Εφαρμογή αυτόματης κατάταξης γνώμης σε δεδομένα του Twitter

Μ. Μπιρμπίλη[1], Γ. Πασχάλης[2], Σ. Κωτσιαντής[3]

[1] Τελειόφοιτη Πληροφορικής ΕΑΠ
matinabirbily@yahoo.gr
[2] Υποψήφιος Διδάκτωρ ΤΗΜΤΥ Παν/μιο Πατρών
gpasxali@upatras.gr
[3] Λέκτορας Μαθηματικό Παν/μιο Πατρών
sotos@math.upatras.gr

Περίληψη

Στην παρούσα εργασία παρουσιάζεται μια εφαρμογή εξόρυξης γνώμης που κατασκευάσαμε για να κατατάσσουμε σαν θετικές ή αρνητικές της γνώμες χρηστών του twitter. Για το σκοπό αυτό, αφού μελετήσαμε και αξιολογήσαμε αλγόριθμους μηχανικής Μάθησης που συνιστώνται για την εξόρυξη γνώμης, βρήκαμε αυτόν που δίνει τα καλύτερα αποτελέσματα και τον παραμετροποιήσαμε κατάλληλα ώστε να βελτιωθεί η απόδοσή του. Πιο συγκεκριμένα, συνδυάσαμε τον NaïveBayesMultinomial με τον αλγόριθμο για stemming της Lovins και τον tokenizer Ngram σε μια προσπάθεια να αντιμετωπίσουμε την υπόθεση του NaïveBayesMultinomial ότι κάθε μεταβλητή είναι ανεξάρτητη με οποιαδήποτε άλλη της δεδομένης κατηγορίας.

Λέξεις κλειδιά: Εξόρυξη γνώμης, Twitter, Μηχανική Μάθηση, Naïve Bayes Multinomial, Εργαλείο για εξόρυξη γνώμης (ToolForOpinionMining).

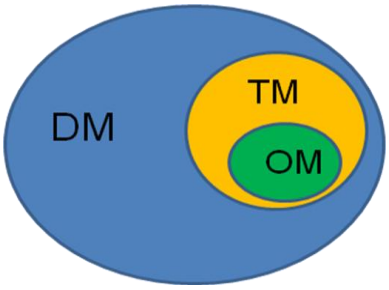
1. Εισαγωγή

Ο ερχομός και η ακμάζουσα δημοτικότητα των σε απευθείας σύνδεση κοινωνικών δικτύων (online social networks) όπως του Twitter και του Facebook έχει οδηγήσει σε μια τεράστια έκρηξη των δεδομένων με επίκεντρο το διαδίκτυο. Τα δεδομένα που υπάρχουν είναι αντικείμενο μελέτης πολλών πεδίων, όπως η κοινωνιολογία, οι επιχειρήσεις, η ψυχολογία, η διασκέδαση, η πολιτική κ.α.

Κοινωνική δικτύωση είναι η δημιουργία μιας διαδικτυακής κοινότητας με κοινά ενδιαφέροντα, και η συγκέντρωση ή συμμετοχή των ατόμων σε συγκεκριμένες ομάδες, όπως για παράδειγμα οι φίλοι, οι συνεργάτες, οι αγροτικές κοινότητες, οι γειτονιές, οι χώροι εργασίας, τα πανεπιστήμια και τα σχολεία. Ο Χτούρης (2004) ορίζει ως κοινωνικά δίκτυα τα «πολυδιάστατα συστήματα επικοινωνίας και διαμόρφωσης της ανθρώπινης πρακτικής και της κοινωνικής ταυτότητας».

Το Twitter (https://twitter.com/) αν και σχετικά καινούριο (δημιουργήθηκε το 2006) έχει πάρα πολλούς χρήστες (554 εκατομμύρια συνολικά εγγεγραμμένους χρήστες από τους οποίους 135.000 συνδέονται σε καθημερινή βάση και ανταλλάσσουν 58 εκατομμύρια μηνύματα σύμφωνα με το Twitter eMarketer, http://www.statisticbrain.com/twitter-statistics, Ιούνιος 2013)

Μια βασική ανάγκη που δημιουργείται είναι να εκμεταλλευτούμε το ασύλληπτα μεγάλο ποσό δεδομένων που είναι διαθέσιμο και να ανακαλύψουμε γνώσεις που κρύβονται πίσω από αυτά. Η Εξόρυξη Γνώσης ή Ανακάλυψη Γνώσης από Βάσεις Δεδομένων ή Εξόρυξη Δεδομένων (Data Mining - DM) (Klosgen & Zytkow, 2002) είναι η εξεύρεση μιας ενδιαφέρουσας, αυτονόητης, μη προφανούς και πιθανόν χρήσιμης πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις.

