tags (Πίνακας 5) που είναι πιο στοχευμένα στο κείμενο του tweet, όπως για παράδειγμα HT για τα hashtags [61].

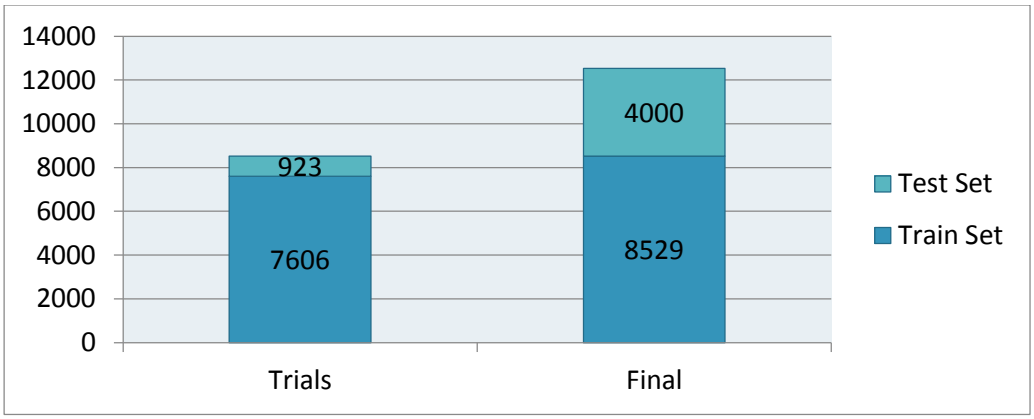
[9] και κατ’ επέκταση του synset στο οποίο ανήκει

#	TAG	DESCRIPTION
1	USR	User
2	URL	Hyperlink
3	HT	Hashtag
4	RT	Retweet

Πίνακας 5 Τα επιπλέον Part-of-Speech tags που υποστηρίζονται από τον Gate POS Tagger

5.3.2 ΔΕΔΟΜΕΝΑ SEMEVAL 2015 TASK 11

Το σύνολο των δεδομένων που παρέχονται από το SemEval 2015 στα πλαίσια του Task 11, αποτελείται από 9000 tweets τα οποία είναι πλούσια σε μεταφορικό λόγο και έχουν συλλεχθεί με τη χρήση της κατηγοριοποίησης των ίδιων των χρηστών μέσω hastags, παραδείγματος χάριν #sarcasm, #irony. Τα δεδομένα που προορίζονται για την φάση εκπαίδευσης, χωρίζονται σε 90% για εκπαίδευση (training data) και 10% δεδομένα ελέγχου (test data). Δεδομένης της εφήμερης φύσης των tweets, ήταν δυνατόν να συλλεχθούν μόνο 8529 tweets, 7606 από τα δεδομένα εκπαίδευσης και 923 από τα δεδομένα ελέγχου. Στο σύνολο των συλλεχθέντων tweets, το 8,2% είναι κατηγοριοποιημένα ως θετικά, το 85% είναι κατηγοριοποιημένα ως αρνητικά και το 6,6% ως ουδέτερα. Ακολουθεί μια μικρή ανάλυση σχετικά με τη σύσταση των δεδομένων και την πολικότητα των tweets.

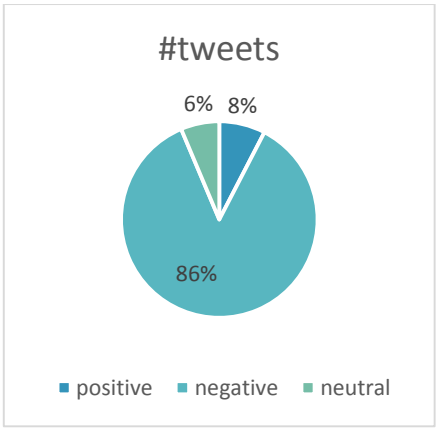


Σχήμα 10. Συνοπτικά ο αριθμός των tweets ανά data set

ΔΕΔΟΜΕΝΑ ΕΚΠΑΙΔΕΥΣΗΣ - TRAIN SET

ΚΑΤΗΓΟΡΙΑ	#TWEETS
POSITIVE	578
NEGATIVE	6545
NEUTRAL	483
	7606

Πίνακας 6. Ο αριθμός των tweet ανά κατηγορία - Δεδομένα εκπαίδευσης

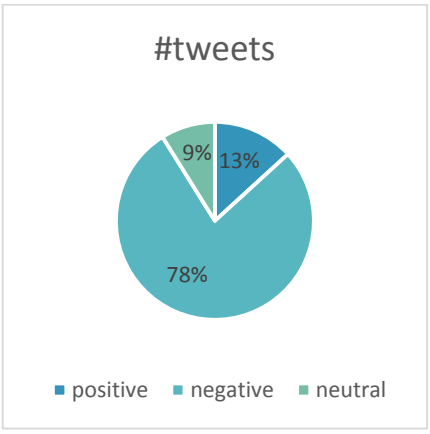


Σχήμα 11. Το ποσοστά των tweets των δεδομένων εκπαίδευσης ανά κατηγορία

ΔΕΔΟΜΕΝΑ ΔΟΚΙΜΗΣ ΣΤΗ ΦΑΣΗ ΕΚΠΑΙΔΕΥΣΗΣ - TEST SET

ΚΑΤΗΓΟΡΙΑ	#TWEETS
POSITIVE	122
NEGATIVE	719
NEUTRAL	82
	923

Πίνακας 7. Ο αριθμός των tweet ανά κατηγορία - Δεδομένα εκπαίδευσης

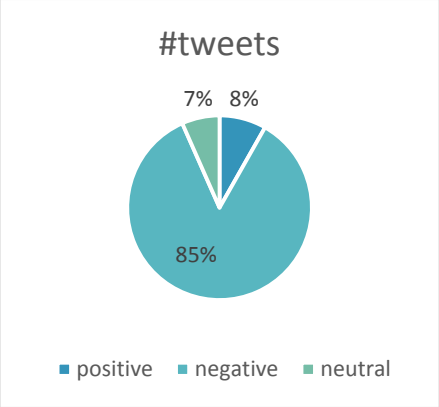


Σχήμα 12. Το ποσοστά των tweets των τελικών δεδομένων δοκιμής στη φάση εκπαίδευσης εκπαίδευσης ανά κατηγορία

ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ ΕΚΠΑΙΔΕΥΣΗΣ – FINAL TRAIN SET

ΚΑΤΗΓΟΡΙΑ	#TWEETS
POSITIVE	700
NEGATIVE	7264
NEUTRAL	565
	8529

Πίνακας 8. Ο αριθμός των tweet ανά κατηγορία - Δεδομένα εκπαίδευσης

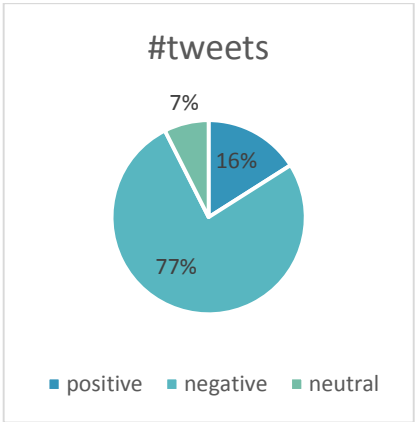


Σχήμα 13. Το ποσοστά των tweets των δεδομένα εκπαίδευσης ανά κατηγορία

ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ ΔΟΚΙΜΗΣ - FINAL TEST SET

ΚΑΤΗΓΟΡΙΑ	#TWEETS
POSITIVE	640
NEGATIVE	3062
NEUTRAL	298
	4000

Πίνακας 9. Τα tweets ανά κατηγορία - Τελικά Δεδομένα Δοκιμής

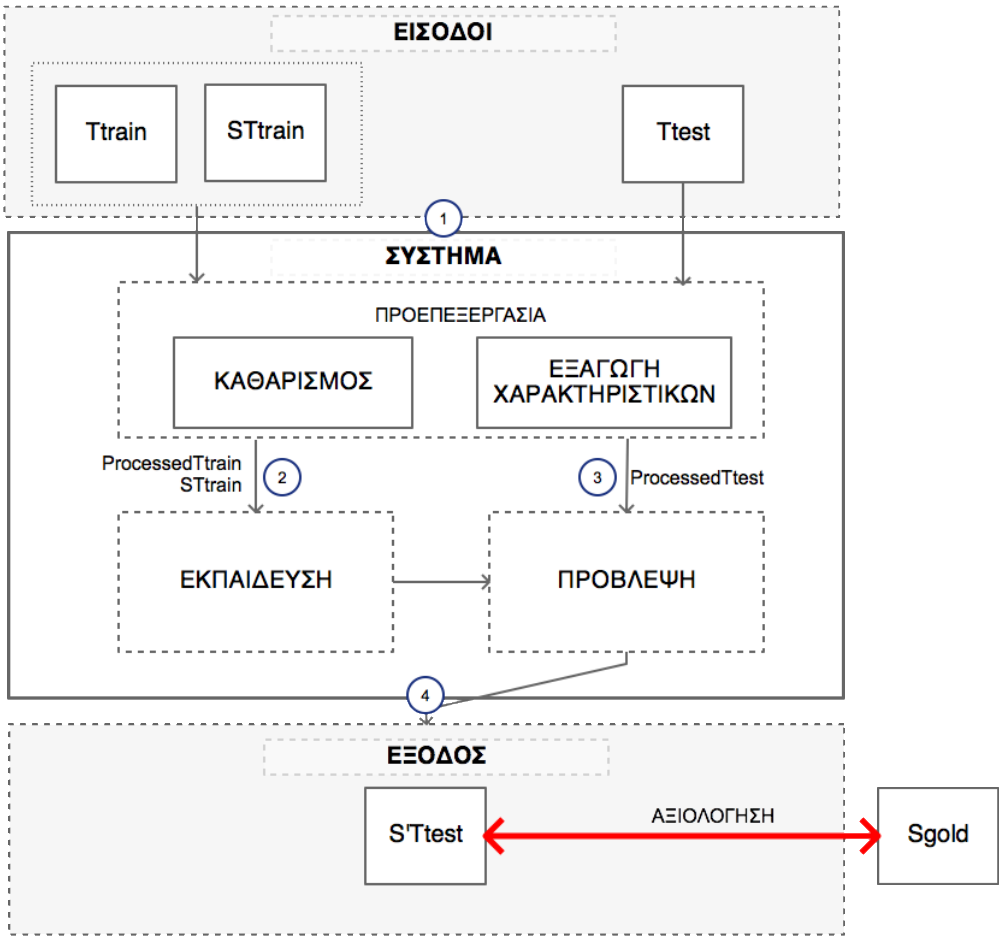


Σχήμα 14. Τα ποσοστά των tweets των τελικών δεδομένων δοκιμής ανά κατηγορία

5.4 ΔΟΜΗ ΣΥΣΤΗΜΑΤΟΣ

Σε αυτό το κεφάλαιο, περιγράφεται η δομή του συστήματος που αναπτύχθηκε, πως είναι οργανωμένος ο κώδικας και οι βασικές λειτουργίες του. Στο Σχήμα 15 φαίνεται η βασική λειτουργία του.

Το σύστημα δέχεται ως είσοδο τα Ttrain, Strain, Ttest. Τα Ttrain και Ttest περνάνε από επεξεργασία, κατά τη διάρκεια της οποίας γίνεται καθαρισμός και εξαγωγή των απαραίτητων για τη διαδικασία χαρακτηριστικών (1). Στη συνέχεια, τα επεξεργασμένα tweets ProcessedTtrain, μαζί με τα αντίστοιχα STtrain χρησιμοποιούνται για την εκπαίδευση ενός classifier (2). Αφού έχει ολοκληρωθεί η εκπαίδευση, το ProcessedTtest σύνολο επεξεργασμένων tweets χρησιμοποιείται στη διαδικασία της πρόβλεψης (3). Το αποτέλεσμα S'Ttest (4) συγκρίνεται με το Sgold που είναι το σύνολο των αναμενόμενων score για να αξιολογηθεί το σύστημα.



Σχήμα 15. Αφαιρετική απεικόνιση της λειτουργίας του συστήματος

5.4.1 ΠΑΚΕΤΑ ΚΑΙ ΚΛΑΣΕΙΣ

Το σύστημα είναι δομημένο σε πακέτα (python packages), για καλύτερη διαχείριση. Το πακέτο FigurativeTextAnalysis περιέχει όλα τα πακέτα που περιγράφονται ακολούθως.

MODELS

TWEET

Η κλάση Tweet αποτελεί την αναπαράσταση ενός tweet (κείμενο και id), μαζί με τα χαρακτηριστικά που είναι απαραίτητα στη διαδικασία ανάλυσης συναισθήματος, όπως για παράδειγμα, τα Part-of-Speech tags του tweet.

TEXTTAGGER

Η κλάση υπεύθυνη για την εξαγωγή μορφολογικών χαρακτηριστικών με τη χρήση regex patterns. Η function *tag_text* εκτελεί ένα loop για όλα τα μορφολογικά features που υπάρχουν ως keys στο *_tag* dictionary και ενημερώνει τις τιμές του αντιστοίχως.

HASHTAGHANDLER

Η κλάση υπεύθυνη για τον χειρισμό των hashtags. Η μέθοδος *handle* παίρνοντας ως είσοδο ένα hashtag και χρησιμοποιούμενη αναδρομικά, υπολογίζει το συνολικό συναίσθημα των hashtags ενός tweet (Εξίσωση 12). Ακολούθως περιγράφεται η διαδικασία με τη χρήση ψευδοκώδικα.

```
# PSEUDOCODE FOR HASHTAG SENTIMENT CALCULATION
IF every letter is capital:
    THEN
    IF is spelled correctly:
        THEN goto calculate
    ELSE spellcheck:
        IF is spelled correctly:
            THEN goto calculate
        ELSE:
            return NEUTRAL
        ENDIF
    ENDIF
ELSE:
    IF can be split by capitals:
        THEN
        FOR each split word:
            IF is spelled correctly:
                THEN goto calculate
            ELSE spellcheck:
                IF is spelled correctly:
                    THEN goto calculate
                ELSE:
                    return NEUTRAL
            ENDIF
ENDIF
return NEUTRAL

calculate:

    FOR each hashtag:
        get score from SentiWordNet

    IF count of negatives >= count of neutrals and count of negatives > 0:
        THEN return NEGATIVE
    ELSE IF count of neutrals > count of negatives:
        THEN return POSITIVE
    ELSE:
        return NEUTRAL
```

Σχήμα 16. Ψευδοκώδικας υπολογισμού του συνολικού συναισθήματος των hashtags των tweets

TEXTCLEANER

Η κλάση αυτή περικλείει όλες τις απαραίτητες functions για τον «καθαρισμό» ενός tweet.

Συγκεκριμένα:
Μέθοδος remove_non_ascii_chars: Απομάκρυνση των χαρακτήρων που δεν μπορούν να τυπωθουν (non-printable).

Μέθοδος remove_RT: Απομάκρυνση της ένδειξης retweet.

Μέθοδος identify_and_remove_laughter: Απομάκρυνση της ένδειξης γέλιου ή παρεμφερών ενδείξεων, π.χ. haha, lol κλπ.

Μέθοδος split_sentences: Το tweet χωρίζεται σε προτάσεις.

Μέθοδος identify_negations: Γίνεται έλεγχος για negations, π.χ. not, don't κλπ.

Μέθοδος has_capitals: Γίνεται έλεγχος για λέξεις με κεφαλαία όλα τα γράμματα, π.χ. NOT

Μέθοδος remove_links: Γίνεται απομάκρυνση των υπερσυνδέσμων.

Μέθοδος store_and_remove_emoticons: Γίνεται απομάκρυνση των βασικών emoticons που λαμβάνουμε υπόψιν.

Μέθοδος remove_reference: Γίνεται απομάκρυνση των @user.

Μέθοδος remove_special_chars: Γίνεται απομάκρυνση όλων των χαρακτήρων που δεν είναι γράμματα, εκτός από τον χαρακτήρα για το κενό μεταξύ των λέξεων.

Μέθοδος fix_space: Λόγω της επεξεργασίας που έχει γίνει έως τώρα, μπορεί να έχουν προκύψει περισσότερο από ένα κενά στο κείμενο του tweet, οπότε γίνεται κανονικοποίηση των κενών για την πιο σωστή επεξεργασία.

Μέθοδος split_words: Το κείμενο του tweet, στην μορφή μετά την επεξεργασία έως αυτό το σημείο, χωρίζεται σε λέξεις.

Μέθοδος convert_to_lower: Στη λίστα λέξεων που προέκυψε, γίνεται μετατροπή όλων των χαρακτήρων σε μικρά.

Μέθοδος remove_multiples: Για κάθε λέξη, γίνεται έλεγχος για το εάν υπάρχουν περισσότεροι από δύο συνεχόμενοι ίδιοι χαρακτήρες με τη χρήση regex pattern, έχοντας τους δύο χαρακτήρες το maximum ώστε να έχουμε μια ορθογραφικά σωστή λέξη. Στην περίπτωση που εντοπιστούν τέτοιες περιπτώσεις, γίνεται επεξεργασία αυτών των λέξεων, ώστε οι πολλαπλοί χαρακτήρες να μειωθούν σε δύο και στη συνέχεια γίνεται προσπάθεια διόρθωσής τους με τη χρήση spell-checker [62].

Μέθοδος remove_stop_words: Γίνεται απομάκρυνση των λέξεων που δεν προσθέτουν αξία στη διαδικασία της ανάλυσης συναισθήματος. Για να επιτευχθεί αυτό χρησιμοποιείται η λίστα από stop words του nltk και μια εξωτερική λίστα.

Μέθοδος set_final_tweet:
Τέλος, το τελικό κείμενο του tweet μετά την επεξεργασία, ανατίθεται στο πεδίο clean_tweet της κλάσης Tweet.

POSTAGGER

Η κλάση αυτή, φορτώνει το GATE Pos-tag μοντέλο, το οποίο έχει επεξεργαστεί και αποθηκευθεί σε μορφή python dictionary σε txt αρχείο, σε έναν custom pos-tagger του nltk[10], για να χρησιμοποιηθεί στη διαδικασία του feature extraction.

CLASSIFIER

Η κλάση που λειτουργεί ως wrapper των classifiers του scikit-learn. Σκοπός της είναι να διευκολύνει τις δοκιμές και τις αλλαγές παραμέτρων με συνέπεια.

TRIAL

Η κλάση αυτή, αποτελεί την κύρια κλάση του προγράμματος. Κάνοντας instantiate μια Trial class, ξεκινάει μια δοκιμή με τις παραμέτρους που έχουν οριστεί.

TRIALSCORE

Ένα TrialScore αποτελεί το αποτέλεσμα μιας δοκιμής για ένα tweet. Συγκεκριμένα, στην βάση κρατείται το Id του tweet, το predicted score και το actual score.

SELECTEDFEATURES

Σκοπός της κλάσης είναι να γίνεται μια εγγραφή στον ομώνυμο πίνακα της βάσης Sentifeed, για κάθε δοκιμή που γίνεται ώστε να καταγράφονται τα features που χρησιμοποιήθηκαν.

BASICMEASURES

Η κλάση αυτή παρέχει μεθόδους για τον υπολογισμό και την αποθήκευση των μετρικών που περιγράφονται στο κεφάλαιο 8.1. Χρησιμοποιεί τις μεθόδους του scikit-learn και δεδομένου μιας λίστας από προβλέψεις και της λίστας των πραγματικών σκορ, μπορεί να υπολογίσει το cosine similarity, το accuracy κλπ.

[10] http://www.nltk.org/api/nltk.tag.html#nltk.tag.sequential.UnigramTagger

LABELNOTFOUND

Χρησιμοποιείται στα πλαίσια του discretization, όταν δεν βρίσκεται η τιμή που ζητήθηκε.

INVALIDCONFIGURATION

Χρησιμοποιείται όταν σε κάποια από τις κλάσεις έχει γίνει λάθος παραμετροποίηση, πράγμα που σημαίνει fatal error.

PROCESSORS

TWEETPROCESSOR

Η κλάση αυτή χρησιμοποιείται για την ενορχήστρωση της διαδικασίας. Είναι υπεύθυνη για την ανάκτηση των δεδομένων από τη βάση και τη μετατροπή τους από κείμενο σε python dictionary (χρησιμοποιώντας την ast.literal_eval της python). Επιπλέον, είναι υπεύθυνη για το post-processing των feature dictionaries, ώστε να χρησιμοποιηθούν μόνο τα επελεγμένα κάθε φορά features στη διαδικασία.

HELPERS

GLOBALS

Περιέχει global μεταβλητές και είναι γενική χρήσης. Παραδείγματος χάριν, έχει όλα τα enums που χρησιμοποιούνται, κάνει instantiate τον logger, συνδέεται με την MySQL μέσω της κλάσης MySQLConnector και κρατάει το instance αυτό στη μεταβλητή myslq_connection, φορτώνει το μοντέλο του Gate-PoS-tagger σε μια μεταβλητή για χρήση από την POSTagger κλάση, γενικότερα φροντίζει για όλα όσα είναι απαραίτητα για την ορθή λειτουργία του προγράμματος, ώστε αυτά να μπορέσουν να χρησιμοποιηθούν από τις υπόλοιπες κλάσεις.

ENCHANTSPELLCHECKER

Η κλάση που λειτουργεί ως wrapper για τον enchant spell-checker [62]. Για τη χρήση του απαιτείται το λεξικό της γλώσσας για την οποία πρόκειται να γίνει ορθογραφικός έλεγχος. Στη παρούσα εργασία χρησιμοποιήθηκε αγγλικό λεξικό[11], το οποίο υπάρχει στον φάκελο data της εφαρμογής[12].

[11] https://wiki.openoffice.org/wiki/Dictionaries

[12] https://pythonhosted.org/pyenchant/tutorial.html#adding-language-dictionaries

DATABASE

MYSQLDATABASECONNECTOR

Η κλάση που λειτουργεί ως wrapper της python-mysql βιβλιοθήκης[13], με όσες μεθόδους χρειάζονται για την αλληλεπίδραση με την MySQL βάση.

REDISCONNECTOR

Η κλάση που λειτουργεί ως wrapper της redis βιβλιοθήκης[14], με όσες functions χρειάζονται για την αλληλεπίδραση με την Redis (χρησιμοποιείται στα πλαίσια του feedback στην δικτυακή εφαρμογή που περιγράφεται σε επόμενο κεφάλαιο.

TESTS

Σε αυτό το πακέτο, περιέχονται κάποια unit tests για τον έλεγχο ορθής λειτουργίας του προγράμματος αλλά και για την διεξαγωγή δοκιμών.

LOGS (ΦΑΚΕΛΟΣ)

Για την αποθήκευση log αρχείων υπό την μορφή {application_name}_date_time.log

DATA (ΦΑΚΕΛΟΣ)

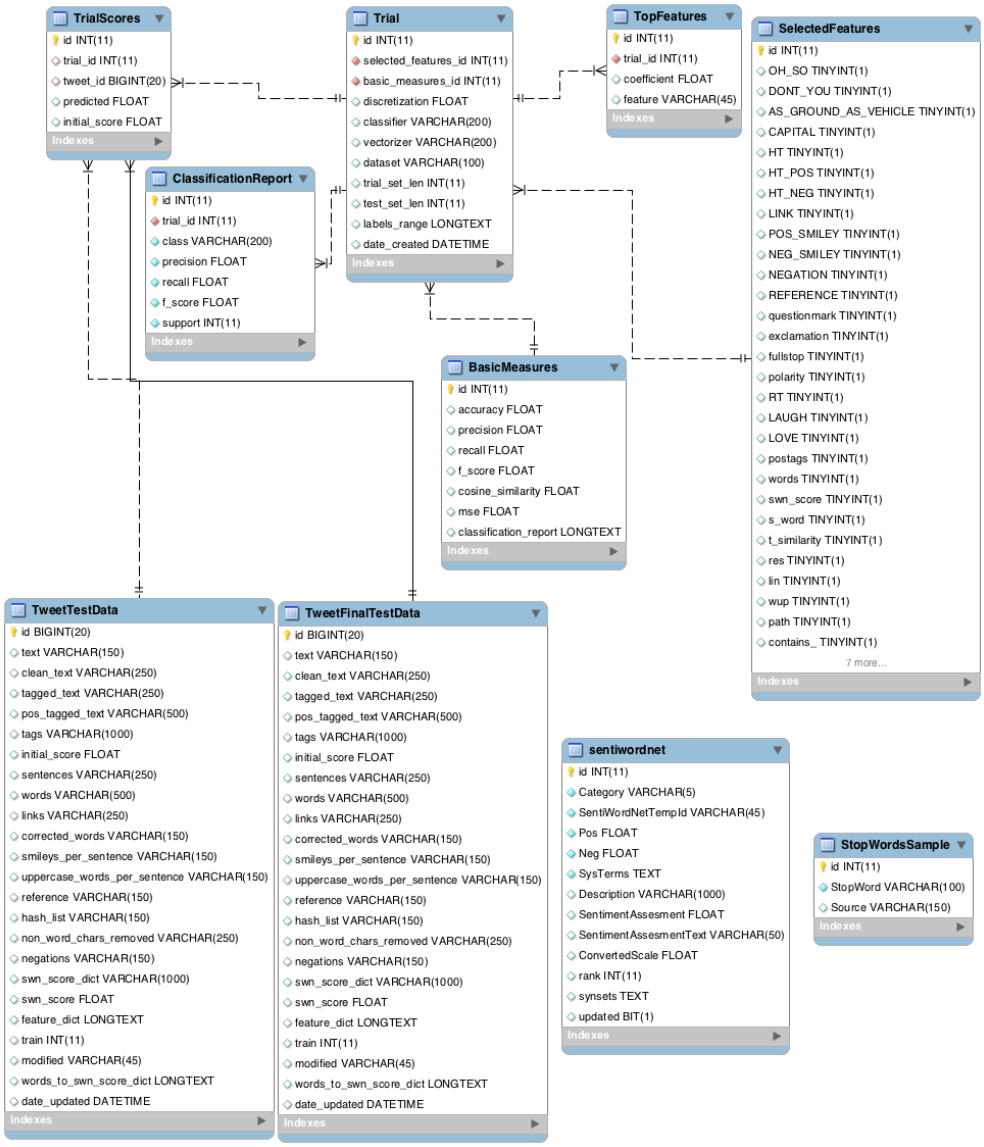
Για την αποθήκευση σχετικών δεδομένων, όπως παραδείγματος χάριν, το μοντέλο του POS tagger.

5.5 ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ

Η βάση της εφαρμογής ονομάζεται *Sentifeed* είναι πολύ απλή και αποτελείται από τους πίνακες που φαίνονται στο παρακάτω διάγραμμα Οντοτήτων-Συσχετίσεων (ER). Σε γενικές γραμμές, είναι απαραίτητος ένας πίνακας για την αποθήκευση του SentiWordNet, ένας για τα Stop words, ένας για την αποθήκευση των Tweets και οι υπόλοιποι πίνακες αφορούν στην αποθήκευση των δεδομένων των πειραμάτων, όπως π.χ. τα επελεγμένα χαρακτηριστικά για κάθε δοκιμή, τον classifier και τον vectorizer και τα αποτελέσματα. Πρακτικά χρησιμοποιούνται δύο πίνακες, ένας για την αποθήκευση των 8526 tweets για την φάση δοκιμών και ένας που περιέχει τα 4000 τελικά tweets ελέγχου.

[13] https://pypi.python.org/pypi/MySQL-python/1.2.5

[14] https://pypi.python.org/pypi/redis/2.10.5

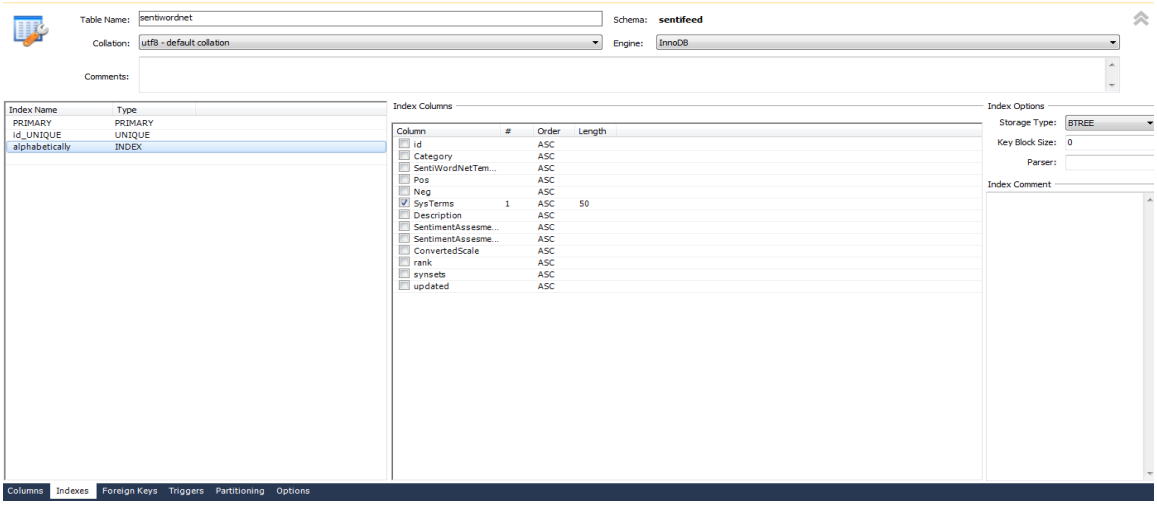


Σχήμα 17. ER Διάγραμμα ER της Sentifeed βάσης

Για την βελτίωση της απόδοσης του συστήματος, δημιουργήθηκε ένα index στον πίνακα SentiWordNet, στο πεδίο SysTerms, όπως φαίνεται στις ακόλουθες εικόνες.

id	Category	SentiWordNetTempId	Pos	Neg	SysTerms	Description	SentimentAssesment	ConvertedScale	rank
00000000001	a	00001740	0.125	0	able	(usually followed by `to') having the necessary means or skill or know-how o…	1.125	0.625	1
00000000002	a	00002098	0	0.75	unable	(usually followed by `to') not having the necessary means or skill or know-ho…	0.25	-3.75	1
00000000003	a	00002312	0	0	dorsal	facing away from the axis of an organ or organism; "the abaxial surface of a…	1	0	2
00000000004	a	00002312	0	0	abaxial	facing away from the axis of an organ or organism; "the abaxial surface of a…	1	0	1
00000000005	a	00002527	0	0	ventral	nearest to or facing toward the axis of an organ or organism; "the upper sid…	1	0	2
00000000006	a	00002527	0	0	adaxial	nearest to or facing toward the axis of an organ or organism; "the upper sid…	1	0	1
00000000007	a	00002730	0	0	acroscopic	facing or on the side toward the apex	1	0	1
00000000008	a	00002843	0	0	basiscopic	facing or on the side toward the base	1	0	1
00000000009	a	00002956	0	0	abducting	especially of muscles; drawing away from the midline of the body or from an …	1	0	1
00000000010	a	00002956	0	0	abducent	especially of muscles; drawing away from the midline of the body or from an …	1	0	1
00000000011	a	00003131	0	0	adductive	especially of muscles; bringing together or drawing toward the midline of the…	1	0	1

Σχήμα 18. Ο πίνακας που περιέχει τα δεδομένα του SentiWordNet



Σχήμα 19. Η δημιουργία του index στο πεδίο SysTerms

5.6 ΕΞΩΤΕΡΙΚΕΣ ΒΙΒΛΙΟΘΗΚΕΣ ΚΑΙ ΕΡΓΑΛΕΙΑ ΑΝΑΠΤΥΞΗΣ

Για την ανάπτυξη του προγράμματος χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python (έκδοση 2.7.10, 32bit)[15] και τα εργαλεία και οι βιβλιοθήκες που περιγράφονται ακολούθως.

