Μελέτη και αξιολόγηση τεχνικών εξόρυξης πολιτικής γνώμης σε tweets

Γιάννης Θηβαίος
Α.Μ. 346

Αύγουστος 2017

Πανεπιστήμιο Πατρών - Τμήμα Μαθηματικών

Τριμελής Εξεταστική Επιτροπή: *Κωτσιαντής Σ., Γράψα Θ., Καββαδίας Δ.*

Περιεχόμενα

- Το 80% περίπου των δεδομένων που βρίσκονται στο διαδίκτυο αποτελούν αδόμητη πληροφορία σε μορφή κειμένου, συνεπώς δημιουργούνται ανάγκες και προκλήσεις επεξεργασίας αυτής της πληροφορίας.

- Επεξεργασία κειμένου: Η εξόρυξη κειμένου είναι η διαδικασία αναζήτησης ή εξαγωγής των χρήσιμων πληροφοριών από τα δεδομένα κειμένου.

- Η αναπαράσταση των εγγράφων γίνεται με το Μοντέλο Διανυσματικού Χώρου, με την οποία τα κείμενα αναπαριστώνται σε μορφή διανυσμάτων. [1]

[1] Raghavan, V. V. and Wong, S. K. M. (1986)

- Έχουμε ένα σύνολο από κείμενα και θεωρούμε το καθένα από αυτά ως ένα bag-of-words, μια σακούλα που περιλαμβάνει όλες τις λέξεις μες στο κείμενο.
- Υπάρχουν δύο βασικά στάδια προκειμένου να μετατραπούν τα δεδομένα στην κατάλληλη μορφή προς επεξεργασία:
  - **Στάθμιση όρων (Term Weighting)**: Η στάθμιση όρων είναι μια σημαντική έννοια που καθορίζει την επιτυχία ή την αποτυχία του συστήματος ταξινόμησης. Εκφράζει τη σημασία της λέξης στο έγγραφο. Χρησιμοποιείται η μετρική TF-IDF, όπου TF η συχνότητα του κάθε όρου μες στο κείμενο και IDF είναι μια τιμή που δηλώνει τη σημαντικότητα ενός όρου στο κείμενο, σε σχέση με ολόκληρη τη συλλογή κειμένων.
  - **Μετρικές ομοιότητας**: Το εσωτερικό γινόμενο του διανύσματος εγγράφων και του διανύσματος ερωτήματος, όπου η αλληλοεπικάλυψη λέξεων υποδεικνύει ομοιότητα.

- Η φάση της προεπεξεργασίας χρειάζεται και μια γλωσσολογική επεξεργασία, έτσι ώστε τα αρχικά δεδομένα σε μια δομή έτοιμη για αξιοποίηση από τους αλγόριθμους μάθησης. [2]
  - **Αφαίρεση των stopwords**: Αφαίρεση λέξεων όπως 'the', 'of', 'to', 'and', που δε φέρουν χρήσιμη πληροφορία.
  - **Αποκατάληξη (stemming)**: Μετατροπή των λέξεων στη ρίζα από την οποία προέρχονται.
  - **Μορφοσυντακτική ανάλυση(POS Tagging)**: Κάθε λέξη κατηγοριοποιείται στο μέρος του λόγου το οποίο ανήκει (επίθετο, ουσιαστικό, ρήμα κ.α).
  - **Μείωση Διαστασιμότητας (Dimensionality Reduction)**: Αφαίρεση λέξεων υψηλής συχνότητας που δε σχετίζονται με τη διαδικασία της ταξινόμησης και σπάνιων λέξεων.

[2] Rashmi Agrawal, Mridula Batra, (2013)

Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. [3]

Είδη Μηχανικής Μάθησης

- Επιβλεπόμενη Μάθηση (Supervised Learning)
- Μη Επιβλεπόμενη Μάθησης (Unsupervised Learning)

Αλγόριθμοι Μηχανικής Μάθησης

Naive Bayes, SVM, k-NN classification, Decision Trees, k-Means, Random Forest κ.α

[3] Mitchell, T. (1997)

Ορισμός

Πρόκειται για την επεξεργασία ενός συνόλου αποτελεσμάτων αναζήτησης για ένα συγκεκριμένο στοιχείο, συγκεντρώνοντας απόψεις για το ίδιο το στοιχείο ή κάποια χαρακτηριστικά του(κακή, ουδέτερη, καλή) [4]