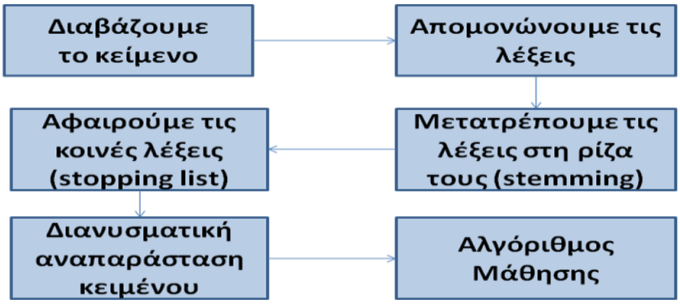


Μια υποκατηγορία της εξόρυξης γνώσης είναι η εξόρυξη κειμένου (Text Mining - TM), η οποία ορίζεται σαν μια διαδικασία εξαγωγής με αυτόματο τρόπο νέας, έγκυρης, και χρήσιμης γνώσης από διαφορετικούς γραπτούς πόρους, καθώς επίσης και η όσο το δυνατόν καλύτερη οργάνωση αυτής της νέας γνώσης-πληροφορίας για την όποια μελλοντική αναφορά (Hearst Marti, 1999; Sehgal, 2004). Μια υποκατηγορία της εξόρυξης κειμένου είναι η εξόρυξη γνώμης (opinion mining -OM) ή ανάλυση συναισθήματος (sentimental analysis) η οποία στοχεύει να εξάγει γνωρίσματα και συστατικά ενός αντικειμένου που έχει σχολιαστεί και να καθορίσει αν το σχόλιο είναι θετικό, αρνητικό ή ουδέτερο (Liu, 2012).

Εφαρμόζοντας τις τεχνικές εξόρυξης γνώσης στα Κοινωνικά Δίκτυα μπορούμε να ανακαλύψουμε ενδιαφέρουσες πλευρές της ανθρώπινης συμπεριφοράς και της ανθρώπινης αλληλεπίδρασης, να βελτιώσουμε την αντίληψη που έχουν οι άνθρωποι σχετικά με ένα θέμα, να προσδιορίσουμε ομάδες ανθρώπων ανάμεσα στις μάζες του πληθυσμού, να μελετήσουμε ομάδες που αλλάζουν με το χρόνο, να βρεθούν άνθρωποι με επιρροή, ή ακόμα και να γίνει η σύσταση ενός προϊόντος ή μιας δραστηριότητας σε ένα άτομο.

Επειδή τα ποσά δεδομένων είναι τεράστια, είναι απαραίτητο να εφαρμοστεί μια διαδικασία αυτόματης κατηγοριοποίησης των κειμένων (A.K.K.). Μια τέτοια διαδικασία απεικονίζεται σχηματικά στην παρακάτω Εικόνα 1. Αρχικά συλλέγεται ένα σύνολο κειμένων, το οποίο κάποιος ειδικός ή μια ομάδα ανθρώπων έχουν

σημασιοδοτήσει (labeled) σε σχέση με κάποιες κατηγορίες. Το σύνολο των κειμένων χωρίζεται σε ένα σύνολο εκπαίδευσης (training set), και ένα σύνολο ελέγχου (test set) το οποίο χρησιμοποιείται για επικύρωση (validation) της απόδοσης του αλγορίθμου κατηγοριοποίησης. Στο σύνολο εκπαίδευσης, το συνολικό κείμενο μετατρέπεται σε ξεχωριστές λέξεις. Όλες οι λέξεις που περιλαμβάνονται στα κείμενα μπαίνουν σε «σακούλα λέξεων» (bag-of-words). Παράλληλα, αγνοούνται όλοι οι μη αλφαριθμητικοί όροι, π.χ. αριθμοί, ημερομηνίες, σύμβολα όπως '<', '>', '=' κ.α. Στη συνέχεια οι λέξεις μετατρέπονται στη ρίζα τους (stemming). Έπειτα αφαιρούνται οι κοινές λέξεις (άρθρα, επιρρήματα, κτλ). Έχει παρατηρηθεί πως οι 10 συχνότερες λέξεις της αγγλικής γλώσσας αποτελούν το 20-30% των λεκτικών μονάδων σε ένα κείμενο (Francis, 1982). Ανάμεσά τους οι λέξεις {is, the, to, for, and, it…} που συναντάμε σχεδόν σε κάθε πρόταση. Η ποσότητα των όρων που αποφασίζουμε πως δε μεταφέρουν πληροφορία επηρεάζει και την ποιότητα των δεδομένων που θα αναπαραστήσουμε.



