ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΡΟΒΛΕΨΗ ΕΚΛΟΓΙΚΟΥ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΜΕ ΧΡΗΣΗ ΣΥΝΑΙΣΘΗΜΑΤΙΚΗΣ ΑΝΑΛΥΣΗΣ ΣΤΑ ΔΕΔΟΜΕΝΑ ΤΟΥ TWITTER

Διπλωματική Εργασία

ΚΩΝΣΤΑΝΤΙΝΟΣ ΜΟΥΛΟΣΙΩΤΗΣ

Επιβλέπουσα: Τουσίδου Ελένη

Βόλος 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΡΟΒΛΕΨΗ ΕΚΛΟΓΙΚΟΥ ΑΠΟΤΕΛΕΣΜΑΤΟΣ ΜΕ ΧΡΗΣΗ ΣΥΝΑΙΣΘΗΜΑΤΙΚΗΣ ΑΝΑΛΥΣΗΣ ΣΤΑ ΔΕΔΟΜΕΝΑ ΤΟΥ TWITTER

Διπλωματική Εργασία

ΚΩΝΣΤΑΝΤΙΝΟΣ ΜΟΥΛΟΣΙΩΤΗΣ

Επιβλέπουσα: Τουσίδου Ελένη

Βόλος 2021

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

PREDICTING ELECTION RESULT

BY USING SENTIMENT ANALYSIS OF TWITTER DATA

Diploma Thesis

KONSTANTINOS MOULOSIOTIS

Supervisor: Tousidou Eleni

Volos 2021

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπουσα	**ΤΟΥΣΙΔΟΥ ΕΛΕΝΗ** Ε.ΔΙ.Π, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας
Μέλος	**ΒΑΣΙΛΑΚΟΠΟΥΛΟΣ ΜΙΧΑΗΛ** Αναπληρωτής καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας
Μέλος	**ΦΕΥΓΑΣ ΑΘΑΝΑΣΙΟΣ** Ε.ΔΙ.Π, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ημερομηνία έγκρισης: 21-09-2021

Ευχαριστίες

Σε αυτό το σημείο, με αφορμή την ολοκλήρωση της διπλωματικής μου εργασίας θα ήθελα να ευχαριστήσω θερμά την κυρία Τουσίδου Ελένη για την άψογη συνεργασία καθ' όλη τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας, καθώς η βοήθεια της ήταν πολύτιμη και καθοριστική για την ολοκλήρωση της.

Ακόμη, θα ήθελα να ευχαριστήσω την οικογένεια μου για όλη την υποστήριξη και εμπιστοσύνη που έδειξε σε εμένα καθ' όλη την διάρκεια της φοιτητικής μου πορείας, και όχι μόνο.

Τέλος, με μεγάλη χαρά και συγκίνηση, καθώς αυτός ο κύκλος τελειώνει, ένα τεράστιο ευχαριστώ στους φίλους μου. Στους φίλους μου, με τους οποίους μοιραστήκαμε τα καλύτερα χρόνια της ζωής μας και ήταν δίπλα μου οποιαδήποτε στιγμή χρειαζόμουν τη βοήθεια τους. Η παρουσία τους ήταν πολύ σημαντική για εμένα, καθώς οι στιγμές που μοιραστήκαμε θα μείνουν χαραγμένες στην μνήμη μου για πάντα.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο/Η Δηλών/ούσα

(Υπογραφή)

Κωνσταντίνος Μουλοσιώτης

18/9/2021

Περίληψη

Στις 3 Νοεμβρίου 2020, στις Ηνωμένες Πολιτείες Αμερικής, έλαβαν χώρα οι Προεδρικές εκλογές, με τους δύο κύριους αντιπάλους να είναι ο Ντόναλντ Τραμπ και ο Τζο Μπάιντεν. Τόσο οι ίδιοι όσο και οι απλοί πολίτες είχαν ενεργή συμμετοχή στα μέσα κοινωνικής δικτύωσης. Το Twitter είναι η πιο δημοφιλής πλατφόρμα διαμοιρασμού απόψεων πολιτικού περιεχομένου, και όχι μόνο.

Στην παρούσα εργασία θα παρουσιαστεί μια ολοκληρωμένη ανάλυση των συναισθημάτων των tweets και με την βοήθεια αυτής θα γίνει πρόβλεψη του εκλογικού αποτελέσματος. Ακόμη θα γίνει διασταύρωση αυτής της πρόβλεψης με τα πραγματικά αποτελέσματα για να διαπιστώσουμε κατά πόσο μπορούμε να βασιστούμε σε παρόμοιες αναλύσεις ώστε να προβλέπουμε τέτοιου είδους αποτελέσματα. Αρχικά, γίνεται περιγραφή των δεδομένων που έχουμε στην διάθεση μας καθώς και εφαρμογή συναισθηματικής ανάλυσης σε αυτά. Στη συνέχεια, με την χρήση μαθηματικής φόρμουλας γίνεται η πρόβλεψη του αποτελέσματος. Έπειτα αξιοποιούνται αλγόριθμοι μηχανικής μάθησης στην διαδικασία της συναισθηματικής ανάλυσης και γίνεται εφαρμογή τους σε νέα δεδομένα. Τέλος, παρουσιάζεται μελέτη των διαθέσιμων tweets, για την εξερεύνηση θεματικών ενοτήτων, που βρισκόταν στην επικαιρότητα κατά την περίοδο των εκλογών.

Abstract

On November 3, 2020, in the United States of America, the Presidential elections took place, with the two main rivals being Donald Trump and Joe Biden. Both candidates, along with ordinary citizens, have been actively involved in social media. Twitter is the most popular platform for sharing political views, and more.

In this dissertation, a complete analysis of the tweets' emotions will be presented and with the contribution of this, the election result will be predicted. Following, a comparison of the actual and predicted results will be given in order to verify whether we can rely on similar analyzes to predict such results. Initially, the data we have at our disposal are described as well as the application of emotional analysis to them. Following, the result is predicted by using a mathematical formula. Then, machine learning algorithms are utilized in the process of emotional analysis and are applied on a new set of data. Finally, a study of the available tweets is presented, aiming at the exploration of thematic units which were in the news during the election period.

Πίνακας Περιεχομένων

Κατάλογος Σχημάτων

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Λόγω της ευρείας χρήσης των μέσων κοινωνικής δικτύωσης πλέον οι χρήστες έχουν την ευκαιρία να αλληλοεπιδρούν μεταξύ τους, να ενημερώνονται καθώς και να μοιράζονται την γνώμη τους πάνω σε επίκαιρα θέματα. Επιπλέον, έχουν την δυνατότητα να συμμετέχουν σε συζητήσεις σοβαρών γεγονότων, όπως για παράδειγμα σε πολιτικά θέματα.

Αυτή η ευκαιρία που δόθηκε στους χρήστες με το πέρασμα των χρόνων αναγνωρίστηκε από τους πολιτικούς και από τα πολιτικά κόμματα παγκοσμίως[1]. Ο πιθανός ρόλος που μπορούν να διαδραματίσουν τα μέσα κοινωνικής δικτύωσης σε πολιτικές εκδηλώσεις τονίστηκε για πρώτη φορά κατά τη διάρκεια των προεδρικών εκλογών το 2008 στις ΗΠΑ. Το Twitter έπαιξε σημαντικό ρόλο στην εκστρατεία του Μπαράκ Ομπάμα, όπου χρησιμοποιήθηκε με αποδοτικό τρόπο για να δημοσιεύει τις ενημερώσεις της καμπάνιας του και να ενημερώνει τους οπαδούς με ευκαιρίες εθελοντισμού για την εκστρατεία[2]. Η κίνηση αυτή φάνηκε να έχει μεγάλη ανταπόκριση από τον κόσμο και από τότε ολοένα και περισσότεροι πολιτικοί και πολιτικά κόμματα άρχισαν να εντάσσουν τα μέσα κοινωνικής δικτύωσης στις εκλογικές του καμπάνιες.

Αυτή η ραγδαία ανάπτυξη της χρήσης του Twitter οδήγησε και στην αύξηση της έρευνας της συναισθηματικής ανάλυσης που αφορά τα μηνύματα του Twitter, αλλά και γενικότερα της ανάλυσης των δεδομένων με στόχο να εντοπίσουν την δημοτικότητα του κάθε πολιτικού και να εξάγουν συμπεράσματα για το αποτέλεσμα των εκλογών[3].

1.1 Η διπλωματική

Σκοπός της συγκεκριμένης διπλωματικής εργασίας είναι να αναλύσει δεδομένα του Twitter κατά την περίοδο των προεδρικών εκλογών των ΗΠΑ στις 3 Νοεμβρίου 2020 και να βγάλει συμπεράσματα για το κατά πόσο μπορούμε να βασιστούμε σε μεθόδους εξόρυξης δεδομένων και ανάλυσης αυτών για να εξάγουμε το πιθανό αποτέλεσμα των εκλογών. Στη μελέτη που γίνεται εφαρμόζεται κατά κύριο λόγο συναισθηματική ανάλυση των tweets που έχουν συλλεχθεί από την πλατφόρμα του Twitter και γίνεται σύγκριση με τα πραγματικά αποτελέσματα των εκλογών όπως αυτά δημοσιεύτηκαν με το πέρας τους.

Ακόμη, γίνεται εκπαίδευση μοντέλων μηχανικής μάθησης, με την βοήθεια των οποίων θα εξάγεται το πιθανό αποτέλεσμα, καθώς και βελτίωση αυτών για την παραγωγή ακριβέστερου αποτελέσματος. Η επιλογή του καταλληλότερου μοντέλου μηχανικής μάθησης καθώς και τεχνικών επεξεργασίας κειμένου γίνεται ύστερα από σειρά πειραμάτων, τα οποία περιγράφονται αναλυτικά στην συνέχεια της παρούσας εργασίας.

Τέλος, η εργασία αυτή εστιάζει στην εξερεύνηση θεματικών ενοτήτων οι οποίες έπαιξαν σημαντικό ρόλο στην εξαγωγή του αποτελέσματος των εκλογών.

ΚΕΦΑΛΑΙΟ 2

Συναφείς εργασίες

Σε αυτήν την ενότητα αναφέρονται, συνοπτικά, άρθρα σχετικά με το αντικείμενο αυτής της διπλωματικής εργασίας, που βοήθησαν σημαντικά στην διεκπεραίωση της.

Στο άρθρο των Ussama Yaqub, Soon Ae Chun, Vijayalakshmi Atluri και Jaideep Vaidya[4], γίνεται συναισθηματική ανάλυση και αναπτύσσεται ένα υποθετικό μοντέλο για την ανάλυση των δεδομένων. Συγκεκριμένα γίνονται τέσσερις υποθέσεις και κάποιες από αυτές είναι ότι «οι χρήστες συνήθως δεν δημιουργούν νέο περιεχόμενο σχετικό με τις εκλογές, αλλά προτιμούν να βρίσκουν ενδιαφέρουσες και χρήσιμες πληροφορίες τις οποίες επανακοινοποιούν» και ότι «η συχνότητα των δημοφιλών όρων στις συζητήσεις στο Twitter μπορεί να χρησιμοποιηθεί για τον εντοπισμό σημαντικών πραγματικών γεγονότων και ειδήσεων που λαμβάνουν χώρα σχετικά με τις εκλογές» κλπ. Οι Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe στο άρθρο τους [5] εστιάζουν σε δώδεκα διαστάσεις για να προβλέψουν το συναίσθημα. Αυτές είναι οι: μελλοντικός προσανατολισμός, προσανατολισμός παρελθόντος, θετικό συναίσθημα, αρνητικό συναίσθημα, θλίψη, άγχος, θυμός, ετοιμότητα, βεβαιότητα, εργασία, επίτευγμα και χρήματα. Ακολουθώντας την μεθοδολογία που χρησιμοποίησαν οι Yu, Kaufmann, and Diermeier (2008), βλέπουμε ότι επεξεργάστηκαν τα tweets με το LIWC αγγλικό λεξικό. Στο άρθρο των Satish Mahadevan Srinivasan, Raghvinder Sangwan, Colin Neill, Tianhai Zu[6] γίνεται χρήση του NRC ταξινομητή για τον εντοπισμό των συναισθημάτων και με την χρήση αυτών γίνεται η πρόβλεψη των αποτελεσμάτων των δεκαεννέα Πολιτειών των ΗΠΑ προς τους υποψηφίους των προεδρικών εκλογών του 2016. Οι Jyoti Ramteke, Samarth Shah, Darshan Godhia, Aadil Shaikh στο άρθρο τους [33] κάνουν πρόβλεψη του εκλογικού αποτελέσματος στις Προεδρικές εκλογές του 2016 των ΗΠΑ, χρησιμοποιώντας δύο αλγόριθμους επιβλεπόμενης μηχανικής μάθησης. Τέλος προτείνουν ότι νικητής αναδεικνύεται εκείνος που έχει το μεγαλύτερο ποσοστό θετικών tweets από το σύνολο των σχολίων που αφορούν τον ίδιο, και όχι από το σύνολο όλων των tweets. Επομένως, καταλήγουν πως η συγκεκριμένη

μέθοδος δεν μπορεί να εφαρμοστεί για να γίνει η πρόβλεψη, καθώς το ποσοστό των θετικών tweets μπορεί να είναι δυσανάλογο μεταξύ των υποψηφίων. Στο άρθρο τους [34] οι A. Tsakalidis, S. Papadopoulos, A. I. Cristea και Y. Kompatsiaris εστιάζουν στα δεδομένα του Twitter για να προβλέψουν το αποτέλεσμα των εκλογών του 2014 σε ευρωπαϊκές χώρες, όπως η Γερμανία, η Ολλανδία και η Ελλάδα. Η συναισθηματική ανάλυση βασίζεται στην προσέγγιση του λεξικού. Λόγω της αδυναμίας εύρεσης λεξικών στις συγκεκριμένες γλώσσες, κάνουν μετάφραση τρία αγγλικά λεξικά με την βοήθεια του Google Translate. Σε κάθε ένα σύνολο δεδομένων κάνουν εφαρμογή τριών αλγορίθμων, όπως η γραμμική παλινδρόμηση, η διαδικασία του Gauss και η διαδοχική ελάχιστη βελτιστοποίηση για την παλινδρόμηση. Οι Ali Hasan, Sana Moin, Ahmad Karim και Shahaboddin Shamshirband στο άρθρο τους [35] εξάγουν συμπεράσματα όσον αφορά τη δημοτικότητα των δημοκρατικών κομμάτων κατά την περίοδο των εκλογών της Ινδίας το 2013. Η συλλογή των δεδομένων έγινε με την χρήση του API του Twitter σε λέξεις κλειδιά με γνωστά ονόματα πολιτικών και πολιτικών κομμάτων. Στα δεδομένα αυτά, εφαρμόστηκαν αλγόριθμοι επιβλεπόμενης και μη-επιβλεπόμενης μάθησης για την ανάλυση 100.000 tweets.

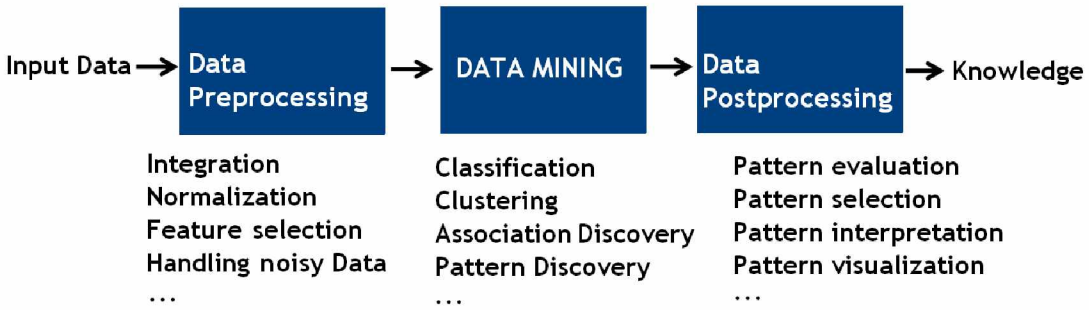
ΚΕΦΑΛΑΙΟ 3

Θεωρητικό υπόβαθρο

Σε αυτό το κεφάλαιο γίνεται μια θεωρητική προσέγγιση της Εξόρυξης Δεδομένων και μία συνοπτική παρουσίαση των αλγορίθμων που χρησιμοποιούνται στη συγκεκριμένη διπλωματική εργασία.

3.1 Εξόρυξη Δεδομένων

Η Εξόρυξη Δεδομένων [7] είναι μια διαδικασία εύρεσης χρήσιμων και άγνωστων πληροφοριών από ένα μεγάλο όγκο δεδομένων. Η εξόρυξη δεδομένων σημαίνει επίσης ανακάλυψη γνώσης από δεδομένα, η οποία περιγράφει την τυπική διαδικασία εξαγωγής χρήσιμων πληροφοριών από ακατέργαστα δεδομένα. Η διαδικασία αυτή αποτελείται γενικά από τις ακόλουθες εργασίες: προεπεξεργασία δεδομένων, εξόρυξη δεδομένων και μεταεπεξεργασία. Στο σχήμα της εικόνας 3.1 παρουσιάζεται η διαδικασία μετασχηματισμού των δεδομένων σε χρήσιμη

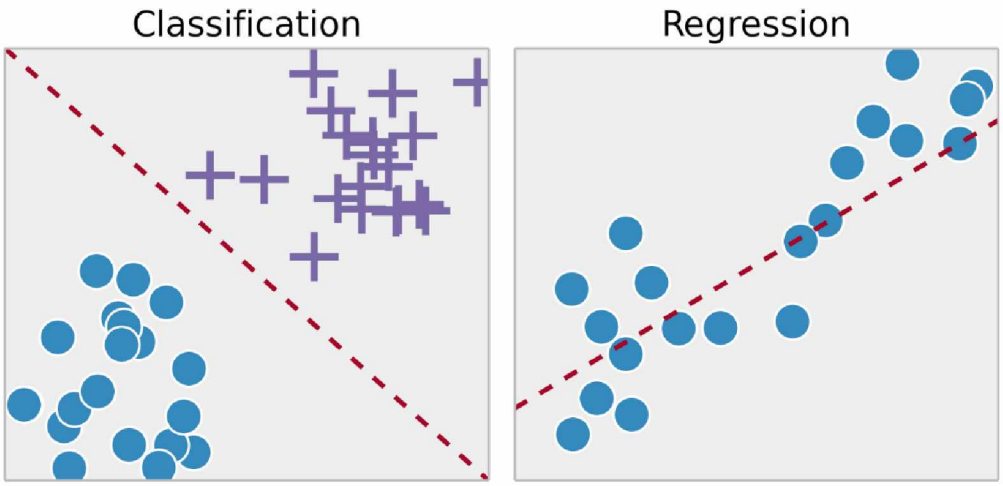


Εικόνα 3.1 Μετατροπή δεδομένων σε χρήσιμη πληροφορία

πληροφορία. Η εξόρυξη δεδομένων αποτελεί αναπόσπαστο μέρος πολλών σχετικών τομέων, όπως η στατιστική, μηχανική μάθηση, αναγνώριση προτύπων, συστήματα βάσεων δεδομένων, οπτικοποίηση, αποθήκη δεδομένων και ανάκτηση πληροφοριών. Οι δύο βασικές κατηγορίες αλγορίθμων Εξόρυξης Δεδομένων είναι η

επιβλεπόμενη και μη-επιβλεπόμενη μάθηση [8]. Η κύρια διαφορά μεταξύ αυτών των δύο κατηγοριών είναι ότι στην επιβλεπόμενη μάθηση η τιμή εξόδου είναι ήδη γνωστή. Ως εκ τούτου, ο στόχος της είναι να προσεγγίζει καλύτερα τη σχέση μεταξύ εισόδου και εξόδου που παρατηρείται στα δεδομένα. Η μη-επιβλεπόμενη μάθηση, από την άλλη πλευρά, δεν έχει ετικέτες εξόδου, οπότε στόχος της είναι να συμπεράνει τη φυσική δομή που υπάρχει σε ένα σύνολο σημείων δεδομένων[9].

Χαρακτηριστικά παραδείγματα της επιβλεπόμενης μάθησης είναι η Κατηγοριοποίηση (Classification) και η Παλινδρόμηση (Regression). Όπως παρουσιάζεται και στα σχήματα της εικόνας 3.2 η κατηγοριοποίηση είναι η αντιστοίχιση της εισόδου στις ετικέτες εξόδου, ενώ η παλινδρόμηση είναι η απεικόνιση της εισόδου σε μία συνεχή έξοδο.



Εικόνα 3.2 Κατηγοριοποίηση και Παλινδρόμηση [23]

Στους αλγόριθμους μη-επιβλεπόμενης μάθησης δεν υπάρχει ετικέτα εξόδου. Δεδομένου ότι δεν παρέχονται ετικέτες, δεν υπάρχει συγκεκριμένος τρόπος ελέγχου της απόδοσης του μοντέλου. Σκοπός ενός τέτοιου αλγορίθμου είναι να ομαδοποιήσει τα δεδομένα εισόδου. Χαρακτηριστικά παραδείγματα αλγορίθμων μη-επιβλεπόμενης μάθησης είναι η Συσταδοποίηση (Clustering) και η Συσχέτιση (Association) [9].