5.6.1 ΕΡΓΑΛΕΙΑ ΑΝΑΠΤΥΞΗΣ

GIT

Το Git είναι ένα σύστημα ελέγχου εκδόσεων λογισμικού (σύστημα ελέγχου αναθεωρήσεων ή σύστημα ελέγχου πηγαίου κώδικα) με έμφαση στην ταχύτητα, στην ακεραιότητα των δεδομένων και στην υποστήριξη για κατανεμημένες μη γραμμικές ροές εργασίας. Σχεδιάστηκε και αναπτύχθηκε αρχικά από τον Λίνους Τόρβαλντς για τη ανάπτυξη του πυρήνα Linux το 2005 και έχει γίνει από τότε το πιο διαδεδομένο σύστημα ελέγχου εκδόσεων για ανάπτυξη λογισμικού.[5] Είναι Ελεύθερο λογισμικό που διανέμεται κάτω από τους όρους της έκδοσης 2 της Γενικής Άδειας Δημόσιας Χρήσης GNU. Χρησιμοποιήθηκε για τον έλεγχο των εκδόσεων όσων αναπτύχθηκαν στα πλαίσια της εργασίας[16] [63].

[15] https://www.python.org

[16] https://git-scm.com

BITBUCKET – SOURCETREE

Το BitBucket είναι ένα κατανεμημένο σύστημα ελέγχου εκδόσεων λογισμικού που διευκολύνει την συνεργασία μεταξύ μελών ομάδων και βασίζεται στο Git[17]. Το SourceTree αποτελεί μια ελεύθερη έκδοση Git και Mercurial client για Windows και Mac[18].

Χρησιμοποιήθηκαν και τα δύο στα πλαίσια της ανάπτυξης των εφαρμογών της παρούσας εργασίας, ώστε να γίνεται καλύτερη διαχείριση λογισμικού.

PYCHARM IDE

Το PyCharm είναι ένα Ολοκληρωμένο Περιβάλλον Ανάπτυξης (Integrated Development Environment (IDE)) το οποίο υποστηρίζει την ανάπτυξη εφαρμογών με τη γλώσσα προγραμματισμού Python. Παρέχει ανάλυση κώδικα, γραφική διεπαφή για τον debugger, unit tester, ενσωμάτωση με VCS, όπως το Git και υποστήριξης στην ανάπτυξη δικτυακών εφαρμογών με τη χρήση frameworks όπως το Django. Αναπτύσσεται από την Τσέχικη εταιρία JetBrains. Είναι cross-platform, δηλαδή είναι δυνατόν να χρησιμοποιηθεί σε Windows, Mac OS X και Linux. Οι διαθέσιμες εκδόσεις είναι η Professional και η Community[19].

Για την ανάπτυξη των εφαρμογών χρησιμοποιήθηκε η πιο πρόσφατη Professional έκδοση του PyCharm (Pycharm 5.1).

MYSQL SERVER (MYSQL COMMUNITY EDITION)

Η MySQL Community Edition είναι μια ελεύθερη προς χρήση έκδοση της δημοφιλούς open source βάσης δεδομένων MySQL. Είναι διαθέσιμη με άδεια χρήσης GPL και υποστηρίζεται από μεγάλο αριθμό προγραμματιστών ελεύθερου λογισμικού[20].

XAMPP

Το XAMPP είναι μια ελεύθερη προς χρήση, εύκολη στην εγκατάσταση διανομή του Apache, η οποία περιέχει MariaDB, PHP και Perl[21]. Χρησιμοποιήθηκε για την εύκολη εγκατάσταση και διαχείριση του MySQL Server.

---

[17] https://bitbucket.org

[18] https://www.sourcetreeapp.com

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

[20] https://www.mysql.com/products/community/ , https://www.jetbrains.com/pycharm/

[21] https://www.apachefriends.org/index.html

MYSQL WORKBENCH

Το MySQL Workbench είναι ένα εργαλείο για αρχιτέκτονες βάσεων δεδομένων, προγραμματιστές λογισμικού και διαχειριστές βάσεων δεδομένων (DBAs). Προσφέρει μοντελοποίηση δεδομένων, ανάπτυξη σε SQL και διάφορα διαχειριστικά εργαλεία για την παραμετροποίηση του MySLQ server, της διαχείρισης χρηστών, της δημιουργίας αντιγράφων ασφαλείας και πολλά άλλα. Είναι διαθέσιμο για Windows, Linux και Mac OS X [22] . Χρησιμοποιήθηκε για την πρόσβαση στον MySQL Server και τη δημιουργία και τον χειρισμό της βάσης.

5.6.2 ΕΞΩΤΕΡΙΚΕΣ ΒΙΒΛΙΟΘΗΚΕΣ

NATURAL LANGUAGE TOOLKIT (NLTK)

Η βιβλιοθήκη NLTK είναι μια κορυφαία πλατφόρμα για τη δημιουργία προγραμμάτων σε Python για την διευκόλυνση αλληλεπίδρασης και χρήσης δεδομένων ανθρώπινης γλώσσας. Παρέχει πρόσβαση σε πάνω από 50 συλλογές λεξιλογικών πόρων, όπως το WordNet, μαζί με μια σειρά από βιβλιοθήκες επεξεργασίας κειμένου για την ταξινόμηση (classification), κατακερμάτιση (tokenization), οι οποίες προκύπτουν, αφαίρεση προθέματος και κατάληξης (stemming), την ανάλυση, και σημασιολογική εκλογίκευση, wrappers για βιβλιοθήκες NLP, και ένα ενεργό φόρουμ συζήτησης [14].

Η NLTK συμπεριλαμβάνει το αγγλικό WordNet με 155287 λέξεις και 117659 ομάδες συνωνύμων (synsets) [15].

Στα πλαίσια του συστήματος που αναπτύχθηκε, χρησιμοποιήθηκε για tokenization, stop words, εύρεση συνωνύμων (synonyms), υπολογισμό της ομοιότητας και συνάφειας λέξεων (semantic similarity), stemming και Part-of-Speech Tagging.

NUMPY

Η βιβλιοθήκη Numpy αποτελεί βασικό δομικό στοιχείο για τον επιστημονικό προγραμματισμό (scientific computing) στην Python. Περιέχει, μεταξύ άλλων, ένα αντικείμενο απεικόνισης και χειρισμού πολυδιάστατων πινάκων, εργαλεία ενσωμάτωσης C/C++ και Fortran κώδικα, χρήσιμη γραμμική άλγεβρα, μετασχηματισμούς Fourier και random number capabilities. Εκτός από τις προφανείς χρήσεις της, η βιβλιοθήκη αυτή μπορεί να χρησιμοποιηθεί ως ένα αποδοτικό πολυδιάστατο container of generic data, δίνοντας τη δυνατότητα να οριστούν αυθαίρετες δομές δεδομένων. Αυτό σημαίνει ότι το Numpy μπορεί να ενσωματώσει έναν μεγάλο αριθμό και μια μεγάλη ποικιλία βάσεων δεδομένων, με γρήγορο και αδιάλειπτο τρόπο. Η άδεια χρήσης λογισμικού του είναι BSD και επιτρέπει την επαναχρησιμοποίηση της

[22] https://www.mysql.com/products/workbench/

βιβλιοθήκης με λίγους περιορισμούς. Χρησιμοποιήθηκε σε συνδυασμό με SciPy και Scikit-learn στη διαδικασία μετατροπής των feature σε vector arrays[23].

SCIPY

Η βιβλιοθήκη SciPy είναι μια ανοιχτού κώδικα Python βιβλιοθήκη για την υποστήριξη των μαθηματικών, της επιστήμης και της εφαρμοσμένης μηχανικής (engineering)[24].

SCIKIT-LEARN

Ένα απλό και αποδοτικό σύνολο από εργαλεία για εξόρυξη και ανάλυση δεδομένων. Είναι επαναχρησιμοποιήσιμη σε αρκετά πλαισια και σε εμπορικές εφαρμογές με BSD άδεια χρήσης λογισμικού. Η βιβλιοθήκη αυτή είναι χτισμένη πάνω στις βιβλιοθήκες Numpy, Scipy και matplotlib. Χρησιμοποιήθηκε για το vectorization, το feature selection και το classification[25].

PYTHON-MYSQL

Διεπαφή για την πρόσβαση και τον χειρισμό MySQL βάσεων δεδομένων στην γλώσσα προγραμματισμού Python. Ανάμεσα στα χαρακτηριστικά της είναι η συμβατότητα με το PEP-0249 που ορίζει το API για βάσεις της Python, thread safety (τα threads δεν μπλοκάρουν το ένα το άλλο). Οι εκδόσεις MySQL που υποστηρίζονται είναι MySQL-3.23 έως 5.5 και Python-2.4 έως 2.7. Η Python 3.0 θα υποστηριχθεί μελλοντικά. Τέλος, υπάρχει υποστήριξη και για το PyPy. Η βιβλιοθήκη python-mysql είναι ελεύθερο λογισμικό και χρησιμοποιήθηκε για την δημιουργία wrapper για την πρόσβαση και εκτέλεση CRUD λειτουργιών στην SentiFeed βάση δεδομένων[26].

5.7 ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ

5.7.1 ΑΝΑΚΤΗΣΗ ΔΕΔΟΜΕΝΩΝ

Για την ανάκτηση των tweets χρησιμοποιήθηκε το python script που δόθηκε από τους διοργανωτές του SemEval 2015 task 11[27], με μικρές παραλλαγές ώστε να είναι δυνατόν να τρέξει όσες φορές χρειάζεται το script για να μαζευτούν όσο το δυνατόν περισσότερα tweets. Ο λόγος είναι ότι τα tweets μπορούν εύκολα να διαγραφούν ή να γίνουν ιδιωτικά, άρα μη προσβάσιμα. Αυτό σημαίνει ότι είναι πιθανόν να μην βρεθούν όλα τα tweets. Η υλοποίηση βρίσκεται στο αρχείο figurative_data_retriever.py του πακέτου helpers.

---

[23] http://www.numpy.org

[24] http://scipy.org

[25] http://scikit-learn.org/stable/

[26] https://pypi.python.org/pypi/MySQL-python/1.2.5

[27] http://alt.qcri.org/semeval2015/task11/data/uploads/download_tweets.py

5.7.2 ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ - PREPROCESSING

Λόγω του ότι τα tweets περιέχουν πολύ θόρυβο, όπως για παράδειγμα, λέξεις που δεν προσφέρουν καμία πληροφορία στη διαδικασία της ανάλυσης συναισθήματος, υπερσυνδέσμους, λέξεις με ορθογραφικά λάθη, λέξεις με περισσότερα από δύο στη σειρά επαναλαμβανόμενα γράμματα, είναι απαραίτητη μια προ-επεξεργασία δεδομένων ώστε να αφαιρεθούν και να διορθωθούν, όσο το δυνατόν, τέτοιου είδους θέματα. Επίσης, για την χρησιμοποίηση των tweets ως είσοδο σε έναν classifier, πρέπει να γίνει εξαγωγή συγκεκριμένων χαρακτηριστικών τους σε μορφή που να είναι εύκολα μετατρέψιμη στην κατάλληλη είσοδο για τον classifier.

ΔΙΑΔΙΚΑΣΙΑ ΚΑΘΑΡΙΣΜΟΥ

Τα βήματα για τον καθαρισμό ενός tweet είναι τα ακόλουθα:

- Αφαίρεση non-ascii χαρακτήρων
- Αφαίρεση της ένδειξης retweet (RT)
- Αφαίρεση κοινών εκφράσεων γέλιου, όπως π.χ. «haha»
- Διαχωρισμός προτάσεων
- Αφαίρεση αρνήσεων (negations), όπως π.χ.
- Αφαίρεση υπερσυνδέσμων (urls, hyperlinks)
- Αφαίρεση των πιο κοινών emoticons (βλ. Πίνακας 11)
- Αφαίρεση αναφοράς σε χρήστη (@user)
- Αφαίρεση ειδικών χαρακτήρων όπως “/”, “}” κλπ
- Κανονικοποίηση των κενών που έχουν προκύψει στο κείμενο έως αυτό το σημείο είτε λόγω της επεξεργασίας είτε λόγω του κειμένου
- Διαχωρισμός λέξεων (tokenization με τη χρήση του nlkt word tokenizer)
- Μετατροπή όλων των γραμμάτων σε μικρά
- Αφαίρεση πάνω από δύο συνεχόμενων χαρακτήρων – γίνεται η υπόθεση ότι μια λέξη με έως και δύο συνεχόμενους ίδιοι χαρακτήρες είναι πιθανόν να είναι ορθογραφικά σωστοί. Στο σημείο αυτό, οι πολλαπλοί συνεχόμενοι χαρακτήρες σε μια λέξη, όπως π.χ. “looooove” αφαιρούνται ώστε να μείνουν μόνο δύο συνεχόμενοι ίδιοι χαρακτήρες, δηλαδή “loove", και στη συνέχεια χρησιμοποιείται ο spell checker σε μια προσπάθεια εξαγωγής της σωστής λέξης, στο παράδειγμα αυτό της “love”.
- Αφαιρούνται οι stop words (και αυτές που περιέχονται στον πίνακα stopwords και αυτές που δίνονται από το nltk)
- Γίνεται ορθογραφικός έλεγχος στις λέξεις.

Αρχικό κείμενο του Tweet:

> “Oh, life is not fair? I appreciate you texting me this, from your iPhone, while on vacation in Hawaii. You're so right.”

Τελικό κείμενο του Tweet:

> "life not fair appreciate texting this your iphone on vacation hawaii so right"

ΔΙΑΔΙΚΑΣΙΑ ΕΞΑΓΩΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ (FEATURE EXTRACTION)

Τα μορφολογικά χαρακτηριστικά, τα οποία φαίνονται στον Πίνακας 11 με τη σημείωση *(μ)*, εξάγονται από το tweet πρίν τη διαδικασία του καθαρισμού, καθώς μετά τον καθαρισμό δεν θα υπάρχουν, όπως για παράδειγμα, τα emoticons και οι υπερσύνδεσμοι. Για την εξαγωγή τους χρησιμοποιούνται regex patterns και υπεύθυνη κλάση για την διαδικασία αυτή είναι ο POSTagger, ο οποίος χρησιμοποιεί και τον HashTagHandler για τα χαρακτηριστικά που εμπλέκουν hashtags. Η Εξίσωση 12 περιγράφει τον υπολογισμό της συνολικής πολικότητας (θετική, αρνητική, ουδέτερη) των hashtags ενός tweet. Ουσιαστικά, υπολογίζεται ένας μέσος όρος της πολικότητας με βάση την ακόλουθη εξίσωση, λαμβάνοντας υπόψιν την καταμέτρηση των θετικών ($c(htPos)$) και των αρνητικών ($c(htNeg)$) hashtags.

$$HTEm_t = \begin{cases} HT\_pos, & c(htPos) > c(htNeg) > 0 \\ HT\_neu, & c(htPos) = c(htNeg) = 0 \\ HT\_neg, & c(htNeg) \geq c(htPos) > 0 \end{cases}$$

Εξίσωση 12. Ο υπολογισμός της πολικότητας των hashtags ενός tweet

Τα υπόλοιπα χαρακτηριστικά χρειάζονται το "καθαρισμένο" tweet. Παραδείγματος χάριν, τα Part-of-speech tags υπολογίζονται στην λίστα λέξεων που προκύπτει από τον καθαρισμό. Το ίδιο και το SentiWordNet score και τα διάφορα similarity measures.

Σχήμα 20.Τα Part-of-speech tags του παραδείγματος

Οι ακόλουθες εξισώσεις περιγράφουν τον υπολογισμό της εννοιολογικής «ομοιομορφίας», ενός tweet, δηλαδή το πόσο συναφείς είναι οι λέξεις μεταξύ τους. Για κάθε ένα από τα τέσσερα similarity metrics που χρησιμοποιούνται, υπολογίζεται το sim_t, το οποίο αποτελείται από τον μέσο όρο του similarity όλων των λέξεων που ανήκουν στις τέσσερις κύριες κατηγορίες του WordNet, Nouns (N), Verbs (V), Adjectives (A), Adverbs (R). Η Εξίσωση 13 περιγράφει τον υπολογισμό του similarity μεταξύ των λέξεων που ανήκουν στην κατηγορία A $\in$ {N, V, A, R}, $\max(sim(A_i, A_{i+1}))$ είναι η μέγιστη ομοιότητα μεταξύ των λέξεων A_i, A_{i+1} και των συνώνυμών τους και $c(A)$ ο αριθμός των λέξεων που ανήκουν στην κατηγορία A.

$$sim_t = \frac{\sum sim_V + \sum sim_N + \sum sim_A + \sum sim_R}{c(V) + c(N) + c(A) + c(R)}$$

Εξίσωση 13. Υπολογισμός του συνολικού Similarity ενός tweet

$$sim_A = \left[\max\left(sim(A_1, A_2)\right), \ldots, \max\left(sim(A_{n-1}, A_n)\right)\right]$$

Εξίσωση 14. Υπολογισμός Similarity ενός συνόλου λέξεων που ανήκουν στην ίδια κατηγορία.

Στο ακόλουθο παράδειγμα, φαίνεται η διαδικασία για την εύρεση των ζευγαριών για τον υπολογισμό του semantic similarity του tweet *«Oh, life is not fair? I appreciate you texting me this, from your iPhone, while on vacation in Hawaii. You're so right.»*. Στα "Nouns" ανήκουν οι λέξεις life, vacation, Hawaii και phone, για τις οποίες συγκεντρώνονται όλα τα συνώνυμά τους. Το ίδιο ισχύει και για τις άλλες τρεις κατηγορίες λέξεων, "Verbs", "Adjectives", "Adverbs". Στη συνέχεια γίνεται υπολογισμός του $\max\left(sim(A_1, A_2)\right)$, π.χ. του $\max(sim(life,\ vacation),\ sim(animation, vacation)\ \ldots)$, του $\max\left(sim(vacation,\ hawaii)\right)$, του $\max(sim(hawaii,\ phone), sim(hawaii, telephone)\ \ldots)$ κλπ, και συνεπώς του sim_A, το οποίο στην περίπτωσή μας είναι το sim_N, όπου N=Nouns. Οι ίδιοι υπολογισμοί γίνονται για τις υπόλοιπες κατηγορίες. Τέλος, υπολογίζεται το sim_t, το συνολικό similarity του tweet, το οποίο είναι το άθροισμα των sim_A, δια το πλήθος των κατηγοριών, δηλαδή 4. Ως βελτίωση στη διαδικασία αυτή, θα μπορούσε να δημιουργηθεί ένα χαρακτηριστικό που να δείχνει το similarity ανά κατηγορία λέξεων, καθώς κάποιες κατηγορίες είναι πιθανόν να έχουν μικρό similarity, κάτι που με τον υπολογισμό του μέσου όρου χάνεται.

```
"NOUNS":
# --- Pair 1 --- #
life word1 [Synset('life.n.01'), Synset('life.n.02'), Synset('life.n.03'),
          Synset('animation.n.01'), Synset('life.n.05'), Synset('life.n.06'),
          Synset('life.n.07'), Synset('life.n.08'), Synset('liveliness.n.02'),
          Synset('life.n.10'), Synset('life.n.11'), Synset('biography.n.01'),
          Synset('life.n.13'), Synset('life_sentence.n.01')]
vacation word2 [Synset('vacation.n.01'), Synset('vacation.n.02'),
          Synset('vacation.v.01')]

# --- Pair 2 --- #
vacation word1 [Synset('vacation.n.01'), Synset('vacation.n.02'),
          Synset('vacation.v.01')]
Hawaii word2 [Synset('hawaii.n.01'), Synset('hawaii.n.02')]

# --- Pair 3 --- #
Hawaii word1 [Synset('hawaii.n.01'), Synset('hawaii.n.02')]
phone word2 [Synset('telephone.n.01'), Synset('phone.n.02'),
          Synset('earphone.n.01'), Synset('call.v.03')]

"VERBS":
# --- Pair 1 --- #
testing word1 [Synset('testing.n.01'), Synset('testing.n.02'),
               Synset('examination.n.05'), Synset('test.v.01'),
               Synset('screen.v.01'), Synset('quiz.v.01'), Synset('test.v.04'),
               Synset('test.v.05'), Synset('test.v.06'), Synset('test.v.07')]
appreciate word2 [Synset('appreciate.v.01'), Synset('appreciate.v.02'),
               Synset('prize.v.01'), Synset('appreciate.v.04'),
               Synset('appreciate.v.05')]
```

Σχήμα 21. Παράδειγμα ζευγών υπολογισμού semantic similarity

```
"ADJECTIVES":
# --- Pair 1 --- #
right word1 [Synset('right.n.01'), Synset('right.n.02'),
            Synset('right_field.n.01'), Synset('right.n.04'),
            Synset('right.n.05'), Synset('right.n.06'), Synset('right.n.07'),
            Synset('right.n.08'), Synset('right.v.01'), Synset('right.v.02'),
            Synset('right.v.03'), Synset('correct.v.01'), Synset('right.a.01'),
            Synset('correct.a.01'), Synset('correct.s.02'), Synset('right.a.04')
            Synset('right.a.05'), Synset('proper.s.04'), Synset('right.a.07'),
            Synset('right.s.08'), Synset('right.s.09'), Synset('correct.s.03'),
            Synset('right.s.11'), Synset('right.s.12'), Synset('good.s.12'),
            Synset('veracious.s.02'), Synset('right.r.01'), Synset('right.r.02')
            Synset('right.r.03'), Synset('right.r.04'), Synset('properly.r.01'),
            Synset('right.r.06'), Synset('right.r.07'), Synset('mighty.r.01'),
            Synset('justly.r.02'), Synset('correctly.r.01')]
fair word2 [Synset('carnival.n.03'), Synset('fair.n.02'), Synset('fair.n.03'),
            Synset('bazaar.n.03'), Synset('fair.v.01'), Synset('fair.a.01'),
            Synset('fair.s.02'), Synset('bonny.s.01'), Synset('fair.a.04'),
            Synset('average.s.03'), Synset('fair.s.06'), Synset('clean.s.11'),
            Synset('honest.s.07'), Synset('fair.s.09'), Synset('fair.s.10'),
            Synset('fairly.r.03'), Synset('fairly.r.02')]

"ADVERBS":
# --- Pair 1 --- #
so word1 [Synset('sol.n.03'), Synset('so.r.01'), Synset('so.r.02'),
          Synset('so.r.03'), Synset('so.r.04'), Synset('so.r.05'),
          Synset('thus.r.02'), Synset('so.r.07'), Synset('then.r.01'),
          Synset('therefore.r.01'), Synset('indeed.r.01')]
not word2 [Synset('not.r.01')]
```

Σχήμα 22. Παράδειγμα ζευγών υπολογισμού semantic similarity (συνέχεια)

```
SIMILARITY TYPE: Path
Nouns
    Looking for  life  <-- and --> vacation
        life ss1
        [Synset('life.n.01'), Synset('life.n.02'), Synset('life.n.03'),
        Synset('animation.n.01'), Synset('life.n.05'), Synset('life.n.06'),
        Synset('life.n.07'), Synset('life.n.08'), Synset('liveliness.n.02'),
        Synset('life.n.10'), Synset('life.n.11'), Synset('biography.n.01'),
        Synset('life.n.13'), Synset('life_sentence.n.01')]
        vacation ss2
        [Synset('vacation.n.01'), Synset('vacation.n.02'),
        Synset('vacation.v.01')]
    Pair Similarity is:  0.2
    Looking for  vacation  <-- and --> Hawaii
        vacation ss1
        [Synset('vacation.n.01'), Synset('vacation.n.02'),
        Synset('vacation.v.01')]
        Hawaii ss2 [
        Synset('hawaii.n.01'), Synset('hawaii.n.02')]
    Pair Similarity is:  0.111111111111
    Looking for  Hawaii  <-- and --> phone
        Hawaii ss1 [
        Synset('hawaii.n.01'), Synset('hawaii.n.02')]
        phone ss2 [Synset('telephone.n.01'), Synset('phone.n.02'),
        Synset('earphone.n.01'), Synset('call.v.03')]
    Pair Similarity is:  0.1
Verbs
    Looking for  testing  <-- and --> appreciate
        testing ss1 [Synset('testing.n.01'), Synset('testing.n.02'),
        Synset('examination.n.05'), Synset('test.v.01'), Synset('screen.v.01'),
        Synset('quiz.v.01'), Synset('test.v.04'), Synset('test.v.05'),
        Synset('test.v.06'), Synset('test.v.07')]
        appreciate ss2 [Synset('appreciate.v.01'), Synset('appreciate.v.02'),
        Synset('prize.v.01'), Synset('appreciate.v.04'),
        Synset('appreciate.v.05')]
    Pair Similarity is:  0.2
```

Σχήμα 23. Παράδειγμα υπολογισμού Shortest Path semantic similarity

```
Adjectives
    Looking for  right  <-- and --> fair
        right ss1 [Synset('right.n.01'), Synset('right.n.02'),
        Synset('right_field.n.01'), Synset('right.n.04'), Synset('right.n.05'),
        Synset('right.n.06'), Synset('right.n.07'), Synset('right.n.08'),
        Synset('right.v.01'), Synset('right.v.02'), Synset('right.v.03'),
        Synset('correct.v.01'), Synset('right.a.01'), Synset('correct.a.01'),
        Synset('correct.s.02'), Synset('right.a.04'), Synset('right.a.05'),
        Synset('proper.s.04'), Synset('right.a.07'), Synset('right.s.08'),
        Synset('right.s.09'), Synset('correct.s.03'), Synset('right.s.11'),
        Synset('right.s.12'), Synset('good.s.12'), Synset('veracious.s.02'),
        Synset('right.r.01'), Synset('right.r.02'), Synset('right.r.03'),
        Synset('right.r.04'), Synset('properly.r.01'), Synset('right.r.06'),
        Synset('right.r.07'), Synset('mighty.r.01'), Synset('justly.r.02'),
        Synset('correctly.r.01')]
        fair ss2
        [Synset('carnival.n.03'), Synset('fair.n.02'), Synset('fair.n.03'),
        Synset('bazaar.n.03'), Synset('fair.v.01'), Synset('fair.a.01'),
        Synset('fair.s.02'), Synset('bonny.s.01'), Synset('fair.a.04'),
        Synset('average.s.03'), Synset('fair.s.06'), Synset('clean.s.11'),
        Synset('honest.s.07'), Synset('fair.s.09'), Synset('fair.s.10'),
        Synset('fairly.r.03'), Synset('fairly.r.02')]
    Pair Similarity is:  0.25
Adverbs
    Looking for  so  <-- and --> not
        so ss1
        [Synset('sol.n.03'), Synset('so.r.01'), Synset('so.r.02'),
        Synset('so.r.03'), Synset('so.r.04'), Synset('so.r.05'),
        Synset('thus.r.02'), Synset('so.r.07'), Synset('then.r.01'),
        Synset('therefore.r.01'), Synset('indeed.r.01')]
        not ss2
        [Synset('not.r.01')]
    Pair Similarity is:  None
```

Σχήμα 24. Παράδειγμα υπολογισμού Shortest Path semantic similarity (συνέχεια)

```
#Final Similarity
#==============
Resnik 0.0
Lin 0.0
Wu-Palmer 0.477142857143
Path 0.172222222222
```

Σχήμα 25. Τα τελικά αποτελέσματα semantic similarity του παραδείγματος ανά similarity metric

Η Εξίσωση 15 δίνει τον υπολογισμό της βαθμολογίας SentiWordNet ($swnScore_{wi}$) για κάθε λέξη:

$$swnScore_{wi} = \frac{\sum_{k=1}^{j} 1 + wScore(i,\ k)_p - wScore(\ i,\ k)_n}{j}$$

Εξίσωση 15. Ο υπολογισμός του SentiWordNet score για κάθε λέξη w στη θέση i ενός tweet

Όπου wi η λέξη στη θέση i στο tweet, j ο αριθμός εμφανίσεων της λέξης αυτής σε μια αναζήτηση στο SentiWordNet, $wScore(i,\ k)_p$ το θετικό σκορ της k εμφάνισης της λέξης στο SentiWordNet και $wScore(\ i,\ k)_n$ το αρνητικό σκορ της k εμφάνισης της λέξης στο SWN. Εδώ πρέπει να σημειωθεί ότι στον υπολογισμό δεν έχει ληφθεί υπόψιν το rank κάθε λέξης, το οποίο

είναι η κατηγοριοποίηση με βάση τη συχνότητα εμφάνισης μιας λέξης σε σχέση με τα συνώνυμά της, το synset δηλαδή στο οποίο ανήκει. Επιλέχθηκε να γίνει πιο απλουστευμένη χρήση της βαθμολογίας του SentiWordNet, δηλαδή έγινε υπολογισμός του μέσου όρου.