Εικόνα 1. *Διαδικασία εκμάθησης που ακολουθείται σε προβλήματα Α.Κ.Κ.*

Έπειτα γίνεται διανυσματική αναπαράσταση του κειμένου (vector representation). Έτσι, κάθε αρχείο κειμένου από το σύνολο κειμένων που έχουμε είναι και ένα διάνυσμα όρων (term vector) στο οποίο κάθε όρος αποτελεί ένα μοναδικό ανεξάρτητο χαρακτηριστικό (feature). Κάθε στοιχείο σε αυτό το διάνυσμα έχει και μια τιμή η οποία αντιστοιχεί στην εμφάνιση του όρου μέσα στο κείμενο. Μια συχνά χρησιμοποιούμενη μέθοδος που δίνει πολύ καλά αποτελέσματα είναι η συνάρτηση TF-IDF (Term Frequency Inverse Document Frequency) (Sehgal, 2004).

Τέλος χρησιμοποιώντας το σύνολο εκπαίδευσης, ένας επιβλεπόμενος αλγόριθμος εκμάθησης (supervised learning algorithm) προσπαθεί να διαμορφώσει το βέλτιστο ταξινομητή.

2. Σχετικές Εργασίες

Στη βιβλιογραφία μπορεί κανείς να βρει πολλές ενδιαφέρουσες εργασίες που εστιάζουν στην εξόρυξη άποψης σε γνώμες διαφόρων ατόμων ώστε να αποφασίσουν αν ένα κομμάτι κειμένου (πρόταση, παράγραφος ή ολόκληρο έγγραφο) εκφράζει ένα θετικό ή αρνητικό συναίσθημα. Μερικές εργασίες παρουσιάζουν εξόρυξη γνώσης

αναγνωρίζοντας το συναισθηματικό προσανατολισμό του συνολικού κειμένου (Sentiment Analysis), άλλες χρησιμοποιούν μεθόδους μηχανικής μάθησης, ενώ άλλες εστιάζουν στην αναγνώριση και εξαγωγή λέξεων που εκφράζουν τις απόψεις και το σημασιολογικό προσανατολισμό (semantic orientation) του κειμένου. Ενδεικτικά αναφέρουμε τις παρακάτω πολύ ενδιαφέρουσες προσεγγίσεις που αφορούν το twitter:

Οι Agarwal etal (2011) και Kouloumpis etal (2011) εφαρμόζουν ένα συνδυασμό μηχανικής των χαρακτηριστικών (feature engineering), π.χ. χρησιμοποιώντας το βασισμένο στα χαρακτηριστικά και το βασισμένο στα δέντρα kernel μοντέλο για ταξινόμηση συναισθημάτων ή χρησιμοποιώντας τα χαρακτηριστικά των n-gram και των λεξικών (ngram and lexicon features), ή εφαρμόζοντας συνδυασμό τεχνικών μηχανικής μάθησης ώστε να βελτιώσουν στην ακρίβεια ταξινόμησης των μηνυμάτων του twitter.

Οι Bifet & Eibe (2010) προτείνουν μια εφαρμογή της Kappa στατιστικής μεθόδου (sliding window Kappa statistic) για αξιολόγηση σε μεταβαλλόμενα με το χρόνο δεδομένα του twitter.

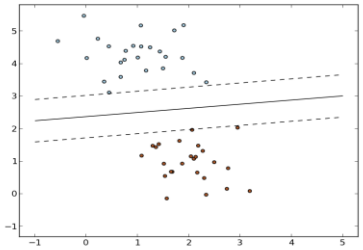
Οι Cha et al (2010) δείχνουν ότι δημοφιλείς χρήστες του twitter που έχουν υψηλή κοινωνική θέση (high in-degree) δεν επηρεάζονται απαραίτητα από τις προωθήσεις μηνυμάτων ή τις αναφορές άλλων χρηστών του twitter, και αυτή η επιρροή κερδίζεται μόνο δια μέσω συντονισμένης προσπάθειας π.χ. περιορισμένων μηνυμάτων στο twitter σε ένα συγκεκριμένο θέμα.