Επίπεδα Συναισθηματικής Ανάλυσης

- Επίπεδο Κειμένου
- Επίπεδο Πρότασης
- Επίπεδο Χαρακτηριστικών

Εφαρμογές Ανάλυσης Συναισθήματος

Social Media, Οικονομία, Επιχειρηματικότητα, Πολιτική, Ιατρική,κ.α

[4] S. ChandraKala and C. Sindhu, (2012)

Λίγα Λόγια για το Twitter

- Πρόκειται για το 2ο δημοφιλέστερο μέσο κοινωνικής δικτύωσης. Επιτρέπει στους χρήστες να επικοινωνούν με σύντομα μηνύματα(140 χαρακτήρες) που ονομάζονται tweets.
- Το Twitter είναι χρήσιμο για την ανάγνωση και εύρεση ενδιαφερόντων θεμάτων σε πραγματικό χρόνο που προσελκύουν την προσοχή του χρήστη.
- Το Twitter επίσης υποστηρίζει αποστολή άμεσων προσωπικών μηνυμάτων μεταξύ των χρηστών, ως εκ τούτου αναπτύσσουν βασικές λειτουργίες ανταλλαγής μηνυμάτων
- Όταν ο χρήστης αποφασίσει να ακολουθήσει άλλο λογαριασμό, το Twitter θα ενημερώσει το προφίλ του ακόλουθου με τα πιο πρόσφατα tweets. Ο χρήστης γίνεται στους όρους του Twitter ≪follower≫.

Η χρήση του Twitter από πολιτικούς

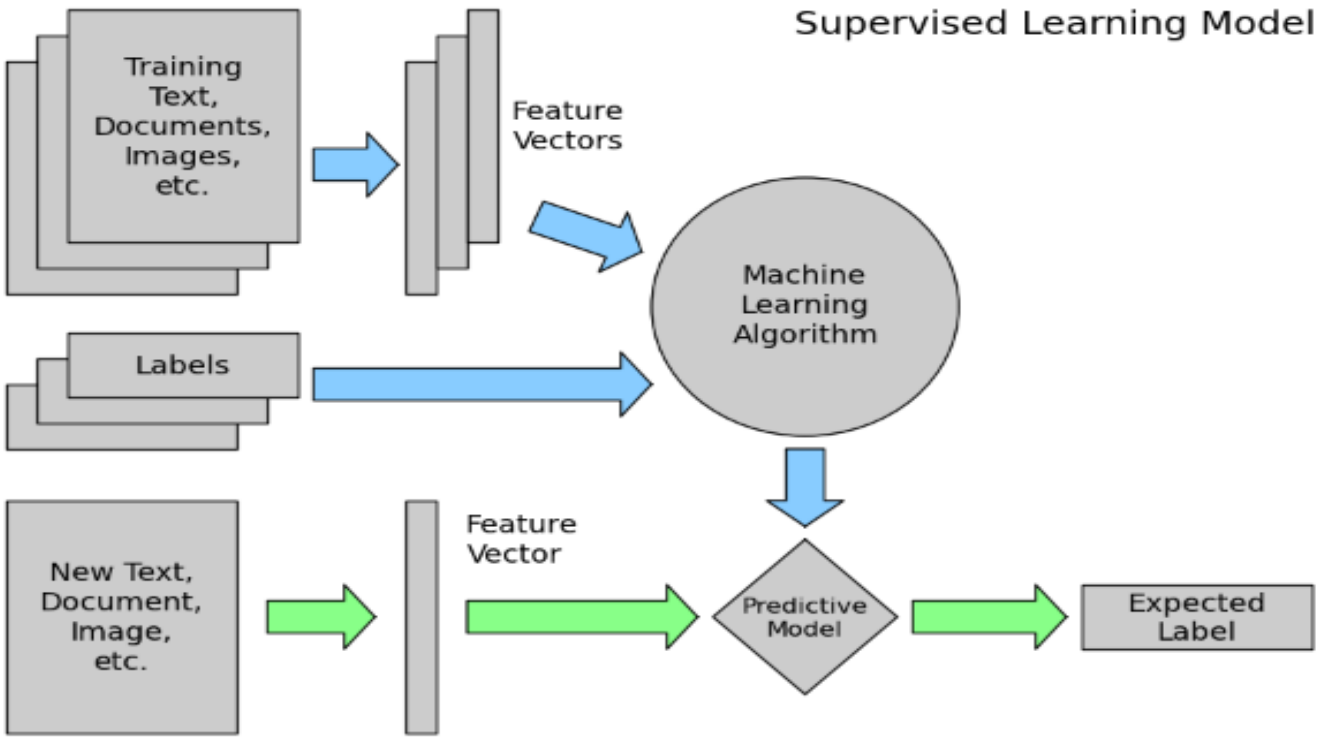
Υπάρχουν τρία βασικά κίνητρα για τη χρήση του Twitter από πολιτικούς: [5]