3.2 Εξόρυξη Κειμένου

Η εξόρυξη κειμένου (text mining) ή εξόρυξη δεδομένων κειμένου, όπως αλλιώς αναφέρεται, είναι η διαδικασία εξαγωγής υψηλής ποιότητας πληροφοριών από διάφορα είδη κειμένων. Τα δεδομένα βρίσκονται σε μορφή κειμένου ανεξαρτήτου μεγέθους. Το κείμενο είναι μία από τις πιο συνηθισμένες δομές δεδομένων και χωρίζεται σε τρείς κατηγορίες οργάνωσης. Υπάρχουν τα:

- **Δομημένα Δεδομένα (Structured Data):** Τα δεδομένα αυτά είναι σε μορφή πίνακα με αριθμημένες γραμμές και στήλες, γεγονός που καθιστά ευκολότερη την διαδικασία αποθήκευσης νέων δεδομένων καθώς και την ανάλυση τους από τους αλγορίθμους μηχανικής μάθησης.
- **Μη-δομημένα Δεδομένα (Unstructured Data):** Στην κατηγορία αυτή όπως μπορεί να υπονοηθεί και από το όνομα της βρίσκονται δεδομένα κάθε μορφής που δεν ακολουθούν κάποια δομή. Αυτά μπορεί να είναι κριτικές προϊόντων ή ταινιών, σχόλια από τα μέσα κοινωνικής δικτύωσης, ακόμα και αρχεία ήχου και εικόνας.
- **Ημι-δομημένα Δεδομένα (Semi-structured Data):** Τα συγκεκριμένα δεδομένα είναι ένα μείγμα δομημένων και μη-δομημένων δεδομένων. Παραδείγματα αυτών είναι τα αρχεία XML, JSON και HTML.

Η εξόρυξη κειμένου συμπεριλαμβάνει αρκετές εφαρμογές στην καθημερινή ζωή μας. Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Process – NLP), είναι η πιο παλιά και μάλιστα το πιο δύσκολο πρόβλημα της τεχνητής νοημοσύνης. Πιο συγκεκριμένα είναι η ανάλυση της φυσικής γλώσσας έτσι ώστε οι υπολογιστές να μπορούν να την κατανοήσουν όπως οι άνθρωποι. Παρόλη τη δυσκολία αυτή, το NLP μπορεί να εντοπίσει τους βασικούς γραμματικούς τύπους μίας πρότασης με υψηλό βαθμό επιτυχίας[10].

3.2.1 Προεπεξεργασία Κειμένου

Προεπεξεργασία κείμενου είναι μια σειρά ενεργειών για να μετατραπεί το κείμενο σε κατάλληλη μορφή, έτσι ώστε η διαδικασία της ανάλυσης να είναι αποτελεσματικότερη[11].

- Μετατροπή όλων των χαρακτήρων σε πεζούς (lowercase).
- Αφαίρεση όλων των σημείων στίξης (punctuations) καθώς και των emoji.
- «Καθάρισμα» του κειμένου από urls, hastags (#), mentions (@), χαρακτήρων αλλαγής γραμμής (\n) και την συμβολοσειρά «RT» που υποδηλώνει ότι ένα tweet έχει γίνει retweet.
- Tokenization ή διαχωρισμός λεκτικών μονάδων. Είναι η διαδικασία που το κείμενο χωρίζεται σε λεκτικές μονάδες, τα λεγόμενα tokens. Αυτή γίνεται για να μπορέσουν να εφαρμοστούν όλες οι παρακάτω τεχνικές επεξεργασίας κειμένου.
- Αφαίρεση των stop-words. Με τον όρο stop-words εννοούμε τις λέξεις μιας γλώσσας που χρησιμοποιούνται πολύ συχνά και συνήθως δεν προσφέρουν χρήσιμη πληροφορία στο νόημα του κειμένου. Μερικά παραδείγματα stop-words της Αγγλικής γλώσσας είναι οι λέξεις as, at, be, because, I κλπ.
- Lemmatization ή λημματοποίηση. Είναι η τεχνική με την οποία μία λέξη μετατρέπεται στη ριζική της μορφή. Για παράδειγμα η λέξη «better» γίνεται «good».
- Stemming ή αποκοπή καταλήξεων. Με αυτή την τεχνική γίνεται αποκοπή των καταλήξεων και μετατρέπει την λέξη σε μια απλούστερη μορφή. Η λέξη που προκύπτει δεν είναι πάντα πραγματική. Για παράδειγμα οι λέξεις trouble, troubled, troubles μετά το stemming μετατρέπονται σε trouble. Η κύρια διαφορά της με την λημματοποίηση είναι ότι η αποκοπή είναι ίδια σε κάθε περίπτωση ανεξάρτητα από το μέρος του λόγου που είναι η λέξη. Αυτή η αποκοπή καταλήξεων μπορεί να μετατρέψει την λέξη σε κάποια άλλη και να αλλάξει εντελώς το νόημα της.

Όλες οι παραπάνω λειτουργίες αποσκοπούν στο να μειωθεί το μέγεθος του κειμένου χωρίς να χαθεί χρήσιμη πληροφορία για την ανάλυση του.

3.2.2 Διανυσματοποίηση

Η διανυσματοποίηση (vectorization) είναι μια αναγκαία μέθοδος για την επεξεργασία δεδομένων κειμένου σε εφαρμογές για την επεξεργασία φυσικής γλώσσας. Η διανυσματοποίηση επιτρέπει στους αλγορίθμους να κατανοήσουν τα περιεχόμενα του κειμένου μετατρέποντας τα σε αριθμητικές αναπαραστάσεις. Στη συνέχεια θα δούμε μία σύντομη περιγραφή από κάποιες τεχνικές διανυσματοποίησης δεδομένων κειμένου [12] [13] [14].

Bag of Words (BoW)

Το μοντέλο BoW είναι η απλούστερη μέθοδος για την αναπαράσταση των δεδομένων κειμένου σε αριθμητική έκφραση. Η ονομασία του προέρχεται από το γεγονός ότι απορρίπτεται η σειρά των λέξεων ενός εγγράφου και δείχνει μόνο αν η λέξη βρίσκεται στο έγγραφο ή όχι. Αρχικά χτίζεται ένα λεξικό με τις μοναδικές τιμές από όλα τα έγγραφα. Στη συνέχεια, για κάθε έγγραφο δημιουργείται ένα διάνυσμα, ίδιου μεγέθους, με 0 αν κάθε λέξη του λεξικού δεν συμπεριλαμβάνεται σε αυτό και με 1 αν υπάρχει. Τέλος, όλα αυτά τα διανύσματα τοποθετούνται σε έναν αραιό πίνακα (sparse matrix) και αποτελούν την αριθμητική αναπαράσταση των εγγράφων κειμένου.

Παρόλο που το συγκεκριμένο μοντέλο χρησιμοποιείται αρκετά παρουσιάζει και σημαντικά μειονεκτήματα και συγκεκριμένα όταν ερχόμαστε αντιμέτωποι με νέα έγγραφα. Κάποια από αυτά είναι τα εξής:

- Αν το νέο έγγραφο περιέχει καινούργιες λέξεις, το μέγεθος του λεξικού θα αυξηθεί και αυτό αναλόγως.

- Αυξάνοντας το μέγεθος του λεξικού τα διανύσματα θα περιέχουν περισσότερα μηδενικά, πράγμα που πρέπει να αποφεύγεται.
- Δεν διατηρείται καμία γραμματική πληροφορία ενώ, επίσης, αγνοείται η σειρά των λέξεων στο κείμενο.

Term Frequency-Inverse Document Frequency (TF-IDF)

Η συγκεκριμένη μέθοδος είναι η πιο συχνά χρησιμοποιούμενη για την διανυσματοποίηση κειμένου και επίσης η τεχνική που χρησιμοποιείται στην συγκεκριμένη εργασία. Ο επίσημος ορισμός όπως δίνεται το Wikipedia είναι:

> *"Term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus."*

Όπως φαίνεται και από το όνομα της η συγκεκριμένη μέθοδος χωρίζεται σε δύο σενάρια. Το πρώτο είναι ο όρος Συχνότητα Λέξης (Term Frequency ή TF) και το δεύτερο ο όρος Αντίστροφη Συχνότητα Εγγράφου (Inverse Document Frequency ή IDF). Η συχνότητα λέξης (TF) δείχνει το πόσο συχνά εμφανίζεται μία λέξη μέσα στο έγγραφο και προκύπτει από την διαίρεση του αριθμού των εμφανίσεων της λέξης στο έγγραφο προς τον συνολικό αριθμό λέξεων του εγγράφου.

$$TF = \frac{αριθμός\ εμφανίσεων\ της\ λέξης\ στο\ έγγραφο}{συνολικός\ αριθμός\ λέξεων\ στο\ έγγραφο}$$

Η αντίστροφη συχνότητα εγγράφου (IDF) χρησιμοποιείται για να κατανοηθεί η σημασία της λέξης. Βασίζεται στο γεγονός ότι οι λιγότερο συχνά εμφανιζόμενες λέξεις είναι πιο χρήσιμες και σημαντικές. Αυτή η έννοια ορίζεται από τον δεκαδικό αλγόριθμο της διαίρεσης του συνολικού αριθμό των εγγράφων προς τον αριθμό των εγγράφων όπου η λέξη εμφανίζεται.

$$\text{IDF} = log_{10} \frac{\text{συνολικός αριθμός εγγράφων}}{\text{αριθμός των εγγράφων όπου εμφανίζεται η λέξη}}$$

Οι λέξεις που εμφανίζουν μεγαλύτερο IDF είναι πιο σπάνιες και πιο σημαντικές για το νόημα του εγγράφου. Το τελικό TF-IDF προκύπτει από τον πολλαπλασιασμό του TF με το IDF και όσο μεγαλύτερο είναι αυτό το νούμερο για κάθε λέξη, συμπεραίνουμε ότι είναι πιο καθοριστικός ο ρόλος της για τη σημασία του κειμένου.

Κλείνοντας την ενότητα της διανυσματοποίησης αξίζει να σημειωθεί ότι η μέθοδος TF-IDF είναι αποτελεσματικότερη καθώς δίνει βαρύτητα στις πιο σημαντικές λέξεις.

3.2.3 Twitter API

Για την συλλογή δεδομένων από το Twitter συχνά χρησιμοποιείται το API του. Αυτό δίνει τη δυνατότητα σε οποιοδήποτε προγραμματιστή να έχει πρόσβαση στα δεδομένα του. Η χρήση του είναι σχετικά απλή και παρέχεται σε όποιον έχει developer account στο Twitter. Μετά τη δημιουργία ενός τέτοιου λογαριασμού παρέχονται στον χρήστη τα access token, access token secret, api key και api secret key. Με τη βοήθεια της βιβλιοθήκης της Python, Tweepy, και με την χρήση του αντικειμένου Cursor συλλέγονται τα δεδομένα με τις επιθυμητές παραμέτρους.

Ενδεικτικές παράμετροι είναι:

- q: Η λέξη κλειδί ή ένας συνδυασμός λέξεων που απαιτείται να υπάρχει στο tweet που συλλέγεται.
- lang: Η γλώσσα στην οποία είναι γραμμένη το tweet που ανακτάται.

Για το κάθε tweet που ικανοποιεί τις προδιαγραφές των παραμέτρων μπορεί να ανακτηθεί το κείμενο του (text), το όνομα του χρήστη που το δημοσίευσε (username), η τοποθεσία του (location) καθώς και άλλες σημαντικές πληροφορίες.

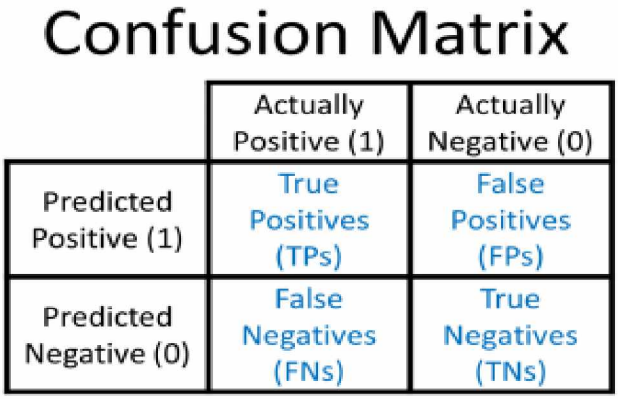
3.3 Κατηγοριοποίηση κειμένου

Η κατηγοριοποίηση κειμένου (text classification) είναι μία τεχνική η οποία κάνει χρήση μοντέλων μηχανικής μάθησης μέσω δεδομένων που έχουν ήδη κατηγοριοποιηθεί, είναι ικανά να ταξινομήσουν άγνωστα δεδομένα. Η συναισθηματική ανάλυση είναι από τα πιο κοινά παραδείγματα κατηγοριοποίησης κειμένου. Τα μοντέλα αυτά εκπαιδεύονται από ένα σύνολο δεδομένων κειμένου τα οποία είναι ταξινομημένα σε «θετικά» και «αρνητικά» και μπορούν να ταξινομήσουν νέα δεδομένα κειμένου που τους είναι εντελώς άγνωστα. Η μέθοδος της κατηγοριοποίησης εφαρμόζεται και σε άλλους τομείς όπως για παράδειγμα στον εντοπισμό ανεπιθύμητων μηνυμάτων κτλ. [18]

3.3.1 Αξιολόγηση απόδοσης κατηγοριοποίησης

Για την αξιολόγηση της απόδοσης κάθε μοντέλου είναι αναγκαίο να γνωρίζουμε τον αριθμό των εγγραφών που ταξινομήθηκαν σωστά και λάθος. Έστω ότι εξετάζουμε το παράδειγμα της συναισθηματικής ανάλυσης και θέλουμε να ταξινομήσουμε το κάθε κείμενο σε «θετικό» και «αρνητικό». Παρατηρώντας τον πίνακα της εικόνας 3.3 φαίνονται τα αποτελέσματα που προέκυψαν μετά την κατηγοριοποίηση και είναι τα εξής [18]:

- True Positives (TP): Οι εγγραφές που ταξινομήθηκαν σωστά ως θετικές.
- False Positives (FP): Οι εγγραφές που ταξινομήθηκαν λάθος ως θετικές.
- False Negatives (FN): Οι εγγραφές που ταξινομήθηκαν λάθος ως αρνητικές.
- True Negatives (TN): Οι εγγραφές που ταξινομήθηκαν σωστά ως αρνητικές.



Εικόνα 3.3 Πίνακας σύγχυσης [19]

Από τον παραπάνω πίνακα σύγχυσης (confusion matrix) υπολογίζονται τα μέτρα απόδοσης των αλγορίθμων [20].

Το πιο συνηθισμένο μέτρο απόδοσης είναι η ακρίβεια (**accuracy**) το οποίο λειτουργεί αποτελεσματικά όταν υπάρχει ίσος αριθμός εγγραφών σε κάθε κλάση. Ορίζεται ως το άθροισμα των σωστών ταξινομήσεων προς το σύνολο όλων των εγγραφών:

$$A = \frac{TP+TN}{TP+TN+FP+FN}$$

Το μέτρο **precision** αντιπροσωπεύει την ικανότητα του ταξινομητή να τοποθετήσει μία εγγραφή στην σωστή κατηγορία σε αντίθεση με όλες τις εγγραφές που βρίσκονται σε αυτή. Το precision ορίζεται ως ο αριθμό των σωστών θετικών ταξινομήσεων προς το άθροισμα των εγγραφών που ταξινομήθηκαν ως σωστές:

$$\pi = \frac{TP}{TP+FP}$$

Υψηλότερο precision σημαίνει λιγότερο ψευδώς θετικές ταξινομήσεις, ενώ χαμηλότερο περισσότερο ψευδώς θετικές.

Το μέτρο απόδοσης **Recall** ορίζεται ως πιθανότητα για το πως μία εγγραφή θα έπρεπε να ταξινομηθεί στην αντίστοιχη κλάση. Ουσιαστικά είναι ο αριθμός σωστών θετικών αποτελεσμάτων προς το άθροισμα όλων των εγγραφών οι οποίες θα έπρεπε να ταξινομηθούν ως θετικές:

$$\rho = \frac{TP}{TP+FN}$$

Τέλος το πιο αξιόπιστο μέτρο απόδοσης είναι το **F1 Score** και είναι ο αρμονικός μέσος ανάμεσα στο precision και στο recall:

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

3.4 Αλγόριθμοι

Σε αυτή την ενότητα θα γίνει μια σύντομη επεξήγηση των αλγορίθμων που πρόκειται να χρησιμοποιηθούν στη συνέχεια για την κατηγοριοποίηση κειμένου.

Multinomial Naïve Bayes

Ο αλγόριθμος Multinomial Naïve Bayes (NMB) είναι γνωστός για την ανίχνευση διάφορων επιθέσεων DDOS. Οι επιθέσεις επιπέδου εφαρμογής DDOS υπερασπίζονται και εξαλείφουν την επίθεση με βάση ένα πολυωνυμικό κατανεμημένο μοντέλο και προτείνεται μια προσέγγιση βάσει ταξινομητή για κακόβουλους επισκέπτες ιστότοπων [24].

Στη συνέχεια γίνεται επεξήγηση πως ο Πολυωνυμικός Naïve Bayes [25] υπολογίζει τις πιθανότητες τάξης για ένα δεδομένο έγγραφο. Το σύνολο των κλάσεων συμβολίζεται με C και με N το μέγεθος του λεξιλογίου. Ο MNB εκχωρεί ένα τεστ

εγγράφου t_i στην κλάση που έχει τη μεγαλύτερη πιθανότητα *Pr(c|ti)*, η οποία δίνεται από τον κανόνα του Bayes:

$$Pr(c|ti) = \frac{Pr(c)\,Pr(ti|c)}{Pr(ti)}, c \in C \ (1)$$

Η προηγούμενη κλάση (prior class) *Pr(c)* μπορεί να εκτιμηθεί διαιρώντας τον αριθμό των εγγράφων που ανήκουν στην κλάση c με το συνολικό αριθμό εγγράφων. *Pr(ti|c)* είναι η πιθανότητα λήψης ενός εγγράφου *ti* στην κλάση c και υπολογίζεται ως εξής:

$$Pr(ti|c) = \left(\sum_n f_{ni}\right)! \prod_n \frac{Pr\,(w_n|c)^{f_{ni}}}{f_{ni}!} \ (2)$$

όπου f_{ni} είναι ο μετρητής της λέξης n στο δοκιμαστικό έγγραφο *ti* και *Pr(wn|c)* η πιθανότητα της λέξης n δεδομένης κλάσης c. Η τελευταία πιθανότητα υπολογίζεται από τα εκπαιδευτικά έγγραφα ως εξής:

$$Pr(wn|c) = \frac{1 + F_{nc}}{N + \sum_{x=1}^{N} F_{xc}} \ (3)$$

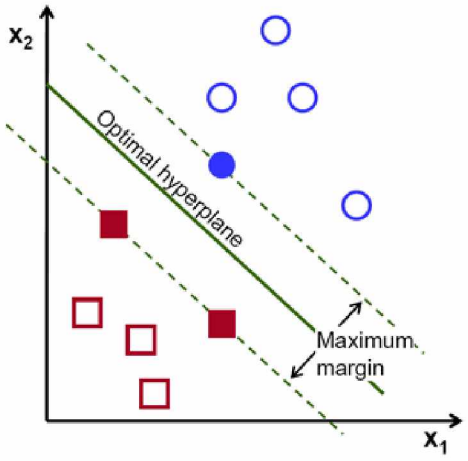
όπου F_{xc} είναι ο αριθμός της λέξης x σε όλα τα έγγραφα εκπαίδευσης που ανήκουν στην τάξη c. Τέλος η πιθανότητα ένα έγγραφο να ανήκει σε μία κλάση, από την εξίσωση (2), μπορεί να απλοποιηθεί, καθώς οι όροι $(\sum_n f_{ni})!$ Και $\prod_n f_{ni}!$ Μπορούν να διαγραφούν χωρίς να αλλάξει το αποτέλεσμα, διότι κανένας από τους δύο δεν εξαρτάται από την κλάση c:

$$Pr(ti|c) = a \prod_n Pr\,(w_n|c)^{f_{ni}},$$

όπου α μία σταθερά η οποία αγνοείται κατά το στάδιο της κανονικοποίησης.