3. Αλγόριθμοι Μηχανικής Μάθησης

Η μηχανική μάθηση έχει σαν σκοπό τη δημιουργία μηχανών, ικανών να μαθαίνουν. Οι μηχανές αυτές, αξιοποιώντας προηγούμενη γνώση και εμπειρία μπορούν να μαθαίνουν και να βελτιώνουν την απόδοσή τους σε κάποιους τομείς. Υπάρχουν πάρα πολλοί αλγόριθμοι μηχανικής μάθησης που δίνουν ικανοποιητικά αποτελέσματα ανάλογα με το είδος της εφαρμογής που θα τους χρησιμοποιήσουμε. Στην περίπτωση της εξόρυξης γνώσης από κείμενο και πιο συγκεκριμένα στην εξόρυξη γνώμης και συναισθηματική ανάλυση κειμένου, την καλύτερη απόδοση έχουν οι αλγόριθμοι που ανήκουν στις κατηγορίες των Επαγωγικών Αλγορίθμων Μηχανικής Μάθησης, και ειδικότερα στις κατηγορίες των Μηχανών Διανυσματικής Υποστήριξης (SVMs) και των Bayesian μέθοδων (Aggarwal & Zhai, 2012; Liu, 2012).

3.1 Αλγόριθμος Στοχαστικής Κλίσης Κατάβασης (SGDText)

Ο SGDText αλγόριθμος, όπως και όλοι οι αλγόριθμοι αυτής της κατηγορίας (SVMs) έχει σαν στόχο την επιλογή ενός μικρού αριθμού στιγμιότυπων εκπαίδευσης από κάθε κλάση,



των διανυσμάτων υποστήριξης (Support Vectors), τα οποία συνορεύουν στο χώρο του προβλήματος με στιγμιότυπα άλλων κλάσεων (Bottou & Bousquet, 2008).

Στο διπλανό σχήμα, με μπλε απεικονίζονται τα θετικά στιγμιότυπα (pos) και με κόκκινο χρώμα τα αρνητικά (neg). Τα στιγμιότυπα τα οποία βρίσκονται πάνω στις έντονες γραμμές αποτελούν τα διανύσματα στήριξης. Ο αλγόριθμος SGD βρίσκει το βέλτιστο υπερ-επίπεδο που διαχωρίζει τα θετικά με αρνητικά στιγμιότυπα.

Τα βήματα του αλγορίθμου είναι τα εξής :

- Διάλεξε ένα αρχικό διάνυσμα παραμέτρων w και βαθμού μάθησης α.
- Επανέλαβε μέχρι ένα κατά προσέγγιση ελάχιστο να επιτευχθεί:
  - Ανάμειξε τυχαία τα παραδείγματα στο σύνολο εκπαίδευσης.
  - Για $i = 1, 2, ..., n$, κάνε: $w := w - \alpha \nabla Q_i(w)$. όπου α είναι ένα μέγεθος βήματος (πολλές φορές λέγεται βαθμός μάθησης - learning rate) στη μηχανική μάθηση και $Q_i(w)$ είναι η τιμή της συνάρτησης απώλειας (loss function) στο i-στο παράδειγμα και w ο αριθμός των επαναλήψεων.

3.2 Naïve Bayes Multinomial αλγόριθμος

Ο Naïve Bayes Multinomial αλγόριθμος λειτουργεί σε σύνολα κειμένων όπου κάθε στιγμιότυπο x αποτελείται από τις τιμές χαρακτηριστικών $\langle a_1, a_2, ..a_i \rangle$ και η συνάρτηση στόχος *f(x)* μπορεί να πάρει οποιαδήποτε τιμή από ένα προκαθορισμένο πεπερασμένο σύνολο $V = (v_1, v_2, ...v_j)$. Η ταξινόμηση αγνώστων στιγμιότυπων περιλαμβάνει τον υπολογισμό της πιο πιθανής τιμής στόχου $v_{\max}$ και ορίζεται ως εξής (McCallum & Nigam, 1998):

$$v_{\max} = \max_{v_j \in V} P(v_j \mid a_1, a_2, ...a_i) \quad \textbf{(1)}$$

Χρησιμοποιώντας το θεώρημα του Bayes το $v_{\max}$ μπορεί να ξαναγραφεί ως εξής:

$$v_{\max} = \max_{v_j \in V} P(a_1, a_2, ...a_i \mid v_j) P(v_j) \quad \textbf{(2)}$$

υποθέτοντας ότι οι τιμές των γνωρισμάτων είναι ανεξάρτητες δοσμένης της τιμής στόχου (target value).

Για τον Naïve Bayes Multinomial αλγόριθμο ισχύει :

$$v_{\max} = \max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_i) \quad \textbf{(3)}$$

όπου V είναι το αποτέλεσμα του ταξινομητή και οι $P(a_i \mid v_i)$ και $P(v_j)$ μπορούν να υπολογιστούν με βάση τη συχνότητά τους στα δεδομένα προς εκπαίδευση.