- Αύξηση της προβολής του προφίλ τους κι επαφή με ηλικιακές κατηγορίες που είναι δύσκολο να έχουν πρόσβαση(πχ 18-24)
- Δημοσιοποίηση δράσεων, προτάσεων και καλέσματα σε κοινωνικοπολιτικές εκδηλώσεις
- Δυνατότητα άμεσης επικοινωνίας και αλληλεπίδρασης με ≪followers≫

[5] Enli, G. S. Skogerbo, E. (2013)

- Έχουμε λοιπόν ένα σύνολο δεδομένων από tweets χρηστών τα οποία αναφέρονται είτε στον Trump είτε στην Clinton, αλλά και ένα σύνολο από θετικές και αρνητικές λέξεις.
- Δημιουργούμε ένα μηχανισμό πρόβλεψης με τη χρήση συγκεκριμένων αλγόριθμων. Εκπαιδεύουμε το σύνολο δεδομένων και προβλέπουμε ποσοστιαία αν ένα tweet κατηγοριοποιείται στον Trump ή στην Clinton και αντίστοιχα αν κατηγοριοποείται θετικά ή αρνητικά.
- Υπολογίζουμε την απόδοση των αλγόριθμων και τις μετρικές αξιολόγησης
- Κατηγοριοποιούμε τυχαία tweets προκειμένου ν' αξιολογήσουμε το μοντέλο πρόβλεψης.
- Χρησιμοποιούμε τη βιβλιοθήκη sklearn της Python.

Συνεπώς, με αυτό το μοντέλο, θα είμαστε σε θέση να κατηγοριοποιούμε αυτόματα μη-κατηγοριοποιημένα tweets σε έναν από τους δύο υποψηφίους και σε μια από τις 2 κατηγορίες πολικότητας.



[6]

[6] Amit Kumar (2015)

Εισαγωγή Συνόλου Δεδομένων

Μπορούμε συνολικά να δούμε το σύνολο των tweets, αλλά και να αποτυπώσουμε σ έναν πίνακα κάποια βασικά στατιστικά στοιχεία.

handle		text
HillaryClinton	Count	3226
	Unique	3224
	top	A man who talks about our veterans and militar...
	freq	2
DonaldTrump	Count	3218
	Unique	3210
	top	MAKE AMERICA GREAT AGAIN!
	freq	8

Προεπεξεργασία Δεδομένων

Χρησιμοποιούμε τη μέθοδο Bag of Words όπου ένα κείμενο θεωρούμε οτι αποτελείται από πληθώρα λέξεων και κάθε λέξη αντιστοιχεί σ΄ έναν αριθμό.

- Βήμα 1: Tokenization

Χωρίζουμε τα tweets στις λέξεις απο τις οποίες αποτελείται. Δίνουμε ένα παράδειγμα από 4 tweets

the, question, in, this, election, who, can, ...

if, we, stand, together, there, 's, nothing, ...

this, election, is, too, important, to, sit, ...

both, candidates, were, asked, about, how, th...

- Βήμα 2: Lemmatization

Χωρίζουμε τα tweets στα λήμματα απο τα οποία αποτελείται. Δίνουμε ένα παράδειγμα από 4 tweets

the, question, in, this, election, who, can, ...

if, we, stand, together, there, 's, nothing, ...

this, election, is, too, important, to, sit, ...

both, **candidate**, were, **ask**, about, how, th...

- Βήμα 3: Μετατροπή των δεδομένων σε διανύσματα

Αυτή η διαδικασία απαιτεί 3 βήματα

- Τη μέτρηση της συχνότητας εμφάνισης μιας λέξης μέσα σ ένα κείμενο
- Τον υπολογισμό του «βάρους» αυτής της μέτρησης. Τα λήμματα που εμφανίζονται συχνότερα θα έχουν μεγαλύτερο βάρος.
- Κανονικοποίηση των διανυσμάτων σε μονάδα μήκους.