Μηχανές Διανυσματικής Υποστήριξης

Οι μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines – SVM) [26] είναι από τους πιο δημοφιλείς αλγόριθμους της μηχανικής μάθησης. Ο SVM είναι ο πιο κατάλληλος αλγόριθμος να εντοπίσει ένα υπερ-πεδίο (hyperplane) στον N-διάστατο χώρο που ταξινομεί ευδιάκριτα τα σημεία δεδομένων. Υπάρχουν αρκετά



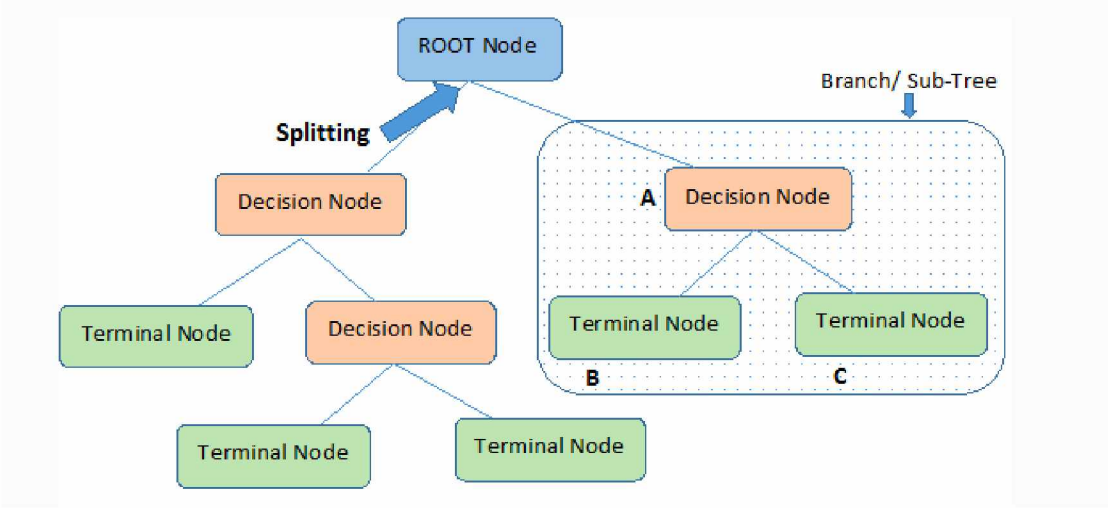
Εικόνα 3.4 Βέλτιστο υπερ-πεδίο SVM [26]

υποψήφια υπερ-πεδία για τον διαχωρισμό των δύο κλάσεων και ο στόχος του συγκεκριμένου αλγορίθμου είναι να εντοπίσει εκείνο το οποίο μεγιστοποιεί το περιθώριο μεταξύ των δύο κλάσεων. Στο διάγραμμα της εικόνας 3.4, φαίνεται η σχηματική αναπαράσταση του διαχωρισμού των δύο κλάσεων από το βέλτιστο υπερ-πεδίο. Η μεγιστοποίηση της απόστασης του περιθωρίου παρέχει κάποια ενίσχυση έτσι ώστε μελλοντικά σημεία δεδομένων να μπορούν να ταξινομηθούν με μεγαλύτερη ακρίβεια.

Δέντρα Απόφασης

Τα Δέντρα Απόφασης (Decision Trees) [27] ανήκουν στην κατηγορία της επιβλεπόμενης μηχανικής μάθησης. Σε αντίθεση με άλλους αλγορίθμους επιβλεπόμενης μηχανικής μάθησης τα δέντρα απόφασης μπορούν να χρησιμοποιηθούν για την επίλυση προβλημάτων κατηγοριοποίησης και παλινδρόμησης. Υπάρχουν δύο τύποι δέντρων απόφασης, τα διακριτών μεταβλητών

(categorical variable) και τα συνεχών μεταβλητών (continuous variable). Στο σχήμα της εικόνας 3.5 παρουσιάζεται ένα στιγμιότυπο του δέντρου του αλγορίθμου.



Εικόνα 3.5 Στιγμιότυπο αλγορίθμου Δέντρων Απόφασης [27]

Τα δέντρα απόφασης αποτελούνται από κόμβους και ακμές.

- Ο κόμβος ρίζα (root node) αντιπροσωπεύει ολόκληρο το δείγμα το οποίο στη συνέχεια χωρίζεται σε δύο ή περισσότερα ομοιογενή σύνολα.
- Οι εσωτερικοί κόμβοι (decision nodes) είναι οι κόμβοι που διαχωρίζονται σε μικρότερα σύνολα.
- Οι κόμβοι φύλλα (leaf/terminal nodes) είναι αυτοί οι οποίοι βρίσκονται πιο βαθιά στο δέντρο και δεν διαχωρίζονται.
- Η υποδιαίρεση ολόκληρου του δέντρου ονομάζεται υπό-δέντρο (sub-tree)
- Ένας κόμβος ο οποίος διαιρείται σε υπό-κόμβους ονομάζεται πατέρας (parent node) και κατά συνέπεια οι υπό-κόμβοι ονομάζονται παιδιά (child node).

Για να γίνει ο διαχωρισμός του συνόλου δεδομένων σε κάθε κόμβο υπάρχουν κάποια μέτρα επιλογής, όπως τα entropy, information gain, gini κ.α. Παρακάτω φαίνονται οι μαθηματικές εκφράσεις υπολογισμού αυτών των μέτρων, όπου c ο αριθμός των κλάσεων και p_i η συχνότητα της κάθε κλάσης.

$$entropy = \sum_{i=1}^{c} - p_i \, log_2 \, p_i$$

$$gini = 1 - \sum_{i=1}^{c} {p_i}^2$$

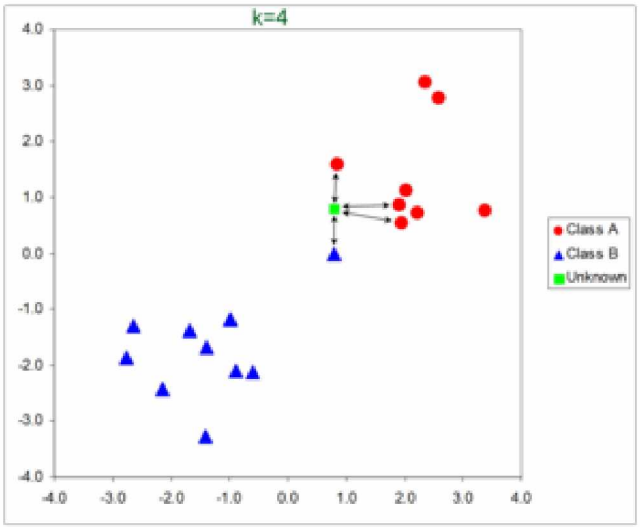
Η επιλογή για τον διαχωρισμό γίνεται με τον υπολογισμό του μεγαλύτερου κέρδους (gain), το οποίο μαθηματικά εκφράζεται ως η διαφορά ενός από τα δυο προηγούμενα μέτρα πριν και μετά από τον διαχωρισμό.

Για παράδειγμα:

$$gain = entropy(before) - entropy(after)$$

Κ Κοντινότεροι Γείτονες

Ο αλγόριθμος των Κ Κοντινότερων Γειτόνων (K Nearest Neighbors) [28] είναι από τους πιο απλούς αλγόριθμους ταξινόμησης και είναι βασισμένος σε συναρτήσεις αποστάσεων για την ταξινόμηση ενός νέου στοιχείου σε κάποια κλάση σύμφωνα με τους πιο κοντινούς γείτονες, όπως φαίνεται και στην εικόνα 3.6.



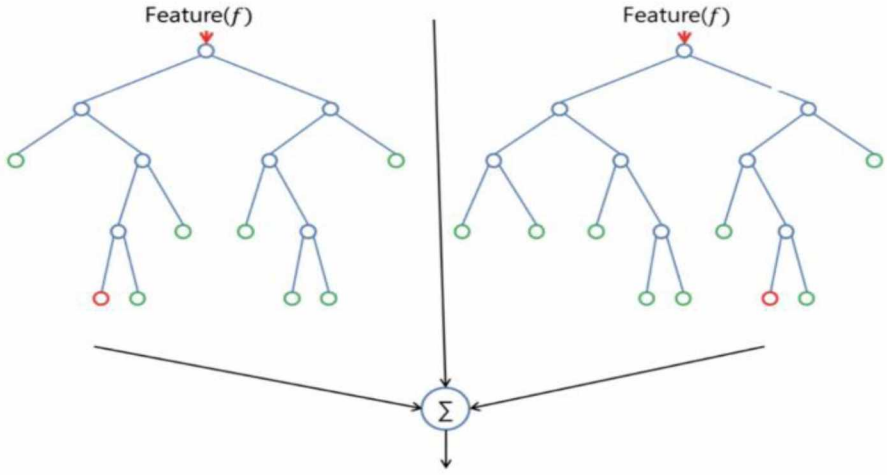
Εικόνα 3.6 Ταξινόμηση στοιχείου με τον KNN [28]

Ο συγκεκριμένος αλγόριθμος είναι κυρίως βασισμένος στην Ευκλείδεια απόσταση, μία από τις πιο δημοφιλείς συναρτήσεις απόστασης, και μαθηματικά εκφράζεται ως εξής:

$$Euclidian\ distance = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Τυχαίο Δάσος

Ο αλγόριθμος του Τυχαίου Δάσους (Random Forest) [29] είναι ένας αλγόριθμος επιβλεπόμενης μηχανικής μάθησης και χρησιμοποιείται, τόσο για την επίλυση προβλημάτων κατηγοριοποίησης όσο και παλινδρόμησης. Το κύριο χαρακτηριστικό του τυχαίου δάσους είναι ότι δεν αποτελεί έναν νέο ταξινομητή, αλλά ένα σύνολο από υπάρχοντες ταξινομητές (ensemble). Πιο απλά, ο αλγόριθμος χτίζει πολλαπλά δέντρα απόφασης (decision trees) και τα συγχωνεύει μεταξύ τους για να πετύχει πιο ακριβείς και σταθερές προβλέψεις. Για την εκπαίδευση και το κτίσιμο ενός δέντρου απόφασης επιλέγεται τυχαία ένα τμήμα από το σύνολο δεδομένων το οποίο είναι ανεξάρτητο από τα υπόλοιπα. Η τελική πρόβλεψη γίνεται για κάθε δέντρο βάσει της πλειοψηφίας των προβλέψεων των δέντρων είτε βάσει του μέσου όρου των προβλέψεων των δέντρων [30]. Στο σχήμα της εικόνας 3.7 παρουσιάζεται ένα στιγμιότυπο κτισίματος του τυχαίου δάσους.



Εικόνα 3.7 Στιγμιότυπο κτισίματος τυχαίου δάσους [29]

Adaboost

Ο αλγόριθμος Adaboost, που αποτελείται από ήδη υπάρχοντες ταξινομητές (ensemble), συνδυάζει πολλούς από αυτούς με σκοπό την αύξηση της ακρίβειας τους. Είναι ένας επαναληπτικός αλγόριθμος ο οποίος ξεκινάει την εκπαίδευση του με ίσα βάρη, επιλέγοντας τυχαία τμήμα του συνόλου δεδομένων. Τα βάρη αυτά τροποποιούνται σε κάθε επανάληψη και επιλέγονται εκείνα που αυξάνονται, έτσι ώστε τελικά να αυξηθεί η ακρίβεια της πρόβλεψης στο σύνολο δεδομένων.

ΚΕΦΑΛΑΙΟ 4

Πρόβλεψη εκλογικού αποτελέσματος με συναισθηματική ανάλυση

Εισαγωγή

Το συγκεκριμένο κεφάλαιο έχει ως σκοπό την ανάλυση του συναισθήματος από ένα σύνολο tweets τα οποία συλλέχθηκαν μια εβδομάδα πριν τις προεδρικές εκλογές των ΗΠΑ. Στη συνέχεια εστιάζει στην δημοτικότητα καθενός υποψηφίου (Ντόναλντ Τραμπ και Τζο Μπάιντεν), τόσο στην επικράτεια όσο και σε επίπεδο Πολιτείας. Τέλος γίνεται εκτίμηση του αποτελέσματος των εκλογών κάνοντας χρήση των δεδομένων που προέκυψαν από την συναισθηματική ανάλυση.

Η προεπεξεργασία κειμένου, η συναισθηματική ανάλυση καθώς και η φόρμουλα που αναπτύχθηκε από τους Wicaksono et al.[16] τα οποία περιγράφονται αναλυτικά στις παρακάτω ενότητες, είναι τα πρώτα βήματα που ακολουθούνται για τη εκτίμηση του αποτελέσματος.

4.1 Τα δεδομένα

Τα δεδομένα που παρουσιάζονται συλλέχθηκαν μία εβδομάδα πριν από την διεξαγωγή των Προεδρικών εκλογών των ΗΠΑ και χρησιμοποιήθηκαν σε μία παρόμοια έρευνα κατά την περίοδο αυτή [15]. Όπως αναφέρεται και στο σχετικό άρθρο τα δεδομένα συλλέχθηκαν με το Twitter API και στόχος της συγκεκριμένης έρευνας ήταν να παρουσιαστεί η διάθεση και η άποψη των πολιτών για τους δύο υποψήφιους εν όψει των εκλογών.

Στη συνέχεια γίνεται μια σύντομη περιγραφή των δεδομένων τα οποία πρόκειται να αναλυθούν στην συγκεκριμένη εργασία.

4.1.1 Περιγραφή δεδομένων

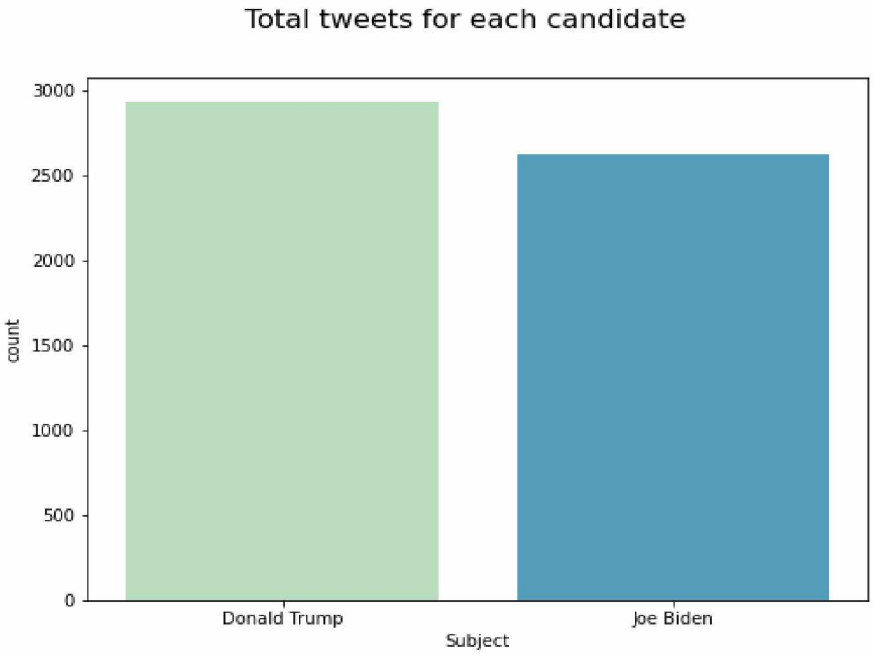
Στον παρακάτω πίνακα της εικόνας 4.1 φαίνεται το σύνολο των δεδομένων (Dataset) το οποίο χρησιμοποιείται στην εργασία για την συναισθηματική ανάλυση

	geo	text	user	location	Subject	state
0	NaN	After the confrontation that cost Walter Wallace Jr. his life set off prote...	wheresestela1	New Jersey	Donald Trump	New Jersey
1	NaN	"COVID-19 is still raging. Trump is still lying."	TheFemaleYungin	The Wrong Address, Texas	Donald Trump	Texas
2	NaN	Donald Trump Jr.: "Why aren't they talking about deaths? Oh, oh, because th...	KathrynTomashu1	Northborough, MA	Donald Trump	Massachusetts
3	NaN	Let's not forget who created these problems. Tell Donald Trump his scheme to...	ajserino	New Jersey, USA	Donald Trump	New Jersey
4	NaN	There is no greater supporter of Donald Trump in the world than Mark Zucker...	SpeakBravely	Twin Cities, Minnesota	Donald Trump	Minnesota

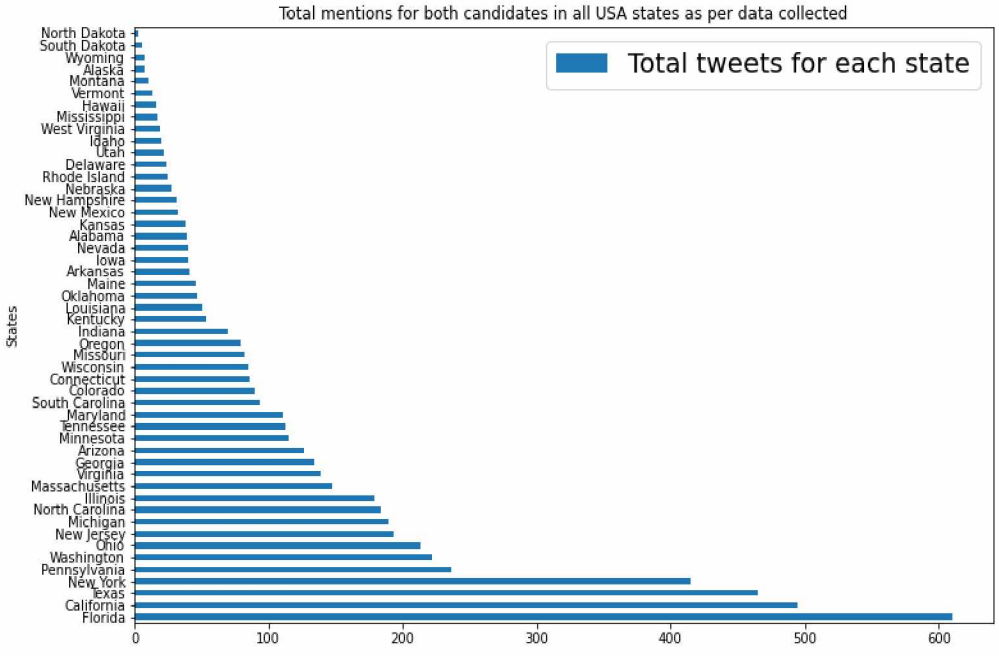
Εικόνα 4.1 Σύνολο δεδομένων

και την πρόβλεψη του εκλογικού αποτελέσματος των προεδρικών εκλογών των ΗΠΑ μέσω αυτής. Η στήλη (geo), η οποία είναι ο γεωγραφικός κωδικός (geocode) του χρήστη που δημιούργησε το κάθε tweet, δεν συμμετέχει στην συνολική διαδικασία της ανάλυσης εφόσον τα δεδομένα της δεν είναι διαθέσιμα. Η δεύτερη στήλη (text) αποτελείται από το αρχικό κείμενο του tweet όπως αυτό συντάχθηκε από τον χρήστη. Όπως παρατηρείται τα κείμενα περιέχουν πολλή, μη χρήσιμη πληροφορία η οποία θα αφαιρεθεί στην συνέχεια κατά την προεπεξεργασία του κειμένου. Η τρίτη στήλη (user) υποδηλώνει το ψευδώνυμο του κάθε χρήστη. Στην τέταρτη (location) και στην τελευταία (state) στήλη υπάρχει η πληροφορία για την τοποθεσία στην οποία συντάχθηκε το tweet. Τέλος, η στήλη Subject δείχνει σε ποιόν από τους δύο υποψήφιους αναφέρεται το tweet.

Συνολικά και για τους δύο υποψήφιους (Ντόναλντ Τραμπ και Τζο Μπάιντεν) έχουν συλλεχθεί 5553 tweets τα οποία, όπως φαίνεται και στο σχήμα της εικόνας 4.2 είναι, 2928 για τον πρώτο και 2625 για τον δεύτερο. Στην εικόνα 4.3 φαίνονται ποσοτικά πώς ταξινομούνται τα σχόλια σε κάθε πολιτεία. Είναι εμφανές ότι μεγαλύτερες σε έκταση πολιτείες έχουν αυξημένο αριθμό tweets, προφανώς λόγω του μεγαλύτερου πληθυσμού.