4. Πειραματικά Αποτελέσματα

4.1 Το λογισμικό WEKA

Για να βρούμε το βέλτιστο αλγόριθμο που θα χρησιμοποιήσουμε στην εφαρμογή μας χρησιμοποιήσαμε το λογισμικό ανάπτυξης εφαρμογών μηχανικής μάθησης και εξόρυξης γνώσης από δεδομένα WEKA (Waikato Environment for Knowledge Analysis), το οποίο είναι ένα από τα πιο δημοφιλής εργαλεία εξόρυξης πληροφορίας κυρίως στον ακαδημαϊκό χώρο (Witten etal, 2011). Το WEKA είναι ελεύθερο λογισμικό / λογισμικό ανοικτού κώδικα γραμμένο σε java και διανέμεται με άδεια χρήσης GNU από το url: http://www.cs.waikato.ac.nz/ml/weka/. Η εφαρμογή περιλαμβάνει υλοποιήσεις διάφορων αλγόριθμων εξόρυξης πληροφορίας καθώς και τεχνικές προεπεξεργασίας, μοντελοποίησης αλλά και τεχνικές οπτικοποίησης των δεδομένων. Η εφαρμογή προϋποθέτει την ύπαρξη των δεδομένων σε ένα απλό αρχείο της μορφής CSV ή arff όπου τα γνωρίσματα (πεδία) κάθε εγγραφής (γραμμής) είναι χωρισμένα με κόμμα.

4.2 Προετοιμάζοντας τα δεδομένα

Για να εισάγουμε τα δεδομένα του Twitter στο Weka πρέπει να δημιουργήσουμε ένα αρχείο .arff το οποίο θα χρησιμοποιήσουμε σαν σύνολο εκπαίδευσης. Για τις ανάγκες της εργασίας μας χρησιμοποιήσαμε ένα έτοιμο αρχείο, twdata.arff (https://dl.dropboxusercontent.com/u/851910/twData.arff) με 17.983 στιγμιότυπα μηνυμάτων του twitter. Τα δεδομένα που χρησιμοποιήσαμε υπάρχουν ελεύθερα στο διαδίκτυο. Θα μπορούσαμε να δημιουργήσουμε μια εφαρμογή που θα τραβά αυτόματα δεδομένα από το twitter, αλλά θα ήταν πολύ χρονοβόρο να ταξινομήσουμε χειροκίνητα τα δεδομένα σε θετικά ή αρνητικά. Γι αυτό το αφήσαμε για μελλοντική εργασία. Στο παρακάτω σχήμα φαίνονται ενδεικτικά οι πρώτες 3 γραμμές του αρχείου twdata.arff που χρησιμοποιήσαμε.

```
@data
'promise! you\'ve a treat to look forward to  and happy chrismukkah!',pos
```

'Are you bringing the sun with you I hope? We\'ve got rain through Wednesday forecast ',neg

'she hates me ',neg

4.3 Επιλογή αλγορίθμων & μεθόδου ταξινόμησης

Όπως προαναφέρθηκε σε προηγούμενη παράγραφο, στην εξόρυξη γνώμης οι πιο αποτελεσματικοί αλγόριθμοι είναι των Μηχανών Διανυσματικής Υποστήριξης (SVMs) και οι Bayesian μέθοδοι. Για τα πειράματά μας χρησιμοποιήθηκαν οι ευρέως γνωστοί SGDText ως εκπρόσωπος των SVMs και ο Naive Bayes Multinomial από την κατηγορία των Bayesian μεθόδων. Επιλέχθηκαν οι συγκεκριμένοι αλγόριθμοι, επειδή είναι ήδη υλοποιημένοι στο WEKA .

Για την μέτρηση της απόδοσης της κατηγοριοποίησης επιλέγουμε τη μέθοδο Διασταυρωμένης επικύρωσης (k-fold Cross Validation) γιατί είναι η πιο αξιόπιστη (McLachlan etal, 2004). Σαν k επιλέγεται το 10 γιατί είναι το πιο διαδεδομένο. Σε αυτή τη μέθοδο το αρχικό σύνολο δεδομένων χωρίζεται σε 10 υποσύνολα. Από τα 10 υποσύνολα, ένα χρησιμοποιείται για επαλήθευση (test set) και τα υπόλοιπα (10-1=9) υποσύνολα χρησιμοποιούνται για την εκπαίδευση (training set). Η διαδικασία αυτή επαναλαμβάνεται 10 φορές (όσες και τα folds), όπου κάθε ένα από τα 10 υποσύνολα χρησιμοποιείται μία φορά σαν δεδομένα επαλήθευσης. Όπως βλέπουμε το πλεονέκτημα αυτής της μεθόδου είναι ότι όλες οι παρατηρήσεις του συνόλου δεδομένων - μέσω τυχαία επιλεγμένων υποσυνόλων δεδομένων - χρησιμοποιούνται και για εκπαίδευσης και για έλεγχο, και το τελικό αποτέλεσμα είναι ο μέσος όρος των παρατηρήσεων Αυτό έχει σαν αποτέλεσμα αυξημένη αξιοπιστία της μεθόδου ανεξάρτητα από το πόσο «καλό» είναι το σύνολο δεδομένων που έχουμε στη διάθεσή μας. Να σημειωθεί εδώ ότι αν επιλέγαμε στο WEKA τη μέθοδο χρησιμοποίησης του ίδιου συνόλου σαν σύνολο εκπαίδευσης και σύνολο δεδομένων (use training set) θα είχε σίγουρα καλύτερα (πιθανώς άριστα) αποτελέσματα, θα υστερούσε όμως φανερά σε αξιοπιστία.