Ας δούμε το κείμενο ενός tweet για παράδειγμα
"When you work hard, you should not be living in poverty."

Words	*TF*	*Number*	*Length*	*IDF*
When	1	8755	0.217922181616	4.46783245647
you	2	8593	0.161469587626	2.85068150539
work	1	8572	0.195824391455	6.22809005432
hard	1	7779	0.183979385543	6.78652231980
should	1	7722	0.196876463794	8.35609126545
not	1	7720	0.162516355764	2.13656784456
be	1	7720	0.161347896547	3.43685423559
living	1	6287	0.321250380345	9.21397659807
in	1	8448	0.154590087421	3.34721321890
poverty	1	6321	0.285789076655	8.89032198990

- Συμπερασματικά από την παραπάνω διαδικασία θέλουμε κάθε λέξη μέσα στο σύνολο που έχουμε δημιουργήσει να έχει ένα αντιπροσωπευτικό βάρος, προκειμένου να κάνουμε την καλύτερη δυνατή κατηγοριοποίηση.
- Η στάθμιση TF-IDF δίνει αρκετά καλά αποτελέσματα, καθώς το βάρος IDF παίρνει μεγάλες τιμές, όταν ένας όρος, υπάρχει σε λίγα κείμενα, ενώ, όταν ο όρος συναντάται σε πολλά από τα κείμενα, τότε το βάρος IDF παίρνει μικρές τιμές.
- Με αυτή τη στάθμιση, οι σπάνιοι όροι έχουν υψηλό IDF, και όροι με μεγάλη συχνότητα βαρύνονται με χαμηλότερο IDF.

Τέλος, έχει δημιουργηθεί ένα αραιό μητρώο με το σύνολο των λέξεων με τα παρακάτω χαρακτηριστικά

- sparse matrix shape: (5722, 9016)
- number of non-zeros: 91236
- sparsity: 0.18

Διαδικασία εκπαίδευσης δεδομένων και κατηγοριοποίησης

- Χωρίζουμε το σύνολο δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου

	Training	*Test*	*Data*
Size	5149	573	5722

- Με τη χρήση των αλγόριθμων Multinomial Naive Bayes και SVM αξιολογούμε το σύνολο εκπαίδευσης και το σύνολο ελέγχου τόσο για την κατηγοριοποίηση υποψηφίου όσο και για την κατηγοριοποίηση συναισθήματος.

Κατηγοριοποίηση υποψηφίου με Multinomial Naive Bayes

Accuracy on Training Set: 96.98%

	Precision	*Recall*	$f1-score$	*support*
Hillary	0.96	0.97	0.97	2629
Trump	0.97	0.97	0.97	3093
avg/total	0.97	0.97	0.97	5722

Accuracy on Test Set: 92.96%

	Precision	*Recall*	$f1-score$	*support*
Hillary	0.92	0.94	0.93	274
Trump	0.97	0.97	0.97	299
avg/total	0.94	0.93	0.94	573

Κατηγοριοποίηση συναισθήματος με Multinomial Naive Bayes

Accuracy on Training Set: 88.94%

	Precision	*Recall*	$f1 - score$	*support*
Negative	0.88	0.89	0.88	2729
Positive	0.90	0.89	0.89	2993
avg/total	0.89	0.89	0.89	5722

Accuracy on Test Set: 77.32%

	Precision	*Recall*	$f1 - score$	*support*
Negative	0.76	0.78	0.77	265
Positive	0.90	0.89	0.89	308
avg/total	0.89	0.89	0.89	573

Κατηγοριοποίηση υποψηφίου με SVM

Accuracy on Training Set: 99.72%

	Precision	*Recall*	$f1-score$	*support*
Hillary	1.00	1.00	1.00	2629
Trump	1.00	1.00	1.00	3093
avg/total	1.00	1.00	1.00	5722

Accuracy on Test Set: 93.47%

	Precision	*Recall*	$f1-score$	*support*
Hillary	0.93	0.94	0.94	243
Trump	0.95	0.94	0.95	330
avg/total	0.94	0.94	0.94	573

Κατηγοριοποίηση συναισθήματος με SVM

Accuracy on Training Set: 98.95%

	Precision	*Recall*	$f1-score$	*support*
Negative	0.99	0.99	0.99	2729
Positive	0.99	0.99	0.99	2993
avg/total	0.99	0.99	0.99	5722