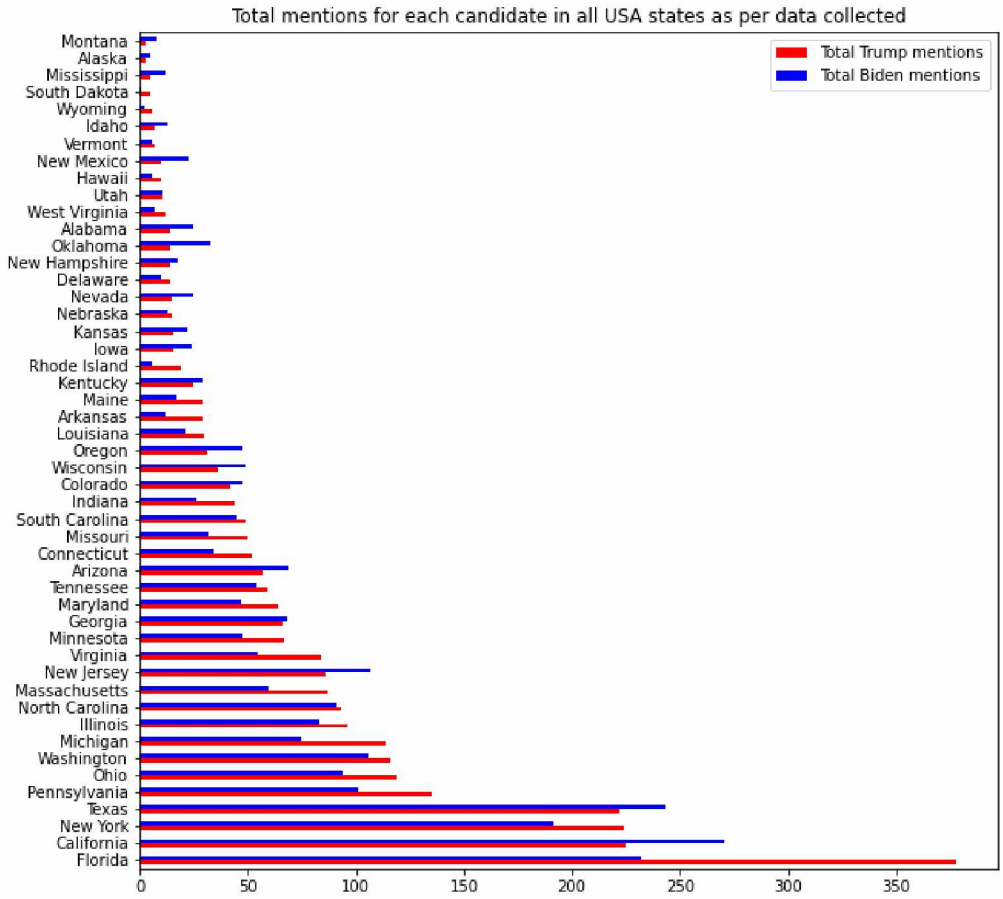


Εικόνα 4.2 Συνολικά tweets για κάθε υποψήφιο



Εικόνα 4.3 Συνολικά tweets ανά Πολιτεία

Πριν γίνει η συναισθηματική ανάλυση των δεδομένων, τόσο σε συνολικό επίπεδο όσο και σε επίπεδο Πολιτείας, αξίζει να επισημανθεί ότι αυτό το dataset μπορεί να δώσει έγκυρα αποτελέσματα, διότι ο αριθμός των tweet για κάθε ένα από τους δύο υποψήφιους είναι σχεδόν ο ίδιος. Όπως φαίνεται και από το διάγραμμα της εικόνας 4.4, οι αναφορές που γίνονται τόσο για τον Ντόναλντ Τραμπ όσο και για τον Τζο Μπάιντεν δεν έχουν μεγάλη απόκλιση μεταξύ τους.



Εικόνα 4.4 Συνολικά tweets για τους δύο υποψήφιους σε κάθε πολιτεία

Ενδιαφέρον παρουσιάζει η Φλόριντα, μία από τις μεγαλύτερες Πολιτείες των ΗΠΑ, που υπερβαίνει κατά πολύ ο αριθμός των tweets τα οποία είναι σχετικά με τον πρόεδρο των Ρεπουμπλικάνων. Σε αυτή την περίπτωση, σημαντικό ρόλο στην εξαγωγή του συμπεράσματος θα παίξει και η ανάλυση του συναισθήματος, όπως θα φανεί στη συνέχεια.

4.2 Προεπεξεργασία Κειμένου

Αποτέλεσμα της προεπεξεργασίας κειμένου είναι να «καθαριστεί» το κείμενο αφαιρώντας λέξεις οι οποίες δεν προσφέρουν κάποια χρήσιμη πληροφορία στο νόημα του . Αυτό θα έχει ως αποτέλεσμα η διαδικασία ανάλυσης του συναισθήματος να είναι γρηγορότερη και αποτελεσματικότερη. Η διαδικασία που ακολουθήθηκε σε αυτήν την εργασία είναι η εξής:

- Αφαίρεση από το κείμενο ασήμαντων για την επεξεργασία πληροφοριών όπως σύνδεσμοι, αναφορές στο όνομα του χρήστη (username) και σύμβολα hastag, κάνοντας χρήση κανονικών εκφράσεων (regular expressions).
- Μετατροπή όλων των χαρακτήρων σε πεζούς.
- Διαγραφή σημείων στίξης και αφαίρεση όλων των emoji.
- Τέλος ακολουθήθηκε μία σειρά βημάτων επεξεργασίας της Φυσικής Γλώσσας με σκοπό την μείωση του μεγέθους των tweets. Αυτή η διαδικασία περιλαμβάνει τον διαχωρισμό του κειμένου σε λεκτικές μονάδες (tokenization), λημματοποίηση (lemmatization), αποκοπή καταλήξεων (stemming) και αφαίρεση των stop-words.

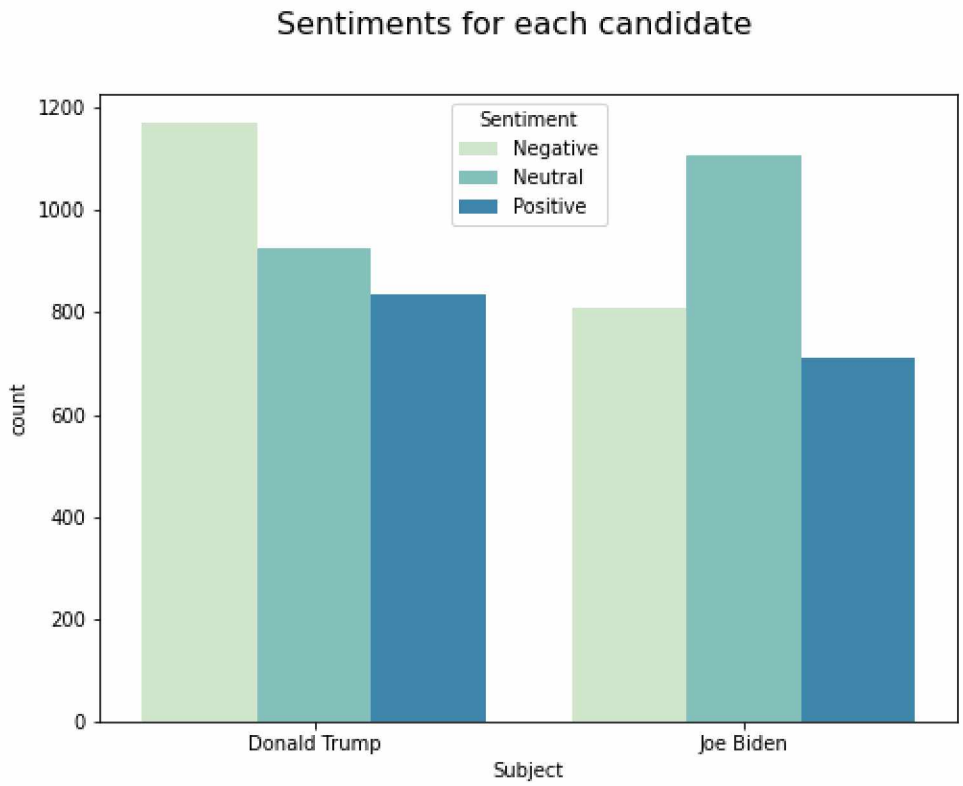
4.3 Πρόβλεψη συνολικού εκλογικού αποτελέσματος και πρόβλεψη ανά πολιτεία

4.3.1 Συναισθηματική ανάλυση

Μετά την προεπεξεργασία του κειμένου, σειρά έχει η εφαρμογή αλγορίθμων για την συναισθηματική ανάλυση των tweets. Για το σκοπό αυτό εφαρμόστηκε η συναισθηματική ανάλυση κειμένου VADER της βιβλιοθήκης NLTK (Natural Language Toolkit) της Python. Το VADER (Valence Aware Dictionary and sentiment Reasoner) είναι ένα λεξικό που χρησιμοποιείται κυρίως για την έκφραση των συναισθημάτων στα μέσα κοινωνικής δικτύωσης. Το συγκεκριμένο λεξικό κατηγοριοποιεί το συναίσθημα σε θετικό, αρνητικό και ουδέτερο, δίνοντας ένα σκορ για το κάθε ένα,

καθώς και ένα κανονικοποιημένο σκορ (compound score) των παραπάνω τριών. Στην ανάλυση αυτή, θετικό θεωρείται το κείμενο με κανονικοποιημένο σκορ μεγαλύτερο ή ίσο από το 0.05, αρνητικό με μικρότερο ή ίσο από το -0.05 και ουδέτερο ανάμεσα στο -0.05 και το 0.05.

Στο σχήμα της εικόνας 4.5 φαίνεται πως έχουν κατανεμηθεί τα συναισθήματα για κάθε έναν από τους δύο υποψήφιους. Παρατηρείται ότι το αρνητικό (negative)

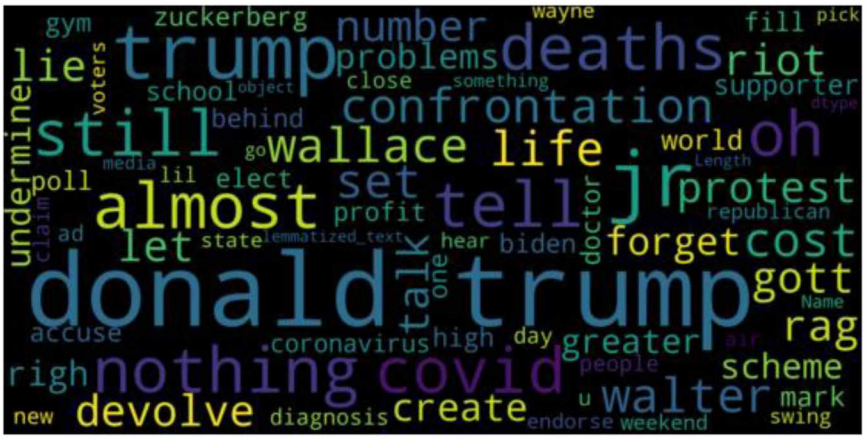


Εικόνα 4.5 Ανάλυση συναισθήματος για τους υποψήφιους

συναίσθημα υπερισχύει για τον Ντόναλντ Τραμπ ενώ, το ουδέτερο (neutral) συναίσθημα για τον Τζο Μπάιντεν. Από την ανάλυση αυτή, προέκυψε ότι το 21.05% του συνόλου των tweets είναι αρνητικά σχετικά με τον Τραμπ, ενώ τα αρνητικά που αναφέρονται στον Μπάιντεν είναι το 14.55%. Ακόμη, προκύπτει ότι το ποσοστό των ουδέτερων σχολίων για τον Μπάιντεν ανέρχεται στο 19.9%, αισθητά υψηλότερο από αυτό του Ντόναλντ Τραμπ, το οποίο είναι στο 16.6%. Υποθετικά, μπορούμε να εξηγήσουμε ότι αυτό συμβαίνει διότι ο Τραμπ ήταν ο τελευταίος πρόεδρος της χώρας και, επομένως, ο κόσμος έχει πιο ολοκληρωμένη άποψη για αυτόν, ενώ ο Τζο

Μπάιντεν δεν έχει εκτελέσει πρόεδρος των ΗΠΑ στο παρελθόν και κατά συνέπεια ο κόσμος κρατάει πιο ουδέτερη στάση απέναντι του.

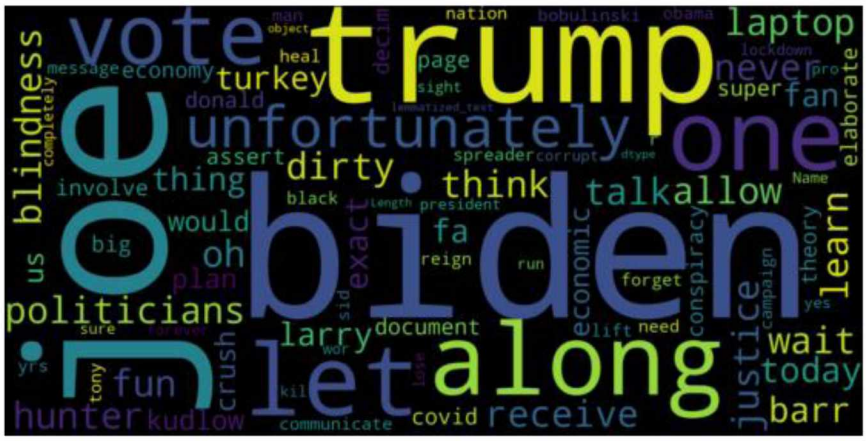
Στην εικόνα 4.6 και στην εικόνα 4.7 παρουσιάζονται τα WordClouds με τις πιο συχνά εμφανιζόμενες λέξεις που αφορούν κάθε έναν από τους δύο υποψήφιους. Παρατηρώντας τις λέξεις που αναφέρονται στον εν ενεργεία, τότε, πρόεδρο των ΗΠΑ, Ντόναλντ Τραμπ, αναμενόμενο είναι να εμφανίζεται πέρα από το όνομά του,



Εικόνα 4.6 Οι πιο συχνά εμφανιζόμενες λέξεις για τον Donald Trump

ενδιαφέρον παρουσιάζουν οι λέξεις «deaths» και «covid», καθώς εκείνη την περίοδο η πανδημία της Covid-19 ήταν ένα από τα πιο φλέγοντα ζητήματα τόσο για τις Ηνωμένες Πολιτείες όσο και για τον υπόλοιπο κόσμο. Επίσης, οι λέξεις «protest (= διαμαρτυρία)» και «confrontation (=αντιπαράθεση)» φαίνονται να είναι αρκετά συχνές στα σχόλια των χρηστών του Twitter, γεγονός που μπορεί να ερμηνευτεί πως ο λαός διαφωνεί με τον πρόεδρο.

Όσον αφορά τις πιο συχνές λέξεις που εμφανίζονται για τον Joe Biden, τον νούμερο ένα αντίπαλο του τότε προέδρου, φαίνεται ξεκάθαρα πως και εδώ πολύ συχνά αναφέρεται το όνομα του προέδρου Ντόναλντ Τραμπ. Μία προφανής υπόθεση που μπορεί να γίνει σχετικά με αυτό είναι ότι οι χρήστες είναι πολύ πιθανό να συγκρίνουν σε μεγάλο βαθμό τον Donald Trump με τον Τζο Μπάιντεν. Αξιοσημείωτο είναι ότι, στο συγκεκριμένο σύνολο λέξεων που αφορά τον Μπάιντεν φαίνονται να



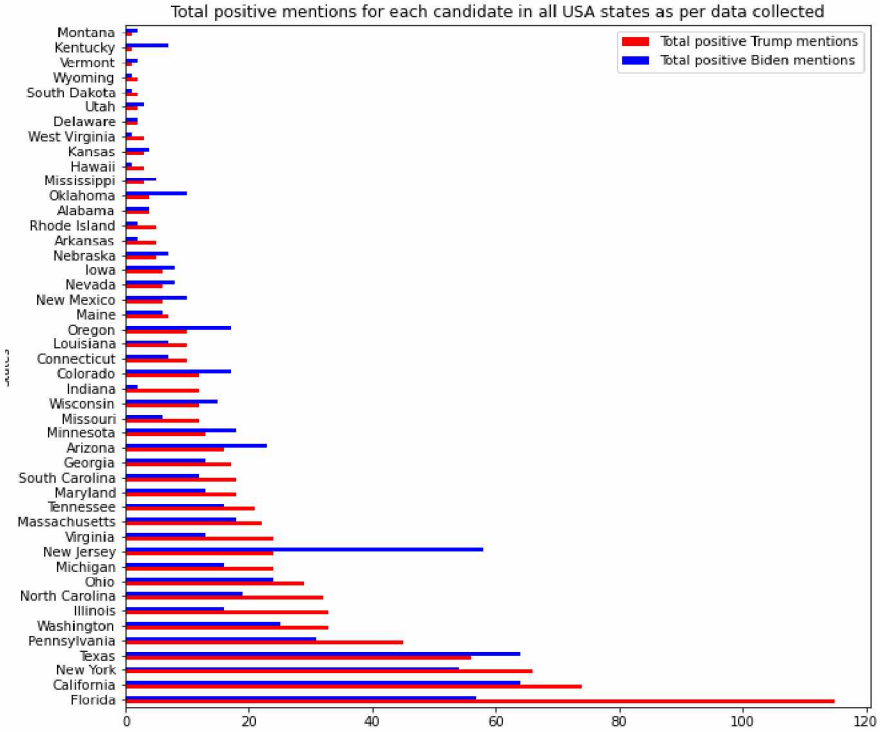
Εικόνα 4.7 Οι πιο συχνά εμφανιζόμενες λέξεις για τον Joe Biden

επικρατούν περισσότερο θετικές λέξεις σε σχέση με το προηγούμενο. Χαρακτηριστική είναι η λέξη «vote» που υποδηλώνει την πρόθεση κάποιου να θέλει να ψηφίσει υπέρ του ή ακόμα και να παροτρύνει άλλους ανθρώπους να τον ψηφίσουν.

4.3.2 Ανάλυση συναισθήματος ανά πολιτεία

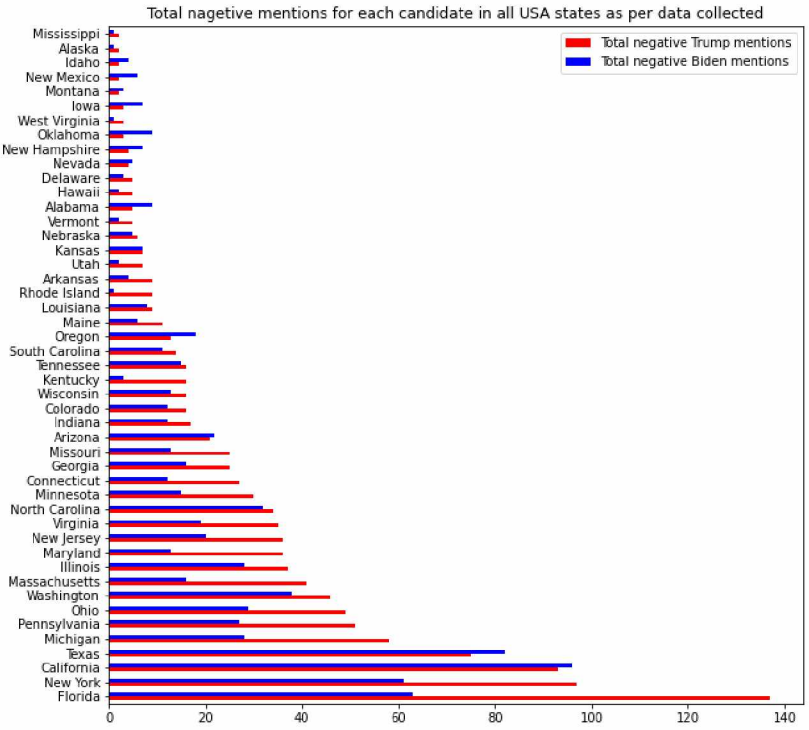
Σημαντική συνεισφορά σε αυτή την έρευνα προσφέρει η ανάλυση της διακύμανσης των συναισθημάτων ανά πολιτεία, εφόσον το συνολικό εκλογικό αποτέλεσμα καθορίζεται από το αποτέλεσμα κάθε πολιτείας ξεχωριστά. Μετά τη συνολική εικόνα της συναισθηματικής ανάλυσης που προηγήθηκε, τώρα θα εστιάσουμε την ανάλυση σε επίπεδο πολιτείας. Στην γραφική παράσταση της εικόνας 4.8, στον κατακόρυφο άξονα φαίνονται όλες οι πολιτείες και στο οριζόντιο το σύνολο των θετικών tweets για κάθε έναν από τους δύο υποψήφιους. Παρατηρώντας το διάγραμμα από κάτω προς τα πάνω διαπιστώνουμε ότι στα πιο χαμηλά επίπεδα βρίσκονται οι μεγαλύτερες πολιτείες σε πληθυσμό και όσο ανεβαίνουμε καταλήγουμε σε μικρότερες. Το συμπέρασμα που μπορεί να βγει από αυτό το διάγραμμα είναι ότι στις μεγαλύτερες πολιτείες ο Trump έχει αισθητά μεγαλύτερο αριθμό θετικών σχολίων ένω όσο κινούμαστε κατακόρυφα προς τα πάνω και σε μικρότερες πολιτείες αυτή η διαφορά ελαχιστοποιείται και ο Biden φαίνεται να έχει περισσότερα θετικά. Ενδιαφέρον εντοπίζεται στην Florida όπου ο αρχηγός των

Ρεπουμπλικάνων έχει αναμφισβήτητα μεγαλύτερη διαφορά υπέρ του σε σχέση με τον αρχηγό των Δημοκρατικών. Το αντίθετο ακριβώς συμβαίνει στο New Jersey. Επομένως, μπορούμε σε αυτές τις δύο πολιτείες να βγάλουμε ένα αρχικό



Εικόνα 4.8 Σύνολο θετικών συναισθημάτων για κάθε υποψήφιο σε όλες τις πολιτείες

συμπέρασμα για τον νικητή. Παρόλα αυτά, μεγάλη σημασία σε αυτήν την απόφαση παίζουν και τα αρνητικά σχόλια για τον κάθε έναν τα οποία μπορεί να ανατρέψουν το αποτέλεσμα. Στο διάγραμμα της εικόνας 4.9 φαίνονται τα αντίστοιχα αρνητικά tweets ανά πολιτεία. Ξεκάθαρα μπορούμε να δούμε ότι στις περισσότερες πολιτείες τα αρνητικά σχόλια για τον Trump είναι περισσότερα από αυτά του αντιπάλου του.