Παραδείγματος χάριν, για το tweet "Oh, you don't like sarcasm? You must be so funny to hang around with." έχουμε τους ακόλουθους υπολογισμούς:

```
Initial Tweet text: "Oh, you don't like sarcasm? You must be so funny to hang around with."
Cleaned tweet text:  "t like sarcasm must so funny hang around"
Tweet Part-of-Speech tags:  {'funny': 'JJ', 'like': 'IN', 'sarcasm': 'NN',
                             'hang': 'VB', 'around': 'RB', 'so': 'RB',
                             't': 'NN', 'must': 'MD'}
For word in tweet clean text:
    **Current word:  "t"
    ## Word length is NOT more than one characters: exclude
    **Current word:  "like"
        Word length is more than one characters: continue processing
        Found an exact match for word
        Score found: word: like position: 1 s_word-1 score: 1.26
    **Current word:  "sarcasm"
        Word length is more than one characters: continue processing
        Found an exact match for word  and Category = "n"
        Score found: word: sarcasm position: 2 s_word-2 score: 1.0
    **Current word:  "must"
        Word length is more than one characters: continue processing
        Found an exact match for word
        Score found: word: must position: 3 s_word-3 score: 1.03
    **Current word:  "so"
        Word length is more than one characters: continue processing
        Found an exact match for word  and Category = "r"
        Score found: word: so position: 4 s_word-4 score: 1.0
    **Current word:  "funny"
        Word length is more than one characters: continue processing
        Found an exact match for word  and Category = "a"
        Score found: word: funny position: 5 s_word-5 score: 0.81
    **Current word:  "hang"
        Word length is more than one characters: continue processing
        Found an exact match for word  and Category = "v"
        Score found: word: hang position: 6 s_word-6 score: 0.99
    **Current word:  "around"
        Word length is more than one characters: continue processing
        Found an exact match for word  and Category = "r"
        Score found: word: around position: 7 s_word-7 score: 1.04

*Set total swn_score  1.02
*Modify total swn_score  0.02
Resnik: 1.30982192069
Lin: 0.126806276146
WuP: 0.4
Path: 0.1
```

Σχήμα 26. Περιγραφή της διαδικασίας υπολογισμού της βαθμολογίας του SentiWordNet για κάθε λέξη

Το τελικό αποτέλεσμα του παραδείγματος που αναφέρθηκε προηγουμένως είναι το ακόλουθο feature dictionary.

```
{
  'wup': 0.4,
  'res': 1.30982192068823,
  's_word-1': 1.26,
  's_word-3': 1.03,
  's_word-2': 1.0,
  'AS_GROUND_AS_VEHICLE': 'False',
  's_word-4': 1.0,
  's_word-7': 1.04,
  's_word-6': 0.99,
  'POS_SMILEY': 'False',
  'LINK': 'False',
  'contains_so': 1.0,
  'exclamation': 0,
  'RT': 'False',
  'funny': 'JJ',
  'swn_score': 0.020000000000000018,
  'DONT_YOU': 'False',
  'REFERENCE': 'False',
  'HT_NEG': 'False',
  'lin': 0.1268062761459509,
  'polarity': u 'negative',
  'contains_around': 1.04,
  'OH_SO': 'True',
  'LOVE': 'False',
  'around': 'RB',
  's_word-5': 0.81,
  'hang': 'VB',
  'word-1': 'like',
  'HT_POS': 'False',
  'word-3': 'must',
  'word-2': 'sarcasm',
  'word-5': 'funny',
  'HT': 'False',
  'word-7': 'around',
  'word-6': 'hang',
  'contains_sarcasm': 1.0,
  'path': 0.1,
  'questionmark': 1,
  'CAPITAL': 'False',
  'must': 'MD',
  'like': 'IN',
```

Σχήμα 27. Παράδειγμα feature dictionary ενός tweet

```
  'like': 'IN',
  'NEG_SMILEY': 'False',
  'word-0': 't',
  'fullstop': 1,
  'NEGATION': 'True',
  'contains_hang': 0.99,
  'so': 'RB',
  't': 'NN',
  'LAUGH': 'False',
  'sarcasm': 'NN',
  't-similarity': 1.30982192068823,
  'contains_funny': 0.81,
  'word-4': 'so'
}
```

Σχήμα 28. Παράδειγμα feature dictionary ενός tweet (συνέχεια)

Το χαρακτηριστικό αυτό μπορεί να περιγραφεί και ως εξής: κατασκευάζεται ένας πίνακας με σειρές όσες και τα tweets (M), στήλες όσες και το μέγιστο των λέξεων από τις οποίες μπορεί να αποτελείται ένα tweet, έστω N και για κάθε tweet καταγράφεται το SentiWordNet score κάθε λέξης. Εάν ένα tweet αποτελείται από λιγότερες από N λέξεις, τότε οι τιμές που παίρνουν οι στήλες για τις λέξεις που δεν υπάρχουν είναι False (0).

TWEETS	S-WORD-1	S-WORD-2	SWORD-3	…	S-WORD-N-1	S-WORD-N
TWEET1	1.3	1.1	0.5	…	0.1	1.6
TWEET2	0.5	1.5	0	…	0	0
…	…	…	…	…	…	…
TWEET-M	0.3	1.5	0		0	1.4

Πίνακας 10. Επεξήγηση του SentiWordNet χαρακτηριστικού για κάθε λέξη

Στην ουσία γίνεται προσπάθεια καταγραφής της αλληλουχίας των συναισθημάτων που φέρουν οι λέξεις, όπως αυτό ορίζεται από το SentiWordNet. Έστω ότι για το tweet1 η λέξη στη θέση 1 (s-word-1) έχει θετική βαθμολογία, ακολουθείται από μια λέξη με λιγότερο θετική βαθμολογία (s-word-2), μια με αρνητική βαθμολογία (s-word-3) κλπ. Στο tweet2 η λέξη στη θέση 1 έχει αρνητική βαθμολογία, η δεύτερη λέξη έχει θετική η τρίτη αρνητική κλπ. Η σκέψη ήταν να εντοπιστούν φαινόμενα που μια έντονα θετική λέξη συσχετίζεται με μια έντονα αρνητική, κάτι που θα μπορούσε να δείξει ότι υπάρχουν φαινόμενα ειρωνείας, τα οποία έχουν τάση προς το αρνητικό συναίσθημα. Ως βελτίωση του συλλογισμού που περιγράφηκε, θα μπορούσε να είναι η εύρεση τέτοιων μοτίβων ανεξαρτήτως θέσης. Για παράδειγμα, εάν έχουμε το μοτίβο «αρνητικό-θετικό-αρνητικό» στις θέσεις 1-2-3 σε ένα tweet και σε ένα άλλο tweet έχουμε το ίδιο μοτίβο αλλά σε άλλες θέσεις, ο classifier δεν θα είναι σε θέση να τα συσχετίσει κάπως καθώς χρειάζεται να είναι τα ίδια μοτίβα στις ίδιες θέσεις. Με διαφορετική χρήση του χαρακτηριστικού αυτού όμως, π.χ. με bigrams ή trigrams μεταξύ των συναισθημάτων θα μπορούσε να προκύψει κάποια ενδιαφέρουσα συσχέτιση μοτίβων-πολικότητας.

ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ	ΤΙΜΗ	ΠΕΡΙΓΡΑΦΗ
Oh so_ *(μ)*(*)	True/ False	Παρουσία/ Απουσία του “Oh so”
Don’t you_*(μ)* (*)	True/ False	Παρουσία/ Απουσία του “Don’t you”
As_As_*(μ)* (*)	True/ False	Παρουσία/ Απουσία του “As … as …”
Questionmark*(μ)*(*)	True/ False	Παρουσία/ Απουσία του “?”
Exclamation - mark*(μ)* (*)	True/ False	Παρουσία/ Απουσία του “!”
Capitals*(μ)* (*)	True/ False	Παρουσία/ Απουσία κεφαλαίων λέξεων
Reference*(μ)* (*)	True/ False	Παρουσία/ Απουσία της “@user” αναφοράς
RT(μ)	True/ False	Παρουσία/ Απουσία της ένδειξης retweet
Negations*(μ)*(*)	True/ False	Παρουσία/ Απουσία της ένδειξης άρνησης
URL*(μ)*	True/ False	Παρουσία/ Απουσία υπερσυνδέσμων
HT_pos*(μ)*(*)	True/ False	Παρουσία/ Απουσία ουδέτερων hashtags
HT_neg*(μ)*(*)	True/ False	Παρουσία/ Απουσία θετικών hashtags (Εξίσωση 12)
HT_neu*(μ)*(*)	True/ False	Παρουσία/ Απουσία αρνητικών hashtags (Εξίσωση 12)
Emoticon Pos*(μ)*(*)	True/ False	Παρουσία/ Απουσία θετικών emoticons
Emoticon Neg*(μ)*(*)	True/ False	Παρουσία/ Απουσία αρνητικών Emoticons
POS-tags(*)	"NN", "VB", "ADJ","RB"	Part Of Speech tags (απλοποιημένα -Εξίσωση 17)
$swnScore_{wi}$ (*)	“positive”, “somewhat positive”, “neutral”, “negative”, “somewhat negative”	SentiWordNet score για κάθε λέξη (Εξίσωση 15
swnScoreTotal	“positive”, “somewhat positive”, “neutral”, “negative”, “somewhat negative”	Μέσος όρος του SentiWordNet score για ένα tweet
sim_t (Resnik*)	Decimal score	Lin, Wu-Palmer, Path, Resnik WordNet semantic similarity measures (Εξίσωση 13)

Πίνακας 11 Το σύνολο των χαρακτηριστικών που δοκιμάστηκαν

ΟΜΑΔΟΠΟΙΗΣΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ - FEATURE DISCRETIZATION

Κατά τη διάρκεια των δοκιμών, έγιναν οι ακόλουθες αλλαγές στις τιμές των χαρακτηριστικών που αφορούν το SentiWordNet σκορ και τα Part-of-speech tags.

ΟΜΑΔΟΠΟΙΗΣΗ ΤΩΝ ΤΙΜΩΝ ΤΟΥ SENTIWORDNET SCORE

Μετά από παρατήρηση των τιμών των λέξεων στο SentiWordNet, αποφασίστηκε η ακόλοθη ομαδοποίησή τους.

$$swnScore_{wi} = \begin{cases} positive, & (> 1.2) \\ somewhat\ positive, & (> 0.05\ \leq 1.2) \\ neutral, & (\leq 0.05\ \geq 0.95) \\ somewhat\ negative, & (< 0.95\ \geq 0.2) \\ negative, & (< 0.2) \end{cases}$$

Εξίσωση 16 Ομαδοποίηση των τιμών του SentiWordNet score για κάθε λέξη w στη θέση i ενός tweet

Ως συνέπεια της ομαδοποίησης, ο Πίνακας 10 μετατρέπεται στον ακόλουθο. Η λογική της συμπλήρωσης των κελιών για τα οποία δεν υπάρχουν τιμές είναι η ίδια (s-word-j : False, εάν δεν υπάρχει j-th λέξη). Η ίδια ομαδοποίηση έγινε και για το συνολικό SentiWordNet score του tweet.

TWEETS	S-WORD-1	S-WORD-2	SWORD-3	…	S-WORD-N-1	S-WORD-N
TWEET1	positive	somewhat _positive	negative		negative	positive
TWEET2	negative	positive	neutral		neutral	neutral
…						
TWEET-M	negative	positive	neutral		neutral	positive

Πίνακας 12. Επεξήγηση του SentiWordNet χαρακτηριστικού για κάθε λέξη

ΟΜΑΔΟΠΟΙΗΣΗ ΤΩΝ POS-TAGS ΑΝΑ ΜΕΡΟΣ ΤΟΥ ΛΟΓΟΥ

Αποφασίστηκε να κρατηθούν ως κατηγορίες μόνο οι τέσσερις κύριες, δηλαδή “Nouns”, “Verbs”, “Adjectives”, “Adverbs”, με την κωδικοποίηση που φαίνεται ακολούθως:

$$groupped\_postag_{wi} = \begin{cases} NN, & pos - tag_{wi}\ in\ NounTags \\ VB, & pos - tag_{wi}\ in\ VerbTags \\ ADJ, & pos - tag_{wi}\ in\ AdjectiveTags \\ ADV, & pos - tag_{wi}\ in\ AdverbTags \end{cases}$$

Εξίσωση 17. Η ομαδοποίηση ενός POS-tag

Όπου NounTags = {N, NP, NN, NNS, NNP, NNPS}, VerbTags= {V, VD, VG, VN, VB, VBD, VBG, VBN, VBP, VBZ}, AdjectiveTags= {ADJ, JJ, JJR, JJS}, AdverbTags={RB, RBR, RBS, WRB}. Οι υπόλοιπες κατηγορίες αγνοούνται και εξαιρούνται από τη διαδικασία.

5.7.3 ΔΙΑΔΙΚΑΣΙΑ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ - CLASSIFICATION

Στο προηγούμενο βήμα έχουμε καταλήξει σε ένα σύνολο από με feature dictionaries $Fd_{Ttrain} = \{fd_{t1}, fd_{t2}, \dots, fd_{tm}\}$, όπου fd_{ti} το λεξικό με τα χαρακτηριστικά που αντιπροσωπεύει ένα tweet, i ∈ [1, m], m <= $c(T_{train})$, δηλαδή το m παίρνει τιμές από 1 εώς και τον αριθμό των tweets στο σύνολο εκπαίδευσης, Το Fd_{Ttrain} θα πρέπει να μετατραπεί σε μορφή κατανοητή προς τον classifier, δηλαδή σε vectors. Για την μετατροπή αυτή, χρησιμοποιείται ο κατάλληλος κάθε φορά vectorizer του scikit-learn. Στην περίπτωση του Fd_{Ttrain}, χρησιμοποιείται ο DictVectorizer), ο οποίος παίρνει ως είσοδο το Fd_{Ttrain} και το μετατρέπει σε ένα σύνολο από διανύσματα (vectors) $VFd_{Ttrain} = \{vfd_{t1}, vfd_{t2}, \dots, tfd_{tm}\}$, όπου$vfd_{ti}$ το διάνυσμα που αντιπροσωπεύει ένα tweet, i ∈ [1, m], m <= $c(T_{train})$. Στη συνέχεια, χρησιμοποιείται ο TfidfTransformer (με την επιλογή use_idf=False), ο οποίος παιρνει το VFd_{Ttrain} και το μετατρέπει σε ένα σύνολο $TF_{Ttrain} = \{tf_{t1}, tf_{t2}, \dots, tf_{tm}\}$, όπου tf_{ti} ο πίνακας με τα term frequencies που αντιστοιχούν στα χαρακτηριστικά ενός tweet. Το TF_{Ttrain} χρησιμοποιείται ως είσοδος για τον classifier, για εκμάθηση και πρόβλεψη. Η διαδικασία είναι παρόμοια και για το train και για το test set κάθε φορά, με μόνη διαφορά ότι το test set δεν γίνεται fit, δεν το «μαθαίνει» δηλαδή ούτε ο vectorizer ούτε ο classifier.

Οι προβλέψεις του classifier, είναι μια λίστα με βαθμολογίες $\boldsymbol{S'_{Ttest}}$ (ή $\boldsymbol{S_{system}}$) σχετικά με το που πιστεύει ότι ανήκει το κάθε tweet. Αυτές συγκρίνονται με το $\boldsymbol{S_{gold}}$, ώστε να αξιολογηθεί η ποιότητα των αποτελεσμάτων που δίνει ο classifier.

Στο ακόλουθο παράδειγμα, το Feature Dictionary ενός tweet που ανήκει στο test set (Σχήμα 29) δίνεται ως είσοδο σε έναν DictVectorizer, ο οποίος αναλαμβάνει να κωδικοποιήσει τα χαρακτηριστικά με τέτοιο τρόπο ώστε να είναι κατάλληλα προς είσοδο στον classifier. Το Σχήμα 30 δείχνει την έξοδο του DictVectorizer για το Feature Dictionary, όπου (0, 11062) -> (θέση tweet – στην συγκεκριμένη περίπτωση το πρώτο tweet, αριθμός feature – όπως αυτό έχει κωδικοποιηθεί από τον vectorizer) = 1.0, δηλαδή υπάρχει το χαρακτηριστικό 11062 στο tweet. Το χαρακτηριστικό 11062, εάν δούμε το vocabulary που έχει δημιουργηθεί, θα δούμε ότι αντιστοιχεί στο χαρακτηριστικό "spy=VB". Θα πρέπει να σημειωθεί ότι τα χαρακτηριστικά που έχουν αριθμητική τιμή, την διατηρούν μετά την κωδικοποίηση από τον DictVectorizer. Παραδείγματος χάριν, εάν το χαρακτηριστικό "__res__", δηλαδή το Resnik Similarity Measure, είχε τιμή άλλη από το μηδέν, π.χ. 1.4, τότε αυτό θα διατηρηθεί στην έξοδο του DictVectorizer. Στη συνέχεια, το αποτέλεσμα του DictVectorizer δέχεται ως είσοδο ο TfidfTransformer, ο οποίος υπολογίζει το Term Frequency, όσον αφορά τα χαρακτηριστικά που υπάρχουν. Δηλαδή, το "spy=VB"=1.0 (True) έχει TF ίσο με 0.182574185835. Γίνεται δηλαδή μια κανονικοποίηση των τιμών των χαρακτηριστικών στο διάστημα [0,1], ώστε να είναι πιο κατάλληλη είσοδος για έναν classifier[28]. Εάν στη διαδικασία ακολουθούσε ο υπολογισμός pairwise cosine similarity, τότε το tweet αυτό θα λάμβανε την τιμή 0.62547063619, η οποία είναι το cosine similarity του από το πρώτο tweet του train set.

[28] http://scikit-learn.org/stable/modules/feature_extraction.html#dict-feature-extraction

Σχήμα 29. Feature Dictionary που αντιστοιχεί σε ένα tweet - είσοδος του DictVectorizer

```
DictVectorizer:
(0, 49)     1.0
(0, 178)    1.0
(0, 1361)   1.0
(0, 1363)   1.0
(0, 1365)   1.0
(0, 1367)   1.0
(0, 1369)   1.0
(0, 1371)   1.0
(0, 1373)   1.0
(0, 1375)   1.0
(0, 1377)   1.0
(0, 1379)   1.0
(0, 1381)   1.0
(0, 1383)   1.0
(0, 1385)   1.0
(0, 1387)   0.0
(0, 1389)   1.0
(0, 1434)   1.0
(0, 1442)   1.0
(0, 1444)   1.0
(0, 1456)   1.0
(0, 1459)   1.0
(0, 1464)   1.0
(0, 1469)   1.0
(0, 1474)   1.0
(0, 3036)   1.0
(0, 6287)   1.0
(0, 9441)   1.0
(0, 9958)   1.0
(0, 10988)  1.0
(0, 11062)  1.0
```

Σχήμα 30. Η έξοδος του DictVectorizer

```
TfidfTransformer:
(0, 49)     0.182574185835
(0, 178)    0.182574185835
(0, 1361)   0.182574185835
(0, 1363)   0.182574185835
(0, 1365)   0.182574185835
(0, 1367)   0.182574185835
(0, 1369)   0.182574185835
(0, 1371)   0.182574185835
(0, 1373)   0.182574185835
(0, 1375)   0.182574185835
(0, 1377)   0.182574185835
(0, 1379)   0.182574185835
(0, 1381)   0.182574185835
(0, 1383)   0.182574185835
(0, 1385)   0.182574185835
(0, 1387)   0.0
(0, 1389)   0.182574185835
(0, 1434)   0.182574185835
(0, 1442)   0.182574185835
(0, 1444)   0.182574185835
(0, 1456)   0.182574185835
(0, 1459)   0.182574185835
(0, 1464)   0.182574185835
(0, 1469)   0.182574185835
(0, 1474)   0.182574185835
(0, 3036)   0.182574185835
(0, 6287)   0.182574185835
(0, 9441)   0.182574185835
(0, 9958)   0.182574185835
(0, 10988)  0.182574185835
(0, 11062)  0.182574185835
```

Σχήμα 31. Η έξοδος του TfidfTransformer

```
{
  '__POS_SMILEY__': 'False',
  '__REFERENCE__': 'False',
  '__OH_SO__': 'False',
  'activists': 'NN',
  '__DONT_YOU__': 'False',
  '__AS_GROUND_AS_VEHICLE__': 'False',
  '__questionmark__': 'False',
  'collectively': 'RB',
  '__s_word-9': 'neutral',
  '__HT_NEG__': 'False',
  '__s_word-5': 'neutral',
  '__s_word-4': 'neutral',
  '__s_word-7': 'neutral',
  '__s_word-6': 'neutral',
  '__s_word-1': 'neutral',
  '__s_word-0': 'neutral',
  '__s_word-3': 'neutral',
  '__s_word-2': 'somewhat_negative',
  'communists': 'NN',
  '__HT_POS__': 'False',
  '__HT__': 'False',
  '__CAPITAL__': 'False',
  '__exclamation__': 'False',
  'animal': 'NN',
  'environmentalists': 'NN',
  '__NEGATION__': 'False',
  'vegetarians': 'NN',
  '__res__': 1.4,
  'rights': 'NN',
  '__NEG_SMILEY__': 'False',
  'referred': 'VB',
  '__s_word-8': 'neutral'
}
```


Σχήμα 33. Παράδειγμα tweet με διαφορετικά αριθμητικά χαρακτηριστικά


```
DictVectorizer:
(0, 1361)   1.0
(0, 1363)   1.0
(0, 1365)   1.0
(0, 1367)   1.0
(0, 1369)   1.0
(0, 1371)   1.0
(0, 1373)   1.0
(0, 1375)   1.0
(0, 1377)   1.0
(0, 1379)   1.0
(0, 1381)   1.0
(0, 1383)   1.0
(0, 1385)   1.0
(0, 1387)   1.4
(0, 1389)   1.0
(0, 1434)   1.0
(0, 1441)   1.0
(0, 1444)   1.0
(0, 1449)   1.0
(0, 1454)   1.0
(0, 1459)   1.0
(0, 1464)   1.0
(0, 1469)   1.0
(0, 1474)   1.0
(0, 1766)   1.0
(0, 9940)   1.0
```


Σχήμα 32. Η έξοδος του DictVectorizer

```
TfidfTransformer:
  (0, 1361)   0.192592803943
  (0, 1363)   0.192592803943
  (0, 1365)   0.192592803943
  (0, 1367)   0.192592803943
  (0, 1369)   0.192592803943
  (0, 1371)   0.192592803943
  (0, 1373)   0.192592803943
  (0, 1375)   0.192592803943
  (0, 1377)   0.192592803943
  (0, 1379)   0.192592803943
  (0, 1381)   0.192592803943
  (0, 1383)   0.192592803943
  (0, 1385)   0.192592803943
  (0, 1387)   0.26962992552
  (0, 1389)   0.192592803943
  (0, 1434)   0.192592803943
  (0, 1441)   0.192592803943
  (0, 1444)   0.192592803943
  (0, 1449)   0.192592803943
  (0, 1454)   0.192592803943
  (0, 1459)   0.192592803943
  (0, 1464)   0.192592803943
  (0, 1469)   0.192592803943
  (0, 1474)   0.192592803943
  (0, 1766)   0.192592803943
  (0, 9940)   0.192592803943
```

Σχήμα 34. Η έξοδος του TfidfTransformer

5.8 ΠΑΡΑΔΕΙΓΜΑΤΑ ΧΡΗΣΗΣ

Το σύστημα που δημιουργήθηκε δεν έχει διεπαφή χρήστη αλλά χρησιμοποιείται μέσα από python scripts. Ακολούθως περιγράφονται κάποια παραδείγματα χρήσης και διεξαγωγής πειραμάτων στα πλαίσια της επίλυσης του προβλήματος.

Το Σχήμα 35 δείχνει τη χρήση της κλάσης Trial και TrialConfig για τη διεξαγωγή ενός πειράματος.

```
# a list containing the features we want to take into account
selected_features =['__OH_SO__',
                    '__DONT_YOU__',
                    '__AS_GROUND_AS_VEHICLE__',
                    '__CAPITAL__',
                    '__HT__',
                    '__HT_POS__',
                    '__HT_NEG__',
                    '__LINK__',
                    '__POS_SMILEY__',
                    '__NEG_SMILEY__',
                    '__NEGATION__',
                    '__REFERENCE__',
                    '__questionmark__',
                    '__exclamation__',
                    '__fullstop__',
                    '__polarity__',
                    '__RT__',
                    '__LAUGH__',
                    '__postags__',
                    '__swn_score__',
                    '__s_word__',
                    '__res__',
                    '__lin__',
                    '__wup__',
                    '__path__',
                    '__contains_'
                    ]

# the configuration of a trial with defaults
trial_config = TrialConfig(selected_features, ds=g.DS_TYPE.Test)
trial = Trial(trial_config)
trial.classify()        # start classification
trial.save_results()    # save results to database
```

Σχήμα 35. Παράδειγμα πειράματος 1

```
Label range is:  ['0.0To1.0', '2.0To3.0', '3.0To4.0', '-2.0To-1.0', '-1.0To0.0',
                  '-4.0To-3.0', '-5.0To-4.0', 'zero', '-3.0To-2.0', '4.0To5.0',
                  '1.0To2.0']
Test Data Length: 923
Train Data Length: 7606
             precision    recall  f1-score   support

       -5.0       0.00      0.00      0.00         4
       -4.0       0.56      0.07      0.12        72
       -3.0       0.49      0.48      0.48       342
       -2.0       0.26      0.44      0.32       209
       -1.0       0.12      0.04      0.06        70
        0.0       0.43      0.63      0.51       117
        1.0       0.20      0.02      0.04        45
        2.0       0.20      0.06      0.10        32
        3.0       0.33      0.08      0.13        25
        4.0       0.33      0.17      0.22         6
        5.0       0.00      0.00      0.00         1

avg / total       0.37      0.37      0.34       923

343/923 PERCENTAGE: 37.1614301192%

Number of Gold entries:: 923
Number of Submitted entries:: 923
Cosine Similarity Score:: 0.7430
Penalty:: 0.0000
Final Score:: 0.7430
accuracy 0.371614301192
```

Σχήμα 36. Αποτελέσματα του παραδείγματος

```
IMPORTANT FEATURES:
=====================
    -0.2784 __LINK__=False
    0.5851  text=ADJ
    -0.2391 s_word-8=neutral
    0.4939  team=VB
    -0.1930 s_word-11=neutral
    0.4939  sports=NN
    -0.1795 s_word-9=neutral
    0.4939  rebels=VB
    -0.1714 s_word-3=somewhat_positive
    0.4873  tweet=NN
    -0.1666 __HT__=False
    0.4600  refined=VB
    -0.1429 __exclamation__=True
    0.4600  fellow=ADJ
    -0.1426 s_word-5=somewhat_positive
    0.4600  educated=VB
    -0.1425 s_word-5=positive
    0.4600  cultured=VB
    -0.1363 s_word-3=somewhat_negative
    0.4549  back=RB
```

Σχήμα 37. Λίστα των πιο σημαντικών χαρακτηριστικών του παραδείγματος με τον αντίστοιχο συντελεστή (coefficient) τους

Για τις επιλογές που φαίνονται στο Σχήμα 38, το feature dictionary για το tweet "Oh, you don't like sarcasm? You must be so funny to hang around with." μετά το post-processing, με τη μορφή δηλαδή που θα δοθεί ως είσοδος σε έναν DictVectorizer, φαίνεται στο Σχήμα 39.

```
selected_features =[
                    '__OH_SO__',
                    '__DONT_YOU__',
                    '__AS_GROUND_AS_VEHICLE__',
                    '__CAPITAL__',
                    '__HT__',
                    '__HT_POS__',
                    '__HT_NEG__',
                    '__POS_SMILEY__',
                    '__NEG_SMILEY__',
                    '__NEGATION__',
                    '__REFERENCE__',
                    '__questionmark__',
                    '__exclamation__',
                    '__postags__',
                    '__s_word__',
                    '__res__',
                    ]

if __name__ == "__main__":
    trial_config = TrialConfig(selected_features, ds=g.DS_TYPE.Test)
    trial = Trial(trial_config)
    trial.classify()
```

Σχήμα 38. Παράδειγμα πειράματος 2

```
Post - Processed feature dictionary: {
  'res': 1.3,
  's_word-1': 'positive',
  's_word-3': 'neutral',
  's_word-2': 'neutral',
  'AS_GROUND_AS_VEHICLE': 'False',
  's_word-4': 'neutral',
  's_word-7': 'neutral',
  's_word-6': 'neutral',
  'POS_SMILEY': 'False',
  'exclamation': 0,
  'funny': 'ADJ',
  'DONT_YOU': 'False',
  'REFERENCE': 'False',
  'HT_NEG': 'False',
  'OH_SO': 'True',
  'around': 'RB',
  's_word-5': 'somewhat_negative',
  'hang': 'VB',
  'HT_POS': 'False',
  'HT': 'False',
  'questionmark': 1,
  'CAPITAL': 'False',
  'NEG_SMILEY': 'False',
  'NEGATION': 'True',
  'so': 'RB',
  't': 'NN',
  'sarcasm': 'NN'
}
```

Σχήμα 39. Παράδειγμα Feature dictionary μετά το Post-processing

6. ΠΕΡΙΓΡΑΦΗ ΒΙΒΛΙΟΘΗΚΗΣ

6.1 ΣΚΟΠΟΣ

Ο στόχος της βιβλιοθήκης επεξεργασίας tweets είναι να παρέχει έναν εύκολο τρόπο μοντελοποίησης ενός tweet και κατ' επέκταση την εύκολη εξαγωγή χαρακτηριστικών για τη χρήση του σε sentiment analysis διαδικασίες. Η βιβλιοθήκη, όπως δομήθηκε, υποστηρίζει τα βασικά χαρακτηριστικά που χρησιμοποιήθηκαν στην εφαρμογή sentiment analysis που περιγράφεται πιο πάνω. Στην ουσία, αντικαθιστά τη λειτουργία της κλάσης TweetProcessor, όσον αφορά το κομμάτι της προεπεξεργασίας.

6.2 ΔΕΔΟΜΕΝΑ

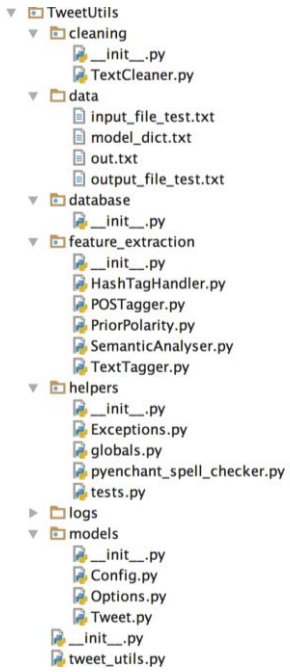
Τα εξωτερικά δεδομένα που χρησιμοποιήθηκαν, είναι τα ίδια που περιγράφονται στο κεφάλαιο 5, εκτός από τα tweets που δόθηκαν από το SemEval στα πλαίσια του Task 11.

6.3 ΕΞΩΤΕΡΙΚΕΣ ΒΙΒΛΙΟΘΗΚΕΣ ΚΑΙ ΕΡΓΑΛΕΙΑ ΑΝΑΠΤΥΞΗΣ

Για τη δημιουργία της, χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python, μαζί με όλες τις βιβλιοθήκες που αναφέρθηκαν στο κεφάλαιο 5.6.