4.3.1 Αποτελέσματα WEKA για το αρχείο twData.arff

Συγκρίνοντας τους δυο προηγούμενους αλγορίθμους, ο NaïveBayesMultinomial δίνει ελαφρώς καλύτερα αποτελέσματα (80,10% έναντι 79.61%) η οποία δεν είναι στατιστικά σημαντική διαφορά. Είναι όμως πολύ πιο γρήγορος, και γι αυτό τον επιλέγουμε για την εφαρμογή μας.

Πίνακας 1. *Αποτελέσματα NaïveBayesMultinomial και SGD Text στο WEKA*

Αλγόριθμος	Ακρίβεια
Naïve Bayes Multinomial	80,10%
SGD Text	79,61%

Για να τον κάνουμε ακόμα πιο αποτελεσματικό τον επιλεγμένο αλγόριθμο, συνδυάσαμε τον NaïveBayesMultinomial με τον αλγόριθμο για stemming της Lovins και τον tokenizer NGram σε μια προσπάθεια να αντιμετωπίσουμε την υπόθεση του NaïveBayesMultinomial ότι κάθε μεταβλητή είναι ανεξάρτητη της άλλης, δεδομένης της κατηγορίας.

Ο αλγόριθμος stemming της Lovins επηρεάστηκε από το τεχνικό λεξικό που ανέπτυξε η Lovins και έχει 294 τελειώματα (endings), 29 καταστάσεις (conditions) και 35 κανόνες μετασχηματισμού (transformation rules). Κάθε τελείωμα σχετίζεται με μια από τις καταστάσεις. Στο πρώτο βήμα, βρίσκεται το μακρύτερο τελείωμα (longest ending) το οποίο ικανοποιεί την συσχετιζόμενη με αυτό κατάσταση και απομακρύνεται. Στο δεύτερο βήμα εφαρμόζονται οι 35 κανόνες μετασχηματισμού για να μετασχηματίσουν το τελείωμα (ending) (Lovins, 1968).

Ο tokenizer N-Gram διαβάζει το πεδίο κειμένου και δημιουργεί n-gram tokens στο δοθέν εύρος. Το N σημαίνει πόσοι όροι εξετάζονται: έχουμε unigrams στην περίπτωση του ενός όρου, biograms στην περίπτωση 2 όρων και trigrams στην περίπτωση 3 όρων (Finke etal, 1999).

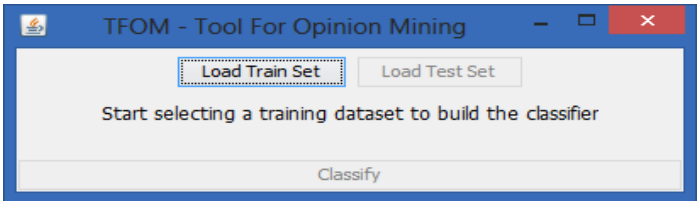
Η ακρίβεια του αλγορίθμου μετά την εφαρμογή του παραπάνω συνδυασμού γίνεται 83.15%, που είναι στατιστικά σημαντική διαφορά (αύξηση 3%).

4.4 Η εφαρμογή «Εργαλείο για Εξόρυξη Γνώμης»

Δημιουργήσαμε την εφαρμογή «Εργαλείο για εξόρυξη Γνώμης (ToolForOpinionMining-TFOM)» χρησιμοποιώντας τον τροποποιημένο Naïve Bayes Multinomial αλγόριθμο, στην οποία θα μπορούμε, όταν εισάγουμε μια νέα πρόταση από δεδομένα του twitter, να την ταξινομεί αυτόματα σαν θετική ή αρνητική.

Η εφαρμογή είναι γραμμένη σε Java και οι κύριες συναρτήσεις της υλοποιήθηκαν παίρνοντας τον κώδικα του τροποποιημένου Naïve Bayes Multinomial Αλγορίθμου από το API του Weka.