Accuracy on Test Set: 80.27%

	Precision	*Recall*	$f1-score$	*support*
Negative	0.81	0.77	0.79	265
Positive	0.80	0.84	0.82	308
avg/total	0.81	0.81	0.81	573

Αξιολόγηση μηχανισμού πρόβλεψης με τον αλγόριθμο Multinomial Naive Bayes

Θα κατηγοριοποιήσουμε τυχαία tweets

- Tweet 1: ' With this election we're simultaneously breaking through the glass ceiling and the rock bottom. We got a really big room now '

  I'm about 76% sure this was tweeted by Hillary and the polarity is about 59% Positive

- Tweet 2: ' Sorry losers and haters, but my I.Q. is one of the highest -and you all know it! Please don't feel so stupid or insecure,it's not your fault '

  I'm about 60% sure this was tweeted by Trump and the polarity is about 67% Negative

Αξιολόγηση μηχανισμού πρόβλεψης με τον αλγόριθμο SVM

Θα κατηγοριοποιήσουμε τυχαία tweets

- Tweet 1: ' With this election we're simultaneously breaking through the glass ceiling and the rock bottom. We got a really big room now '

  I'm about 92% sure this was tweeted by Hillary and the polarity is about 92% Positive

- Tweet 2: ' Sorry losers and haters, but my I.Q. is one of the highest -and you all know it! Please don't feel so stupid or insecure,it's not your fault '

  I'm about 75% sure this was tweeted by Trump and the polarity is about 95% Negative

SVM Vs Multinomial Naive Bayes

- Παρατηρούμε πως μέσα από την εκπαίδευση των δεδομένων, αλλά και με την αξιολόγηση του μηχανισμού πρόβλεψης κατηγοριοποιώντας τυχαία tweets, με το LinearSVC παίρνουμε καλύτερα αποτελέσματα απ' ότι με τον Multinomial Naive Bayes.

- Ο Multinomial Naive Bayes είναι αρκετά πιο γρήγορος από τον SVM, για το λόγω του ότι διαρκεί αρκετή ώρα η εκπαίδευση των συντελεστών για τη γραμμική κατηγοριοποίηση με SVM.

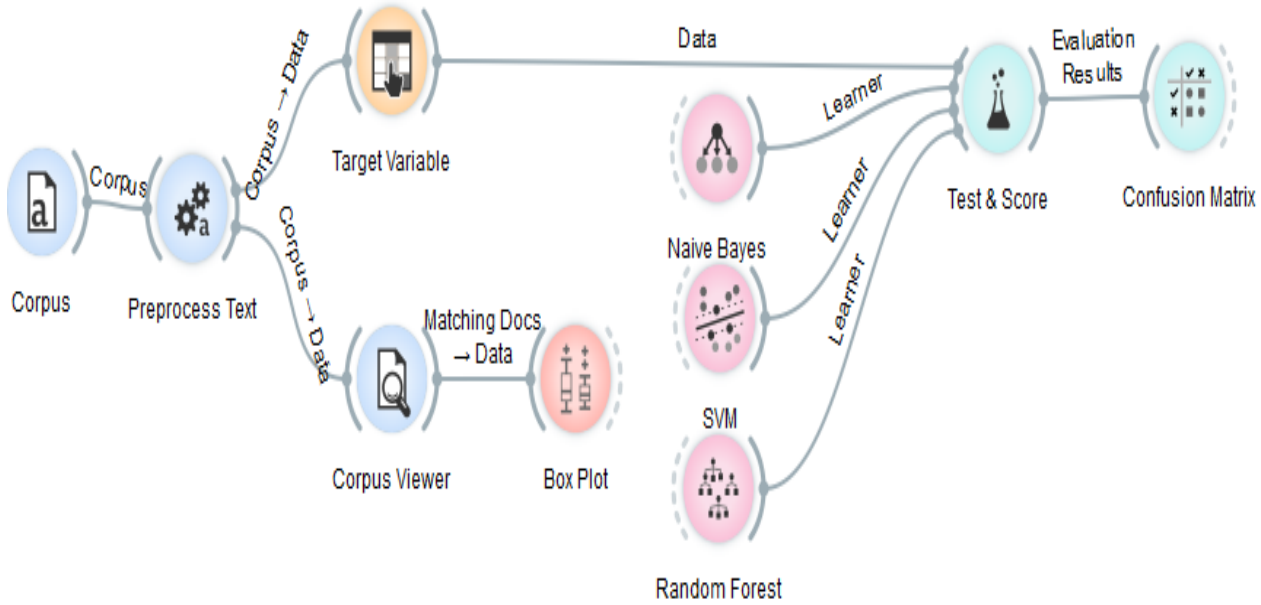
Εισαγωγή στο Orange [7]