6.4 ΔΟΜΗ ΒΙΒΛΙΟΘΗΚΗΣ

Ακολούθως περιγράφεται η δομή του project της TweetUtils, όπως φαίνεται στην εικόνα που ακολουθεί, τα πακέτα και οι κλάσεις.



Σχήμα 40. Η δομή του project της βιβλιοθήκης TweetUtils

6.4.1 ΠΑΚΕΤΑ ΚΑΙ ΚΛΑΣΕΙΣ

Οι βασικές κλάσεις που χρησιμοποιήθηκαν για την εφαρμογή που περιγράφηκε στα προηγούμενα κεφάλαια χρησιμοποιούνται και εδώ με κάποιες αλλαγές όσον αφορά τις εισόδους που παίρνουν, τις εξαρτήσεις μεταξύ τους και τον τρόπο οργάνωσής τους. Περιγράφονται ακολούθως τα πακέτα και οι επιπλέον κλάσεις που χρησιμοποιήθηκαν.

TWEETUTILS

Το κεντρικό πακέτο που περιέχει όλα τα υπόλοιπα και την κύρια κλάση TweetUtils.

CLEANING

Το πακέτο cleaning περιλαμβάνει την κλάση TextCleaner, υπεύθυνη για τον καθαρισμό ενός tweet.

DATA (ΦΑΚΕΛΟΣ)

Φάκελος για την αποθήκευση του μοντέλου GATE για τον POSTagger και των αρχείων που δημιουργεί η TweetUtils στην περίπτωση που ζητηθεί αποθήκευση σε αρχείο.

DATABASE

Περιέχει την κλάση για την αλληλεπίδραση με τη βάση.

FEATURE EXTRACTION

Περιέχει τις κλάσεις για την εξαγωγή χαρακτηριστικών, την HashtagHandler, την POSTagger και την SemanticAnalyser.

HELPERS

Περιέχει τις βοηθητικές κλάσεις Globals και EnchantSpellChecker.

LOGS (ΦΑΚΕΛΟΣ)

Φάκελος για την αποθήκευση των log αρχείων.

MODELS

Περιέχει τις κλάσεις Tweet, Config, Options, CleaningOptions, FeatureOptions και Option.

6.5 ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ

Στα πλαίσια της βιβλιοθήκης, είναι απαραίτητοι δύο από τους πίνακες που αναφέρθηκαν στο κεφάλαιο 5.5, ο ένας για την αποθήκευση του SentiWordNet και ο άλλος για την αποθήκευση των επιπλέον stop words. Το indexing στο πεδίο synset του πίνακα SentiWordNet υπάρχει και σε αυτή την περίπτωση, καθώς είναι απαραίτητο για την καλή ταχύτητα επεξεργασίας των tweets κατά το feature extraction. Μια πιο φορητή λύση θα ήταν η χρήση sqlite βάσης ώστε να μην είναι αναγκαία η χρήση MySQL Server και να απλοποιηθεί η χρήση της βιβλιοθήκης.

6.6 ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ

Οι βασικές επιλογές και η λειτουργία της βιβλιοθήκης περιγράφεται ακολούθως, μαζί με κάποια παραδείγματα χρήσης.

6.6.1 ΕΠΙΛΟΓΕΣ CLEANING

Για τον "καθαρισμό" ενός tweet υπάρχουν οι ακόλουθες επιλογές:

- Αφαίρεση non-ascii χαρακτήρων
- Αφαίρεση ένδειξης RT
- Αφαίρεση της ένδειξης γέλιου
- Αφαίρεση negations
- Αφαίρεση υπερσυνδέσμων
- Αφαίρεση των κοινών emoticons
- Αφαιρεση της αναφοράς σε χρήστη (@user)
- Αφαίρεση σημείων στίξης
- Μετατροπή όλων των γραμμάτων σε μικρά
- Αφαίρεση πολλαπλών συνεχόμενων γραμμάτων
- Αφαίρεση των stop words (από το WordNet και από εξωτερική λίστα που δίνει ο χρήστης)
- Ορθογραφικός έλεγχος των λέξεων

6.6.2 ΕΠΙΛΟΓΕΣ ΕΞΑΓΩΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Τα χαρακτηριστικά που εξάγονται από το tweet από προεπιλογή είναι αυτά που αναφέρονται στον Πίνακας 11, με τους κανόνες που περιγράφονται στο κεφάλαιο 5.

6.6.3 ΠΡΟΣΘΗΚΗ ΝΕΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Στον χρήστη δίνεται η δυνατότητα να προσθέσει δικά του χαρακτηριστικά στη διαδικασία, παρέχοντας μια function η οποία θα πρέπει να δέχεται ως είσοδο κείμενο και να πραγματοποιεί την εξαγωγή του επιθυμητού χαρακτηριστικού από το κείμενο αυτό. Επίσης, μπορεί να οριστεί από τον χρήστη εάν το κείμενο στο οποίο θα γίνει η επεξεργασία θα είναι το αρχικό κείμενο του tweet ή το «καθαρισμένο» κείμενο του tweet.

6.7 ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ

6.7.1 CLEANING

Στο ακόλουθο παράδειγμα, η TweetUtils έχει παραμετροποιηθεί έτσι ώστε να πραγματοποιηθεί μόνο ο «καθαρισμός» του tweet, δηλαδή Config(True, False), όπου το True αφορά τον καθαρισμό και το False την εξαγωγή των χαρακτηριστικών. Στο παράδειγμα φαίνονται και οι διαφορετικές είσοδοι και έξοδοι της βιβλιοθήκης.

```
# examples: Cleaning only
utils = TweetUtils(Config(True, False))

# process a list of tweet texts
tweets = utils.process(["This is lovely!!!#NOT :( :) http://dsgrg.vom/vfda", "I hate Monday mornings... :) :( !!!"])
print "Clean text:", tweets[0].clean_text
print "Features:", tweets[0].feature_dict

# process one tweet text: returns a Tweet object
single_tweet = utils.process("This is lovely!!!#NOT")
print str(single_tweet), single_tweet.clean_text, single_tweet.feature_dict

# the input file must be in form: one tweet text per line
# the output is saved in output_file_test.txt, one tweet per line
tweets_from_file = utils.process(None, "../TweetUtils/data/tweets.txt", "../TweetUtils/data/output_file_test.txt")

```

Σχήμα 41. Παράδειγμα χρήσης της βιβλιοθήκης μόνο για καθαρισμό

6.7.2 CLEANING AND FEATURE EXTRACTION

Στο ακόλουθο παράδειγμα, η TweetUtils έχει παραμετροποιηθεί έτσι ώστε να πραγματοποιηθεί και ο «καθαρισμός» και η εξαγωγή χαραστηριστικών του tweet, δηλαδή Config(True, True), όπου το πρώτο True αφορά τον καθαρισμό και το επόμενο την εξαγωγή των χαρακτηριστικών. Στο παράδειγμα φαίνονται και οι διαφορετικές είσοδοι και έξοδοι της βιβλιοθήκης.

```
# example: Cleaning and Feature Extraction
utils = TweetUtils(Config(True, True))

# process a list of tweet texts: returns a list of Tweet objects
tweets = utils.process(["This is lovely!!!#NOT :( :) http://dsgrg.vom/vfda", "I hate Monday mornings... :) :( !!!"])
print "Clean text:", tweets[0].clean_text
print "Features:", tweets[0].feature_dict

# process one tweet text
single_tweet = utils.process("This is lovely!!!#NOT")
print str(single_tweet), single_tweet.clean_text, single_tweet.feature_dict

# the input file must be in form: one tweet text per line
# the output is saved in output_file_test.txt, one tweet per line
tweets_from_file = utils.process(None, "../TweetUtils/data/tweets.txt", "../TweetUtils/data/output_file_test.txt")
```

Σχήμα 42. Παράδειγμα της βιβλιοθήκης για καθαρισμό και για εξαγωγή χαρακτηριστικών

6.7.3 ΠΡΟΣΘΑΦΑΙΡΕΣΗ FEATURES

Στο ακόλουθο παράδειγμα, χρησιμοποιούνται δυο functions ορισμένες από τον χρήστη, η *pre_clean_function* και η post_clean_function, οποίες δέχονται ως είσοδο κείμενο, δηλαδή το κείμενο του tweet. Προστίθονται ως νέα χαρακτηριστικά και για την δεύτερη ορίζεται post_clean= True, δηλαδή, να χρησιμοποιηθεί το κείμενο του tweet όπως αυτό έχει προκύψει μετά τον καθαρισμό του. Στο Σχήμα 44 βλέπουμε την έξοδο του προγράμματος.

```python
from TweetUtils import TweetUtils
from TweetUtils.models.Config import Config
from TweetUtils.models.Options import FeatureOption

def pre_clean_function(text):
    print "pre_clean_function", text
    return "new pre-clean feature added"


def post_clean_function(text):
    print "post_clean_function", text
    return "new post-clean feature added"

if __name__ == "__main__":

    # examples
    utils_cfg = Config(True, True)
    utils_cfg.feature_options.add_feature(FeatureOption("test", function=pre_clean_function))
    utils_cfg.feature_options.add_feature(FeatureOption("test", post_clean=True, function=post_clean_function))
    utils = TweetUtils(utils_cfg)
    tweet = utils.process(["This is lovely!!!#NOT :( :) http://dsgrg.vom/vfda", "I hate Monday mornings... :) :( !!!"])[0]
    print tweet.clean_text
    print tweet.feature_dict
    print tweet.extra_features[0]
    print tweet.extra_features[1]

```

Σχήμα 43. Παράδειγμα προσθήκης χαρακτηριστικών - FeatureOption

```
/usr/bin/python /Users/mariakaranasou/Projects/large-scale-sentiment-analysis/SentimentAnalysis/TweetUtils/tweet_utils.py
pre_clean_function This is lovely!!!#NOT :( :) http://dsgrg.vom/vfda
post_clean_function is lovely not
pre_clean_function I hate Monday mornings... :) :( !!!
post_clean_function hate monday mornings
is lovely not
{'__POS_SMILEY__': 'True', '__REFERENCE__': 'False', '__OH_SO__': 'False', 's_word-1': 1.56, 's_word-0': 1.0, u'is': 'VBZ', 's_word-2':
new pre-clean feature added
new post-clean feature added

Process finished with exit code 0
```

Σχήμα 44. Η έξοδος (output) του παραδείγματος

6.8 ΑΠΟΔΟΣΗ

Η επεξεργασία των tweets με την βιβλιοθήκη φαίνεται να έχει σχετικά ικανοποιητική απόδοση (~15 λεπτά για 8526 tweets). Για την βελτίωση της απόδοσης, είναι δυνατόν να χρησιμοποιηθεί το multiprocessing πακέτο της Python και ξεκινάνε πολλαπλά processes από το Pool, ώστε να γίνεται παράλληλη επεξεργασία των tweets. Αυτό που θα πρέπει να προσεχθεί εδώ είναι το concurrency σχετικά με την πρόσβαση στη βάση αλλά και με την αποθήκευση σε αρχείο.

Μια ακόμη βελτίωση θα μπορούσε να είναι η χρήση κάποιας in-memory βασης, όπως για παράδειγμα Redis, καθώς bottleneck στη διαδικασία φαίνεται να είναι τα reads στην mysql, και πιο συγκεκριμένα στον πίνακα SentiWordNet. Σηκώνοντας τον πίνακα αυτό στη μνήμη, θα μπορούσε να έχει αρκετά καλύτερα αποτελέσματα όσον αφορά την ταχύτητα επεξεργασίας των tweets.

```python
# create a function to be executed by each thread
def process(tweet_tuple):
    global processed
    try:
        # create a TweetUtils instance and process the tweet
        utils = TweetUtils(Config(True, True))
        tweet_string = tweet_tuple[1].replace("\n", "").replace("'", "\'").replace("\"", "\\\"").replace("\r", "")
        tweet = utils.process(tweet_string)

        # update the tweet in database
        g.mysql_conn.update(q_update.format(str(tweet.feature_dict), tweet_tuple[0]))
    except:
        print tweet_tuple
        g.logger.error(tweet_tuple)
    processed += 1
    if processed % 100 == 0:
        print processed, datetime.datetime.now()

if __name__ == "__main__":

    processed = 0

    # set the queries to retrieve and update the tweets
    q ="SELECT id, text FROM SentiFeed.tweettestdata;"
    q_update = """UPDATE SentiFeed.tweettestdata SET feature_dict="{0}" WHERE id="{1}";"""

    data = g.mysql_conn.execute_query(q)        # get the tweets
    print "start", datetime.datetime.now()
    num_of_threads = 2                          # set the number of threads
    pool = Pool(processes=num_of_threads)
    pool.map(process, data)                     # begin
    print processed
    print "finished", datetime.datetime.now()
```

Σχήμα 45. Παράδειγμα χρήσης της βιβλιοθήκης με Multiprocessing

7. ΠΕΡΙΓΡΑΦΗ ΔΙΚΤΥΑΚΗΣ ΕΦΑΡΜΟΓΗΣ

7.1 ΣΚΟΠΟΣ

Για την οπτικοποίηση των αποτελεσμάτων αλλά και της διαδικασίας, επιλέχθηκε να δημιουργηθεί μια δικτυακή εφαρμογή, η οποία επικοινωνεί με την εφαρμογή *Figurative Text Analysis* και τη βάση δεδομένων *Sentifeed* και οπτικοποιεί πληροφορίες όπως κατανομή θετικών, αρνητικών και ουδέτερων tweets σε κάθε dataset κ.α.

7.2 ΕΞΩΤΕΡΙΚΕΣ ΒΙΒΛΙΟΘΗΚΕΣ ΚΑΙ ΕΡΓΑΛΕΙΑ ΑΝΑΠΤΥΞΗΣ

Εκτός από τα εργαλεία που περιγράφονται σε προηγούμενο κεφάλαιο και αφορούν την κατασκευή του συστήματος και της βιβλιοθήκης, χρησιμοποιήθηκαν και οι ακόλουθες βιβλιοθήκες Python και Javascript.

7.2.1 ΕΡΓΑΛΕΙΑ ΑΝΑΠΤΥΞΗΣ

Τα εργαλεία ανάπτυξης που αναφέρθηκαν στα πλαίσια της περιγραφής του συστήματος χρησιμοποιήθηκαν και για την κατασκευή της δικτυακής εφαρμογής. Το τελικό αποτέλεσμα είναι ένα ενιαίο project όπου συνυπάρχουν και η δικτυακή εφαρμογή και το σύστημα και η βιβλιοθήκη TweetUtils.

7.2.2 ΕΞΩΤΕΡΙΚΕΣ ΒΙΒΛΙΟΘΗΚΕΣ

DJANGO FRAMEWORK

Το Django είναι ένα πλαίσιο web εφαρμογή δωρεάν και ανοιχτού κώδικα, γραμμένο σε Python, το οποίο ακολουθεί το Model-View-Controller (MVC) αρχιτεκτονικό πρότυπο. Συντηρείται από τον Django Ίδρυμα Ελεύθερου Λογισμικού (DSF), μια ανεξάρτητη μη κερδοσκοπική οργάνωση[29].

Πρωταρχικός στόχος Django είναι να διευκολύνει την δημιουργία περίπλοκων, data-driven δικτυακών εφαρμογών. Το Django δίνει βάση στην επαναχρησιμοποίηση και της ανεξαρτησίας ("pluggability") των δομικών στοιχείων, στην γρήγορη ανάπτυξη, καθώς και την αρχή της μη επανάληψης (“don’t repeat yourself”). Η γλώσσα προγραμματισμού Python χρησιμοποιείται παντού, ακόμα και για τις ρυθμίσεις, τα αρχεία, και τα μοντέλα δεδομένων. Το Django παρέχει επίσης μια προαιρετική διαχειριστική εφαρμογή που επιτρέπει CRUD διαδικασίες στα μοντέλα που έχει δημιουργήσει ο χρήστης[30].

[29] https://en.wikipedia.org/wiki/Django_(web_framework)

[30] https://www.djangoproject.com/

D3.JS

H D3.js είναι μια βιβλιοθήκη JavaScript για το χειρισμό των εγγράφων που βασίζονται στα δεδομένα. Βοηθάει στην οπτικοποίηση των δεδομένων χρησιμοποιώντας HTML, SVG, και CSS. Δίνει έμφαση στην τήρηση των web standards ώστε να υποστηρίζονται όλες οι δυνατότητες των σύγχρονων browsers, συνδυάζει ισχυρά συστατικά οπτικοποίησης και μια προσέγγιση με γνώμονα τα δεδομένα για την DOM χειραγώγηση[31].

C3.JS

Javascript βιβλιοθήκη, wrapper της d3.js για τον σχεδιασμό διαγραμμάτων. Επιλέχθηκε γιατί αποτελεί μια εύχρηστη και απλή διεπαφή της d3.js και δίνει την δυνατότητα εύκολης χρήσης των δυνατοτήτων της. Τα διαγράμματα που μπορούν να δημιουργηθούν με την c3.js αποτελούν ένα μικρό υποσύνολο των δυνατοτήτων της d3.js, όμως καλύπτουν τις ανάγκες της συγκεκριμένης εφαρμογής[32].

GOOGLE CHARTS

Τα Google Charts αποτελούν βιβλιοθήκη Javascript για την εύκολη δημιουργία διαδραστικών διαγραμμάτων για φυλλομετρητές και φορητές συσκευές. Επιλέχθηκε για την ευκολία χρήσης της[33].

BOOTSTRAP 3

HTML, CSS, και JS framework για τη δημιουργία responsive, στοχευμένα για οθόνες κινητών web projects[34].

BOOTSTRAP-TOGGLE

Βιβλιοθήκη Javascript για τη μετατροπή των checkboxes σε toggle-buttons[35].

REDIS

---

[31] http://d3js.org/

[32] http://c3js.org

[33] https://developers.google.com/chart/

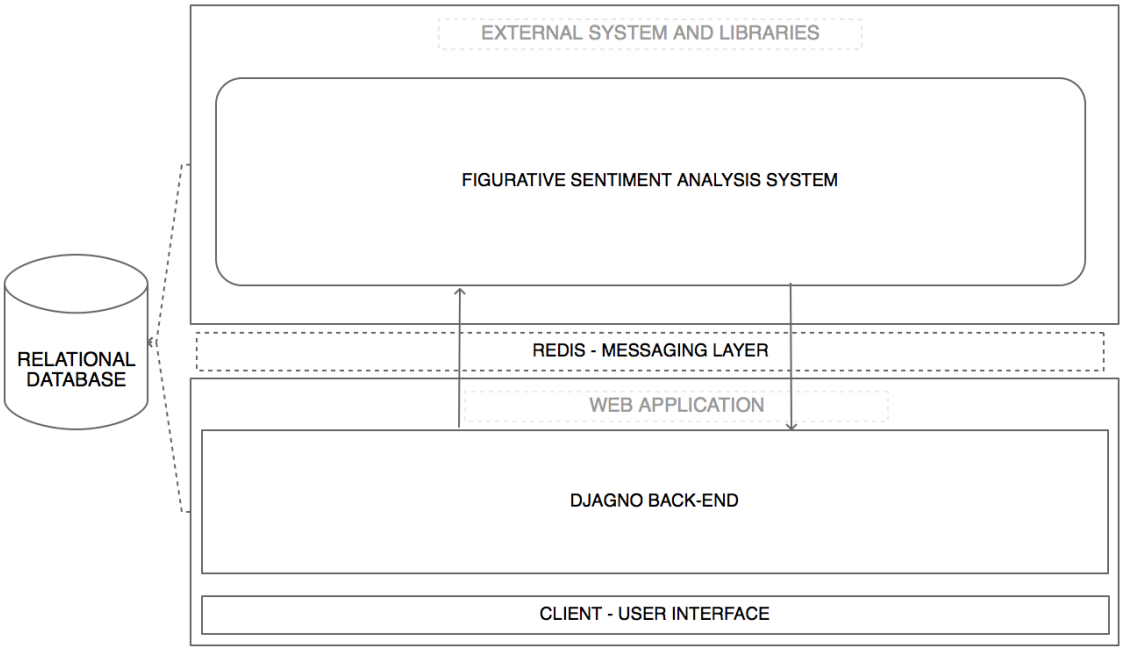
[34] http://getbootstrap.com

[35] http://www.bootstraptoggle.com/

Η Redis είναι μια in-memory δομή δεδομένων ανοιχτού κώδικα (άδεια χρήσης BSD), η οποία χρησιμοποιείται ως βάση δεδομένων, cache και message broker. Υποστηρίζει δομές δεδομένων, όπως strings, hashes, lists, sets, sorted sets μέσω επερωτημάτων εύρους (range queries), bitmatps, hyperlogs και geospatial indexes με επερωτήματα ακτίνας (radius queries). Χρησιμοποιήθηκε ως message broker, μέσω του pubsub (publish–subscribe) που διαθέτει, για την ασύγχρονη επικοινωνία του client με το Django backend, για την ενημέρωση του χρήστη ως προς το στάδιο στο οποίο βρίσκεται το πείραμα που έχει ζητηθεί[36].

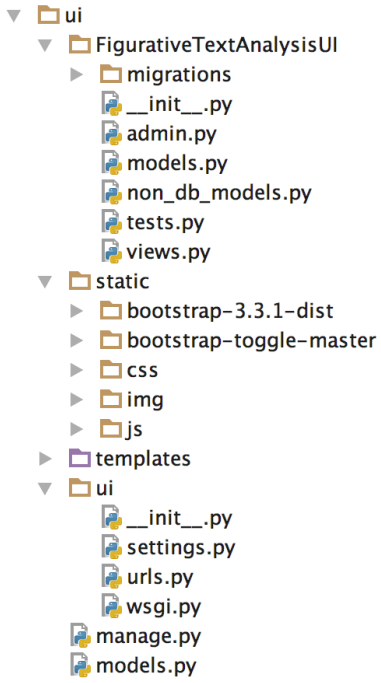
7.3 ΔΟΜΗ

Η δομή της εφαρμογής φαίνεται στις ακόλουθες εικόνες. Το κομμάτι που σημειώνεται ως Web Application είναι υπεύθυνο για την ενορχήστρωση της κάθε διαδικασίας, καθώς αυτό επικοινωνεί με τη βάση και χρησιμοποιεί το Figurative Sentiment Analysis System για την πειραματική διαδικασία. Έχει εισαχθεί και ένα messaging layer, το οποίο το έχει αναλάβει μια Redis, για το feedback προς τον χρήστη, καθώς και η διαδικασία του preprocessing και η διαδικασία των δοκιμών είναι αρκετά χρονοβόρες.



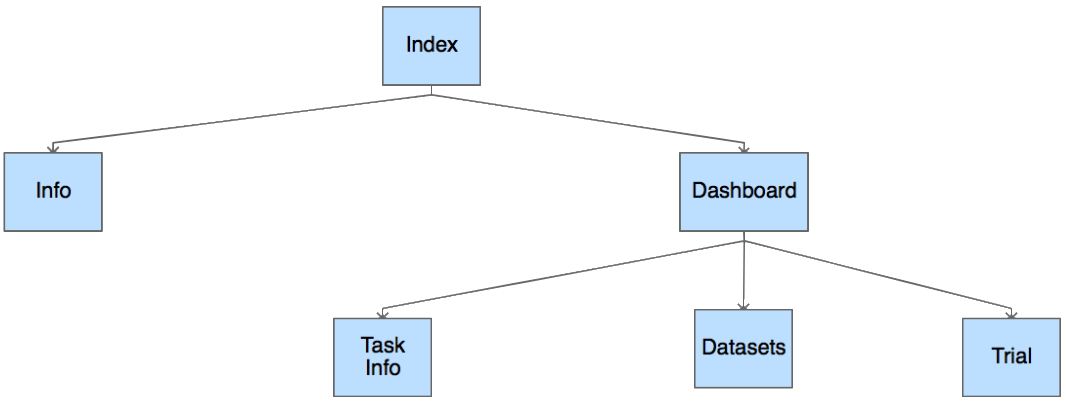
Σχήμα 46. Η αρχιτεκτονική της εφαρμογής

[36] http://redis.io



Σχήμα 47. Η δομή του Web Application Project

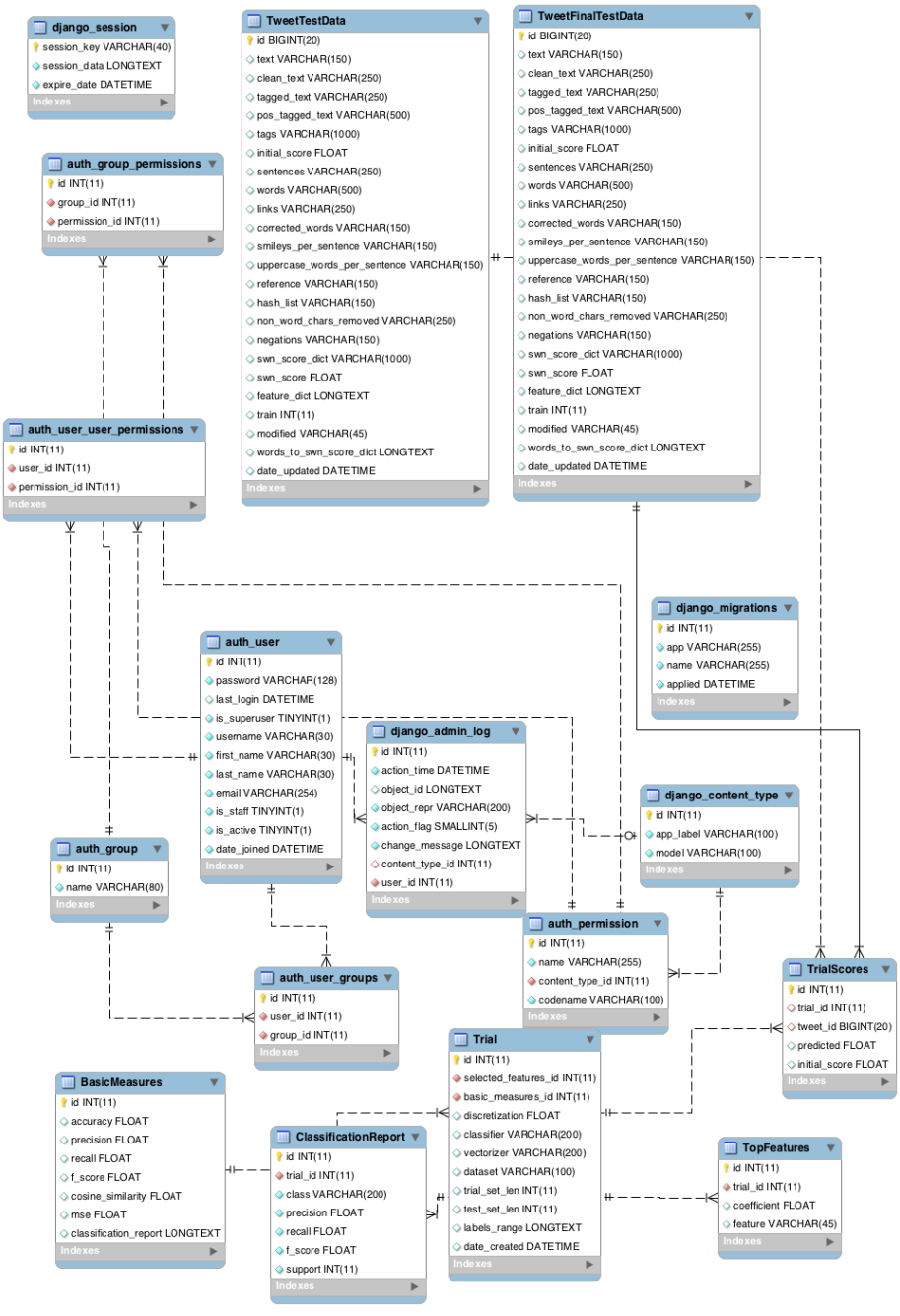
Στο Σχήμα 48 περιγράφεται η δομή των σελίδων της εφαρμογής. Η είσοδος γίνεται από την index σελίδα. Από εκεί ο χρήστης έχει τη δυνατότητα να προχωρήσει στην info για να δει κάποιες γενικές λεπτομέρειες για την εφαρμογή ή να μπει στη σελίδα dashboard και να επιλέξει από το μενού μια από τις ακόλουθες σελίδες: Task Info για να δει λεπτομέρειες για το SemEval Task 11, Datasets για να δει λεπτομέρειες για τα δεδομένα που χρησιμοποιήθηκαν, Results για να δει τα τελικά αποτελέσματα του Task 11, Trial για να κάνει δοκιμές με διάφορες παραμέτρους και να δει τα αποτελέσματα και Utils για τις επιλογές προ-επεξεργασίας των δεδομένων.



Σχήμα 48. Δομή σελίδων

7.4 ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ

Η βάση Sentifeed που περιγράφηκε εμπλουτίστηκε με τους πίνακες που δημιουργεί το Django για τη λειτουργία του. Αυτή η διαδικασία είναι αυτόματη με την εντολή syncdb του Django[37].



Σχήμα 49. Διάγραμμα Οντοτήτων-Συσχετίσεων (ER) της βάσης της εφαρμογής

[37] https://docs.djangoproject.com/en/1.9/ref/django-admin/

7.5 ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ

Το Django διαθέτει έναν ενσωματωμένο server, η χρήση του οποίου προορίζεται μόνο για τα πλαίσια των δοκιμών σε προγραμματιστικό επίπεδο (development server). Οι οδηγίες χρήσης καλύπτουν αυτή την περίπτωση.