Τρέχουμε το αρχείο ToolOpMining.jar (https://dl.dropboxusercontent.com/u/851910/ToolOpMining.jar) και ανοίγει η εφαρμογή TFOM (Εικόνα 2).

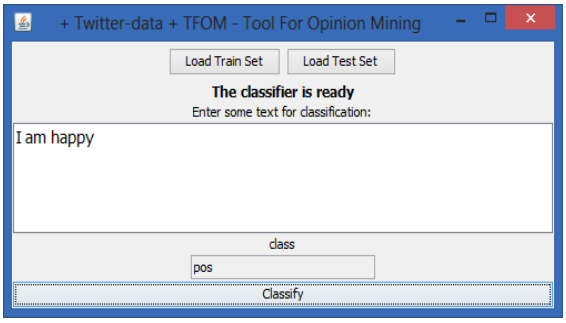


Εικόνα 2. *Η εφαρμογή TFOM*

Φορτώνουμε πατώντας το κουμπί Load Train Set το αρχείο twData.arff. Τώρα ο ταξινομητής μας είναι έτοιμος και περιμένει να εισάγουμε μια νέα πρόταση για να κάνει την ανάλυση κειμένου και να την κατατάξει αυτόματα σε θετική ή αρνητική.

Βάζουμε την πρόταση ”I am happy”. Πατώντας το κουμπί Classify την κατατάσσει αυτόματα ως pos (positive – θετική, Εικόνα 3). Αν δοκιμάσουμε μια πρόταση με άρνηση π.χ. «I am not satisfied with this product», το σύστημα την κατατάσσει αυτόματα ως αρνητική.

Εδώ να σημειώσουμε ότι η εφαρμογή μας μπορεί να χρησιμοποιηθεί για ταξινόμηση κειμένου στα ελληνικά, αρκεί το σύνολο εκπαίδευσης που θα δώσουμε να έχει δεδομένα στα ελληνικά. Μόνο που στην περίπτωση αυτή δεν θα λειτουργεί ο αλγόριθμος stemming της Lovins και κατά συνέπεια θα έχουμε μικρότερη απόδοση. Επίσης μπορεί να χρησιμοποιηθεί και για ταξινόμηση και άλλων κειμένων (εκτός από κείμενα του Twitter), αρκεί να δώσουμε το κατάλληλο σύνολο εκπαίδευσης.



Εικόνα 3. *Θετική (pos) αξιολόγηση από τον ταξινομητή TFOM.*

5. Συμπεράσματα - Μελλοντικές Κατευθύνσεις

Στην παρούσα εργασία θέλοντας να εκμεταλλευτούμε τον τεράστιο όγκο δεδομένων που υπάρχει στο twitter και να ανακαλύψουμε χρήσιμες πληροφορίες που βρίσκονται πίσω από αυτόν, χρησιμοποιήσαμε αλγορίθμους μηχανικής μάθησης και κατασκευάσαμε μια εφαρμογή η οποία δέχεται σαν είσοδο ένα κείμενο του twitter, εξορίσει τη γνώμη που κρύβεται μέσα σε αυτό και την κατατάσσει αυτόματα σαν θετική ή αρνητική.

Επεκτείνοντας την εφαρμογή μας, κατ’ αρχήν θα μπορούσαμε να προσθέσουμε μια ακόμη κατηγορία ταξινόμησης κειμένου, την ουδέτερη (neutral) για δύσκολες περιπτώσεις μηνυμάτων, τα οποία δεν μπορούν να θεωρηθούν ούτε θετικά, ούτε αρνητικά. Το μόνο που χρειάζεται γι’ αυτό είναι η χειροκίνητη ταξινόμηση των μηνυμάτων στο αρχείο .arff να γίνεται σε 3 κατηγορίες (pos, neu, neg). Επίσης θα μπορούσαμε να κινηθούμε προς την κατεύθυνση άλλων έτοιμων συστημάτων

διαχείρισης διαδικτυακής Φήμης (πχ google Alerts, Trackur κτλ) και να ενσωματώσουμε κάποιες δυνατότητές τους.