Εικόνα 4.9 Σύνολο αρνητικών συναισθημάτων για κάθε υποψήφιο σε όλες τις πολιτείες

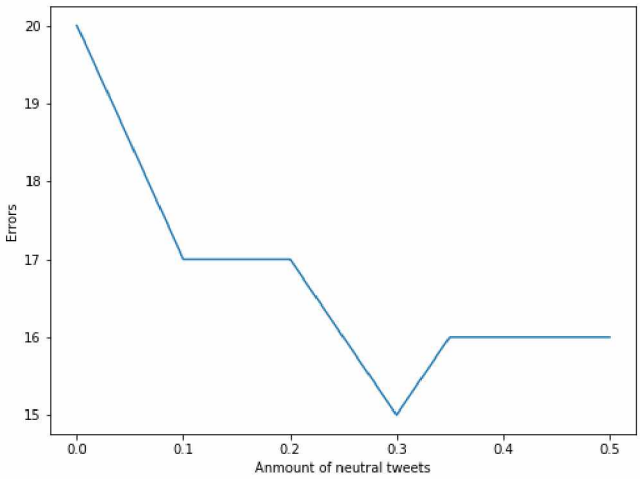
Συνοψίζοντας, αφού ολοκληρώθηκε η συναισθηματική ανάλυση και έχει σχηματιστεί μία εικόνα για την γνώμη των χρηστών του Twitter, όλα αυτά τα στοιχεία θα συμπεριληφθούν στην εκτίμηση του εκλογικού αποτελέσματος τόσο σε επίπεδο χώρας όσο και για κάθε πολιτεία ξεχωριστά.

4.3.3 Πρόβλεψη εκλογικού αποτελέσματος

Για την πρόβλεψη του εκλογικού αποτελέσματος θα βασιστούμε στη φόρμουλα που αναπτύχθηκε από τους Wicaksono et al.[15] Στο άρθρο τους αναφέρουν ότι το κάθε tweet αναπαριστά μία ψήφο. Ένα θετικό tweet προσμετράται σαν μια ψήφος προς το συγκεκριμένο σχήμα, ενώ ένα αρνητικό ως μία ψήφος προς το αντίπαλο. Επομένως, με βάση αυτή τη θεωρία καταλήξαν στον μαθηματικό τύπο που δίνει το ποσοστό νίκης του κάθε κόμματος:

$$\text{Ποσοστό νίκης(Α)} = \frac{\theta\varepsilon\tau\iota\kappa\acute{\alpha}\ tweets(A)+\alpha\rho\nu\eta\tau\iota\kappa\acute{\alpha}\ tweets(B)}{\sigma\upsilon\nu o\lambda\iota\kappa\acute{\alpha}\ tweets(A+B)}$$

Παρατηρώντας τα δεδομένα μας διαπιστώνουμε ότι ένας σημαντικός αριθμός των tweets είναι ουδέτερα. Μετά από αρκετές πειραματικές δοκιμές καταλήξαμε στο συμπέρασμα ότι καλύτερες προβλέψεις προκύπτουν όταν ενσωματώσουμε το 30% των ουδέτερων σχολίων του κάθε υποψήφιου στις ψήφους υποστήριξής του. Η επιλογή του συγκεκριμένου ποσοστού εγίνε ύστερα από τη σύγκριση των σωστών προβλέψεων του αποτελέσματος σε κάθε Πολιτεία. Παρατηρώντας το διάγραμμα της εικόνας 4.10, για το ποσοστό 30% είχαμε 15 λάθος προβλέψεις, ενώ κάνοντας χρήση του μαθηματικού τύπου, όπως αυτός αναφέρεται στην έρευνα, είχαμε 20 λάθος προβλέψεις.

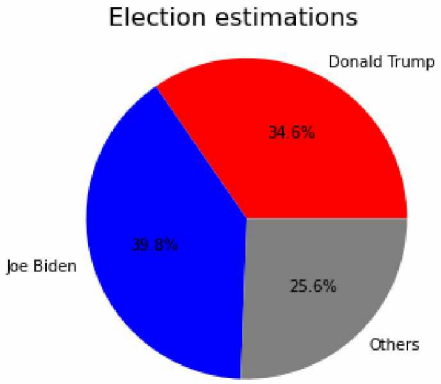


Εικόνα 4.10 Ενσωμάτωση ουδέτερων tweets

Έτσι καταλήξαμε στον ακόλουιο τύπο για την πρόβλεψη του αποτελέσματος:

$$\text{Ποσοστό νίκης(Α)} = \frac{\theta\varepsilon\tau\iota\kappa\acute{\alpha}\ tweets(A)+\alpha\rho\nu\eta\tau\iota\kappa\acute{\alpha}\ tweets(B)}{\sigma\upsilon\nu o\lambda\iota\kappa\acute{\alpha}\ tweets(A+B)} + 0.3*\text{ουδέτερα tweets(Α)}$$

Εφαρμόζοντας τον παραπάνω μαθηματικό τύπο και για τους δύο υποψήφιους λαμβάνουμε την εκτίμηση του εκλογικού αποτελέσματος, όπως φαίνεται στο σχήμα της εικόνας 4.11. Σύμφωνα με αυτό παρατηρείται ότι προηγείται ο Τζο Μπάιντεν με



Εικόνα 4.11 Εκτίμηση εκλογικού αποτελέσματος

39.8%, ένω ο Ντόναλντ Τραμπ μένει πίσω με 34.6%. Η διαφορά τους κυμαίνεται περίπου στις 5 μονάδες. Τα πραγματικά αποτελέσματα με το πέρας των εκλογών διαμορφώθηκαν ως εξής: 51.4% για τον πρώτο και 46.9% για τον δεύτερο. Κι εδώ βλέπουμε μια διαφορά περίπου στο 5%. Παρόλο δηλαδή που τα ποσοστά της εκτίμησής μας είναι αρκετά διαφορετικά, η μεταξύ τους διαφορά εκτιμήθηκε σωστά. Μια αιτία της σημαντικής διαφοράς στην πρόβλεψη μπορεί να αποδοθεί στο 25.6% των ουδέτερων tweets που υποδεικνύουν τους αναποφάσιστουν χρήστες.

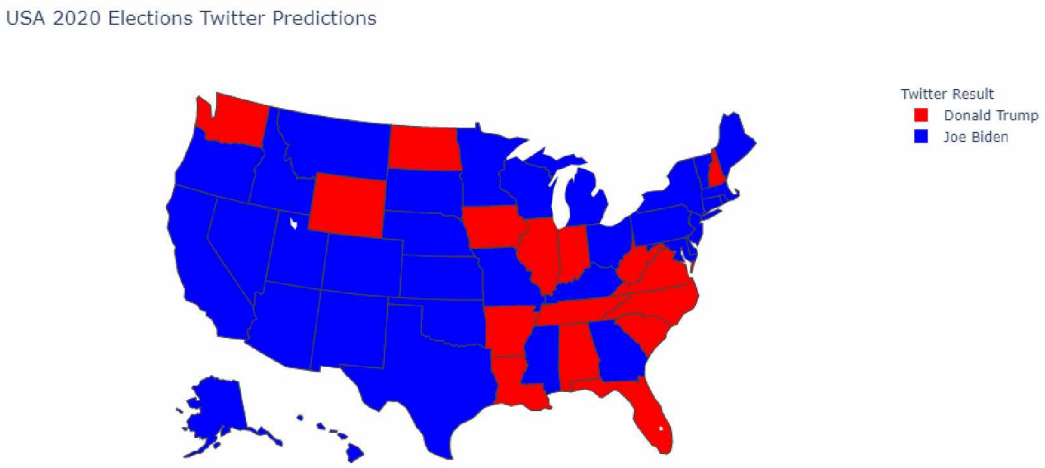
Στον πίνακα της εικόνας 4.12 φαίνονται τα δεδομένα που προέκυψαν για τις πενήντα πολιτείες των ΗΠΑ μετά την συναισθηματική ανάλυση. Τα πραγματικά αποτελέσματα μετά το πέρας των εκλογών παρουσιάζονται στην στήλη «Real Result», ενώ οι προβλέψεις από την εφαρμογή της μαθηματικής φόρμουλας που είδαμε παραπάνω φαίνονται στην στήλη «Twitter Result». Η ανάλυση μας φτάνει να προβλέψει σωστά το αποτέλεσμα για 33 από τις 50 πολιτείες, το οποίο είναι ένα καλό ποσοστό επειδή για αρκετές από αυτές τις πολιτείες, όπως φαίνεται και στον πίνακα, τα δεδομένα είναι πολύ περιρισμένα καθιστώντας την ανάλυση λιγότερο αποτελεσματική. Παρόλα αυτά μπορούμε να θεωρήσουμε ότι τα αποτελέσματα μας είναι αρκετά ικανοποιητικά, καθώς και σε έρευνα που πραγματοποιήθηκε από τους Tim Hamling, Ankur Agrawal [17] σε κολέγιο της Νέας Υόρκης για τις αντίστοιχες εκλογές του 2016 παρατηρούμε ότι σημειώνεται το ίδιο ποσοστό σωστών προβλέψεων.

	state	Trump Positive	Trump Neutral	Trump Negative	Biden Positive	Biden Neutral	Biden Negative	% Result (Trump)	% Result (Biden)	Twitter Result	Real Result
0	Alabama	4	5	5	4	12	9	30.256	28.462	Donald Trump	Donald Trump
1	Alaska	0	1	2	0	4	1	12.500	32.500	Joe Biden	Donald Trump
2	Arizona	16	20	21	23	24	22	29.683	35.635	Joe Biden	Joe Biden
3	Arkansas	5	15	9	2	6	4	30.000	24.634	Donald Trump	Donald Trump
4	California	74	58	93	64	110	96	32.040	32.747	Joe Biden	Joe Biden
5	Colorado	12	14	16	17	19	12	27.333	37.667	Joe Biden	Joe Biden
6	Connecticut	10	15	27	7	15	12	26.628	35.349	Joe Biden	Joe Biden
7	Delaware	2	7	5	2	5	3	25.833	29.167	Joe Biden	Joe Biden
8	Florida	115	126	137	57	112	63	32.279	30.574	Donald Trump	Donald Trump
9	Georgia	17	24	25	13	39	16	26.418	31.493	Joe Biden	Joe Biden
10	Hawaii	3	2	5	1	3	2	31.250	33.750	Joe Biden	Joe Biden
11	Idaho	0	5	2	1	8	4	21.500	24.000	Joe Biden	Donald Trump
12	Illinois	33	26	37	16	39	28	33.743	29.944	Donald Trump	Joe Biden
13	Indiana	12	15	17	2	12	12	35.571	25.000	Donald Trump	Donald Trump
14	Iowa	6	7	3	8	9	7	32.500	32.000	Donald Trump	Donald Trump
15	Kansas	3	6	7	4	11	7	25.526	32.105	Joe Biden	Donald Trump
16	Kentucky	1	8	16	7	19	3	10.185	44.259	Joe Biden	Donald Trump
17	Louisiana	10	11	9	7	6	8	37.059	29.608	Donald Trump	Donald Trump
18	Maine	7	11	11	6	5	6	31.522	33.043	Joe Biden	Joe Biden
19	Maryland	18	10	36	13	21	13	27.117	40.090	Joe Biden	Joe Biden
20	Massachusetts	22	24	41	18	26	16	27.483	37.075	Joe Biden	Joe Biden
21	Michigan	24	32	58	16	31	28	28.148	34.868	Joe Biden	Joe Biden
22	Minnesota	13	24	30	18	15	15	26.696	37.826	Joe Biden	Joe Biden
23	Mississippi	3	0	2	5	6	1	21.765	48.235	Joe Biden	Donald Trump
24	Missouri	12	13	25	6	13	13	30.488	33.415	Joe Biden	Donald Trump
25	Montana	1	0	2	2	3	3	28.182	39.091	Joe Biden	Donald Trump
26	Nebraska	5	4	6	7	1	5	34.643	41.071	Joe Biden	Donald Trump
27	Nevada	6	5	4	8	12	5	27.500	36.000	Joe Biden	Joe Biden
28	New Hampshire	4	6	4	0	11	7	33.438	19.062	Donald Trump	Joe Biden
29	New Jersey	24	26	36	58	29	20	23.731	47.617	Joe Biden	Joe Biden
30	New Mexico	6	2	2	10	7	6	32.727	40.909	Joe Biden	Joe Biden
31	New York	66	61	97	54	76	61	30.602	34.867	Joe Biden	Joe Biden
32	North Carolina	32	27	34	19	40	32	33.967	29.783	Donald Trump	Donald Trump
33	North Dakota	2	0	1	0	0	0	66.667	23.333	Donald Trump	Donald Trump
34	Ohio	29	41	49	24	41	29	28.920	33.146	Joe Biden	Donald Trump
35	Oklahoma	4	7	3	10	14	9	26.383	34.681	Joe Biden	Donald Trump
36	Oregon	10	8	13	17	13	18	31.646	37.975	Joe Biden	Joe Biden
37	Pennsylvania	45	39	51	31	43	27	32.034	33.729	Joe Biden	Joe Biden
38	Rhode Island	5	5	9	2	3	1	28.800	36.800	Joe Biden	Joe Biden
39	South Carolina	18	17	14	12	22	11	32.766	30.213	Donald Trump	Donald Trump
40	South Dakota	2	1	2	1	0	0	38.333	40.000	Joe Biden	Donald Trump
41	Tennessee	21	22	16	16	23	15	33.717	30.177	Donald Trump	Donald Trump
42	Texas	56	91	75	64	97	82	30.258	31.312	Joe Biden	Donald Trump
43	Utah	2	2	7	3	6	2	18.182	44.091	Joe Biden	Donald Trump
44	Vermont	1	1	5	2	2	2	20.769	46.923	Joe Biden	Joe Biden
45	Virginia	24	25	35	13	23	19	32.230	31.942	Donald Trump	Joe Biden
46	Washington	33	37	46	25	43	38	31.847	31.577	Donald Trump	Joe Biden
47	West Virginia	3	6	3	1	5	1	28.947	24.211	Donald Trump	Donald Trump
48	Wisconsin	12	8	16	15	21	13	27.647	38.235	Joe Biden	Joe Biden
49	Wyoming	2	4	0	1	0	1	48.750	12.500	Donald Trump	Donald Trump

Εικόνα 4.12 Πρόβλεψη αποτελέσματος ανά πολιτεία

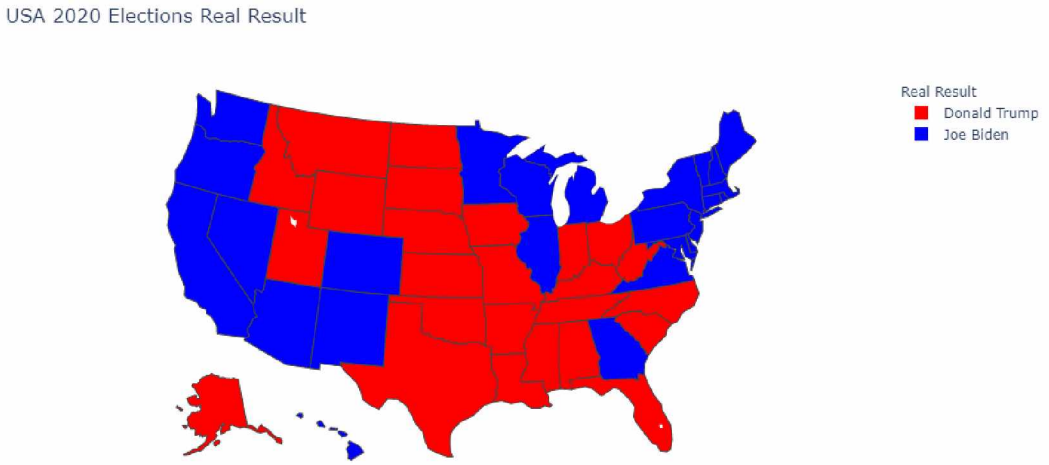
Στον χάρτη της εικόνας 4.13 παρουσιάζονται χρωματικά οι προβλέψεις που προέκυψαν από τα δεδομένα του Twitter για κάθε πολιτεία ξεχωριστά, ενώ στον χάρτη της εικόνας 4.14 φαίνονται τα πραγματικά αποτελέσματα μετά την λήξη των

εκλογών. Με μπλε χρώμα βάφεται κάποια πολιτεία όταν νικητής της αναμέτρησης είναι ο Τζο Μπάιντεν, ενώ με κόκκινο όταν είναι ο Ντόναλντ Τραμπ.



Εικόνα 4.13 Χάρτης προβλέψεων ανά πολιτεία

Παρατηρώντας τις προβλέψεις στη πλειοψηφία των πολιτειών νικητής είναι ο Joe Biden. Συγκρίνοντας τους χάρτες, διαφορές φαίνονται να υπάρχουν στις βορειοδυτικές και στις κεντρώες νότιες πολιτείες. Αυτό συμβαίνει επειδή τα δεδομένα που αναλύονται σε αυτήν την εργασία δεν είναι αρκετά για να βγει ένα εγκυρότερο



Εικόνα 4.14 Χάρτης πραγματικών αποτελεσμάτων ανά πολιτεία

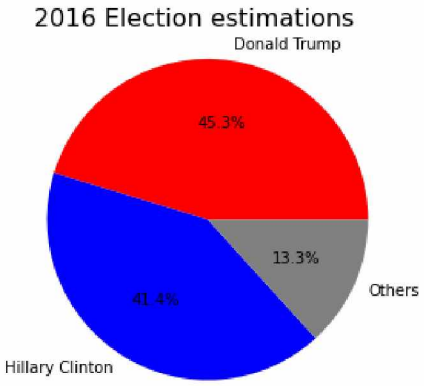
αποτέλεσμα. Ενδιαφέρον φαίνεται να έχει το Texas, όπου παρά τον ικανοποιητικό αριθμό από tweets βγαίνει λάθος αποτέλεσμα με ελάχιστη όμως διαφορά όπως μπορούμε να δούμε και στον πίνακα των αποτελεσμάτων παραπάνω.

4.3.4 Πρόβλεψη αποτελέσματος στις εκλογές του 2016

Για να ελέγξουμε την εγκυρότητα και την αποτελεσματικότητα του μαθηματικού τύπου θα εφαρμόσουμε την ίδια διαδικασία σε δεδομένα που συλλέχθηκαν κατά την περίοδο των Αμερικάνικων εκλογών του 2016. Τα δεδομένα που θα χρησιμοποιηθούν βρίσκονται δημοσιευμένα στην διαδικτυακή κοινότητα επιστημόνων δεδομένων και επαγγελματιών μηχανικής μάθησης – Kaggle [32] με τίτλο «Hillary Clinton and Donald Trump Tweets».

Το σύνολο δεδομένων περιλαμβάνει για κάθε tweet, το «id», το «handle» (σε ποιόν από τους δύο υποψήφιους αναφέρεται το tweet), το «text», καθώς και πληροφορίες αν το σχόλιο είναι ανακοινοποιημένο ή απάντηση σε άλλο tweet.

Μετά την εκτέλεση της διαδικασίας (προεπεξεργασία κειμένου, συναισθηματική ανάλυση και εφαρμογή του μαθηματικού τύπου), λαμβάνουμε την πρόβλεψη του εκλογικού αποτελέσματος, όπως φαίνεται στο διάγραμμα της εικόνας 4.15.



Εικόνα 4.15 Πρόβλεψη εκλογικού αποτελέσματος στις εκλογές του 2016

Από το παραπάνω διάγραμμα διαπιστώνουμε ότι η πρόβλεψη μας, ότι νικητής των εκλογών θα είναι ο Ντόναλντ Τραμπ, αποδεικνύεται σωστή, καθώς στην πραγματικότητα αυτός ήταν ο νικητής αυτής της εκλογικής αναμέτρησης. Τα πραγματικά αποτελέσματα ήταν η νίκη του Ντόναλντ Τραμπ με 56.5%, ενώ το ποσοστό των ψήφων που κατάφερε να λάβει η Χίλαρι Κλίντον ήταν το 42.1%. Επομένως παρατηρούμε ότι η πρόβλεψη μας έχει μια μικρή απόκλιση από τα

πραγματικά δεδομένα όσον αφορά τον Ντόναλντ Τραμπ ενώ είναι πολύ κοντά στα αποτελέσματα της Κλίντον και καταφέρνει να εντοπίσει τον νικητή των εκλογών.

Συνοψίζοντας και κλείνοντας αυτή την ενότητα μπορούμε να πούμε ότι από τα μέσα κοινωνικής δικτύωσης καταλήγουμε σε αρκετά έγκυρα αποτελέσματα όσων αφορά εκλογικές αναμετρήσεις και αυτό φαίνεται λεπτομερώς στην συγκεκριμένη ανάλυση. Μπορούμε αυτό να το διαπιστώσουμε αν παρατηρήσουμε και τις προβλέψεις των στοιχηματικών εταιριών που έδιναν ξεκάθαρο προβάδισμα στο Ρεπουμπλικανικό κόμμα στις Προεδρικές εκλογές 2020, ενώ στις εκλογές του 2016 προέβλεπαν ως νικήτρια, με μεγάλη άνεση, την Χίλαρι Κλίντον.