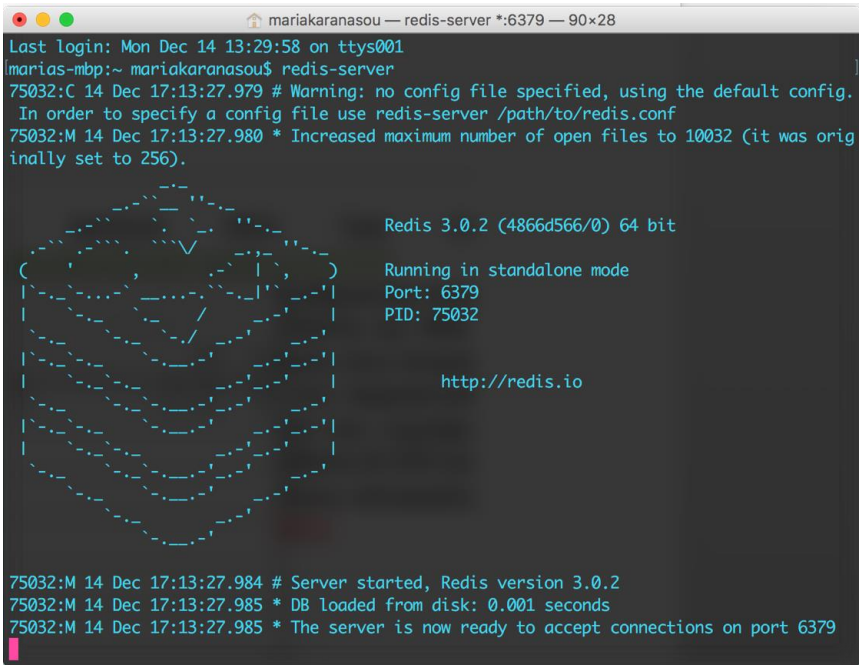
Για να ξεκινήσει ο server του Django τρέχουμε την ακόλουθη εντολή:

```
/usr/bin/python /Users/mariakaranasou/Projects/figurative-text-analysis/ui/manage.py runserver
Performing system checks...

System check identified no issues (0 silenced).
November 15, 2015 - 13:39:05
Django version 1.8.4, using settings 'ui.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CONTROL-C.
```

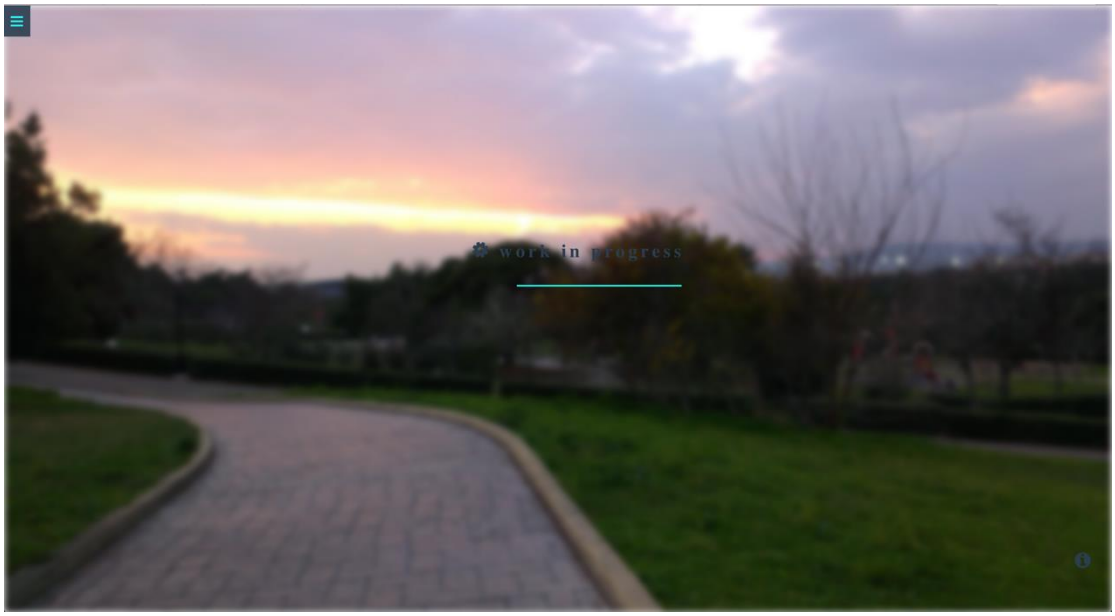
Σχήμα 50. Εκκίνηση του Django Development server

Προϋπόθεση για να τρέξει σωστά η εφαρμογή είναι η Redis να τρέχει, οπότε την ξεκινάμε σε μια γραμμή εντολων με την εντολή “redis-server” (δεδομένου ότι έχει γίνει η προεπιλεγμένη εγκατάσταση).

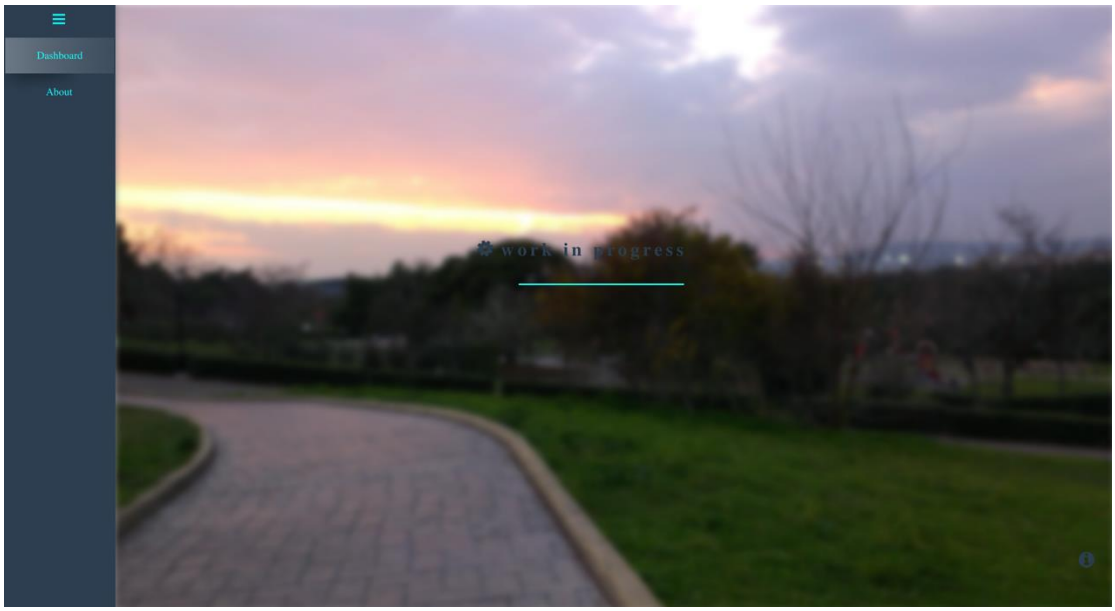


Σχήμα 51. Έναρξη redis-server

Ανοίγοντας σε έναν browser την διεύθυνση http://127.0.0.1:8000/ βλέπουμε την ακόλουθη εικόνα:

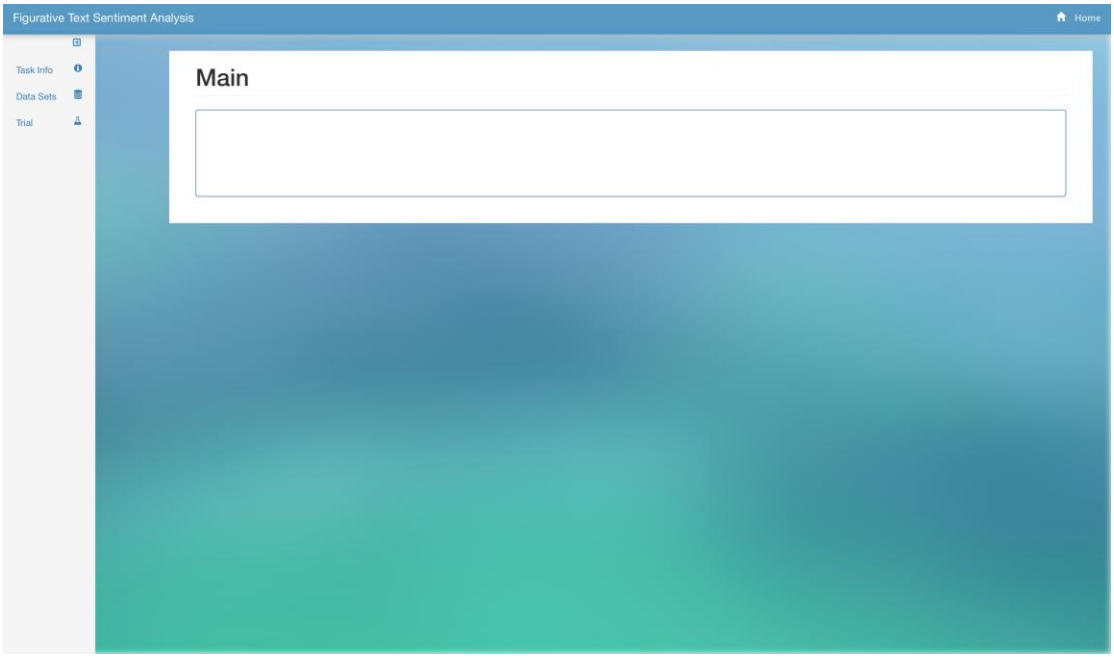


Σχήμα 52. Αρχική σελίδα (index)



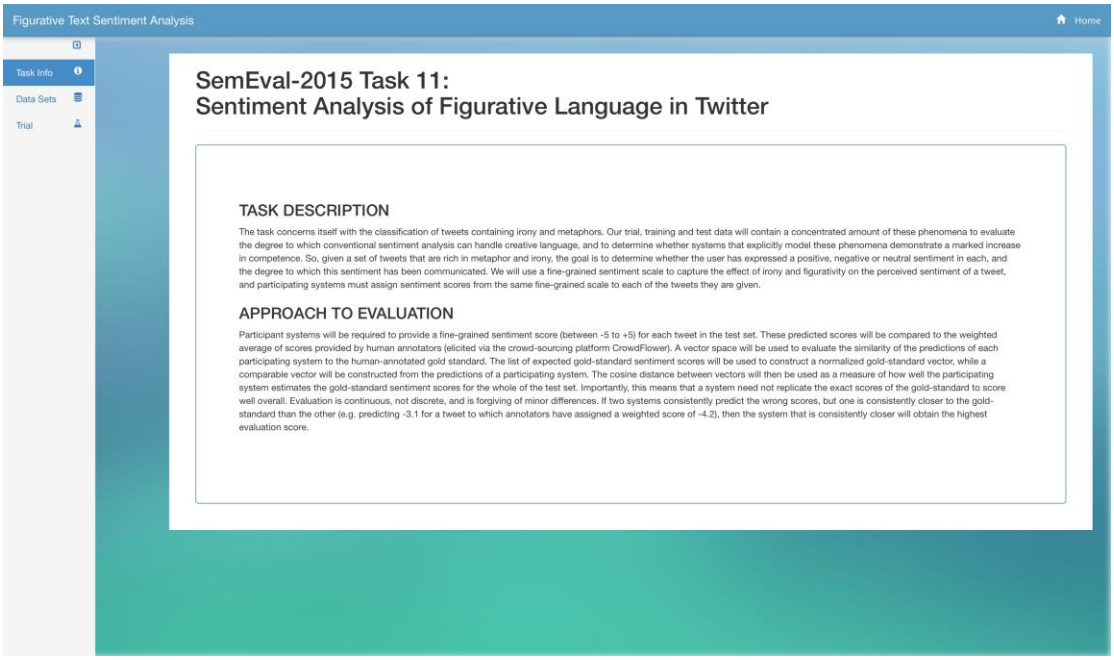
Σχήμα 53. Το μενού της αρχικής σελίδας

Από το μενού στα δεξιά επιλέγοντας Dashboard βρισκόμαστε στην αρχική σελίδα όπως φαίνεται στην ακόλουθη εικόνα:



Σχήμα 54. Η σελίδα dashboard με τις επιλογές της

Task Info: μια σύντομη περιργαφή του προβλήματος όπως αυτό τίθεται από το SemEval 2015 Task 11.

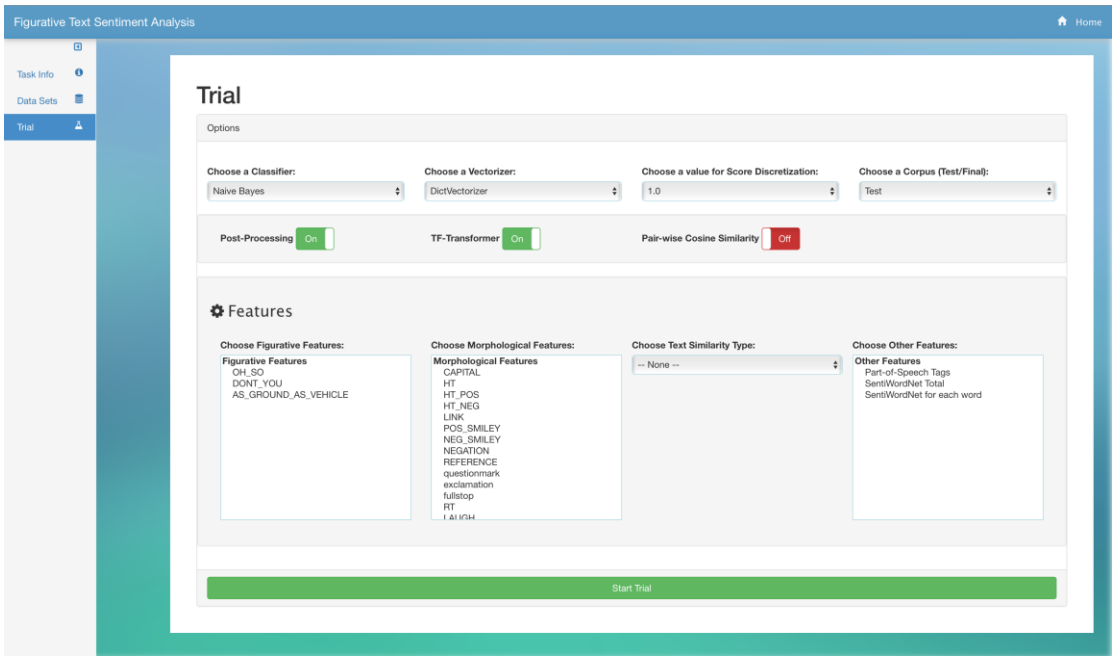


Σχήμα 55. Η σελίδα Task Info

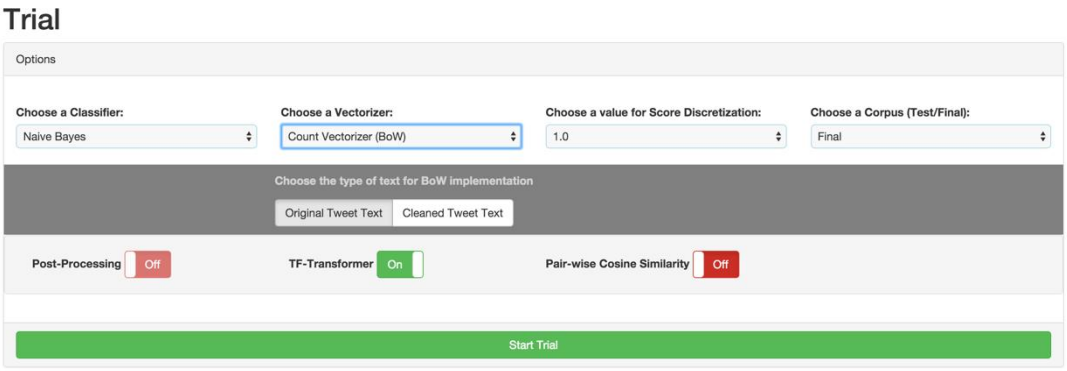
Στην επιλογή του μενού Data Sets γίνεται η ανάλυση των βασικών χαρακτηριστικών των διαθέσιμων δεδομένων του προβλήματος. Από τα σχήματα φαίνεται ότι η κατηγορία «negative» είναι η κυρίαρχη σε όλα τα σύνολα των δεδομένων που χρησιμοποιήθηκαν.

Σχήμα 56. Η σελίδα Datasets

Στην επιλογή του μενού Trial δίνεται η δυνατότητα να τρέξει κανείς δοκιμές επιλέγοντας Classifier, Vectorizer, Discretization, Dataset (Test/Final), εάν θα γίνει post-processing, εάν θα χρησιμοποιηθεί TF transformer, εάν θα γίνει υπολογισμός του pairwise cosine similariy και ποια από τα χαρακτηριστικά (features) των tweets θα ληφούν υπόψιν στη δοκιμή. Επίσης, στην περίπτωση που επιλεχθεί ο CountVectorizer, η δοκιμή θα γίνει με Bag-of-Words μοντέλο, και εμφανίζεται η επιλογή εάν θα χρησιμοποιηθεί το αρχικό κείμενο του tweet ή το «καθαρισμένο» κείμενο (Σχήμα 58).

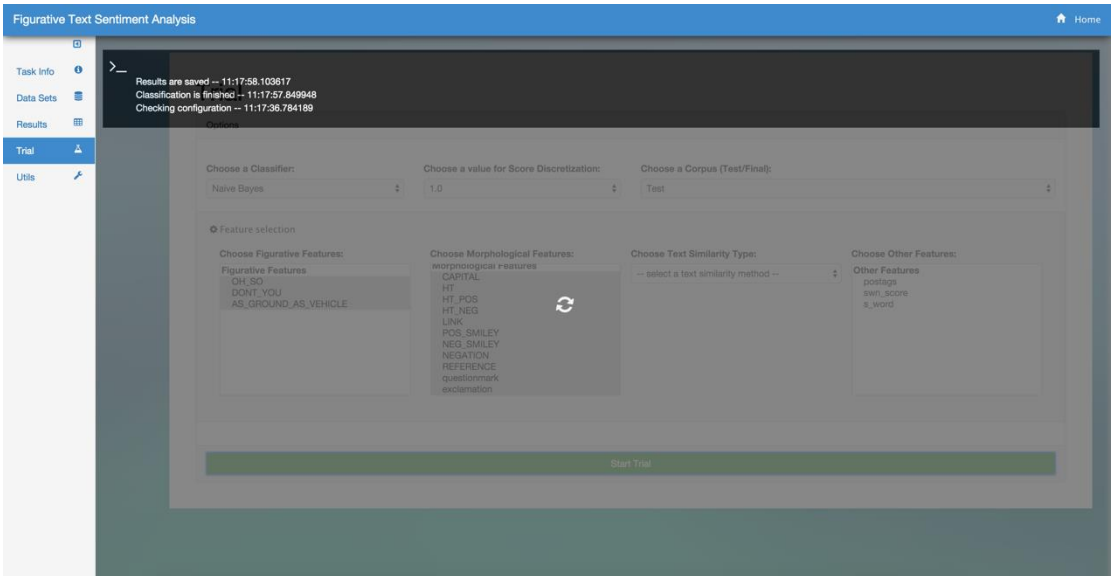


Σχήμα 57. Η σελίδα των δοκιμών (Trial) με τις επιλογές ενός πειράματος



Σχήμα 58. Οι επιλογές του BoW μοντέλου

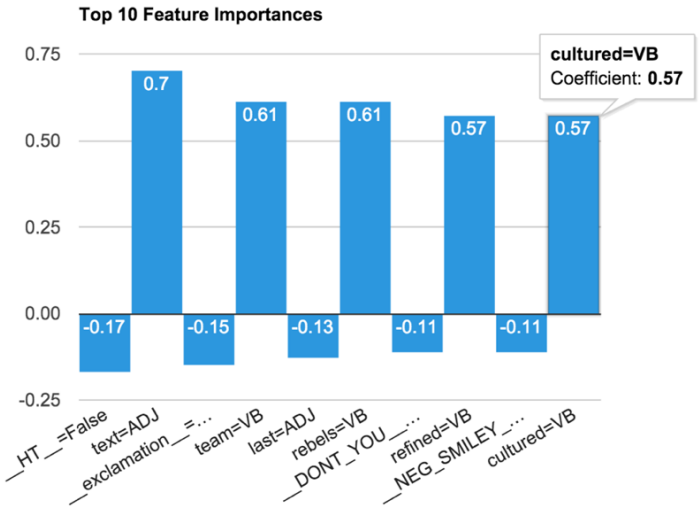
Κατά τη διάρκεια ενός πειράματος, γίνονται publish μηνύματα στην redis ώστε να ενημερώνεται ο χρήστης για το στάδιο στο οποίο βρίσκεται το πείραμα. Παράδειγμα τέτοιων μηνυμάτων φαίνονται στο Σχήμα 59.



Σχήμα 59. Η ανάδραση (feedback) κατά τη διάρκεια του πειράματος (Trial)

Όταν το πείραμα ολοκληρωθεί, παρουσιάζονται στην ίδια σελίδα τα αποτελέσματα, όπως φαίνονται στις ακόλουθες εικόνες. Αρχικά, παρουσιάζονται οι βασικές μετρικές cosine similarity, accuracy, precision, recall, f-score και mse. Το mse παρουσιάζεται σε ξεχωριστό διάγραμμα καθώς οι τιμές του δεν έχουν τα ίδια όρια με τις υπόλοιπες μετρικές οι οποίες βρίσκονται στο διάστημα [0,1]. Ακολουθεί μια γραφική παράσταση με τα δέκα πιο σημαντικά για τον classifier του πειράματος χαρακτηριστικά, μαζί με τα βάρη που τους έχουν δοθεί. Για κάποιους classifiers όπως ο Decision Tree, δεν είναι διαθέσιμο το διάγραμμα αυτό καθώς δεν υπάρχει η έννοια του συντελεστή (coefficient).

Σχήμα 60. Οι βασικές μετρικές για την αξιολόγηση του αποτελέσματος



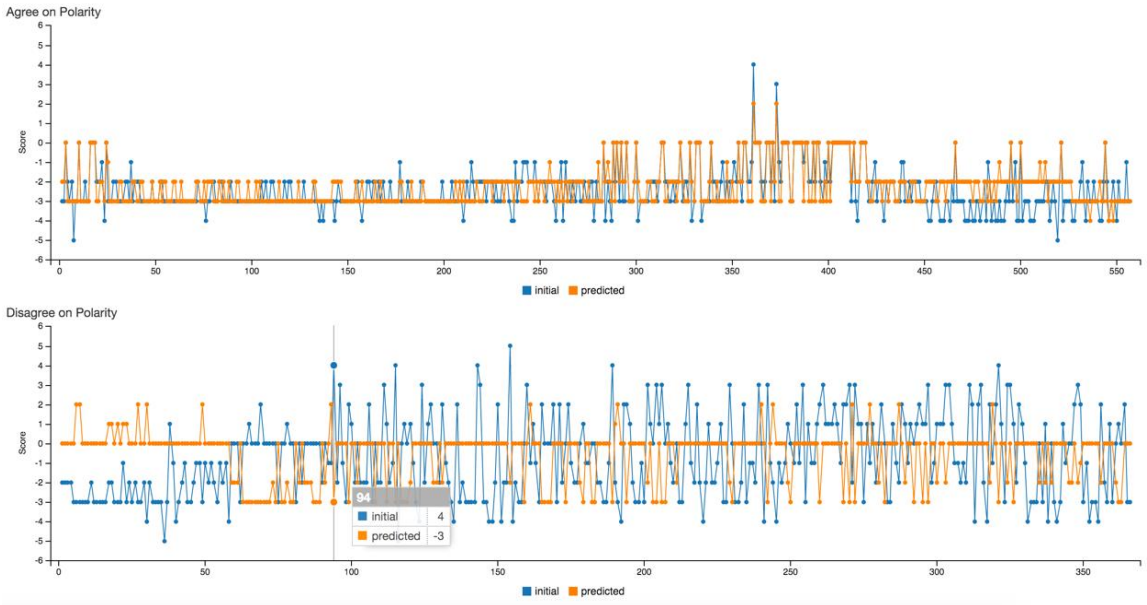
Σχήμα 61. Τα 10 πιο σημαντικά χαρακτηριστικά για τον classifier του πειράματος

Επιπλέον, παρουσιάζεται ένας πίνακας που περιέχει αναλυτικά για κάθε κατηγορία τα αποτελέσματα για precision, recall, f-score. Η στήλη Support περιέχει τον αριθμό των tweets ανά κατηγορία. Για παράδειγμα, τα tweets που έχουν βαθμολογηθεί με -5 είναι 4. Ο πίνακας αυτός βοηθάει στην καλύτερη εικόνα για το μοντέλο που έχει χτίσει ο classifier. Εάν δηλαδή ο classifier έχει ένα cosine similarity 0.6 αλλά αυτό οφείλεται στο ότι έχει προβλέψει ότι το 90% των tweets ως -3 τότε αυτό θα φανεί στον πίνακα ώστε να είναι πιο κατανοητό τι σημαίνει 0.60 cosine και πόσο καλό είναι το μοντέλο.

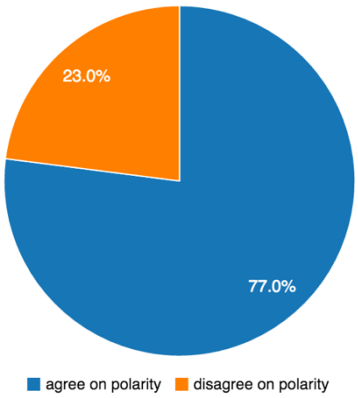
Class	Precision	Recall	F-Score	Support
-5	0	0	0	4
-4	0.41	0.11	0.17	100
-3	0.26	0.47	0.33	737
-2	0.46	0.43	0.45	1541
-1	0.18	0.05	0.07	680
0	0.11	0.27	0.16	298
1	0.08	0.05	0.06	169
2	0.08	0.04	0.05	155
3	0.32	0.11	0.16	201
4	0.25	0.01	0.02	111
5	0	0	0	4

Σχήμα 62. Πίνακας των αποτελεσμάτων ενός classifier ανά κατηγορία

Για τον ίδιο σκοπό παρουσιάζονται και δύο διαγράμματα, ένα για τις προβλέψεις που συμφωνούν σε πολικότητα με τα αρχικά scores και ένα για αυτές που διαφωνούν. Τα διαγράμματα αυτά δείχνουν πόσο σωστά ή λάθος είναι οι προβλέψεις του classifier, φαίνεται δηλαδή με μια πρώτη ματιά εάν υπάρχουν πολλά ακραία φαινόμενα, παραδείγματος χάριν, εάν μια συγκεκριμένη κατηγορία έστω τη -3 την κατηγοριοποιεί συνεχώς ως 5, ή εάν οι προβλέψεις που διαφωνούν έχουν μεγάλη απόσταση από τις πραγματικές βαθμολογίες, π.χ. -5 και 5. Το διάγραμμα που ακολουθεί, δίνει μια σύνοψη της κατάστασης όσον αφορά την πολικότητα, δηλαδή τα ποσοστά επιτυχίας, αποτυχίας, όσον αφορά τις κατηγορίες “positive”, “negative”, “neutral”.

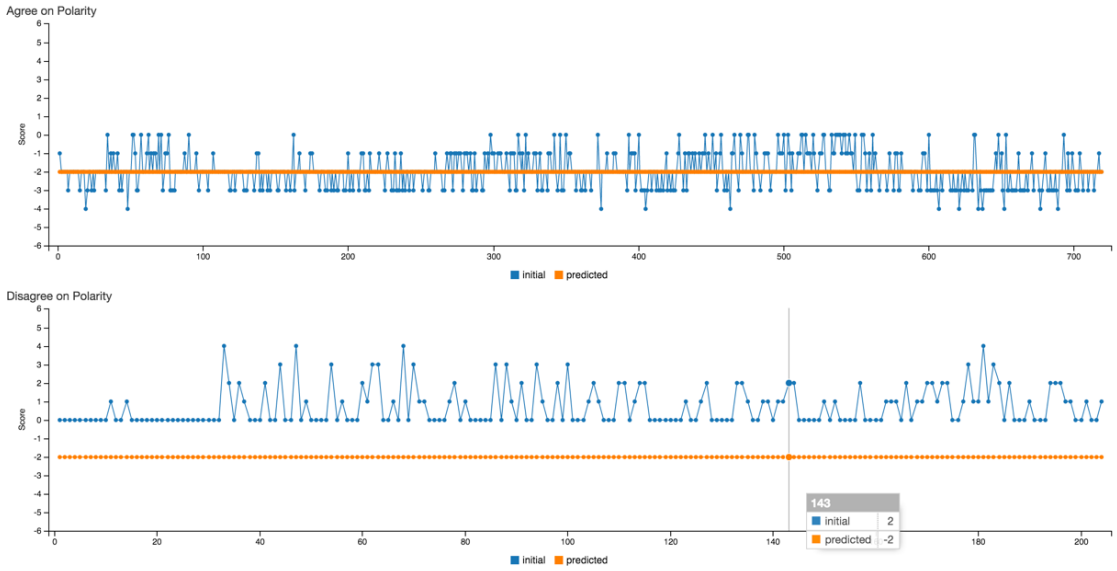


Σχήμα 63. Τα διαγράμματα δείχνουν το πόσο σωστές (Agree on Polarity) ή λάθος (Disagree on Polarity) είναι οι προβλέψεις του classifier του πειράματος



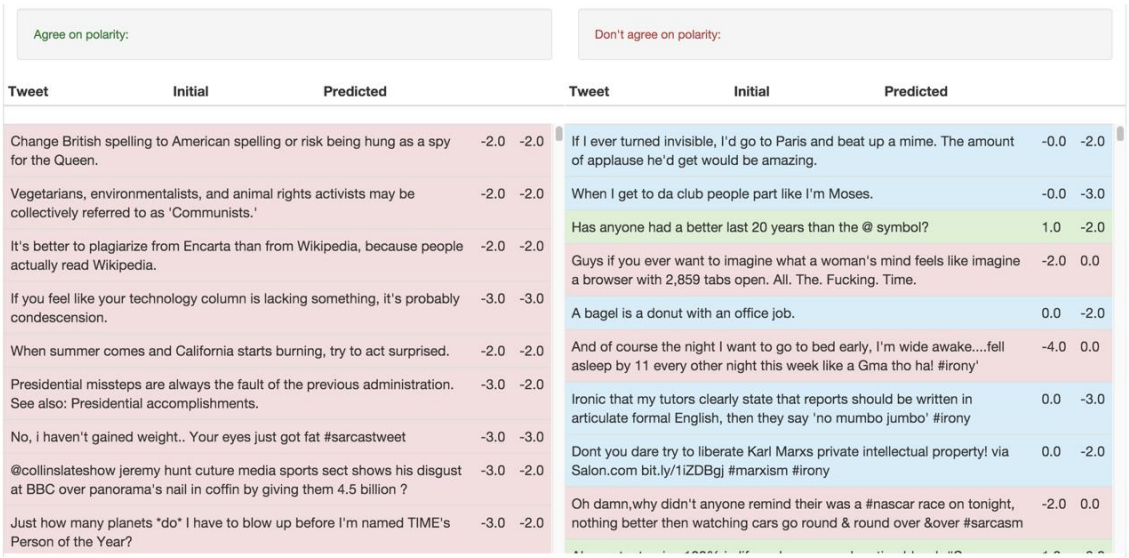
Σχήμα 64. Το ποσοστό σωστών (agree) και λάθος (disagree) προβλέψεων όσον αφορά την πολικότητα των tweets

Η χρησιμότητα των δύο τελευταίων απεικονίσεων φαίνεται στο ακόλουθο παράδειγμα όπου το cosine similarity είναι 0.694 αλλά ο classifier έχει κατατάξει όλα τα tweets στην κατηγορία -2 όπως φαίνεται στο ακόλουθο σχήμα.