- Το Orange είναι μια ανοικτού κώδικα οπτικοποίηση δεδομένων, μηχανικής μάθησης και εργαλείο εξόρυξης δεδομένων.
- Αποτελείται από μικροεφαρμογές(widgets), τα οποία μπορεί να είναι από την απεικόνιση απλών συνόλων δεδομένων, υποσύνολα επιλογής και προεπεξεργασίας μέχρι εμπειρική αξιολόγηση αλγόριθμων μάθησης και προγνωστική μοντελοποίηση.
- Αποτελείται επίσης από μια επιφάνεια γραφικών, πάνω στην οποία ο χρήστης τοποθετεί τα widgets και δημιουργεί μια ροή ανάλυσης δεδομένων (data analysis workflow).

[7] Marinka Žitnik; Blaž Zupan (2013)

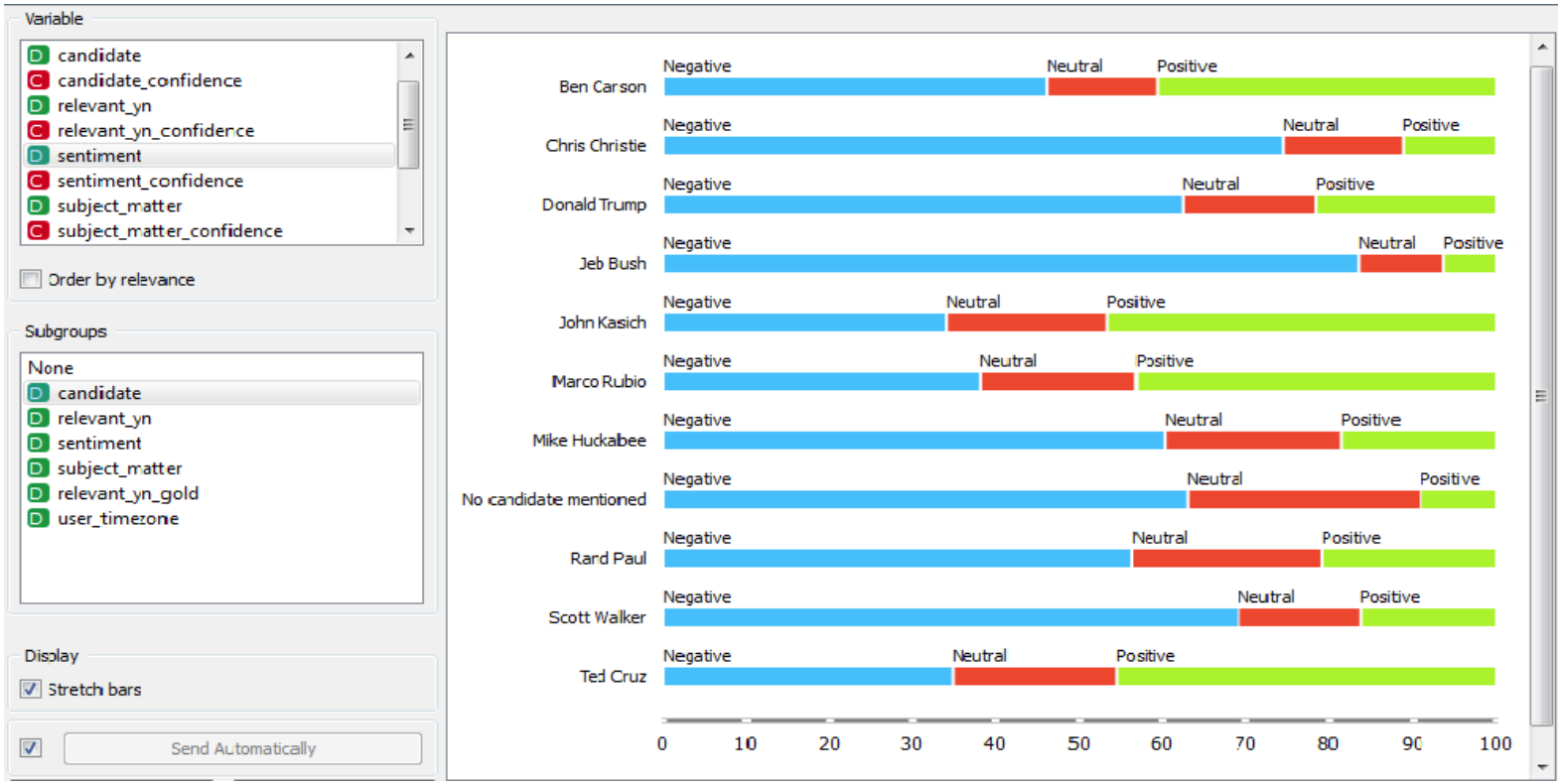
Παράδειγμα Υλοποίησης

- **Σύνολο Δεδομένων**: Είναι συλλογή χιλιάδων tweets μετά το πρώτο debate μεταξύ των υποψηφίων του ρεπουμπλικανικού κόμματος πριν τις Αμερικανικές εκλογές. Τα tweets έχουν αναφορά σε κάποιον υποψήφιο, ενώ υπάρχει και η άποψη για τον συγκεκριμένο υποψήφιο (θετική, αρνητική, ουδέτερη).
- **Ροή εργασίας (Workflow Process)**:



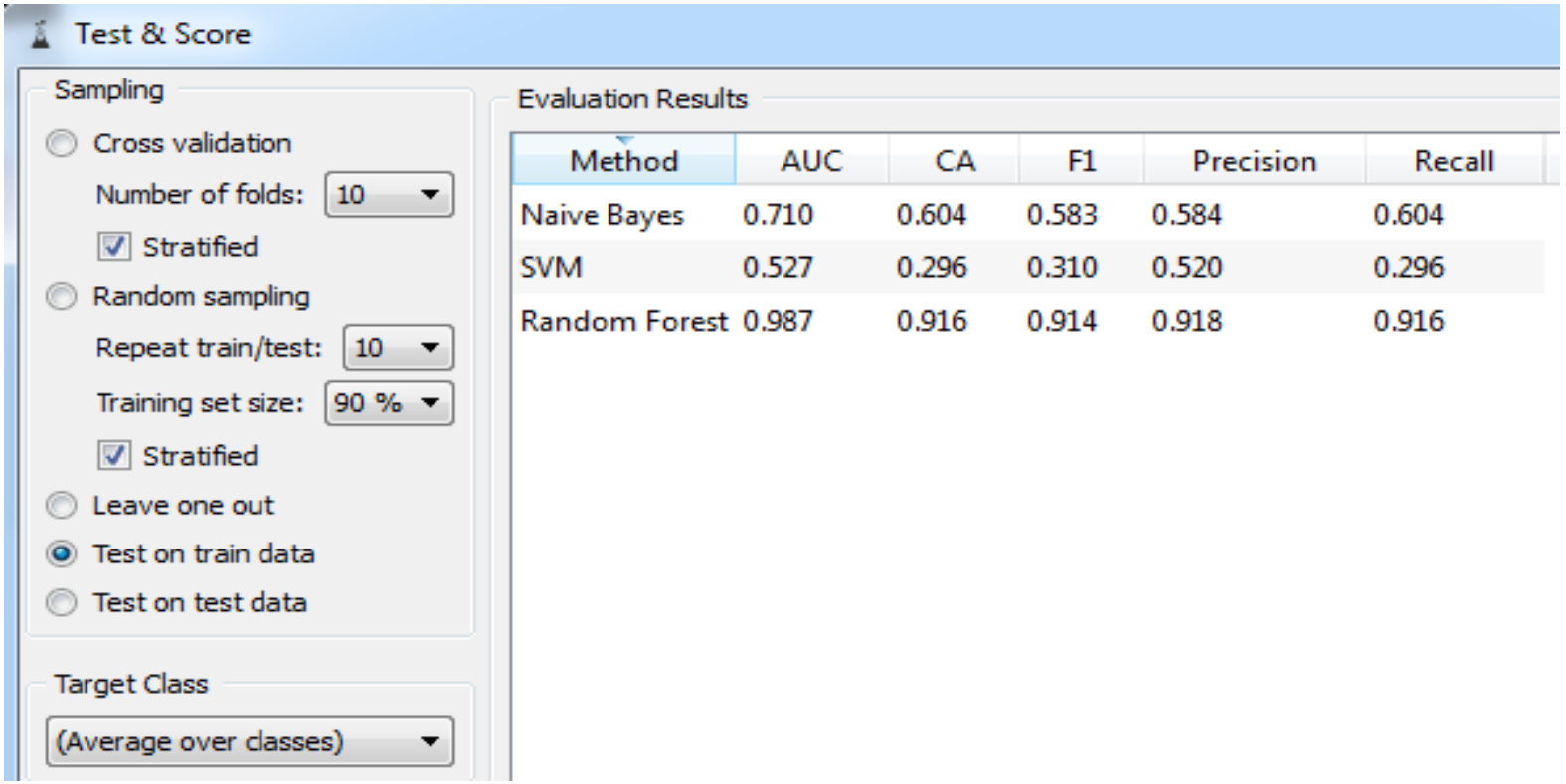
Ανάλυση Δεδομένων με τη χρήση του Box Plot