ΚΕΦΑΛΑΙΟ 5

Μοντέλα μηχανικής μάθησης και αποτέλεσμα εκλογών

Εισαγωγή

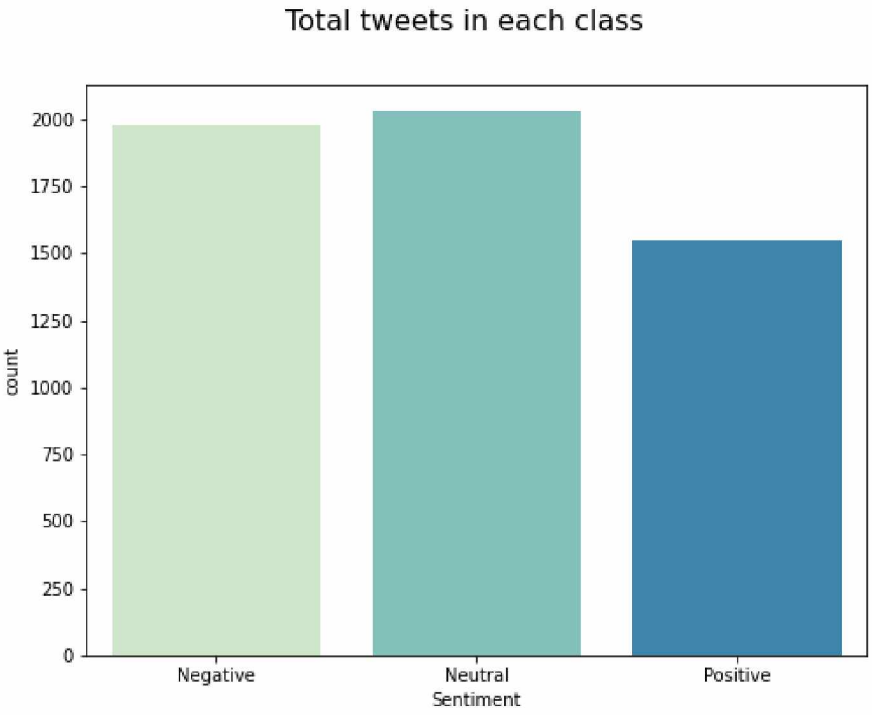
Στο κεφάλαιο 5 θα γίνει εφαρμογή μοντέλων κατηγοριοποίησης (classification) της μηχανικής μάθησης, για την πρόβλεψη των συναισθημάτων των tweets, και μέσω αυτής θα εκτιμηθεί το αποτέλεσμα των εκλογών όπως στο προηγούμενο κεφάλαιο. Ακόμη, θα γίνει μια βελτίωση των πιο αποδοτικών αλγορίθμων για την εξαγωγή πιο έγκυρου αποτελέσματος.

5.1 Η μέθοδος της Μηχανικής Μάθησης

Προηγουμένως εκφράσαμε το συναίσθημα βασισμένοι στο λεξικό VADER της βιβλιοθήκης NLTK. Το κύριο πρόβλημα είναι ότι οι προσεγγίσεις που βασίζονται σε λεξικό δεν προσαρμόζονται καλά σε διαφορετικούς τομείς ή διαφορετικές γλώσσες. Μια λέξη μπορεί να εκφράζει θετικό συναίσθημα σε έναν τομέα αλλά αρνητικό σε κάποιον άλλο. Η έρευνα [21] των Hailong Zhang, Wenyan Gan, Bo Jiang έδειξε ότι προσεγγίσεις της επιβλεπόμενης μηχανικής μάθησης πετυχαίνουν υψηλότερα ποσοστά πρόβλεψης σε σχέση με τα λεξικά.

Η διαδικασία που ακολουθείται σε αυτό το κεφάλαιο είναι αρχικά η προεπεξεργασία των δεδομένων κειμένου που αποτελούν το σύνολο δεδομένων (dataset), προκειμένου η διαδικασία της κατηγοριοποίησης να γίνει αποτελεσματικότερη και λιγότερο χρονοβόρα. Στη συνέχεια, τα δεδομένα μας χωρίζονται σε ποσοστό 70%-30%, τα οποία αποτελούν το σύνολο εκπαίδευσης και ελέγχου αντίστοιχα. Από τα δεδομένα που έχουμε στην διάθεση μας, για την εκπαίδευση των μοντέλων, θα αξιοποιήσουμε τα προεπεξεργασμένα δεδομένα κειμένου (text) καθώς και το αποτέλεσμα του συναισθήματος (sentiment). Τέλος, για κάθε αλγόριθμο θα ελέγξουμε την απόδοση του με τη χρήση του F1-score επειδή οι

κλάσεις του συγκεκριμένου συνόλου είναι δυσανάλογες μεταξύ τους, με τα περισσότερα tweets να είναι ουδέτερα, όπως φαίνεται και στο διάγραμμα της εικόνας 5.1. Τα δεδομένα που θα χρησιμοποιηθούν είναι τα ίδια με αυτά του προηγούμενου κεφαλαίου και η αντιστοιχία του κάθε tweet σε ένα από τα τρία συναισθήματα (για την εκπαίδευση των μοντέλων) λαμβάνεται υπόψιν όπως ακριβώς εξήχθησαν με τη χρήση του λεξικού VADER.



Εικόνα 5.1 Κατανομή των συναισθημάτων

5.2 Αλγόριθμοι κατηγοριοποίησης

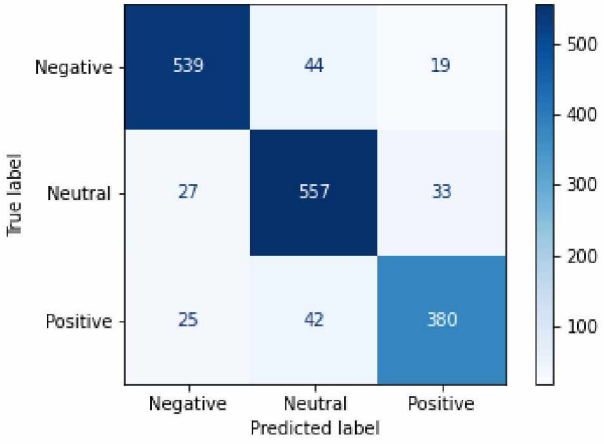
Οι αλγόριθμοι κατηγοριοποίησης που θα εφαρμοστούν στο σύνολο των δεδομένων είναι ο Multinomial Naïve Bayes, οι Μηχανές Διανυσματικής Υποστήριξης (SVC), τα Δέντρα Απόφασης, οι Κ-κοντινότεροι γείτονες (KNN) καθώς και οι μετά-αλγόριθμοι, Τυχαία Δάση (Random Forests) και Adaboost. Στα μοντέλα που θα σημειώσουν το υψηλότερο F1-score θα γίνει βελτίωση των υπερ-παραμέτρων τους με τη βοήθεια της βιβλιοθήκης GridSearchCV, με σκοπό την μεγιστοποίηση του F1-score.

Για κάθε αλγόριθμο δημιουργήθηκε ο πίνακας σύγχυσης (confusion matrix) και με τα αποτελέσματα αυτού, υπολογίστηκαν τα μέτρα απόδοσης accuracy, precision, recall και f1-score. Για την σύγκριση της απόδοσης των μοντέλων που θα εξετάσουμε στην συνέχεια, λαμβάνεται υπόψιν το μέτρο απόδοσης f1-score, το οποίο είναι ο αρμονικός μέσος ανάμεσα στο precision και το recall, καθώς δεν υπάρχει ισορροπία στις κλάσεις (τα ουδέτερα tweets είναι περισσότερα).

5.2.1 Αλγόριθμος Multinomial Naïve Bayes

Από τον αλγόριθμο Multinomial Naïve Bayes προκύπτει ο πίνακας σύγχυσης της εικόνας 5.2. Ο αλγόριθμος αυτός έχει προβλέψει:

- σωστά 539 αρνητικά tweets
- λανθασμένα 44 ως ουδέτερα και 19 ως θετικά, ενώ στην πραγματικότητα είναι αρνητικά
- σωστά 557 ουδέτερα tweets
- λανθασμένα 27 ως αρνητικά και 33 ως θετικά, ενώ στην πραγματικότητα είναι ουδέτερα
- σωστά 380 θετικά tweets
- λανθασμένα 25 ως αρνητικά και 42 ως ουδέτερα, ενώ στην πραγματικότητα είναι θετικά



Εικόνα 5.2 Πίνακας σύγχυσης για τον Multinomial Naive Bayes

Τα μέτρα απόδοσης του αλγορίθμου Multinomial Naïve Bayes είναι τα εξής:

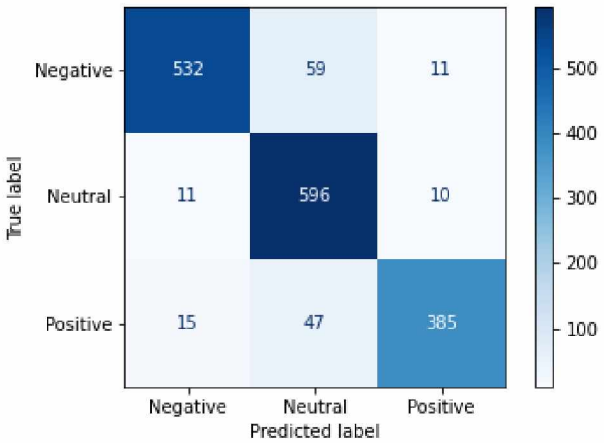
Accuracy	0.89
Precision	0.88
Recall	0.88
F1-score	0.88

Πίνακας 5.1 Μέτρα απόδοσης του αλγορίθμου Multinomial Naïve Bayes

5.2.2 Αλγόριθμος Μηχανών Διανυσματικής Υποστήριξης (SVM)

Από τον αλγόριθμο SVM προκύπτει ο πίνακας σύγχυσης της εικόνας 5.3. Ο αλγόριθμος αυτός έχει προβλέψει:

- σωστά 532 αρνητικά tweets
- λανθασμένα 59 ως ουδέτερα και 11 ως θετικά, ενώ στην πραγματικότητα είναι αρνητικά
- σωστά 596 ουδέτερα tweets
- λανθασμένα 11 ως αρνητικά και 10 ως θετικά, ενώ στην πραγματικότητα είναι ουδέτερα
- σωστά 385 θετικά tweets
- λανθασμένα 15 ως αρνητικά και 47 ως ουδέτερα, ενώ στην πραγματικότητα είναι θετικά



Εικόνα 5.3 Πίνακας σύγχυσης για τον SVM

Τα μέτρα απόδοσης του αλγορίθμου SVM είναι τα εξής:

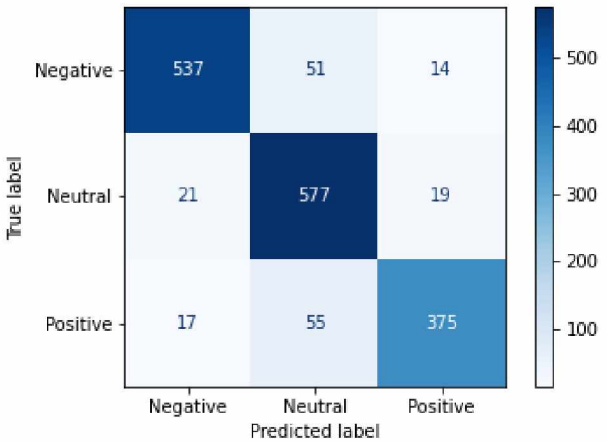
Accuracy	0.91
Precision	0.92
Recall	0.91
F1-score	0.91

Πίνακας 5.2 Μέτρα απόδοσης του αλγορίθμου SVM

5.2.3 Αλγόριθμος Δέντρων Απόφασης

Από τον αλγόριθμο των Δέντρων Απόφασης προκύπτει ο πίνακας σύγχυσης της εικόνας 5.4. Ο αλγόριθμος αυτός έχει προβλέψει:

- σωστά 537 αρνητικά tweets
- λανθασμένα 51 ως ουδέτερα και 14 ως θετικά, ενώ στην πραγματικότητα είναι αρνητικά
- σωστά 577 ουδέτερα tweets
- λανθασμένα 21 ως αρνητικά και 19 ως θετικά, ενώ στην πραγματικότητα είναι ουδέτερα
- σωστά 375 θετικά tweets
- λανθασμένα 17 ως αρνητικά και 55 ως ουδέτερα, ενώ στην πραγματικότητα είναι θετικά



Εικόνα 5.4 Πίνακας σύγχυσης για τα Δ.Α

Τα μέτρα απόδοσης του αλγορίθμου των Δέντρων Απόφασης είναι τα εξής:

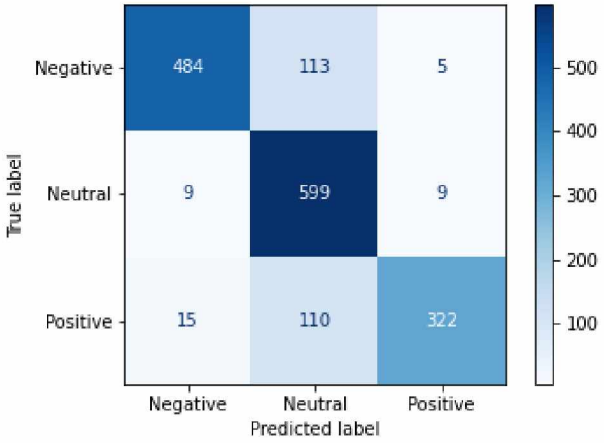
Accuracy	0.89
Precision	0.90
Recall	0.89
F1-score	0.89

Πίνακας 5.3 Μέτρα απόδοσης του αλγορίθμου των Δέντρων Απόφασης

5.2.4 Αλγόριθμος Κ-Κοντινότερων Γειτόνων (KNN)

Από τον αλγόριθμο των Κ-Κοντινότερων Γειτόνων προκύπτει ο πίνακας σύγχυσης της εικόνας 5.5. Ο αλγόριθμος αυτός έχει προβλέψει:

- σωστά 484 αρνητικά tweets
- λανθασμένα 113 ως ουδέτερα και 5 ως θετικά, ενώ στην πραγματικότητα είναι αρνητικά
- σωστά 599 ουδέτερα tweets
- λανθασμένα 9 ως αρνητικά και 9 ως θετικά, ενώ στην πραγματικότητα είναι ουδέτερα
- σωστά 375 θετικά tweets
- λανθασμένα 15 ως αρνητικά και 110 ως ουδέτερα, ενώ στην πραγματικότητα είναι θετικά



Εικόνα 5.5 Πίνακας σύγχυσης για τον KNN

Τα μέτρα απόδοσης του αλγορίθμου των Κ-Κοντινότερων Γειτόνων είναι τα εξής:

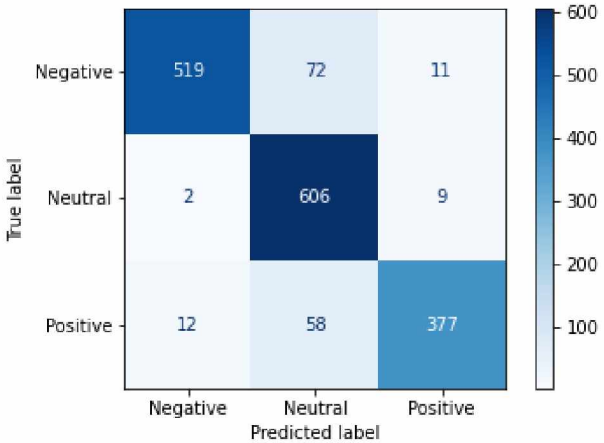
Accuracy	0.84
Precision	0.88
Recall	0.83
F1-score	0.84

Πίνακας 5.4 Μέτρα απόδοσης του αλγορίθμου των Κ-Κοντινότερων Γειτόνων

5.2.5 Αλγόριθμος Τυχαίου Δάσους

Από τον αλγόριθμο του Τυχαίου Δάσους προκύπτει ο πίνακας σύγχυσης της εικόνας 5.6. Ο αλγόριθμος αυτός έχει προβλέψει:

- σωστά 519 αρνητικά tweets
- λανθασμένα 72 ως ουδέτερα και 11 ως θετικά, ενώ στην πραγματικότητα είναι αρνητικά
- σωστά 606 ουδέτερα tweets
- λανθασμένα 2 ως αρνητικά και 9 ως θετικά, ενώ στην πραγματικότητα είναι ουδέτερα
- σωστά 377 θετικά tweets
- λανθασμένα 12 ως αρνητικά και 58 ως ουδέτερα, ενώ στην πραγματικότητα είναι θετικά



Εικόνα 5.6 Πίνακας σύγχυσης για τα Τυχαία Δάση

Τα μέτρα απόδοσης του αλγορίθμου του Τυχαίου Δάσους είναι τα εξής:

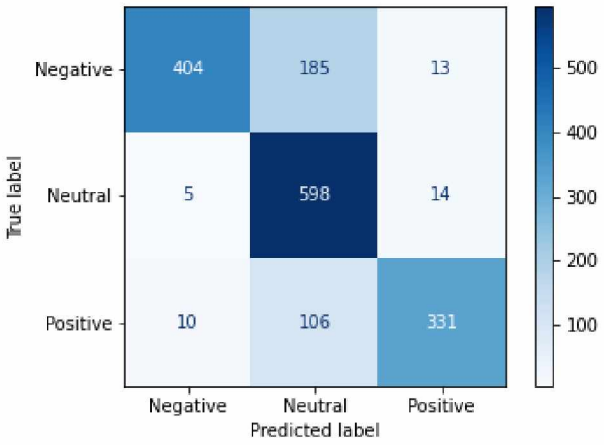
Accuracy	0.90
Precision	0.92
Recall	0.90
F1-score	0.90

Πίνακας 5.5 Μέτρα απόδοσης του αλγορίθμου του Τυχαίου Δάσους

5.2.6 Αλγόριθμος Adaboost

Από τον αλγόριθμο Adaboost προκύπτει ο πίνακας σύγχυσης της εικόνας 5.7. Ο αλγόριθμος αυτός έχει προβλέψει:

- σωστά 414 αρνητικά tweets
- λανθασμένα 185 ως ουδέτερα και 13 ως θετικά, ενώ στην πραγματικότητα είναι αρνητικά
- σωστά 598 ουδέτερα tweets
- λανθασμένα 5 ως αρνητικά και 14 ως θετικά, ενώ στην πραγματικότητα είναι ουδέτερα
- σωστά 331 θετικά tweets
- λανθασμένα 10 ως αρνητικά και 106 ως ουδέτερα, ενώ στην πραγματικότητα είναι θετικά



Εικόνα 5.7 Πίνακας σύγχυσης για τον Adaboost

Τα μέτρα απόδοσης του αλγορίθμου Adaboost είναι τα εξής:

Accuracy	0.80
Precision	0.85
Recall	0.79
F1-score	0.80

Πίνακας 5.6 Μέτρα απόδοσης του αλγορίθμου Adaboost

5.2.7 Σύγκριση όλων των αλγορίθμων

Λαμβάνοντας υπόψιν όλα τα μέτρα απόδοσης, όπως αυτά φαίνονται στις προηγούμενες υπο-ενότητες, ο αλγόριθμος των Μηχανών Διανυσματικής Υποστήριξης (SVM) πετυχαίνει το υψηλότερο F1-score με 0.91, όπως φαίνεται και στον παρακάτω πίνακα. Επίσης, ικανοποιητικά είναι τα σκορ και των υπόλοιπων μοντέλων μηχανικής μάθησης, όπως για παράδειγμα ο αλγόριθμος των Τυχαίων Δασών με σκορ 0.90.

Αλγόριθμος	**F1-score**
Multinomial Naïve Bayes	0.88
SVM	0.91
Δέντρα Απόφασης	0.89
KNN	0.84
Τυχαία Δάση	0.90
Adaboost	0.80

Πίνακας 5.7 F1-scores των αλγορίθμων

Παρατηρώντας τους πίνακες σύγχυσης των αλγορίθμων, όλοι τους καταφέρνουν να ταξινομήσουν με ελάχιστο σφάλμα τα tweets, ως θετικά και αρνητικά. Επομένως, τα σκορ τους καθορίζονται σε πολύ μεγάλο βαθμό από τον αριθμό των tweets που πρόβλεψαν ως ουδέτερα.