Σχήμα 65. Παράδειγμα προβληματικού μοντέλου

Τέλος, παρουσιάζονται σε δύο πίνακες τα αναλυτικά αποτελέσματα, δηλαδή το κείμενο του tweet, η σωστή βαθμολογία και η πρόβλεψη. Ο ένας πίνακας παρουσιάζει τα tweets που ο classifier κατηγοριοποίησε σωστά όσον αφορά την πολικότητα και ο άλλος πίνακας αυτά που δεν κατηγοριοποιήθηκαν σωστά.



Agree on polarity:

Tweet	Initial	Predicted
Change British spelling to American spelling or risk being hung as a spy for the Queen.	-2.0	-2.0
Vegetarians, environmentalists, and animal rights activists may be collectively referred to as 'Communists.'	-2.0	-2.0
It's better to plagiarize from Encarta than from Wikipedia, because people actually read Wikipedia.	-2.0	-2.0
If you feel like your technology column is lacking something, it's probably condescension.	-3.0	-3.0
When summer comes and California starts burning, try to act surprised.	-2.0	-2.0
Presidential missteps are always the fault of the previous administration. See also: Presidential accomplishments.	-3.0	-2.0
No, i haven't gained weight.. Your eyes just got fat #sarcastweet	-3.0	-3.0
@collinslateshow jeremy hunt cuture media sports sect shows his disgust at BBC over panorama's nail in coffin by giving them 4.5 billion ?	-3.0	-2.0
Just how many planets *do* I have to blow up before I'm named TIME's Person of the Year?	-3.0	-2.0

Don't agree on polarity:

Tweet	Initial	Predicted
If I ever turned invisible, I'd go to Paris and beat up a mime. The amount of applause he'd get would be amazing.	-0.0	-2.0
When I get to da club people part like I'm Moses.	-0.0	-3.0
Has anyone had a better last 20 years than the @ symbol?	1.0	-2.0
Guys if you ever want to imagine what a woman's mind feels like imagine a browser with 2,859 tabs open. All. The. Fucking. Time.	-2.0	0.0
A bagel is a donut with an office job.	0.0	-2.0
And of course the night I want to go to bed early, I'm wide awake....fell asleep by 11 every other night this week like a Gma tho ha! #irony'	-4.0	0.0
Ironic that my tutors clearly state that reports should be written in articulate formal English, then they say 'no mumbo jumbo' #irony	0.0	-3.0
Dont you dare try to liberate Karl Marxs private intellectual property! via Salon.com bit.ly/1iZDBgj #marxism #irony	0.0	-2.0
Oh damn,why didn't anyone remind their was a #nascar race on tonight, nothing better then watching cars go round & round over &over #sarcasm	-2.0	0.0

Σχήμα 66. Αναλυτικά αποτελέσματα χωρισμένα σε αυτά που η πολικότητα που προβλέφθηκε συμφωνεί ή διαφωνεί με την πραγματική πολικότητα

8. ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΣΥΣΤΗΜΑΤΟΣ

8.1 ΜΕΤΡΙΚΕΣ - METRICS

Σε αυτό το κεφάλαιο περιγράφονται οι μετρικές που χρησιμοποιήθηκαν για να αξιολογηθεί η απόδοση του συστήματος, όσον αφορά την σωστή πρόβλεψη του συναισθήματος. Το πρώτο και κύριο κριτήριο στο οποίο δόθηκε το μεγαλύτερο βάρος Cosine Similarity, καθώς αυτό χρησιμοποιήθηκε από τους διοργανωτές του SemEval 2015 Task 11 για την αξιολόγηση των συστημάτων. Οι υπόλοιπες μετρικές παίζουν υποστηρικτικό ρόλο, στο να γίνουν πιο κατανοητά τα αποτελέσματα και η αξία αυτών.

8.1.1 COSINE SIMILARITY

Cosine Similarity είναι ένα μέτρο ομοιότητας μεταξύ δύο διανυσμάτων (vectors) ενός inner product space το οποίο μετράει το συνημίτονο της γωνίας μεταξύ τους. Το συνημίτονο των 0° είναι ίσο με 1, και είναι μικρό από 1 για οποιαδήποτε άλλη γωνία. Είναι ένα μέτρο προσανατολισμού και όχι μεγέθους, δηλαδή δύο διανύσματα με τον ίδιο προσανατολισμό έχουν cosine similarity ίση με 1, δύο κάθετα διανύσματα στις 90° δηλαδή, έχουν cosine similarity ίση με 0. Εάν δύο διανύσματα είναι αντιδιαμετρικά έχουν cosine similarity ίση με -1.Το μέτρο αυτό χρησιμοποιείται συνήθως στον θετικό άξονα, όπου το αποτέλεσμα περιορίζεται στο κλειστό διάστημα [0,1]. Η τεχνική αυτή χρησιμοποιείται και για την μέτρηση συνοχής σε clusters στον τομέα της εξόρυξης δεδομένων.

Συνεπώς, δεδομένου ενός διανύσματος $\hat{X}$ των προβλεφθέντων από το σύστημα labels και ενός διανύσματος X των πραγματικών labels των N tweets, το cosine similarity δίνεται από την ακόλουθη εξίσωση:

$$cos(X, \hat{X}) = \frac{\sum_{i=1}^{N} X_i \cdot \hat{X}_i}{\sqrt{\sum_{i=1}^{N} X_i^2} \cdot \sqrt{\sum_{i=1}^{N} \hat{X}_i^2}}$$

Εξίσωση 18. Cosine Similarity

8.1.2 ΜΕΣΟ ΤΕΤΡΑΓΩΝΙΚΟ ΣΦΑΛΜΑ - MEAN SQUARED ERROR (MSE)

Ως μέσο τετραγωνικό σφάλμα (MSE) ορίζεται ο μέσος όρος των τετραγώνων των λαθών που κάνει ένας classifier, δηλαδή των λάθος προβλέψεων. Ορίζεται από την Εξίσωση 19, όπου N ο αριθμός των tweets, $\hat{X}_i$ είναι η κλάση που προβλέφθηκε ότι ανήκει το tweet i και X_i είναι η πραγματική κλάση στην οποία ανήκει το tweet i. Η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος ταυτίζεται με την αποτελεσματικότητα και η βέλτιστη τιμή για αυτή τη μετρική είναι μηδέν.

$$MSE = \frac{\sum_{i=1}^{N} (X_i - \hat{X}_i)^2}{N}$$

Εξίσωση 19. Μέσο Τετραγωνικό Σφάλμα - Mean Squared Error

8.1.3 ACCURACY

Ως Accuracy ορίζεται ως ο αριθμός των σωστών προβλέψεων προς τον αριθμό του συνόλου που εξετάζεται. Πιο συγκεκριμένα, accuracy είναι το ποσοστό που προκύπτει από την διαίρεση του αριθμού των σωστών προβλέψεων των scores για ένα σύνολο από tweets δια το πλήθος των tweets.

$$accuracy = \frac{number\ of\ correctly\ classified\ items}{number\ of\ items}$$

Εξίσωση 20. Accuracy

8.1.4 PRECISION

Ως Precision ορίζεται η δυνατότητα ενός classifier να μην προβλέπει ως θετικό ένα δείγμα που είναι αρνητικό, να μην κατηγοριοποιεί δηλαδή ψευδώς ως θετικό το αρνητικό. Στην περίπτωσή μας, ως Precision ορίζεται το ποσοστό

$$tp/(tp + fp)$$

Εξίσωση 21. Precision

όπου tp είναι ο αριθμός των πραγματικών θετικών (true positive) και fp ο αριθμός των λάθος θετικών (false positive). Υψηλό precision σημαίνει ότι ο αλγόριθμος έχει επιστρέψει πολλά περισσότερα σχετικά αποτελέσματα από ότι μη σχετικά[38].

8.1.5 RECALL

Recall είναι το ποσοστό που περιγράφεται από την Εξίσωση 22 όπου tp είναι ο αριθμός των true positive και fn είναι ο αριθμός των false negatives. Διαισθητικά μπορεί να πει κανείς ότι recall είναι η ικανότητα ενός αλγορίθμου να βρεί όλα τα θετικά δείγματα, οπότε υψηλό recall σημαίνει ότι βρέθηκαν όλα τα πιο σχετικά αποτελέσματα[39].

$$tp/(tp + fn)$$

Εξίσωση 22. Recall

8.1.6 F-SCORE

Το F-score (F-measure) είναι μια μετρική που δείχνει την ακρίβεια ενός πειράματος καθώς εξαρτάται και από το precision και από το recall του πειράματος για τον υπολογισμό του, όπως

[38] https://en.wikipedia.org/wiki/Accuracy_and_precision

[39] https://en.wikipedia.org/wiki/Precision_and_recall

φαίνεται στην Εξίσωση 23 [40]. Οι πιθανές τιμές βρίσκονται στο διάστημα [0, 1], όπου 1 η καλύτερη τιμή και 0 η χειρότερη. Το F-score μπορεί να ερμηνευτεί ως ένας σταθμισμένος μέσος όρος του precision και του recall.

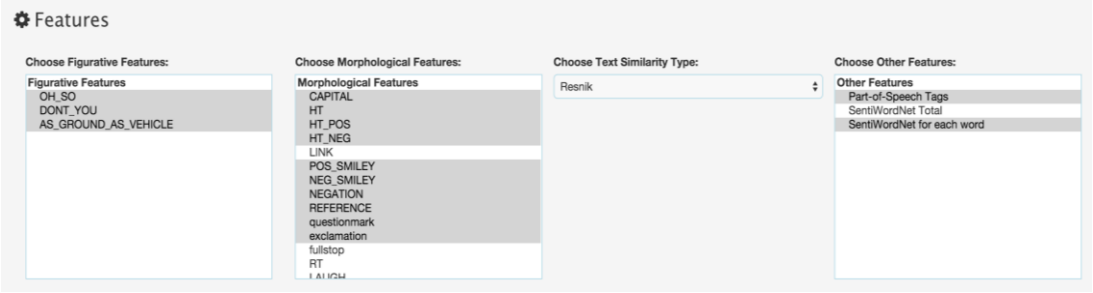
$$F - score = \frac{precision * recall}{precision + recall}$$

Εξίσωση 23. F-score

8.2 ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ

Ακολουθεί μια σειρά πειραμάτων τα οποία έγιναν και στα δεδομένα δοκιμής (αυτά που είχαν δοθεί για την εκπαίδευση και αξιολόγηση των συστημάτων) και στα τελικά δεδομένα. Συνοπτικά, έγιναν εκτιμήσεις για το πόση αξία έχουν στη διαδικασία τα διάφορα χαρακτηριστικά, το discretization της βαθμολογίας ώστε να αποφασιστεί εάν θα χρησιμοποιηθούν οι συνεχείς ή οι διακριτές τιμές, το Pairwise Cosine Similarity, η χρήση του TfidfTransformer, η χρήση της ομαδοποίησης (discretization) των τιμών των χαρακτηριστικών και τελικά η χρήση των διάφορων classifiers. Παρουσιάζονται στη συνέχεια αυτά που έχουν περισσότερο ενδιαφέρον.

Για την πλειοψηφία των πειραμάτων που ακολουθούν, έχουν χρησιμοποιηθεί τα χαρακτηριστικά που χρησιμοποιήθηκαν για την υποβολή των αποτελεσμάτων στο Task 11 και φαίνονται στο Σχήμα 67. Σε αντίθετη περίπτωση, αναφέρονται οι παραμέτροι των πειραμάτων αναλυτικά.



Σχήμα 67. Τα χαρακτηριστικά των πειραμάτων

8.2.1 ΑΞΙΟΛΟΓΗΣΗ FEATURES

Για την αξιολόγηση των χαρακτηριστικών έγιναν κάποιες δοκιμές με προσθαφαίρεση, λαμβάνοντας υπόψιν αυτά που αναφέρονται στη βιβλιογραφία ότι προσφέρουν περισσότερο στη διαδικασία. Από τις δοκιμές προκύπτουν κάποια χαρακτηριστικά που φαίνεται να παίζουν

[40] https://en.wikipedia.org/wiki/F1_score

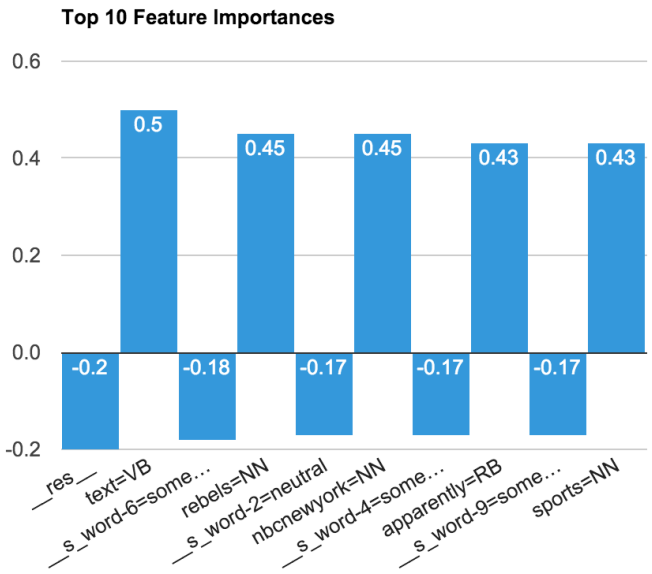
μεγάλο ρόλο πάντα και κάποια άλλα που η χρησιμότητά τους είναι μικρή και εξαρτάται και από τα υπόλοιπα χαρακτηριστικά.

Στα χαρακτηριστικά που φαίνεται να παίζουν ρόλο πάντα είναι τα ακόλουθα: Part-of-Speech tags και SentiWordNet score για κάθε λέξη.

Τα χαρακτηριστικά που φαίνεται να παίζουν κάποιο μικρό ρόλο στην διαμόρφωση του αποτελέσματος είναι τα ακόλουθα:

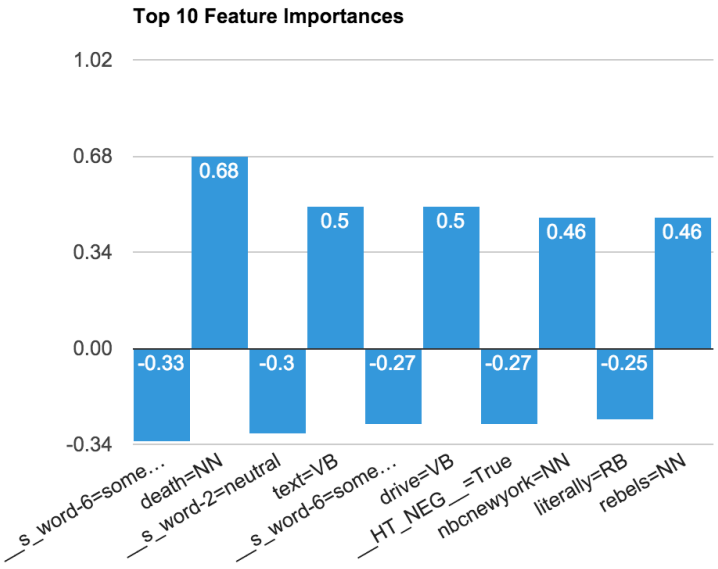
- Τα hashtags
- Το Semantic Similarity (Resnik)[41]
- Τα σημεία στίξης: "!", "?"
- Τα μοτίβα "Oh so", "Don't you", "As * as *", "Reference", "Negation", "Capitals"
- Τα emoticons

Τα υπόλοιπα χαρακτηριστικά που φαίνεται να δίνουν ασαφή αποτελέσματα.



Σχήμα 68. Τα 10 πιο σημαντικά χαρακτηριστικά - Δεδομένα Δοκιμής, Linear SVM

[41] Το Resnik Similarity Measure είχε περισσότερη σημασία στο Trial παρά στο Test set. Πιθανόν γιατί το Test set περιέχει και μη μεταφορικά tweets.



Σχήμα 69.Τα 10 πιο σημαντικά χαρακτηριστικά - Τελικά Δεδομένα, Linear SVM

COEFFICIENT	FEATURE
-0.203	__res__
0.495	text=VB
-0.180	__s_word-6=somewhat_negative
0.448	rebels=NN
-0.174	__s_word-2=neutral
0.448	nbcnewyork=NN
-0.172	__s_word-4=somewhat_positive
0.435	apparently=RB
-0.170	__s_word-9=somewhat_negative
0.427	sports=NN
-0.161	__HT__=False
0.424	national=ADJ
-0.140	__s_word-3=somewhat_negative
0.424	refined=VB
-0.140	__s_word-7=somewhat_negative
0.424	educated=ADJ
-0.139	__exclamation__=True
0.424	cultured=ADJ
-0.138	__s_word-3=neutral
0.424	cnbjftv=NN
-0.130	__s_word-5=somewhat_negative
0.416	__s_word-3=positive
-0.128	__questionmark__=False
0.407	tweet=NN

-0.128	__REFERENCE__=False
0.406	thoughts=NN
-0.126	__s_word-6=somewhat_positive
0.406	subnational=ADJ
-0.126	__s_word-1=somewhat_negative

Πίνακας 13. Τα 30 πιο σημαντικά χαρακτηριστικά - Δεδομένα Δοκιμής, Linear SVM

COEFFICIENT	FEATURE
-0.325	__s_word-6=somewhat_negative
0.678	death=NN
-0.298	__s_word-2=neutral
0.504	text=VB
-0.272	__s_word-6=somewhat_positive
0.498	drive=VB
-0.266	__HT_NEG__=True
0.464	nbcnewyork=NN
-0.251	literally=RB
0.458	rebels=NN
-0.233	__s_word-6=neutral
0.449	others=NN
-0.199	__exclamation__=True
0.432	refined=VB
-0.197	__HT__=True
0.432	educated=ADJ
-0.187	__HT_POS__=True
0.432	cultured=ADJ
-0.174	__s_word-1=somewhat_negative
0.432	cnbjftv=NN
-0.173	__s_word-6=positive
0.430	obviously=RB
-0.169	speak=VB
0.428	results=NN
-0.169	__s_word-8=somewhat_negative
0.428	Vogel=NN
-0.162	__s_word-10=somewhat_negative
0.428	sports=NN
-0.161	__s_word-3=somewhat_negative

Πίνακας 14.Τα 30 πιο σημαντικά χαρακτηριστικά - Τελικά Δεδομένα, Linear SVM

8.2.2 ΟΜΑΔΟΠΟΙΗΣΗ ΤΙΜΩΝ - DISCRETIZATION

Τα δεδομένα ελέγχου που δόθηκαν έχουν δύο επιλογές ως προς τη βαθμολόγησή τους, διακριτές και μη τιμές για scores. Για τη χρήση classifiers όπως ο N. Bayes και ο SVM που δεν δέχονται δεκαδικές τιμές ως είσοδο, έγιναν δοκιμές χωρίζοντας τις μη διακριτές τιμές σε διαστήματα, όπως φαίνεται στο Σχήμα 70. Για τον υπολογισμό της απόδοσης, χρησιμοποιείται η μέση τιμή του διαστήματος, δηλαδή εάν ένα tweet έχι προβλεφθεί ότι ανήκει στην κατηγορία "-5.0To-4.5", τότε ο μέσος όρος -4.75 και στη συνέχεια η στρογγυλοποίηση -5.0 χρησιμοποιείται για την σύγκριση με την διακριτή τιμή του S_{gold} για το tweet αυτό.

```
# 0.2
['-5.0To-4.8','-4.8To-4.6', ..., 'zero', ..., '4.6To4.8','4.8To5.0']
'-5.0To-4.8' = [-5.0, -4.8) -> ((-5.0) + (-4.8))/2.0 = -4.9
'zero' = 0
'4.8To5.0' = (4.8, 5.0] -> ((5.0) + (4.8))/2.0 = 4.9

# 0.5
['-5.0To-4.5', '-4.5To-4.0', ...,'zero', ..., '4.0To4.5','4.5To5.0']
'-5.0To-4.5' = [-5.0, -4.5) -> ((-5.0) + (-4.5))/2.0 = -4.75
'zero' = 0
'4.5To5.0' = (4.5, 5.0] -> ((5.0) + (4.5))/2.0 = 4.75

# 1.0
[-5, -4, ... , 0, ..., 4, 5]
```

Σχήμα 70. Ομαδοποίηση τιμών

Από τις δοκιμές που έγιναν, και όπως φαίνεται από τον Πίνακας 15, έχουμε κατά κανόνα καλύτερα αποτελέσματα όταν παίρνουμε τις ακέραιες τιμές (discretization=1.0).

COSINE SIMILARITY	ΔΕΔΟΜΕΝΑ ΔΟΚΙΜΩΝ			ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ		
CLASSIFIER	0.2	0.5	1	0.2	0.5	1
NAÏVE BAYES	0.699	0.687	0.703	0.545	0.548	0.555
DECISION TREE	0.667	0.675	0.687	0.46	0.458	0.436
LINEAR SVM	0.754	0.755	0.781	0.547	0.575	0.603

Πίνακας 15. Αξιολόγηση της χρήσης του discretization τιμών (cosine similarity)

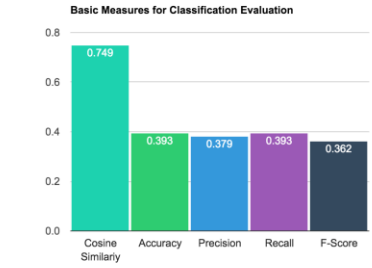
ACCURACY	ΔΕΔΟΜΕΝΑ ΕΛΕΓΧΟΥ			ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ		
CLASSIFIER	0.2	0.5	1	0.2	0.5	1
NAÏVE BAYES	0.372	0.345	0.334	0.185	0.191	0.23
DECISION TREE	0.324	0.294	0.319	0.212	0.231	0.222
LINEAR SVM	0.391	0.368	0.389	0.245	0.279	0.294

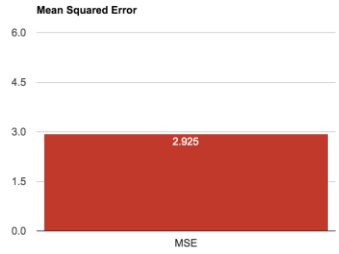
Πίνακας 16. Αξιολόγηση της χρήσης του discretization τιμών (accuracy)

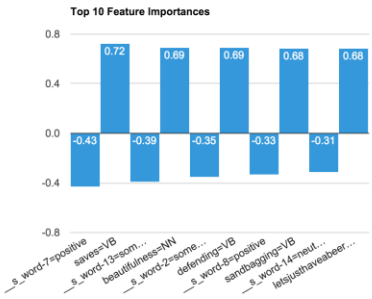
MSE	ΔΕΔΟΜΕΝΑ ΕΛΕΓΧΟΥ			ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ		
CLASSIFIER	0.2	0.5	1	0.2	0.5	1
NAÏVE BAYES	4.766	4.483	3.862	6.626	6.575	5.979
DECISION TREE	3.947	3.767	3.649	5.478	5.277	5.622
LINEAR SVM	2.863	2.792	2.519	4.636	4.167	3.929

Πίνακας 17. Αξιολόγηση της χρήσης του discretization τιμών (MSE)

Ακολουθούν τα αναλυτικά αποτελέσματα του classifier με τα καλύτερα αποτελέσματα στις δοκιμές (Linear SVM) για κάθε είδους discretization.

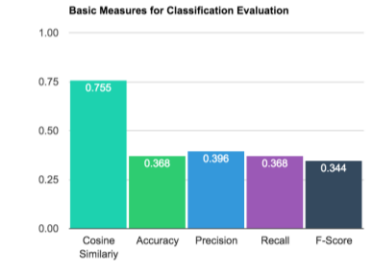


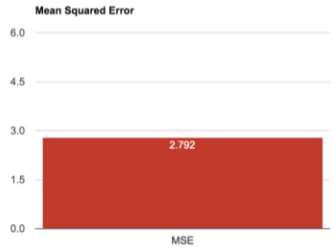


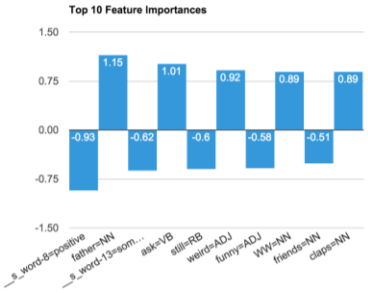


Class	Precision	Recall	F-Score	Support
-5	0	0	0	4
-4	0.44	0.1	0.16	72
-3	0.49	0.54	0.51	342
-2	0.27	0.37	0.31	209
-1	0.13	0.06	0.08	70
0	0.44	0.71	0.54	117
1	0.33	0.07	0.11	45
2	0.08	0.03	0.05	32
3	0.5	0.12	0.19	25
4	0.5	0.17	0.25	6
5	0	0	0	1

Σχήμα 71. Linear SVM – 0.2 – Δεδομένα Δοκιμών

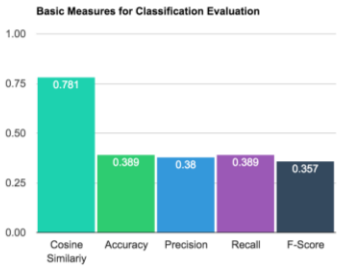


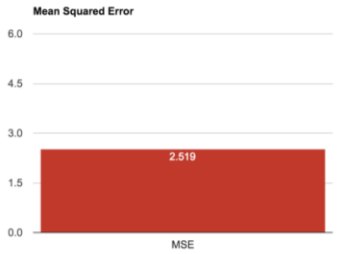


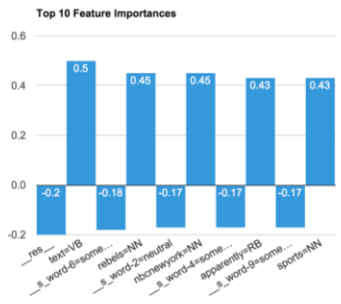


Class	Precision	Recall	F-Score	Support
-5	0	0	0	4
-4	0.75	0.04	0.08	72
-3	0.5	0.42	0.45	342
-2	0.27	0.52	0.36	209
-1	0.17	0.11	0.14	70
0	0.49	0.62	0.55	117
1	0.08	0.02	0.03	45
2	0.09	0.03	0.05	32
3	0.25	0.12	0.16	25
4	0.5	0.33	0.4	6
5	0	0	0	1

Σχήμα 72.Linear SVM – 0.5 – Δεδομένα Δοκιμών

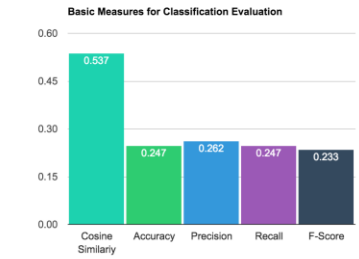


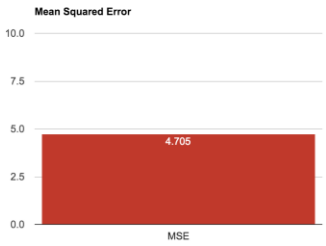


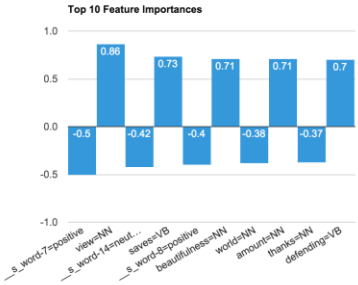


Class	Precision	Recall	F-Score	Support
-5	0	0	0	4
-4	0.4	0.03	0.05	72
-3	0.49	0.45	0.47	342
-2	0.3	0.53	0.38	209
-1	0.17	0.07	0.1	70
0	0.49	0.67	0.56	117
1	0.2	0.07	0.1	45
2	0.08	0.03	0.05	32
3	0.4	0.16	0.23	25
4	0.5	0.17	0.25	6
5	0	0	0	1

Σχήμα 73. Linear SVM – 1.0 – Δεδομένα Δοκιμών

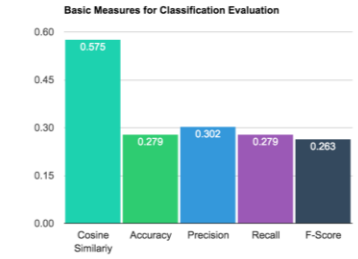


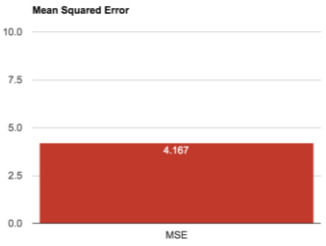


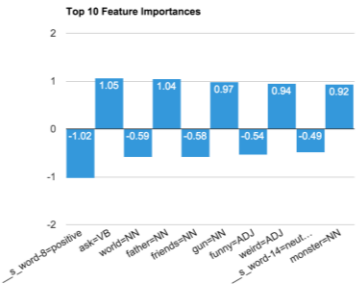


Class	Precision	Recall	F-Score	Support
-5	0	0	0	4
-4	0.1	0.09	0.09	100
-3	0.23	0.44	0.3	737
-2	0.4	0.32	0.35	1541
-1	0.18	0.05	0.08	680
0	0.12	0.3	0.17	298
1	0.13	0.08	0.1	169
2	0.06	0.04	0.05	155
3	0.22	0.1	0.14	201
4	0.25	0.01	0.02	111
5	0	0	0	4

Σχήμα 74. Linear SVM – 0.2 – Τελικά Δεδομένα

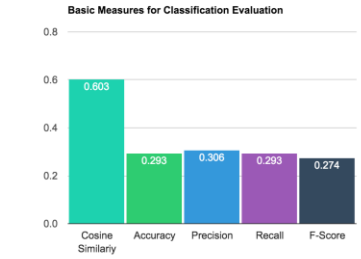


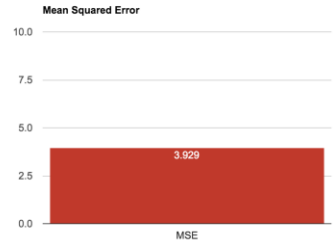


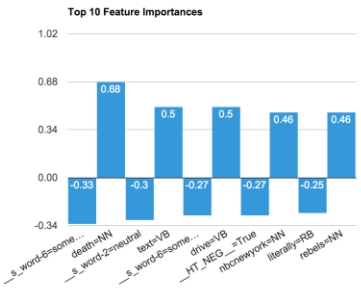


Class	Precision	Recall	F-Score	Support
-5	0	0	0	4
-4	0.23	0.09	0.13	100
-3	0.24	0.43	0.31	737
-2	0.44	0.42	0.43	1541
-1	0.24	0.07	0.1	680
0	0.1	0.24	0.14	298
1	0.07	0.04	0.05	169
2	0.04	0.02	0.03	155
3	0.22	0.07	0.11	201
4	0.67	0.02	0.04	111
5	0	0	0	4

Σχήμα 75.Linear SVM – 0.5 – Τελικά Δεδομένα







Class	Precision	Recall	F-Score	Support
-5	0	0	0	4
-4	0.36	0.12	0.18	100
-3	0.26	0.48	0.34	737
-2	0.46	0.42	0.44	1541
-1	0.19	0.05	0.07	680
0	0.11	0.27	0.15	298
1	0.11	0.07	0.09	169
2	0.04	0.02	0.03	155
3	0.34	0.12	0.18	201
4	0.33	0.02	0.03	111
5	0	0	0	4

Σχήμα 76. Linear SVM – 1.0 – Τελικά Δεδομένα

8.2.3 ΔΟΚΙΜΕΣ ΜΕ BOW

Καθώς το Bag-of-Words μοντέλο, θεωρείται από τα πιο απλά και βασικά σε τέτοιου είδους πειράματα, έγιναν δοκιμές δίνοντας ως είσοδο για τη δημιουργία του BoW είτε το κείμενο του tweet ως έχει, χωρίς καμία επεξεργασία, είτε το κείμενο που προκύπτει μετά τον καθαρισμό του tweet, είτε το feature dictionary που προκύπτει, αλλά σε κειμενική μορφή. Στους ακόλουθους πίνακες παρουσιάζονται τα αποτελέσματα των δοκιμών. Όπου BoW + TF είναι το BoW μοντέλο το οποίο έχει δημιουργηθεί με τη χρήση ενός CountVectorizer, ακολουθούμενο από έναν TfidfTransformer (χρήση TF μόνο) πριν αυτό δοθεί ως είσοδος στον classifier.