Τέλος η εφαρμογή TFOM που περιγράψαμε παραπάνω θα μπορούσε να ενσωματωθεί σε ένα ενιαίο ολοκληρωμένο περιβάλλον που θα δουλεύει ως εξής:

- Θα τραβάει με αυτόματο τρόπο σε καθημερινή βάση δεδομένα από το twitter πάνω σε ένα θέμα που θα καθορίζουμε χειροκίνητα εμείς, για ένα συγκεκριμένο χρονικό διάστημα.
- Τις γνώμες που περιέχουν τα δεδομένα μιας ημέρας θα τις ταξινομήσουμε χειροκίνητα εμείς σε θετικές, αρνητικές και ουδέτερες και θα τις χρησιμοποιήσουμε σαν σύνολο εκπαίδευσης για το σύστημα.
- Στα συγκεντρωτικά δεδομένα για το προς μελέτη χρονικό διάστημα, θα εφαρμόζεται ο κώδικας της εφαρμογής TFOM για να προσδίδει σε κάθε κείμενο του twitter το χαρακτηρισμό θετικό ή αρνητικό.
- Το σύστημα θα βγάζει το συνολικό ποσοστό των θετικών/αρνητικών γνωμών πάνω στο συγκεκριμένο θέμα ανά τακτά χρονικά διαστήματα.

Ένα τέτοιο σύστημα θα μπορούσε να βρει άμεση εφαρμογή σε πολλά πεδία: π.χ. θα μπορούσε να βγαίνει αυτόματα κάθε μήνα μια δημοσκόπηση βασισμένη στα σχόλια του twitter για τις γνώμες των Ελλήνων πολιτών για τους πολιτικούς.

Αναφορές

Agarwal, A., Xie B., Vovsha I., Rambow O., and Passonneau R. (2011). *Sentiment Analysis of Twitter Data*, Proc. ACL 2011 Workshop on Languages in Social Media, pp. 30-38.

Aggarwal C., Zhai ChengXiang (2012). A *Survey of text classification Algorithms*, Mining Text Data Boοκ, pp 163-213, ISBN 978-1-4614-3222-7 e-ISBN 978-1-4614-3223-4 Springer.

Bifet A. & Eibe F. (2010). *Sentiment knowledge discovery in Twitter streaming dat*', In Proc 13th International Conference on Discovery Science, Canberra, Australia, Springer, pp 1-15.

Bottou L. & Bousquet, O. (2008). *The Tradeoffs of Large Scale Learning*. Advances in Neural Information Processing Systems 20. pp. 161–168.

Cha M., Haddadi H., Benevenuto F., and Gummadi K. P. (2010). *Measuring User Inuence in Twitter: The Million Follower Fallacy*, In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, pp 10-17.

Finke M., Fritsch J., Koll D., and Waibel A. (1999). *Modeling and efficient decoding of large vocabulary conversational speech*, in proceedings Eurospeech'99, Budapest, Hungary, pp. 467–470.

Francis W. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*, Houghton Mifflin.

Hearst Marti A. (1999). *Untangling Text Data Mining*, Proceedings of ACL '99: the 37th Annual Meeting of the Association for computational Linguistics, University of Maryland, June 20-26.

Klosgen W., Zytkow J. (2002). *Handbook of data mining and knowledge discovery*, Oxford University Press, New York.

Kouloumpis, E., Wilson T., and Moore J. (2011). *Twitter Sentiment Analysis: The Good the Bad and the OMG!*, Proc. ICWSM.

McLachlan, G., Do Kim-Anh, Ambroise C. (2004). *Analyzing microarray gene expression data*, Wiley.

Liu B. (2012). *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, May 2012 (165 pages).

Lovins Julie Beth (1968). *Development of a stemming algorithm*, Mechanical Translation and Computational Linguistics. 11:22-31,.

McCallum A. and Nigam K. (1998). *A Comparison of Event Models for Naive Bayes Text Classification.* In AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48.

Sehgal A.K. (2004). *Text Mining: The Search for Novelty in Text*, Ph.D. Comprehensive Examination Report, Dept. of Computer Science, The University of Iowa.

Witten I., Frank E., Hall M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, Third Edition (The Morgan Kaufmann Series in Data Management Systems).

Χτούρης Σ. (2004). Ορθολογικά Συμβολικά Δίκτυα. Αθήνα: Νήσος, Retrieved 30 January 2013 from http://www.sae-europe.eu/2008-02-15-13-39-24/2008-02-15-13-42-00.

Abstract

In this paper we deal with the use of machine learning techniques to implement an Opinion Mining java application. The application takes as entrance a twitter message, mines the opinion hiding in it and classifies it as positive or negative. We tested the state of the art opinion mining algorithms using WEKA application, we found the one with the best results, we modified it to improve the given results and finally we implement the proposed modified algorithm. In detail, we combined NaïveBayesMultinomial with stemming algorithm of Lovins and the tokenizer NGram in an attempt to handle the hypothesis of NaïveBayesMultinomial that every variable is independent of each other given the class.

Keywords: Opinion Mining, Twitter, Machine Learning, Naïve Bayes Multinomial, Tool For Opinion Mining -TFOM)