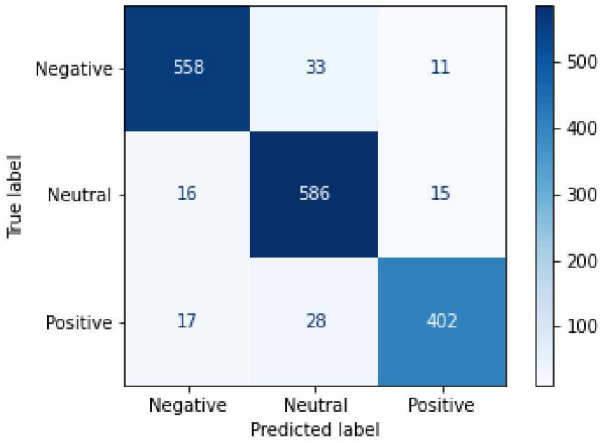
5.3 Βελτίωση υπερ-παραμέτρων

Σε αυτή την ενότητα θα γίνει βελτίωση των υπερ-παραμέτρων των αλγορίθμων Μηχανών Διανυσματικής Υποστήριξης και Τυχαίων Δασών, οι δύο αλγόριθμοι με τα υψηλότερα μέτρα. Στόχος αυτής της διαδικασίας είναι, μέσα από ένα σύνολο παραμέτρων, να εντοπιστεί ο καλύτερος συνδυασμός αυτών έτσι ώστε να επιτύχουμε ένα βελτιωμένο F1-score. Η διαδικασία αυτή εφαρμόστηκε για κάθε έναν από τους δύο αλγόριθμους και ο καλύτερος συνδυασμός εξήχθη με την βοήθεια της βιβλιοθήκης GridSearchCV.

Ο καλύτερος συνδυασμός παραμέτρων που επέλεξε η βιβλιοθήκη GridSearchCV για τον αλγόριθμο των Μηχανών Διανυσματικής Υποστήριξης είναι:

```
Best estimator is: {'clf__C': 10, 'clf__class_weight': 'balanc
ed', 'clf__kernel': 'linear'}
```

Μετά την εφαρμογή των παραμέτρων, ο αλγόριθμος κατάφερε να πετύχει F1-score 0.93 (το αρχικό ήταν 0.91). Στην εικόνα 5.8 φαίνεται ο πίνακας σύγχυσης του αλγορίθμου με βελτιωμένες τις υπερ-παραμέτρους.

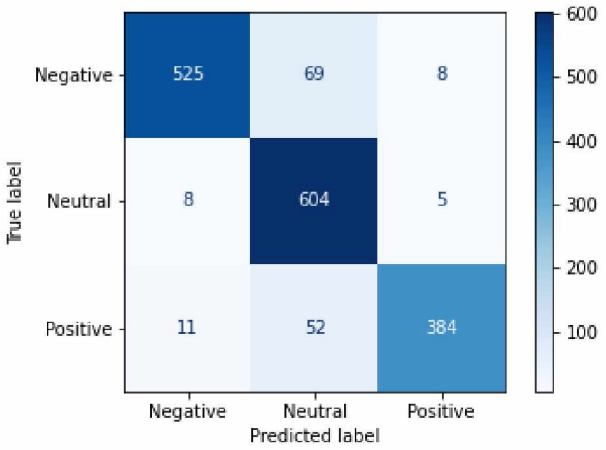


Εικόνα 5.8 Πίνακας σύγχυσης SVM με βελτιωμένες υπερ-παραμέτρους

Ο καλύτερος συνδυασμός παραμέτρων που επέλεξε η βιβλιοθήκη GridSearchCV για τον αλγόριθμο των Τυχαίων Δασών είναι:

```
Best estimator is: {'clf__max_depth': 200, 'clf__max_features'
: 15, 'clf__n_estimators': 200}
```

Μετά την εφαρμογή των παραμέτρων, ο αλγόριθμος των Τυχαίων Δασών κατάφερε να πετύχει F1-score 0.91 (το αρχικό ήταν 0.90). Στην εικόνα 5.9 φαίνεται ο πίνακας σύγχυσης του αλγορίθμου με βελτιωμένες τις υπερ-παραμέτρους.



Εικόνα 5.9 Πίνακας σύγχυσης Τυχαίων Δασών με βελτιωμένες υπερ-παραμέτρους

Επομένως, καταλήγουμε στο συμπέρασμα ότι ο αλγόριθμος των Μηχανών Διανυσματικής Υποστήριξης πετυχαίνει το υψηλότερο F1-score, τόσο πριν όσο και μετά την αυτοματοποιημένη βελτίωση των υπερ-παραμέτρων του.

5.4 Πρόβλεψη εκλογικού αποτελέσματος με τον αποδοτικότερο αλγόριθμο

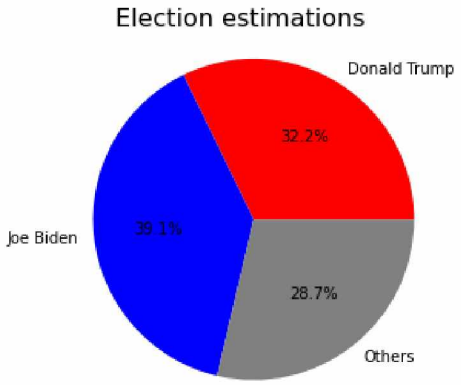
Σε αυτήν την ενότητα θα γίνει εφαρμογή του αποδοτικότερου αλγορίθμου (Μηχανές Διανυσματικής Υποστήριξης), όπως αυτό προέκυψε προηγουμένως, με σκοπό τη σύγκριση των προβλέψεων με τα πραγματικά αποτελέσματα των εκλογών. Τα δεδομένα που θα χρησιμοποιηθούν βρίσκονται δημοσιευμένα στην διαδικτυακή κοινότητα επιστημόνων δεδομένων και επαγγελματιών μηχανικής μάθησης – Kaggle [22] με τίτλο «US Election 2020 Tweets» και είναι διαφορετικά από αυτά που έγινε η εκπαίδευση των μοντέλων.

Κατά την προεπεξεργασία των δεδομένων παρατηρείται ένα μεγάλο ποσοστό κενών τιμών καθώς και tweets που συλλέχθηκαν από χώρες εκτός των Ηνωμένων Πολιτειών και σε άλλες γλώσσες. Τα σχετικά δεδομένα δεν προσφέρουν κάποια χρήσιμη πληροφορία για την εκτίμηση του αποτελέσματος και αφαιρούνται από το σύνολο. Στη συνέχεια, εφαρμόζονται οι τεχνικές της λημματοποίησης, αποκοπής καταλήξεων, αφαίρεσης των stopwords και συμβόλων, έτσι ώστε τα δεδομένα να είναι έτοιμα να δοθούν ως είσοδος στον αλγόριθμο. Τέλος, ο αλγόριθμος των Μηχανών Διανυσματικής Υποστήριξης κάνει την εκτίμηση του συναισθήματος για όλα τα σχόλια και τα αποτελέσματα που λαμβάνουμε φαίνονται στον παρακάτω πίνακα.

	Θετικά	**Ουδέτερα**	**Αρνητικά**
Ντόναλντ Τραμπ	16842	20692	20789
Τζο Μπάιντεν	13282	21834	10391

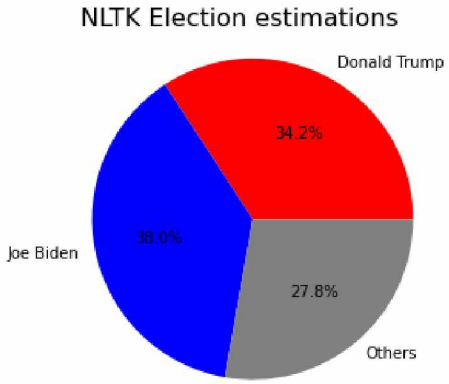
Πίνακας 5.8 Σύνολο συναισθημάτων για κάθε υποψήφιο

Εφαρμόζοντας την μαθηματική φόρμουλα που αναπτύχθηκε από τους Wicaksono et al.[15], όπως και στο προηγούμενο κεφάλαιο, νικητής των εκλογών αναδεικνύεται και με αυτή την μέθοδο ο Τζο Μπάιντεν με 39,1%. Στο διάγραμμα της εικόνας 5.10 φαίνονται τα συνολικά ποσοστά της πρόβλεψης μας, βασισμένοι στα συναισθήματα των tweets που εξήχθησαν με την χρήση του αλγορίθμου SVM με βελτιωμένες τις υπέρ-παραμέτρους.



Εικόνα 5.10 Πρόβλεψη εκλογικού αποτελέσματος με τον αποδοτικότερο αλγόριθμο

Στο διάγραμμα της εικόνας 5.11 φαίνονται τα ποσοστά της πρόβλεψης, βασισμένα στα συναισθήματα των tweets που εξήχθησαν με την χρήση την βιβλιοθήκης NLTK της Python.



Εικόνα 5.11 Πρόβλεψη εκλογικού αποτελέσματος με την βιβλιοθήκη NLTK

Παρατηρώντας τις προβλέψεις που κάναμε στο καινούργιο σύνολο δεδομένων, εύκολα μπορούμε να δούμε ότι τα αποτελέσματα είναι πολύ κοντά μεταξύ τους. Επίσης, η διαφορά μεταξύ του Τζο Μπάιντεν και του Ντόναλντ Τραμπ διαπιστώνεται ότι είναι σχεδόν ίση και στις δύο περιπτώσεις.

Τέλος, αξίζει να σημειωθεί ότι λόγω του αρχικού «προβληματικού» συνόλου δεδομένων που είχαμε στην διάθεση μας, καθώς και η αδυναμία να ανακτήσουμε πληροφορίες που δεν ήταν αρχικά διαθέσιμες, καθιστά την ανάλυση μας λιγότερη αποτελεσματική από το αναμενόμενο. Παρόλα αυτά, σημαντικό είναι να τονισθεί ότι ο σκοπός αυτού του κεφαλαίου, ο οποίος ήταν η εκπαίδευση αλγορίθμων μηχανικής μάθησης, επετεύχθη σε πολύ ικανοποιητικό βαθμό καθώς καταφέραμε να φτάσουμε το μοντέλο των Μηχανών Διανυσματικής Υποστήριξης να πετυχαίνει ποσοστό σωστών προβλέψεων στο 93%.

ΚΕΦΑΛΑΙΟ 6

Εξερεύνηση θεματικών ενοτήτων

Εισαγωγή

Σκοπός του κεφαλαίου 6 είναι να αναλύσει τα δύο σύνολα δεδομένων που χρησιμοποιήθηκαν στα κεφάλαια 4 και 5 και να εξηγήσει τους λόγους για τους οποίους καταλήξαμε, μέσω των προβλέψεων μας, να θεωρήσουμε τον Τζο Μπάιντεν ως τον επόμενο πρόεδρο των Ηνωμένων Πολιτειών. Η ανάλυση αυτή είναι βασισμένη στις απόψεις των πολιτών, όπως αυτοί τις δημοσιοποίησαν στην πλατφόρμα κοινωνικής δικτύωσης του Twitter. Η διαδικασία περιλαμβάνει την τεχνική της συσταδοποίησης (clustering) με τον αλγόριθμο K-means, για τον εντοπισμό των θεματικών ενοτήτων. Επίσης, εφαρμόζεται συναισθηματική ανάλυση για κάθε tweet, με την χρήση του NRC λεξικού. Μέσω των συναισθημάτων που αποτελούν την κάθε συστάδα (cluster), θα γίνει προσπάθεια επεξήγησης του τρόπου με τον οποίο αυτή, συνέβαλε στην διαμόρφωση των προβλέψεων μας.

Το πρώτο σύνολο δεδομένων [15] που πρόκειται να αναλυθεί, συλλέχθηκε μία βδομάδα πριν από την ημέρα των εκλογών, ενώ το δεύτερο [22] συλλέχθηκε σε περίοδο ενός μήνα και είναι αρκετά μεγαλύτερο από το πρώτο.

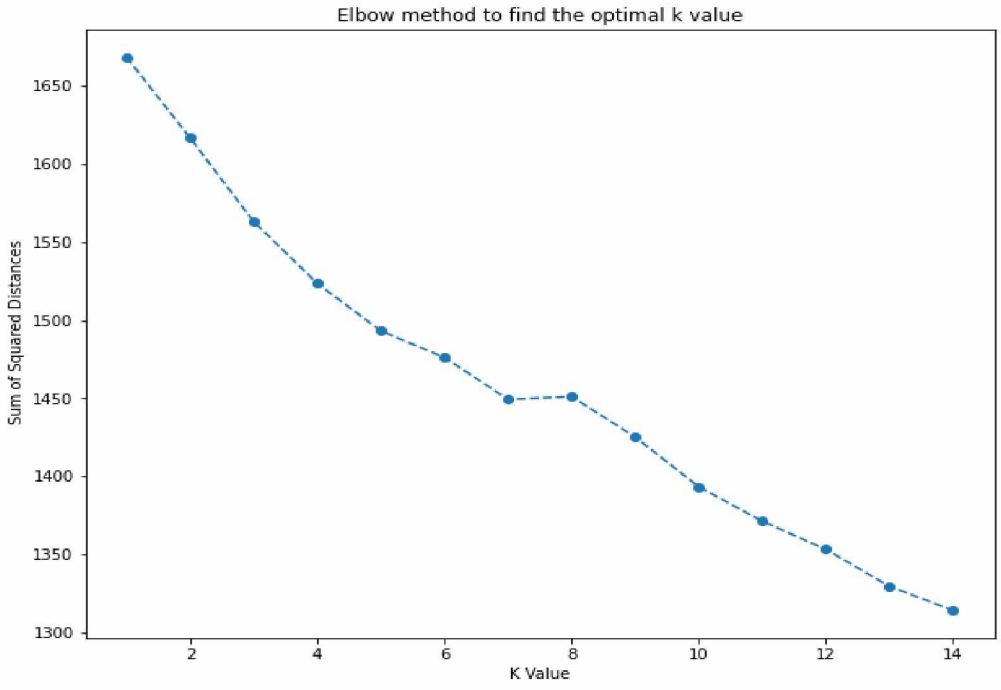
6.1 Tweets τελευταίας εβδομάδας

Σε αυτή την ενότητα θα εξετάσουμε το πρώτο σύνολο δεδομένων που έχουμε στην διάθεση μας το οποίο συλλέχθηκε μία εβδομάδα πριν από τις εκλογές των Ηνωμένων Πολιτειών. Για τον σκοπό της συσταδοποίησης (clustering) των δεδομένων κειμένου δίνουμε έμφαση μόνο στο περιεχόμενο των tweets. Η διαδικασία που ακολουθείται για να δοθούν τα δεδομένα στον αλγόριθμο K-means ως είσοδος είναι η εξής:

- Διαγραφή διπλότυπων tweets, καθώς το περιεχόμενο του κειμένου είναι ακριβώς το ίδιο και δεν προσφέρει κάτι στην διαδικασία, με σκοπό την μείωση του χρόνου για την υλοποίηση της εκτέλεσης του αλγορίθμου.
- Καθαρισμός του κειμένου όπως εφαρμόστηκε σε προηγούμενα κεφάλαια.
- Εφαρμογή των τεχνικών λημματοποίησης, αποκοπής καταλήξεων και διαχωρισμός του κειμένου σε λεκτικές μονάδες.
- Διανυσματοποίηση με την μέθοδο TF-IDF.

6.1.1 Συσταδοποίηση

Για να γίνει η συσταδοποίηση των tweets θα πρέπει αρχικά να αποφασιστεί ο αριθμός των συστάδων (clusters). Για την επιλογή του αριθμού των συστάδων στην μη-επιβλεπόμενη μάθηση δεν υπάρχει κάποιος γενικός κανόνας, ο οποίος να λειτουργεί εγγυημένα για όλες τις περιπτώσεις. Ένα απλό και πρακτικό τέχνασμα, το οποίο μπορεί να βοηθήσει σε ορισμένες περιπτώσεις, είναι «ο κανόνας του αγκώνα» (elbow rule). Μετά από διαδοχικές επαναλήψεις του αλγορίθμου, ο κανόνας υποδεικνύει ως βέλτιστο σημείο, εκείνο στο οποίο υπάρχει απότομη αλλαγή της



Εικόνα 6.1 Η μέθοδος του αγκώνα για τα tweets της τελευταίας εβδομάδας

κλίσης της καμπύλης και σταθεροποίηση του σφάλματος. Εκτός από το σημείο που διακρίνει ο κανόνας, γίνεται δοκιμή συσταδοποίησης με κοντινούς αριθμούς συστάδων, με σκοπό την ανάδειξη καλύτερων συστάδων. Στο διάγραμμα της εικόνας 6.1 ο κανόνας του αγκώνα υποδεικνύει ότι η επιλογή k=7, είναι αρκετά καλή, γεγονός που επαληθεύτηκε και από τις πειραματικές δοκιμές.

Στη συνέχεια γίνεται εφαρμογή του αλγορίθμου K-means για τις 7 συστάδες. Παρακάτω παρουσιάζεται η λίστα με τις δέκα πιο συχνά εμφανιζόμενες λέξεις κάθε συστάδας. Από τις λέξεις αυτές θα ορίσουμε μία θεματική ενότητα για κάθε συστάδα, ώστε να κατανοήσουμε τους λόγους για τους οποίους το αποτέλεσμα των εκλογών διαμορφώθηκε έτσι όπως παρουσιάστηκε στα προηγούμενα κεφάλαια.

Οι δέκα πιο συχνά εμφανιζόμενες λέξεις κάθε συστάδας:

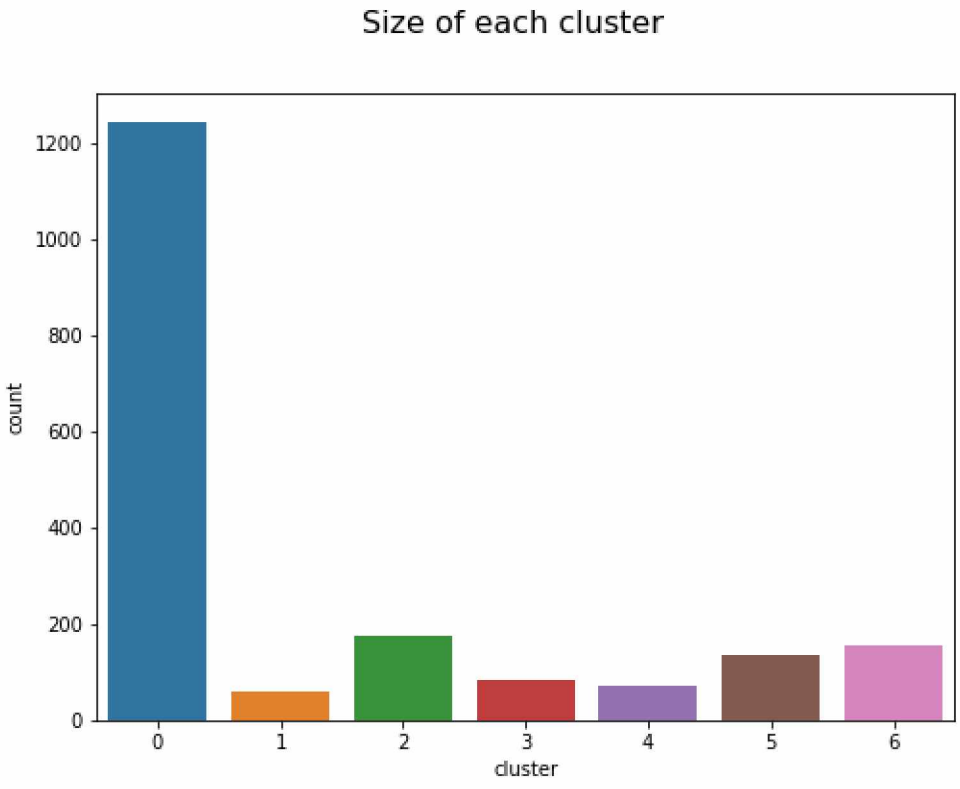
- **Συστάδα #0:** 'hunter', 'go', 'hunter biden', 'campaign', 'know', 'via', 'make', 'election', 'win', 'support'
- **Συστάδα #1:** 'americans', 'kamala', 'harris', 'day', 'vote', 'vote joe', 'take', 'election', 'run', 'campaign'
- **Συστάδα #2:** 'president', 'president donald', 'president trump', 'election', 'campaign', 'via', 'former', 'obama', 'rally', 'america'
- **Συστάδα #3:** 'get', 'us', 'use', 'back', 'like', 'right', 'president', 'say', 'try', 'think'
- **Συστάδα #4:** 'people', 'vote', 'black', 'like', 'call', 'time', 'say', 'president', 'many', 'go'
- **Συστάδα #5:** 'say', 'jr', 'trump jr', 'nothing', 'almost', 'almost nothing', 'deaths', 'covid', 'day', 'go'
- **Συστάδα #6:** 'vote', 'vote joe', 'vote donald', 'go', 'want', 'election', 'let', 'president', 'country', 'right'

Από τις σημαντικότερες λέξεις κάθε συστάδας, καθώς και από μελέτη των tweets που περιέχονται μέσα σε αυτές, εξάγονται οι παρακάτω θεματικές ενότητες:

- **Συστάδα #0:** tweets που αφορούν τον Χάντερ Μπάιντεν, γιο του Τζο Μπάιντεν

- **Συστάδα #1:** tweets που αφορούν την προεκλογική καμπάνια της αντιπροέδρου των ΗΠΑ, Καμάλα Χάρις
- **Συστάδα #2:** αντιπαράθεση του Ντόναλντ Τραμπ με τον προκάτοχο του, Μπαράκ Ομπάμα
- **Συστάδα #3:** -
- **Συστάδα #4:** εκλογές και τα δικαιώματα των μαύρων
- **Συστάδα #5:** tweets σχετικά με την πανδημία της Covid-19 και η εσφαλμένη αντιμετώπιση της από τον πρώην πρόεδρο, Ντόναλντ Τραμπ
- **Συστάδα #6:** tweets υποστήριξης των δύο κύριων υποψηφίων

Στο διάγραμμα της εικόνας 6.2 μπορούμε να δούμε το μέγεθος της κάθε συστάδας. Η συστάδα 0 (tweets σχετικά με τον Χάντερ Μπάιντεν) όπως φαίνεται έχει και τον μεγαλύτερο αριθμό από tweets.