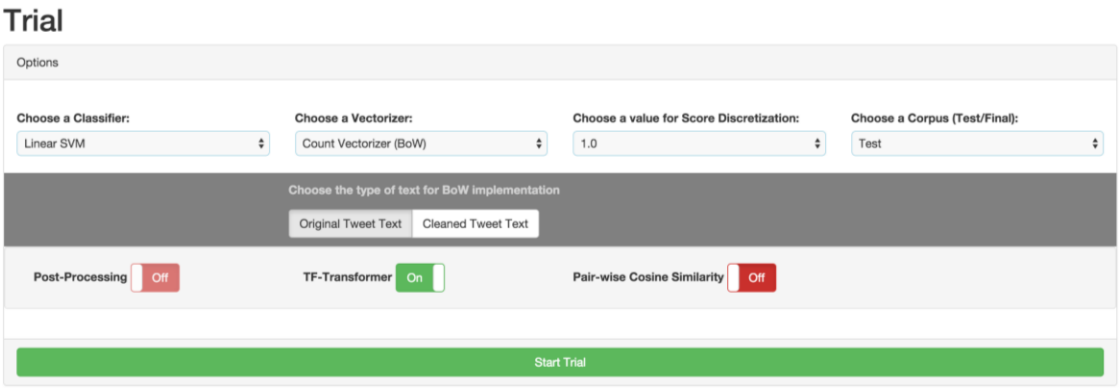
COSINE SIMILARITY	ΔΕΔΟΜΕΝΑ ΔΟΚΙΜΩΝ		ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ	
CLASSIFIER	BoW	BoW + TF	BoW	BoW + TF
NAÏVE BAYES	0.712	0.706	0.551	0.56
LINEAR SVM	0.73	0.746	0.5443	0.579
DECISION TREE	0.74	0.74	0.524	0.524

Πίνακας 18. BoW με τη χρήση του κειμένου του tweet (cosine similarity)

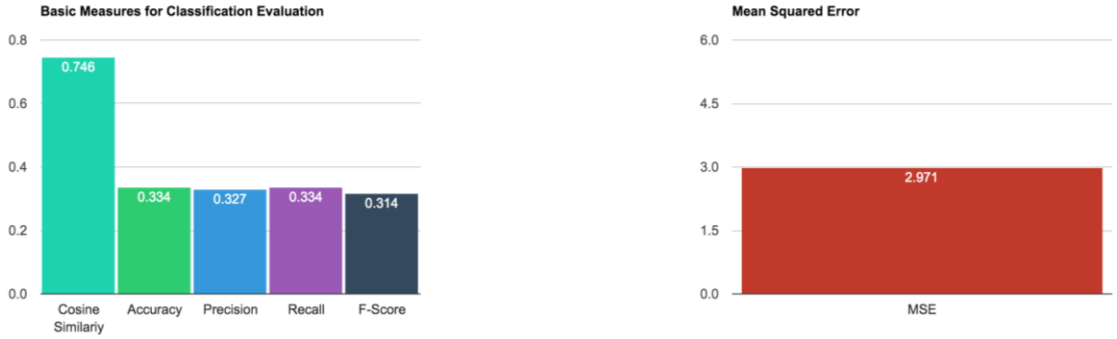
COSINE SIMILARITY	ΔΕΔΟΜΕΝΑ ΔΟΚΙΜΩΝ		ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ	
CLASSIFIER	BoW	BoW + TF	BoW	BoW + TF
NAÏVE BAYES	0.71	0.706	0.547	0.558
LINEAR SVM	0.726	0.716	0.513	0.561
DECISION TREE	0.735	0.745	0.525	0.542

Πίνακας 19. BoW με τη χρήση του "καθαρισμένου" κειμένου του tweet (cosine similarity)

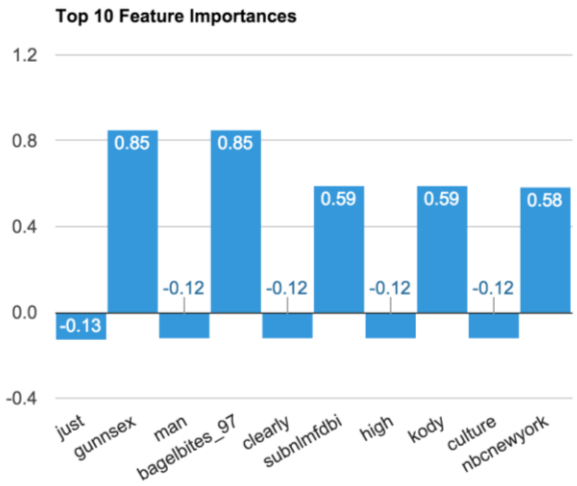
Αναλυτικά αποτελέσματα του classifier με τα καλύτερα αποτελέσματα της δοκιμής (Linear SVM, BoW + TF):



Σχήμα 77. Επιλογές δοκιμής BoW με TF - Linear SVM



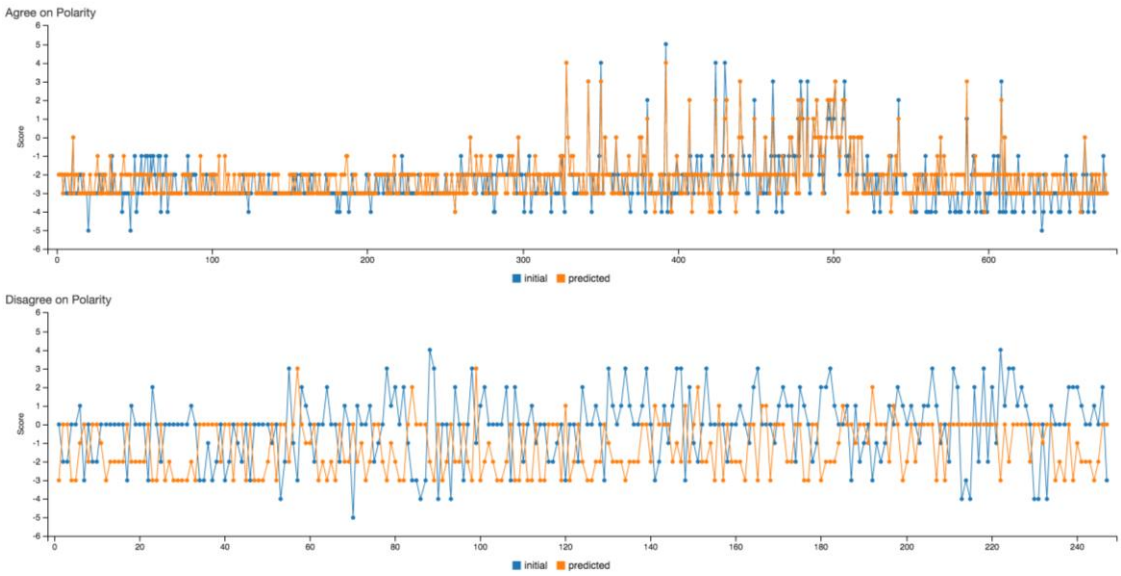
Σχήμα 78. Τα αποτελέσματα των βασικών μετρικών της δοκιμής BoW και TF - Linear SVM



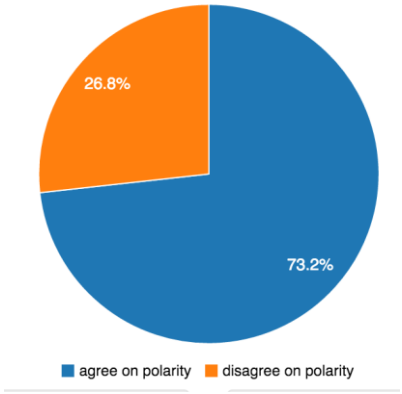
Σχήμα 79.Τα 10 πιο σημαντικά χαρακτηριστικά της δοκιμής BoW και TF - Linear SVM

Class	Precision	Recall	F-Score	Support
-5	0	0	0	4
-4	0.31	0.06	0.09	72
-3	0.49	0.5	0.5	342
-2	0.26	0.44	0.33	209
-1	0.06	0.04	0.05	70
0	0.23	0.25	0.24	117
1	0.06	0.02	0.03	45
2	0.25	0.13	0.17	32
3	0.43	0.12	0.19	25
4	0.5	0.17	0.25	6
5	0	0	0	1

Πίνακας 20.Τα αναλυτικά αποτελέσματα ανά κατηγορία της δοκιμής BoW και TF - Linear SVM



Σχήμα 80. Ανάλυση των αποτελεσμάτων σχετικά με τη πολικότητα της δοκιμής BoW και TF - Linear SVM



Σχήμα 81. Το ποσοστό των σωστών (agree) και λάθος (disagree) προβλέψεων σε σχέση με την πολικότητα της δοκιμής BoW και TF - Linear SVM

8.2.4 ΔΟΚΙΜΕΣ ΜΕ PAIRWISE COSINE SIMILARITY

Ο πίνακας που ακολουθεί, περιλαμβάνει τα αποτελέσματα των δοκιμών που έγιναν στα πλαίσια της αξιολόγησης του Pairwise Cosine Similarity (PwCS) στην διαδικασία της ανάλυσης συναισθήματος. Οι δοκιμές έγιναν με και χωρίς τη χρήση TF.

COSINE SIMILARITY	ΔΕΔΟΜΕΝΑ ΔΟΚΙΜΩΝ		ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ	
CLASSIFIER	Χωρίς PwCS	Με PwCS	Χωρίς PwCS	Με PwCS
NAÏVE BAYES	0.703	0.698	0.555	0.547
DECISION TREE	0.679	0.674	0.41	0.505
LINEAR SVM	0.781	0.684	0.603	0.546

Πίνακας 21. Αποτελέσματα δοκιμών αξιολόγησης του Pairwise Cosine Similarity – με TF (cosine similarity)

8.2.5 ΔΟΚΙΜΕΣ ΜΕ TF

Στα ακόλουθα αποτελέσματα, φαίνεται η αξία της χρήσης του TF, όπου η διαφορά μεταξύ των πειραμάτων χωρίς TF και με TF είναι αρκετά αισθητή. Τα πειράματα αυτής της κατηγορίας έγιναν με επιλογές και τα χαρακτηριστικά που φαίνονται στις εικόνες πιο κάτω.

COSINE SIMILARITY	ΔΕΔΟΜΕΝΑ ΔΟΚΙΜΩΝ		ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ	
CLASSIFIER	Χωρίς TF	Με TF	Χωρίς TF	Με TF
NAÏVE BAYES	0.643	0.703	0.41	0.555
LINEAR SVM	0.744	0.781	0.545	0.603

Πίνακας 22. Αποτελέσματα δοκιμών αξιολόγησης του TF (cosine similarity)

ACCURACY	ΔΕΔΟΜΕΝΑ ΕΛΕΓΧΟΥ		ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ	
CLASSIFIER	Χωρίς TF	Με TF	Χωρίς TF	Με TF
NAÏVE BAYES	0.14	0.334	0.128	0.23
LINEAR SVM	0.346	0.389	0.255	0.294

Πίνακας 23.Αποτελέσματα δοκιμών αξιολόγησης του TF (accuracy)

MSE	ΔΕΔΟΜΕΝΑ ΕΛΕΓΧΟΥ		ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ	
CLASSIFIER	Χωρίς TF	Με TF	Χωρίς TF	Με TF
NAÏVE BAYES	4.369	3.862	4.121	5.979
LINEAR SVM	2.95	2.519	4.5095	3.929

Πίνακας 24.Αποτελέσματα δοκιμών αξιολόγησης του TF (mse)

8.2.6 ΔΟΚΙΜΕΣ ΟΜΑΔΟΠΟΙΗΣΗΣ – FEATURE VALUE DISCRETIZATION

Καθώς η ομαδοποίηση που περιγράφηκε σε προηγούμενο κεφάλαιο έγινε εμπειρικά και με βάση τις παρατηρήσεις από διάφορες δοκιμές, έπρεπε να αξιολογηθεί το κατά πόσο χρήσιμο είναι τελικά στη διαδικασία. Αυτός είναι και ο στόχος των πειραμάτων που ακολουθούν.

Από τα αποτελέσματα φαίνεται πως η ομαδοποίηση προσθέτει κάποια αξία στη διαδικασία όσον αφορά τον Linear SVM. Οι υπόλοιποι classifiers της δοκιμής δεν φαίνεται να δίνουν θετικά αποτελέσματα σχετικά με τη χρήση του.

COSINE SIMILARITY	ΔΕΔΟΜΕΝΑ ΔΟΚΙΜΩΝ		ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ	
CLASSIFIER	Χωρίς ομαδοποίηση	Με ομαδοποίηση	Χωρίς ομαδοποίηση	Με ομαδοποίηση
NAÏVE BAYES	0.705	0.703	0.555	0.555
DECISION TREE	0.644	0.683	0.45	0.429
LINEAR SVM	0.77	0.781	0.579	0.603

Πίνακας 25. Αποτελέσματα δοκιμών αξιολόγησης της ομαδοποίησης (cosine similarity)

ACCURACY	ΔΕΔΟΜΕΝΑ ΔΟΚΙΜΩΝ		ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ	
CLASSIFIER	Χωρίς ομαδοποίηση	Με ομαδοποίηση	Χωρίς ομαδοποίηση	Με ομαδοποίηση
NAÏVE BAYES	0.34	0.334	0.221	0.23
DECISION TREE	0.31	0.341	0.226	0.219
LINEAR SVM	0.389	0.389	0.281	0.294

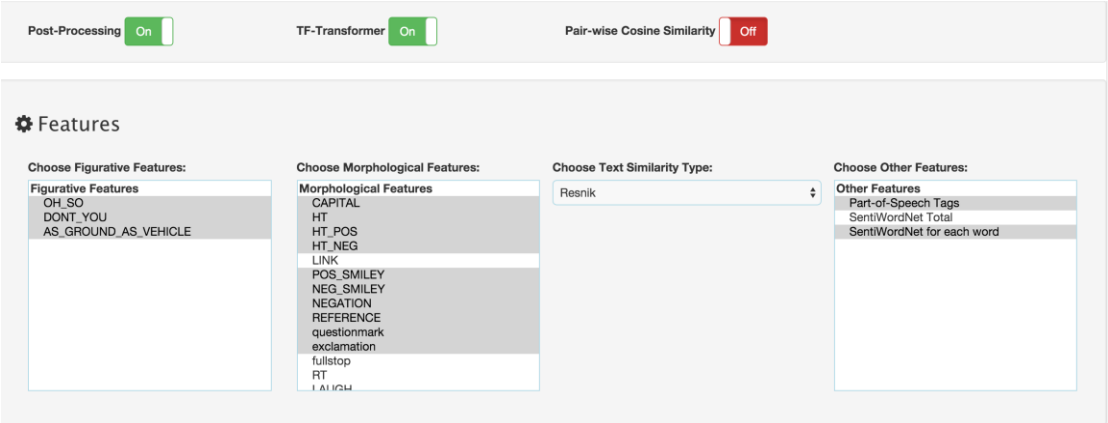
Πίνακας 26. Αποτελέσματα δοκιμών αξιολόγησης της ομαδοποίησης (accuracy)

MSE	ΔΕΔΟΜΕΝΑ ΔΟΚΙΜΩΝ		ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ	
CLASSIFIER	Χωρίς ομαδοποίηση	Με ομαδοποίηση	Χωρίς ομαδοποίηση	Με ομαδοποίηση
NAÏVE BAYES	3.883	3.862	6.046	5.979
DECISION TREE	4.163	3.725	5.379	5.648
LINEAR SVM	2.634	2.519	4.138	3.929

Πίνακας 27. Αποτελέσματα δοκιμών αξιολόγησης της ομαδοποίησης (mse)

8.2.7 ΠΕΙΡΑΜΑΤΑ ΜΕ ΔΙΑΦΟΡΕΤΙΚΟΥΣ CLASSIFIERS

Αφού από τις προηγούμενες δοκιμές έχουμε καταλήξει σχετικά σε μια διαμόρφωση πειραμάτων η οποία φαίνεται να έχει τα καλύτερα αποτελέσματα, γίνονται δοκιμές με διαφορετικούς classifiers, ώστε να συμπεράνουμε ποιος είναι ο πιο κατάλληλος για την επίλυση του προβλήματος μας.



Σχήμα 82. Τα χαρακτηριστικά των ακόλουθων δοκιμών

Από τα ακόλουθα αποτελέσματα, δύο classifiers ξεχωρίζουν, ο Linear SVM και ο SVR (Support Vector Regressor). Όπως φαίνεται από τα αναλυτικά αποτελέσματα, ο SVR, παρόλο που στα τελικά δεδομένα δίνει πολύ καλά αποτελέσματα, στα δεδομένα δοκιμών δεν δίνει τόσο καλά αποτελέσματα, πράγμα που σημαίνει ότι, καθώς η αξιολόγηση στα τελικά αποτελέσματα δεν είναι γνωστή εκ των προτέρων, δεν θα είχε επιλεγεί για να σταλούν τα αποτελέσματα στο SemEval Task 11. Επιπλέον, παρατηρώντας τις τιμές που προβλέπονται από τον SVR, βλέπουμε ότι κυμαίνονται στο διάστημα [-1, 1], πράγμα που σημαίνει ότι χρειάζεται πολύ καλύτερη διερεύνηση για την ορθότητά τους. Τέλος, οι τιμές που δίνονται από τον SVR αλλάζουν μεταβάλοντας τη μεταβλητή max_iterations, οπότε στην παρούσα φάση δεν είναι δυνατόν να βγάλουμε ασφαλή αποτελέσματα για το ποιο είναι το βέλτιστο μοντέλο του. Ο Linear SVM, έχει σταθερά καλή απόδοση και στις δοκιμές και στα τελικά δεδομένα, οπότε και επιλέχθηκε για την αποστολή των τελικών αποτελεσμάτων του Task 11.

CLASSIFIER	ΔΕΔΟΜΕΝΑ ΔΟΚΙΜΩΝ	ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ
ΝΑΪVE BAYES	0.703	0.555
DECISION TREE	0.676	0.435
LINEAR SVM	0.781	0.603
SVM RBF KERNEL	0.698	0.547
SVR [42]	0.761	0.672
SGD	0.734	0.526
PERCEPTRON	0.689	0.415

Πίνακας 28. Αποτελέσματα δοκιμών αξιολόγησης των classifiers με τα τελικά χαρακτηριστικά (cosine similarity)

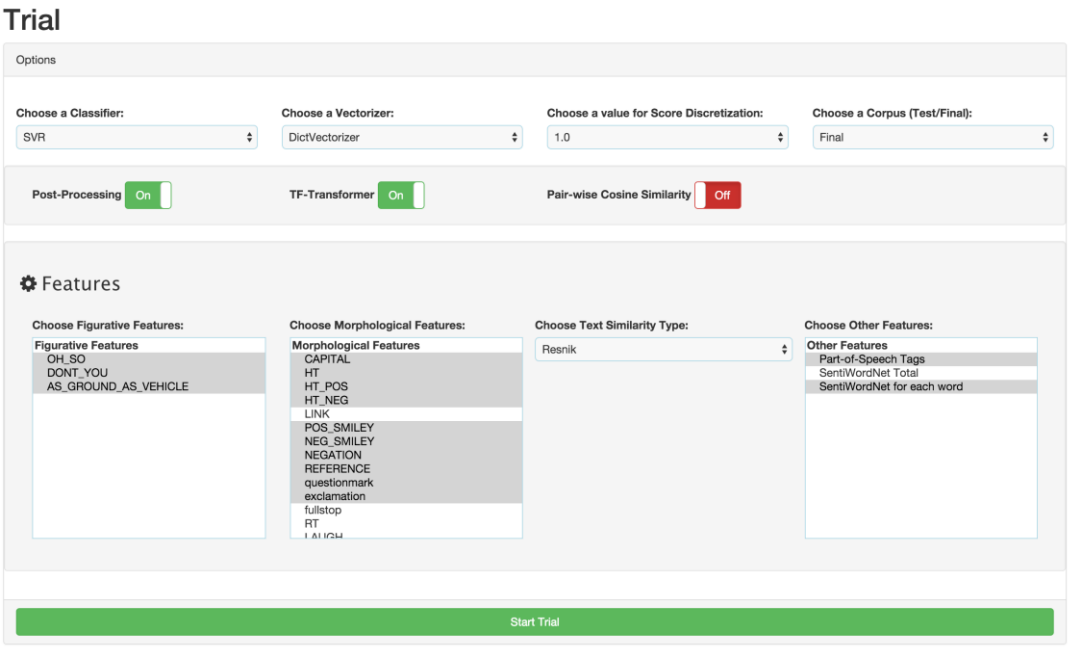
CLASSIFIER	ΔΕΔΟΜΕΝΑ ΔΟΚΙΜΩΝ	ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ
ΝΑΪVE BAYES	0.334	0.23
DECISION TREE	0.327	0.221
LINEAR SVM	0.389	0.334
SVM RBF KERNEL	0.371	0.184
SVR	0.236	0.236
SGD	0.364	0.282
PERCEPTRON	0.314	0.215

Πίνακας 29. Αποτελέσματα δοκιμών αξιολόγησης των classifiers με τα τελικά χαρακτηριστικά (accuracy)

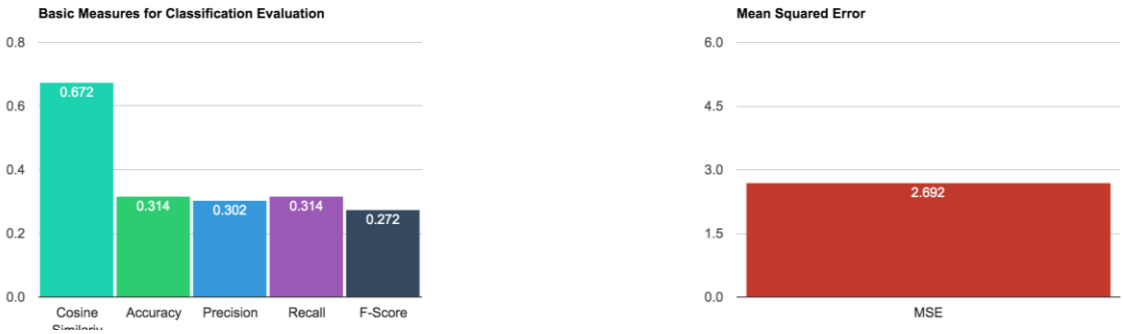
CLASSIFIER	ΔΕΔΟΜΕΝΑ ΔΟΚΙΜΩΝ	ΤΕΛΙΚΑ ΔΕΔΟΜΕΝΑ
ΝΑΪVE BAYES	3.862	5.979
DECISION TREE	3.803	5.627
LINEAR SVM	2.519	3.929
SVM RBF KERNEL	4.776	6.642
SVR	2.69	2.69
SGD	2.939	4.783
PERCEPTRON	3.484	5.645

Πίνακας 30.Αποτελέσματα δοκιμών αξιολόγησης των classifiers με τα τελικά χαρακτηριστικά (mse)

[42] Ο SVR (http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html) χρησιμοποιεί το συνεχόμενο εύρος βαθμολογίας των tweets και οι προβλέψεις του στρογγυλοποιούνται για να αξιολογηθούν. (max_iterations=550)



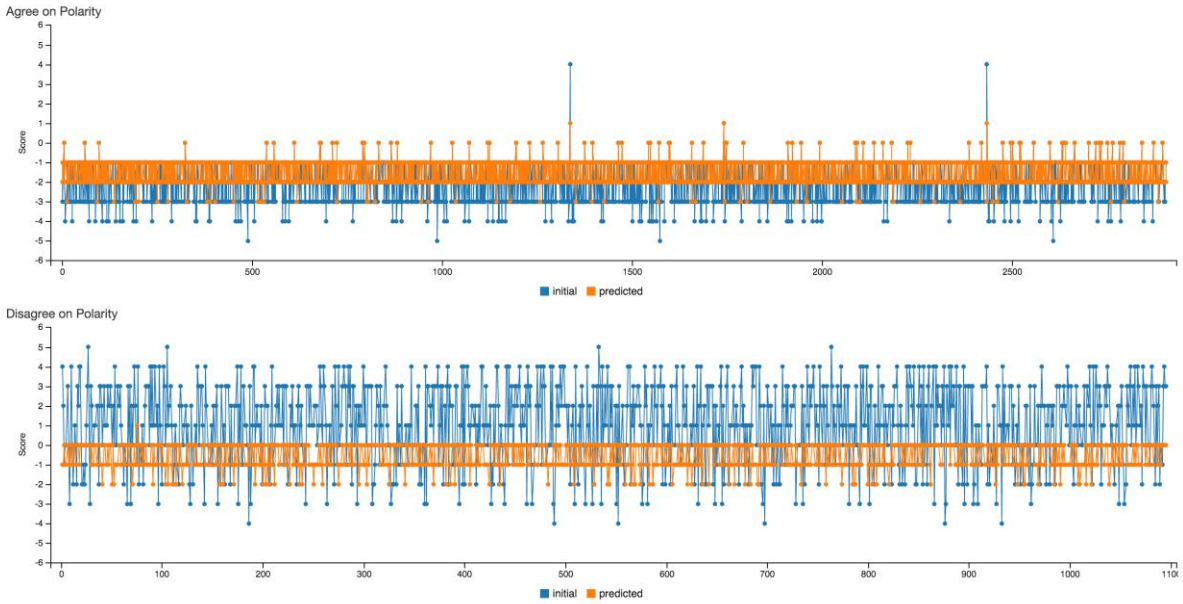
Σχήμα 83. SVR: Οι παράμετροι της δοκιμής



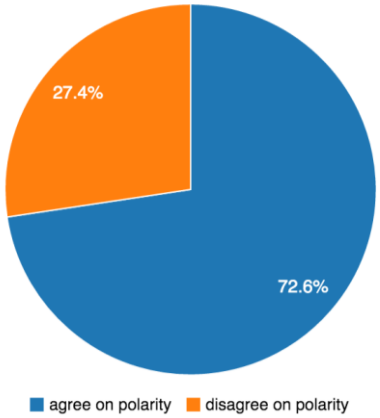
Σχήμα 84.Τα αναλυτικά αποτελέσματα – SVR

Class	Precision	Recall	F-Score	Support
-5	0	0	0	4
-4	0	0	0	100
-3	0.27	0.02	0.04	737
-2	0.52	0.57	0.54	1541
-1	0.18	0.43	0.25	680
0	0.12	0.26	0.16	298
1	0.25	0.01	0.01	169
2	0	0	0	155
3	0	0	0	201
4	0	0	0	111
5	0	0	0	4

Πίνακας 31.Τα αποτελέσματα ανά κατηγορία - SVR

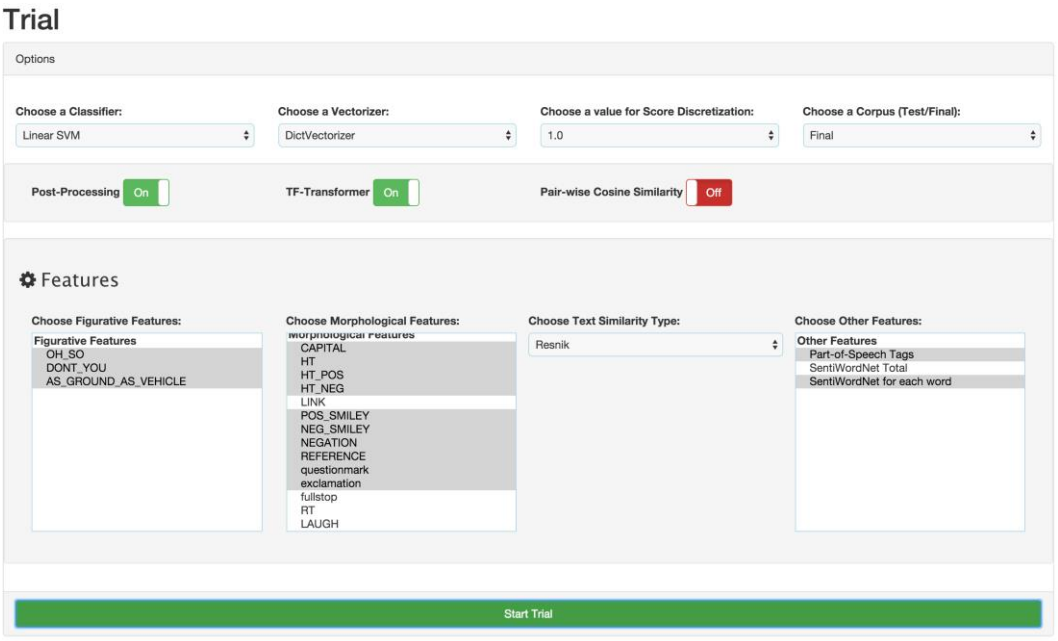


Σχήμα 85. Ανάλυση των αποτελεσμάτων σχετικά με τη πολικότητα – SVR

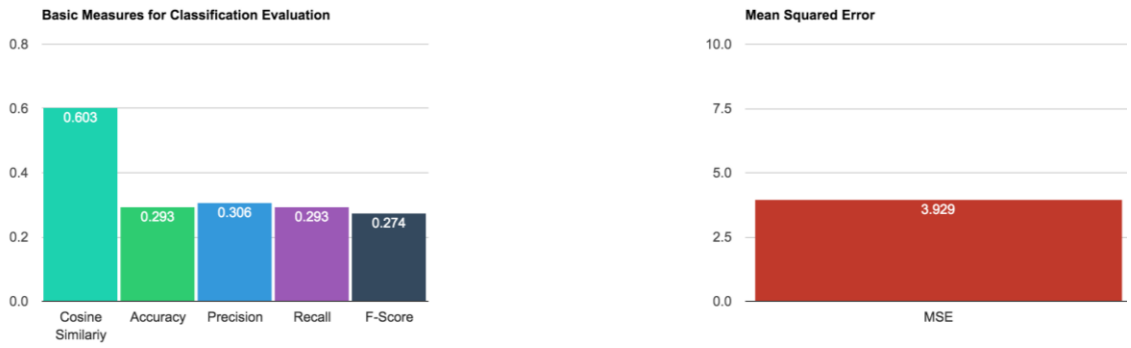


Σχήμα 86.Το ποσοστό των σωστών (agree) και λάθος (disagree) προβλέψεων σε σχέση με την πολικότητα - SVR