Αποτυπώνουμε με τη βοήθεια του Box Plot τη σχέση μεταξύ του χαρακτηριστικού 'sentiment' και του χαρακτηριστικού 'candidate', προκειμένου να δούμε γραφικά την κατανομή σε σχέση με το συναίσθημα(θετικό, αρνητικό, ουδέτερο) των tweets για κάθε υποψήφιο.



Εκπαίδευση δεδομένων και απόδοση αλγόριθμων

Με τη χρήση αλγόριθμων μάθησης (Naïve Bayes, SVM, Random Forest) εκπαιδεύουμε τα δεδομένα μας και φτιάχνουμε ένα μοντέλο πρόβλεψης.



Συμπεράσματα

- Οι δύο αλγόριθμοι παρουσιάζουν υψηλά ποσοστά ακρίβειας τόσο για την πρόβλεψη του υποψηφίου όσο και για την πρόβλεψη του συναισθήματος(πάνω από 95%). Το γεγονός αυτό οφείλεται στο ότι έγινε σωστή προεπεξεργασία των δεδομένων, που αποτελεί και το σημαντικότερο παράγοντα για τη σωστή ταξινόμηση.
- Η εισαγωγή συνόλου λέξεων με θετική και αρνητική πολικότητα, μας βοήθησε πολύ και μας έδωσε πολύ καλά αποτελέσματα σχετικά με τη διαμόρφωση γνώμης για τα tweets.
- Χρησιμοποιήσαμε κάποια τυχαία tweets τα οποία ήταν αρκετά δημοφιλή κατά την προεκλογική περίοδο. Τα αποτελέσματα που πήραμε μέσα από το μηχανισμό πρόβλεψης, σχετικά με την πολικότητα του tweet είναι αρκετά ενθαρρυντικά σχετικά με την ακρίβεια και την αξιοπιστία.

Μελλοντική επέκταση

- Τη βελτιστοποίηση του υπάρχοντος μοντέλου, προκειμένου να μπορούμε να επιτυγχάνουμε μεγαλύτερη ακρίβεια. Αυτό μπορεί να γίνει τόσο με την αξιοποίηση περισσότερων χαρακτηριστικών γνωρισμάτων που μπορούν να εξαχθούν από tweets, όσο και από τη βελτιστοποίηση των παραμέτρων των αλγόριθμων.

- Την επεκτασιμότητα του υπάρχοντος μοντέλου, προκειμένου να μπορούμε να πάρουμε περισσότερα στοιχεία για το χρήστη(φύλο, ηλικία, μορφωτικό επίπεδο) και να αξιολογήσουμε το προφίλ του.

1. Raghavan, V. V. and Wong, S. K. M. A critical analysis of vector space model for information retrieval. Journal of the American Society for Information Science, Vol.37 (5), p. 279-87, 1986
2. Rashmi Agrawal, Mridula Batra, "A Detailed Study on Text Mining Techniques", IJSCE, ISSN: 2231-2307, Vol. 2, Issue-6, January 2013.
3. Mitchell, T. (1997). Machine Learning, McGraw Hill, Machine Learning, McGraw Hill, p.2
4. S. ChandraKala and C. Sindhu, (2012), Opinion Mining And Sentiment Classification: A Survey, ICTACT Journal on Soft Computing, Vol- 03, ISSUE: 01, ISSN: 2229-6956
5. Enli, G. S. Skogerbo, E. (2013). Personalized Campaigns In Party-Centred Politics.
6. Introduction-to-machine-learning (May 2015), Amit Kumar.
7. Marinka Žitnik; Blaž Zupan (2013). "Orange: data mining toolbox in Python" (PDF). JMLR. 14 (1): 2349–2353.

Ευχαριστώ για την προσοχή σας