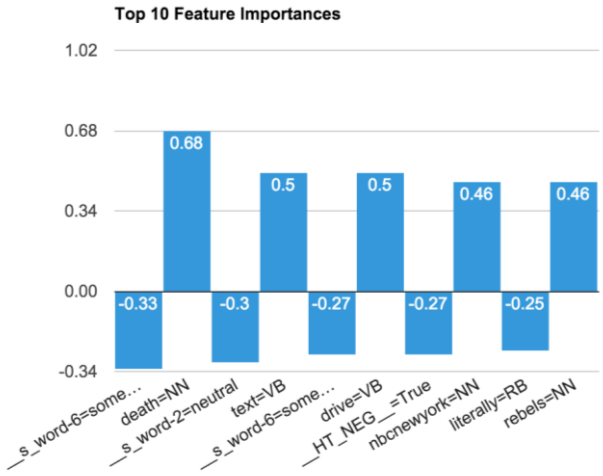
ΑΝΑΛΥΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΟΥ LINEAR SVM



Σχήμα 87. Linear SVM: Οι παράμετροι της δοκιμής



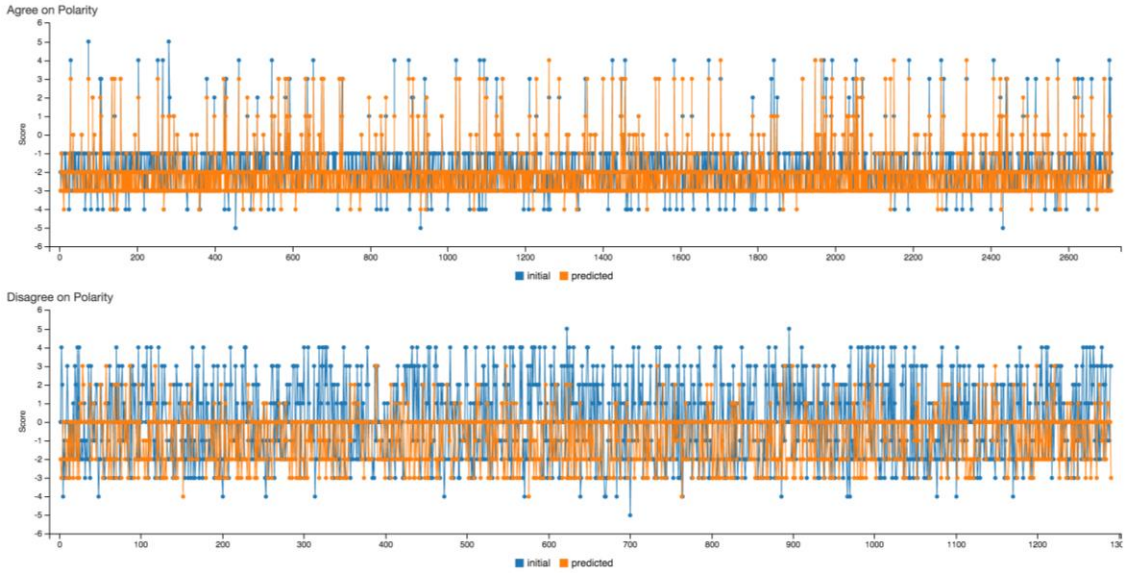
Σχήμα 88. Linear SVM: Τα αναλυτικά αποτελέσματα



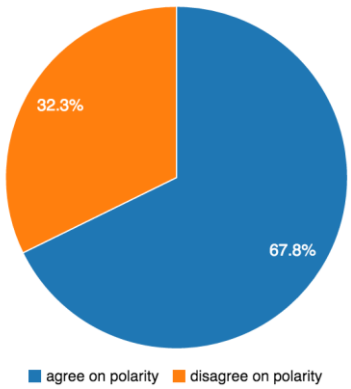
Σχήμα 89. Linear SVM: Τα πιο 10 σημαντικά χαρακτηριστικά της δοκιμής

Class	Precision	Recall	F-Score	Support
-5	0	0	0	4
-4	0.36	0.12	0.18	100
-3	0.26	0.48	0.34	737
-2	0.46	0.42	0.44	1541
-1	0.19	0.05	0.07	680
0	0.11	0.27	0.15	298
1	0.11	0.07	0.09	169
2	0.04	0.02	0.03	155
3	0.34	0.12	0.18	201
4	0.33	0.02	0.03	111
5	0	0	0	4

Πίνακας 32. Linear SVM: Τα αποτελέσματα ανά κατηγορία



Σχήμα 90. Linear SVM: Ανάλυση των αποτελεσμάτων σχετικά με τη πολικότητα



Σχήμα 91. Linear SVM: Το ποσοστό των σωστών (agree) και λάθος (disagree) προβλέψεων σε σχέση με την πολικότητα

9. ΑΠΟΤΕΛΕΣΜΑΤΑ

9.1 ΤΕΛΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ SEMEVAL 2015 TASK 11

Τα τελικά αποτελέσματα του SemEval 2015 Task 11[43] παρουσιάζονται ακολούθως αναλυτικά. Η DsUniPi ομάδα [64] βρίσκεται στην 10η θέση από τους 15, έχοντας σχετικά καλό ποσοστό Cosine Similarity και MSE. Επίσης, στην 10η θέση βρίσκεται και όσον αφορά το MSE, όπως φαίνεται στον Πίνακας 33. Σε γενικές γραμμές, τα αποτελέσματα στο ειρωνικό και στο σαρκαστικό κείμενο είναι αρκετά καλά (0.839 και 0.87 αντιστοίχως), σε κείμενο που περιέχει μεταφορά και σε μη μεταφορικό κείμενο όμως δεν υπήρχε καλή απόδοση (0.359 και 0.251 αντιστοίχως). Τα αναλυτικά αποτελέσματα για τις τέσσερις κύριες κατηγορίες tweets που υπήρχαν στα τελικα δεδομένα φαίνονται στον Πίνακας 34.

Team	Cosine	MSE
ClaC	0.758	2.117
UPF	0.710	2.458
LLT_PolyU	0.678	2.600
LT3	0.6581	3.398
elirf	0.6579	3.096
ValenTo	0.634	2.999
HLT	0.630	4.088
CPH	0.625	3.079
prhlt	0.623	3.023
DsUniPi	0.601	3.925
PKU	0.574	3.746
KELabTeam	0.552	6.090
RGU	0.523	8.602
SHELLFBK	0.431	7.701
BUAP	0.059	6.785

Πίνακας 33. Τα επίσημα συγκεντρωτικά αποτελέσματα του SemEval 2015 Task 11

[43] http://alt.qcri.org/semeval2015/task11/index.php?id=task-results-and-initial-analysis-1

	Cosine Similarity	MSE
Overall	0.601	3.925
Sarcasm	0.87	1.499
Irony	0.839	1.656
Metaphor	0.359	7.106
Other	0.271	5.744
Rank	10	10

Πίνακας 34. Τα τελικά αποτελέσματα ανά κατηγορία

Στον ακόλουθο πίνακα, φαίνεται η απόδοση των τριών κύριων classifiers στις δοκιμές και στα τελικά δεδομένα, όσον αφορά το cosine similarity και το accuracy. Ο SVM (Linear) έχει τα καλύτερα αποτελέσματα, και είναι και αυτός που χρησιμοποιήθηκε για να δώσει τα τελικά αποτελέσματα στο Task 11. Στον Πίνακας 36 απεικονίζονται τα αποτελέσματα όλων των ομάδων που συμμετείχαν στο Task 11.

Classifier	D. Tree		Naïve Bayes		SVM	
Trials/ Final	Train	Final	Train	Final	Train	Final
Cosine Similarity	0.68	0.45	0.70	0.55	0.78	0.60
Accuracy	0.31	0.21	0.33	0.23	0.38	0.29

Πίνακας 35. Οι διαφορές των αποτελεσμάτων μεταξύ του train και του final dataset

	Mean Squared Error measure						Cosine Similarity measure					
Team	Rank	Overall	Sarcasm	Irony	Metaphor	Other	Rank	Overall	Sarcasm	Irony	Metaphor	Other
ClaC	1	2.117	1.023	0.779	3.155	3.411	1	0.758	0.892	0.904	0.655	0.584
UPF-Dec-19	2	2.458	0.934	1.041	4.186	3.772	2	0.711	0.903	0.873	0.52	0.486
UPF-Dec-19		2.458	0.934	1.041	4.186	3.772		0.711	0.903	0.873	0.52	0.486
LLT_PolyU-Dec-20_7_31_46		2.602	0.997	0.671	3.917	4.617	3	0.687	0.896	0.918	0.535	0.29
LLT_PolyU-Dec-20_7_10_29		2.673	1.021	0.702	4.102	4.685		0.677	0.892	0.914	0.506	0.293
LLT_PolyU-Dec-20_14_42_31	3	2.6	1.018	0.673	3.917	4.587		0.687	0.893	0.917	0.535	0.301
LT3-dec-19-10-21-28-run1		3.398	1.287	1.224	5.67	5.444	4	0.6581	0.891	0.897	0.443	0.346
LT3-dec-19-10-21-28-run2	4	2.912	1.286	1.083	4.793	4.503		0.648	0.872	0.861	0.355	0.357
LT3-dec-19-12-11-44-run1		3.398	1.287	1.224	5.67	5.444		0.6581	0.891	0.897	0.443	0.346
LT3-dec-19-12-11-44-run2		2.912	1.286	1.083	4.793	4.503		0.648	0.872	0.861	0.355	0.357
elirf	8	3.096	1.349	1.034	4.565	5.235	5	0.6579	0.904	0.905	0.411	0.247
ValenTo	5	2.999	1.004	0.777	4.73	5.315	6	0.634	0.895	0.901	0.393	0.202
HLT	11	4.088	1.327	1.184	6.589	7.119	7	0.63	0.887	0.907	0.379	0.365
CPH-ridge		3.079	1.041	0.904	4.916	5.343	8	0.625	0.897	0.886	0.325	0.218
CPH-esemble	7	3.078	0.971	0.774	5.014	5.429		0.623	0.9	0.903	0.308	0.226
CPH-specialesemble		11.274	19.267	9.124	7.806	7.027		0.298	-0.148	0.281	0.535	0.612
PrhIt-ETR-ngram	6	3.023	1.028	0.784	5.446	4.888	9	0.623	0.891	0.901	0.167	0.218
PrhIt-ETR-word		3.112	1.041	0.791	5.031	5.448		0.611	0.89	0.901	0.294	0.129
PrhIt-RFR-word		3.107	1.06	0.809	5.115	5.345		0.613	0.888	0.898	0.282	0.17
PrhIt-RFR-ngram		3.229	1.059	0.811	5.878	5.243		0.597	0.888	0.898	0.135	0.192
PrhIt-BRR-word		3.299	1.146	0.934	5.178	5.773		0.592	0.883	0.88	0.28	0.11
PrhIt-BRR-ngram		3.266	1.1	0.941	5.925	5.205		0.593	0.886	0.879	0.119	0.186
DsUniPi	10	3.925	1.499	1.656	7.106	5.744	10	0.601	0.87	0.839	0.359	0.271
PKU	9	3.746	1.148	1.015	5.876	6.743	11	0.574	0.883	0.877	0.35	0.137
KELabTeam		5.552	1.198	1.255	7.264	9.905		0.531	0.883	0.895	0.341	0.117
KELabTeam-content based		6.09	1.756	1.811	8.707	11.526	12	0.552	0.896	0.915	0.341	0.115
KELabTeam-emotiona pattern based	12	4.177	1.189	0.809	6.829	7.628		0.533	0.874	0.9	0.289	0.135
RGU-testsentfinal	13	5.143	1.954	1.867	8.015	8.602	13	0.523	0.829	0.832	0.291	0.165
RGU-testsentwarppred		5.323	1.855	1.541	8.033	9.505		0.509	0.842	0.861	0.28	0.09
RGU-testsentpredictions		5.323	1.855	1.541	8.033	9.505		0.509	0.842	0.861	0.28	0.09
SHELLFBK-run3	15	7.701	4.375	4.516	9.219	12.16	14	0.431	0.669	0.625	0.35	0.167
SHELLFBK-run2		9.265	5.183	5.047	11.058	15.055		0.427	0.681	0.652	0.346	0.146
SHELLFBK-run1		10.486	12.326	9.853	10.649	8.957		0.145	-0.013	0.104	0.167	0.308
SHELLFBK-run1_Dec_9		10.486	12.326	9.853	10.649	8.957		0.145	-0.013	0.104	0.167	0.308
BUAP	14	6.785	4.339	7.609	8.93	7.253	15	0.058	0.412	-0.209	-0.023	-0.025

Πίνακας 36. Η ανάλυση των τελικών αποτελεσμάτων για όλες τις ομάδες και τις δοκιμές

10. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Η παρούσα εργασία ασχολήθηκε με το πολύπλευρο πρόβλημα της ανάλυσης συναισθήματος σε κοινωνικά δίκτυα και συγκεκριμένα, με τον μεταφορικό λόγο στο Twitter. Αρχικά ορίστηκε το πρόβλημα και έγινε η βιβλιογραφική επισκόπηση των τομέων που εμπλέκονται στην προσπάθεια επιλυσής του. Ακολούθησε η περιγραφή της προσέγγισης που ακολουθήθηκε και του συστήματος που αναπτύχθηκε για την αντιμετώπιση του προβλήματος της κατηγοριοποίησης συναισθήματος.

Η μέθοδος που παρουσιάστηκε και χρησιμοποιήθηκε στα πλαίσια του συστήματος συνδυάζει δομημένες πηγές πληροφορίας, όπως το SentiWordNet, κοινά χαρακτηριστικά των tweets, όπως τα hashtags και χαρακτηριστικά που υποδεικνύουν την παρουσία και σχετίζονται με τον μεταφορικό λόγο. Η προσέγγιση ακολουθεί τις αρχές της επιβλεπόμενης μάθησης, έχοντας ως στόχο την κατηγοριοποίηση των tweets που περιέχουν ειρωνεία, μεταφορά και σαρκασμό.

Τα αποτελέσματα του συστήματος στα πλαίσια του SemEval 2015 Task 11, κατέταξαν το σύστημα που αναπτύχθηκε στη $10^{η}$ θέση σε σύνολο 15 συστημάτων, χρησιμοποιώντας ως κριτήριο τις μετρικές cosine similarity και mse.

Αρκετά καλά αποτελέσματα, με υψηλό cosine similarity (> 0.8) προέκυψαν στις κατηγορίες της ειρωνείας και του σαρκασμού. Στα tweets που περιέχουν μεταφορές, τα αποτελέσματα ήταν αρκετά χαμηλά, και ακόμα χαμηλότερα στα tweets που δεν περιέχουν μεταφορικό λόγο.

Ως πιο χρήσιμα χαρακτηριστικά σε όλες τις δοκιμές που έγιναν, αναδείχθηκαν τα Part-of-Speech tags και η βαθμολογία του SentiWordNet, με τον τρόπο που χρησιμοποιήθηκαν, ακολουθούμενα από τα hashtags και το semantic similarity.

Για την αξιολόγηση του συστήματος πραγματοποιήθηκαν δοκιμές με διαφορετικές παραμέτρους, όπως για παράδειγμα, με διαφορετικούς classifiers, με discretization κλπ. Ως classifier με την καλύτερη απόδοση στα πλαίσια του προβλήματος φάνηκε ότι είναι ο Linear SVM.

Από τα τελικά αποτελέσματα, φάνηκε διαισθητικά πως η απόδοση του συστήματος θα μπορούσε να βελτιωθεί με τη χρήση χαρακτηριστικών που να καλύπτουν τις περιπτώσεις των tweets που περιέχουν μεταφορές και των tweets που δεν περιέχουν καθόλου μεταφορικό λόγο. Επίσης, ο καλύτερος χειρισμός των hashtags και των negations φαίνεται να είναι ένα ακόμα σημείο προς βελτίωση. Τέλος, η καλύτερη χρήση του SentiWordNet όσον αφορά το prior polarity των λέξεων σε σχέση με τη θέση τους, η αναγνώριση δηλαδή μοτίβων που αφορούν την αλληλουχία των συναισθημάτων, φαίνεται ότι μπορεί να οδηγήσει σε καλύτερα αποτελέσματα.

Κατά την κατασκευή του συστήματος προέκυψε η ανάγκη για μια βιβλιοθήκη που να χειρίζεται την προεπεξεργασία των tweets και μια εφαρμογή για την οπτικοποίηση των αποτελεσμάτων. Καθώς η προεπεξεργασία είναι απαραίτητο κομμάτι της ανάλυσης συναισθήματος, δημιουργήθηκε μια βιβλιοθήκη σε Python, η οποία περιλαμβάνει όλα τα βήματα που χρησιμοποιήθηκαν στο σύστημα που περιγράφηκε, με αρκετές βελτιώσεις και προσθήκες. Η βιβλιοθήκη αυτή δίνει τη δυνατότητα προσθήκης επιπλέον βημάτων είτε για την εξαγωγή χαρακτηριστικών είτε για τη διαδικασία καθαρισμού. Ως μελλοντικές βελτιώσεις σε αυτό το κομμάτι μπορούν να αναφερθούν οι ακόλουθες: η διαδικασία καθαρισμού και εξαγωγής

χαρακτηριστικών θα μπορούσε να είναι πιο παραμετρική, να δίνεται η δυνατότητα στον χρήστη να ορίζει τη σειρά των βημάτων. Αυτή τη στιγμή αυτό δεν γίνεται, καθώς πρέπει να οριστούν οι ομάδες χαρακτηριστικών και βημάτων οι οποίες είναι άρρηκτα συνδεδεμένες μεταξύ τους και εάν δεν έχουν ακολουθηθεί όλα τα βήματα, δεν είναι δυνατόν να γίνει σωστά η διαδικασία. Παραδείγματος χάριν, εάν δεν έχει γίνει tokenization, τότε δεν είναι δυνατόν να γίνει αφαίρεση των stop words. Ένα ακόμα βήμα για την βελτίωση της βιβλιοθήκης θα ήταν η απεξάρτηση της από την MySQL βάση, στην οποία είναι αποθηκευμένο το SentiWordNet και κάποιες stop words.

Η οπτικοποίηση των αποτελεσμάτων με διαγράμματα και ο χειρισμός των πειραμάτων από γραφική διεπαφή είναι σημαντική βοήθεια στην εξαγωγή συμπερασμάτων. Η δικτυακή εφαρμογή που χρησιμοποιήθηκε, βοήθησε στη διαδικασία αυτή. Στα πλαίσια της χρήσης της δικτυακής εφαρμογής φάνηκε η ανάγκη να αυξηθούν οι επιλογές στα κριτήρια των πειραμάτων και να υπάρχει η δυνατότητα αναζήτησης παλαιών αποτελεσμάτων. Οι τεχνικές αλλαγές που θα μπορούσαν να βελτιώσουν το σύστημα, συμπεριλαμβάνουν τη χρήση websockets για το feedback, κάτι που στην παρούσα φάση υλοποιείται με ajax calls) και η χρήση authentication, όπως αυτό παρέχεται από το Django, ώστε να υπάρχει δυνατότητα διαχωρισμού των αποτελεσμάτων διαφορετικών χρηστών και γενικότερα της ανεξαρτητοποίησης της διαδικασίας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] B. Pang και L. Lee, «Opinion mining and sentiment analysis,» *Foundations and Trends in Information Retrieval,* τόμ. 2, αρ. 1-2, 01 2008.

[2] Twitter, Inc., [Ηλεκτρονικό]. Available: https://about.twitter.com/company . [Πρόσβαση 12 2015].

[3] B. Palace, «Data Mining: What is Data Mining?,» 1996. [Ηλεκτρονικό]. Available: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/index.htm.

[4] F. Provost και T. Fawcett, Data Science for Business, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2013.

[5] «Big Data,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Big_data .

[6] D. Murthy, Twitter: Social Communication in the Twitter Age, Cambridge: Polity Press, 2013.

[7] «SemEval Portal,» [Ηλεκτρονικό]. Available: http://aclweb.org/aclwiki/index.php?title=SemEval_Portal . [Πρόσβαση 12 2015].

[8] «Wikipedia - SemEval,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/SemEval . [Πρόσβαση 12 2015].

[9] C. Clifton, «Data mining,» [Ηλεκτρονικό]. Available: http://www.britannica.com/technology/data-mining. [Πρόσβαση 12 2015].

[10] «DATA MINING CURRICULUM: A PROPOSAL,» [Ηλεκτρονικό]. Available: http://www.kdd.org/curriculum/index.html. [Πρόσβαση 12 2015].

[11] «Data Mining,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Data_mining. [Πρόσβαση 12 2015].

[12] J. Dean, Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, Wiley, 2014.

[13] J. Alpert και N. Hajaj, «googleblog,» Google Inc., 25 7 2008. [Ηλεκτρονικό]. Available: https://googleblog.blogspot.gr/2008/07/we-knew-web-was-big.html. [Πρόσβαση 12 2015].

[14] S. Bird, E. Klein και E. Loper, Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit, O'Reilly Media, 2008.

[15] N. Hardeniya, NLTK Essentials, Birmingham: Packt Publishing Ltd., 2015.

[16] B. Santorini, «Part-of-speech tagging guidelines for the Penn Treebank Project,» 1990.

[17] «Alphabetical list of part-of-speech tags used in the Penn Treebank Project,» [Ηλεκτρονικό]. Available: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.

[18] M. Lingling, H. Runqing και G. Junzhong, «A Review of Semantic Similarity Measures in WordNet,» *International Journal of Hybrid Information Technology,* τόμ. 6, αρ. 1, p. 1, 1 2013.

[19] T. Pedersen, S. Patwardhan και J. Michelizzi, «WordNet::Similarity - Measuring the Relatedness of Concepts,» σε *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, San Jose, 2004.

[20] Princeton University, «"About WordNet.",» Princeton University, 2010. [Ηλεκτρονικό]. Available: http://wordnet.princeton.edu. [Πρόσβαση 12 2015].

[21] C. Fellbaum, WordNet: An Electronic Lexical Database, Cambridge: MA: MIT Press., 1998.

[22] «Bag-of-words model,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Bag-of-words_model . [Πρόσβαση 12 2015].

[23] P. Harrington, Machine Learning in Action, J. Bleiel, Επιμ., Manning Publications, 2012.

[24] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[25] T. Mitchell, Machine Learning, McGraw Hill, 1997.

[26] P. Berka και J. Rauch, «Machine Learning and Association Rules,» σε *In 19th Int. Conf. On Computational Statistics COMPSTAT*, 2010.

[27] M. Radovanović και M. Ivanović, «TEXT MINING: APPROACHES AND APPLICATIONS,» *Novi Sad Journal of Mathematics,* τόμ. 38, αρ. 3, pp. 227-234, 2008.

[28] scikit-learn, [Ηλεκτρονικό]. Available: http://scikit-learn.org/stable/tutorial/machine_learning_map/ . [Πρόσβαση 12 2015].

[29] J. R. Quinlan, «Induction of Decision Trees.,» *Machine Learning,* τόμ. 1, αρ. 1, pp. 81-106 , 1986.

[30] J. R. Quinlan, «Combining Instance-Based and Model-Based Learning.,» *ICML,* pp. 236-243, 1993.

[31] R. Feldman, «Techniques and Applications for Sentiment Analysis,» *Communications of the ACM,* αρ. 56(4), p. 82–89, 04 2013.

[32] M. Taboada, J. Brooke, M. Tofiloski, K. Voll και M. Stede, «Lexicon- Based Methods for Sentiment Analysis,» *Computational Linguistics,* τόμ. 37, αρ. 2, pp. 267-307, 2011.

[33] B. Pang και L. Lee, «Opinion mining and sentiment analysis,» *Foundations and Trends in Information Retrieval,* τόμ. 2, αρ. 1-2, pp. 1-135, 2008.

[34] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.

[35] N. Indurkhya και F. J. Damerau, «Sentiment Analysis and Subjectivity,» σε *Handbook of Natural Language Processing Second Edition*, CRC Press.

[36] N. Farra, E. Challita, R. Assi και H. Hajj, «Sentence-level and Document-level Sentiment Mining for Arabic Texts,» σε *Data Mining Workshops (ICDMW), IEEE International Conference*, 2010.

[37] B. Liu και M. Hu, «Opinion Mining, Sentiment Analysis, and Opinion Spam Detection,» 15 5 2004. [Ηλεκτρονικό]. Available: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. [Πρόσβαση 12 2015].

[38] M. Ganapathibhotla και B. Liu, σε *COLING '08 - Proceedings of the 22nd International Conference on Computational Linguistics*, 2008.

[39] P. Turney, «Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,» σε *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[40] D. C. D. M. S. S. D. M. O. Philip J. Stone, The General Inquirer: A Computer Approach to Content Analysis, MIT Press , 1966.

[41] T. Wilson, J. Wiebe και P. Hoffmann, «Recognizing Contextual Polarity in Phrase- Level Sentiment Analysis,» σε *HLT-EMNLP*, 2005.

[42] E. Riloff και J. Wiebe, «Learning Extraction Patterns for Subjective Expressions,» σε *In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, 2003.

[43] M. Hu και B. Liu, «Mining and Summarizing Customer Reviews,» σε *ACM SIGKDD*, 2004.

[44] S. Baccianella, A. Esuli και F. Sebastiani, «SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.,» σε *LREC*, 2010.

[45] C. Potts, «Sentiment Symposium Tutorial,» 2011. [Ηλεκτρονικό]. Available: http://sentiment.christopherpotts.net/lexicons.html. [Πρόσβαση 12 2015].

[46] D. Derks, A. E. R. Bos και J. Von Grumbkow, «Emoticons and online message interpretation,» *Social Science Computer Review,* τόμ. 26, αρ. 3, pp. 379-388, 2007.

[47] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai και A. Kappas, «Sentiment Strength Detection in Short Informal Text,» *Journal of the American Society for In- formation Science and Technology,* τόμ. 61, αρ. 12, p. 2544–2558, 12 2010.

[48] A. Go, R. Bhayani και L. Huang, «Twitter sentiment classification using distant supervision,» σε *In: Proceeding LSM '11 Proceedings of the Workshop on Languages in Social Media*, 2009.

[49] E. Kouloumpis, T. Wilson και J. Moore, «Twitter sentiment analysis: The Good the Bad and the OMG!,» σε *In: Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, Proceedings of the Fifth Inter- national Conference on Weblogs and Social Media, ICWSM' 11*, Barcelona, 2010.

[50] A. Agarwal, B. Xie, I. Vovsha, O. Rambow και R. Passonneau, «Sentiment Analysis of Twitter Data,» σε *In: LSM'11 Pro- ceedings of the Workshop on Languages in Social Media*, 2011.

[51] P. R. T. V. Antonio Reyes, «A Multidimensional Approach for Detecting Irony in Twitter.,» *Languages Resources and Evaluation,* αρ. 47(1), pp. 239-268., 2013.

[52] A. Reyes, P. Rosso και D. Buscaldi, «From Humor Recognition to Irony Detection: The Figurative Language of Social Media.,» *Data & Knowledge Engineering,* αρ. 73, pp. 1-12, 2012.

[53] Y. Hao και T. Veale, «An Ironic Fist in a Velvet Glove: Creative Mis-Representation in the Construction of Ironic Similes,» *Minds and Machines,* τόμ. 20, αρ. 4, p. 635–650, 2010.

[54] D. Davidov, O. Tsur και A. Rappoport, «Semi-supervised recognition of sarcastic sentences in twitter and amazon.,» σε *In Proceedings of the Fourteenth Conference on Computational Natural Language Learning, CoNLL 2010*, 2010.

[55] E. Riloff, A. Qadir, P. Surve, L. D. Silva, N. Gilbert και R. Huang, «Sarcasm as Contrast between a Positive Sentiment and Negative Situation.,» σε *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.

[56] E. Shutova, L. Sun και A. Korhonen, «Metaphor Identification Using Verb and Noun Clustering.,» σε *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics*, 2010.

[57] CrowdFlower, [Ηλεκτρονικό]. Available: http://www.crowdflower.com.

[58] SemEval 2015, [Ηλεκτρονικό]. Available: http://alt.qcri.org/semeval2015/task11/. [Πρόσβαση 12 2015].

[59] «Emoticon Analysis in Twitter,» [Ηλεκτρονικό]. Available: http://datagenetics.com/blog/october52012/. [Πρόσβαση 8 2014].

[60] T. Pedersen, S. Patwardhan και J. Michelizzi, «Wordnet::similarity - measuring the relatedness of concepts.,» σε *In: Demonstration papers at HLT-NAACL*, 2004.

[61] L. Derczynski, A. Ritter, S. Clark και K. Bontcheva, «Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data.,» σε *In: Proceedings of the International Con- ference on Recent Advances in Natural Lan- guage Processing, ACL*, 2013.

[62] R. Kelly, «PyEnchant a spellchecking library for Python,» [Ηλεκτρονικό]. Available: https://pythonhosted.org/pyenchant/.

[63] Wikipedia, «Git λογισμικό,» [Ηλεκτρονικό]. Available: https://el.wikipedia.org/wiki/Git_(λογισμικό). [Πρόσβαση 12 2015].

[64] L. Derczynski, A. Ritter, S. Clarke και K. Bontcheva, «Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data,» σε *In Proceedings of the International Conference on Recent Advances in Natural Language Processing, ACL*, 2013.

[65] F. Rosenblatt, «The perceptron: A probabilistic model for information storage and organization in the brain,» *Psychological Review,* τόμ. 65, αρ. 6, pp. 386-408, 11 1958.