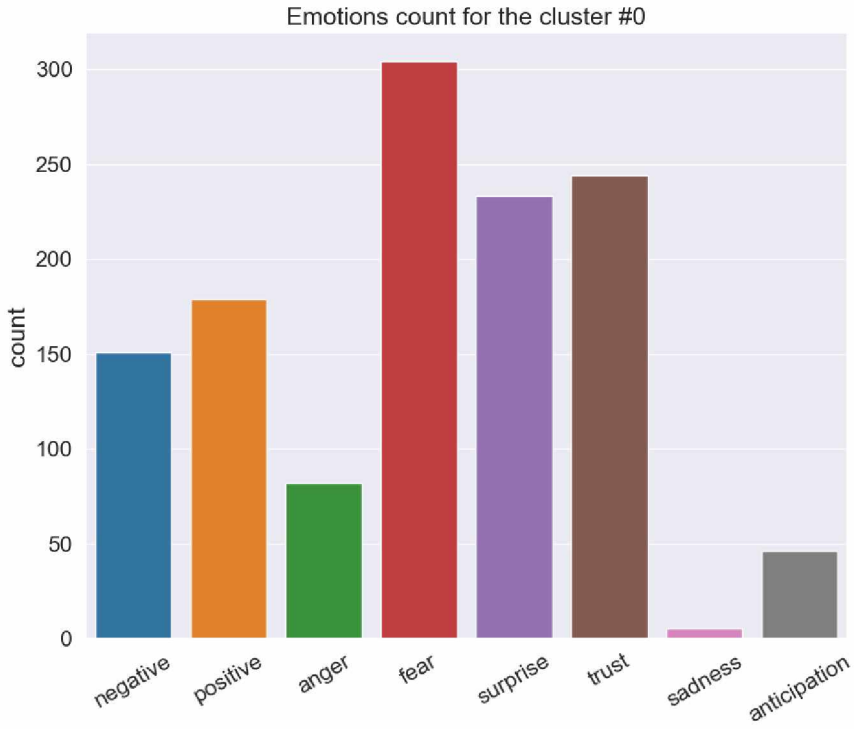


Εικόνα 6.2 Μέγεθος συστάδων

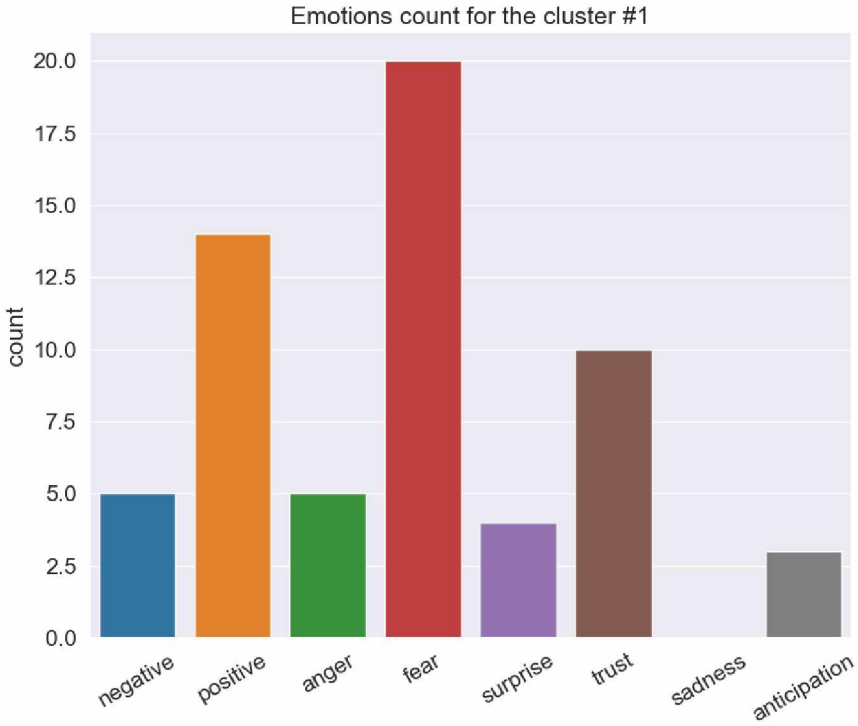
6.1.2 Συναισθηματική ανάλυση συστάδων

Μέχρι στιγμής η συναισθηματική ανάλυση γινόταν με την βοήθεια της βιβλιοθήκης NLTK (Natural Language Toolkit) της Python και την χρήση του VADER Lexicon. Σε αυτή την ενότητα επιλέχθηκε ένα διαφορετικό λεξικό της Python στην Αγγλική γλώσσα, το NRC Lexicon. Το συγκεκριμένο λεξικό αντιστοιχεί κάθε tweet σε οκτώ διαφορετικά συναισθήματα (emotions), τα οποία είναι: negative (=αρνητικό), positive (=θετικό), anger (=θυμός), fear (=φόβος), surprise (=έκπληξη), trust (=εμπιστοσύνη), sadness (=λύπη), anticipation (=προσδοκία).

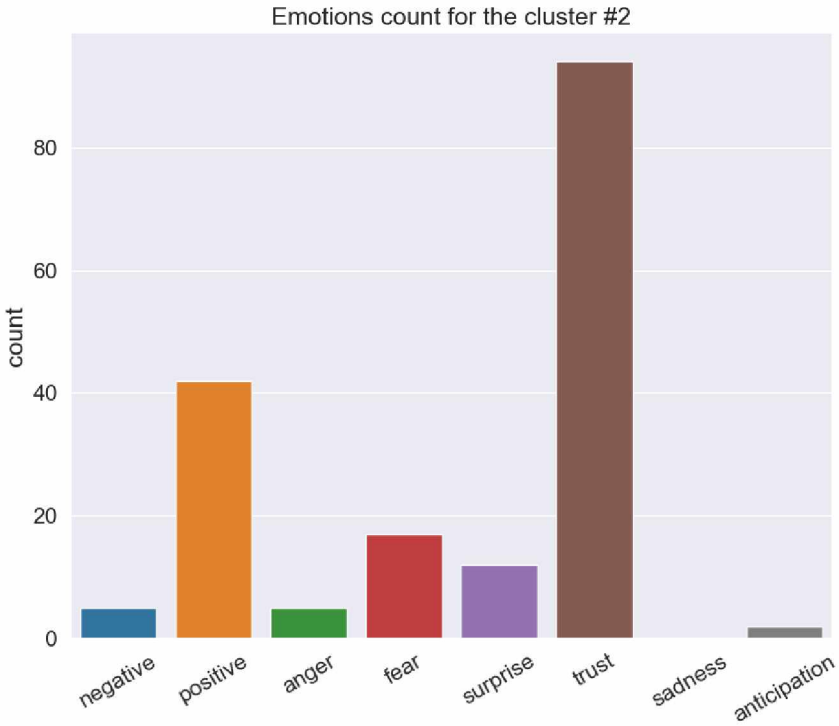
Στα διαγράμματα των εικόνων 6.3 έως 6.9 φαίνονται πως κατανέμονται τα συναισθήματα για κάθε συστάδα. Η συστάδα 0 που αφορά τον Hunter Biden, η οποία είναι η μεγαλύτερη σε μέγεθος συστάδα, χαρακτηρίζεται κυρίως από tweets που υποδηλώνουν φόβο. Αυτό συμβαίνει διότι εκείνη την περίοδο ο ίδιος είχε κάποιες μυστικές οικονομικές δραστηριότητες με την Κίνα. Η συστάδα 1 η οποία σχετίζεται με την καμπάνια της αντιπροέδρου των Ηνωμένων Πολιτειών χαρακτηρίζεται και αυτή από σχόλια φόβου, ενώ επίσης το θετικό συναίσθημα είναι το αμέσως μεγαλύτερο. Η συστάδα 2 αποτελείται από tweets σχετικά με την αντιπαράθεση του Ντόναλντ Τραμπ με τον Μπαράκ Ομπάμα. Η συστάδα αυτή χαρακτηρίζεται εξ ολοκλήρου από σχόλια εμπιστοσύνης. Σημαντικό είναι ότι στις δηλώσεις του ο Μπαράκ Ομπάμα υποστήριξε ανοιχτά τον πρόεδρο του Δημοκρατικού κόμματος, Τζο Μπάιντεν. Η συστάδα 3, για την οποία δεν μπορεί να οριστεί κάποια θεματική ενότητα, αποτελείται κυρίως από σχόλια εμπιστοσύνης. Η συστάδα 4 είναι από τις πιο ενδιαφέρουσες θεματικές ενότητες, η οποία είναι σχετική με τα δικαιώματα των μαύρων. Από μία σύντομη ανασκόπηση στα δεδομένα αυτής της συστάδας παρατηρήθηκε ότι ο κόσμος παραμένει έκπληκτος από τον γνωστό ράπερ Lil Wayne, ο οποίος υποστήριξε ανοικτά τον Ντόναλντ Τραμπ. Η συστάδα 5 που αποτελείται από σχόλια σχετικά με τον Covid-19 και τον εσφαλμένο τρόπο αντιμετώπισης του από τον Ντόναλντ Τραμπ, χαρακτηρίζεται από σχόλια φόβου. Η τελευταία συστάδα η οποία αποτελείται από tweets υποστήριξης των δύο αντιπάλων, χαρακτηρίζεται από σχόλια που εκφράζουν εμπιστοσύνη καθώς και θυμό.



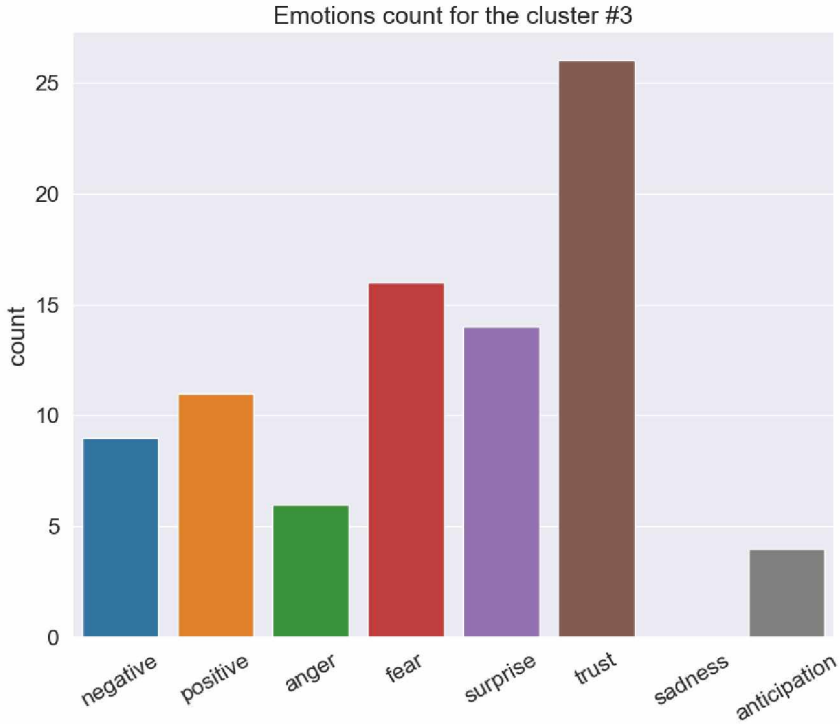
Εικόνα 6.3 Συστάδα 0 (Hunter Biden)



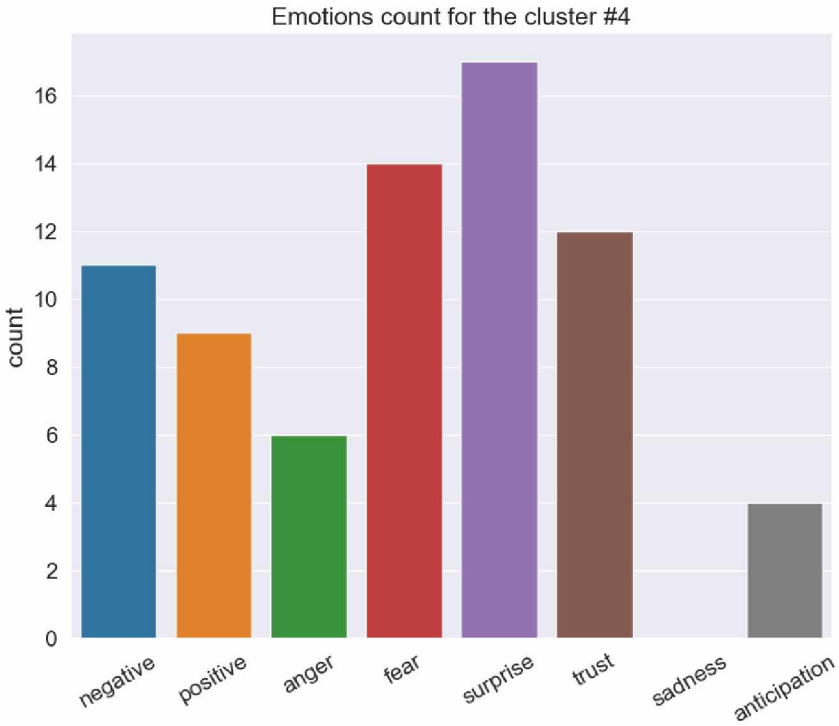
Εικόνα 6.4 Συστάδα 1 (Kamala Harris)



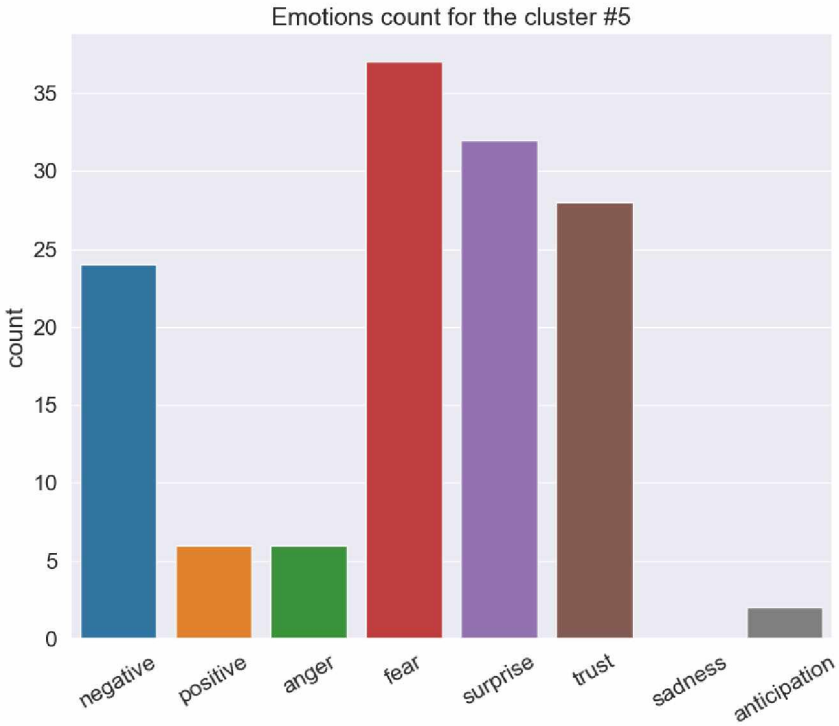
Εικόνα 6.5 Συστάδα 2 (Αντιπαράθεση Τραμπ – Μπάιντεν)



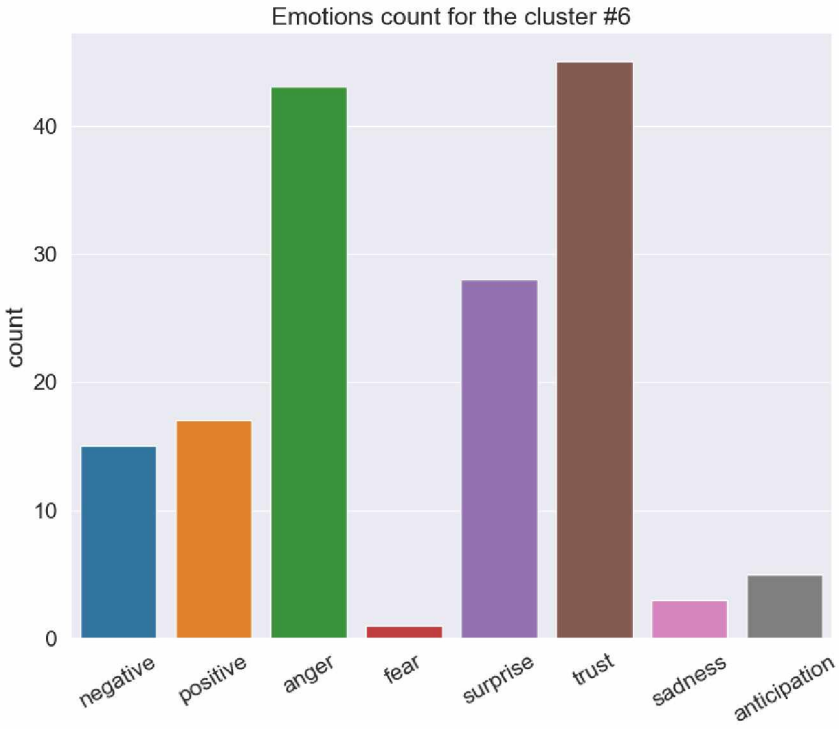
Εικόνα 6.6 Συστάδα 3 (-)



Εικόνα 6.7 Συστάδα 4 (Δικαιώματα μάυρων)



Εικόνα 6.8 Συστάδα 5 (Covid-19)



Εικόνα 6.9 Συστάδα 6 (tweets υποστήριξης)

6.1.3 Ανάλυση υπό-συστάδων

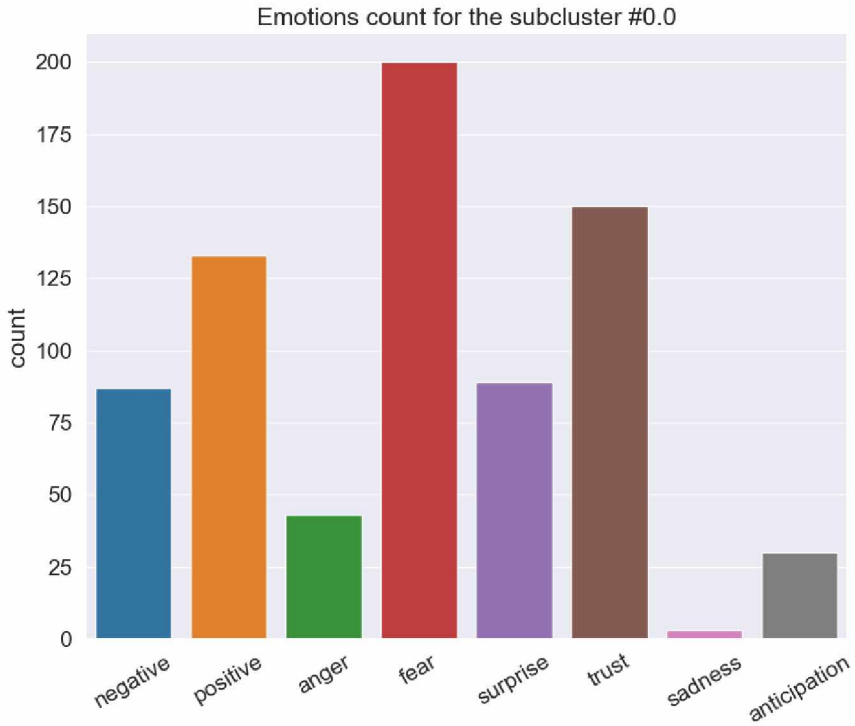
Αφού ολοκληρώθηκε η συναισθηματική ανάλυση των επτά συστάδων που προέκυψαν από την διαδικασία της συσταδοποίησης παρατηρείται ότι η συστάδα 0 είναι αρκετά μεγαλύτερη από τις υπόλοιπες. Ρίχνοντας μια ματιά στις δέκα σημαντικότερες λέξεις της, εκτός από το όνομα του γιού του Τζο Μπάιντεν, υπάρχουν και άλλες λέξεις, γενικότερες. Για τον λόγο αυτό θα γίνει ανάλυση αυτής της συστάδας, με σκοπό την εξερεύνηση νέων θεματικών ενοτήτων.

Ακολουθώντας την ίδια διαδικασία συσταδοποίησης με την βοήθεια του κανόνα του αγκώνα, καταλήγουμε στις πέντε υπο-ενότητες της συστάδας 0:

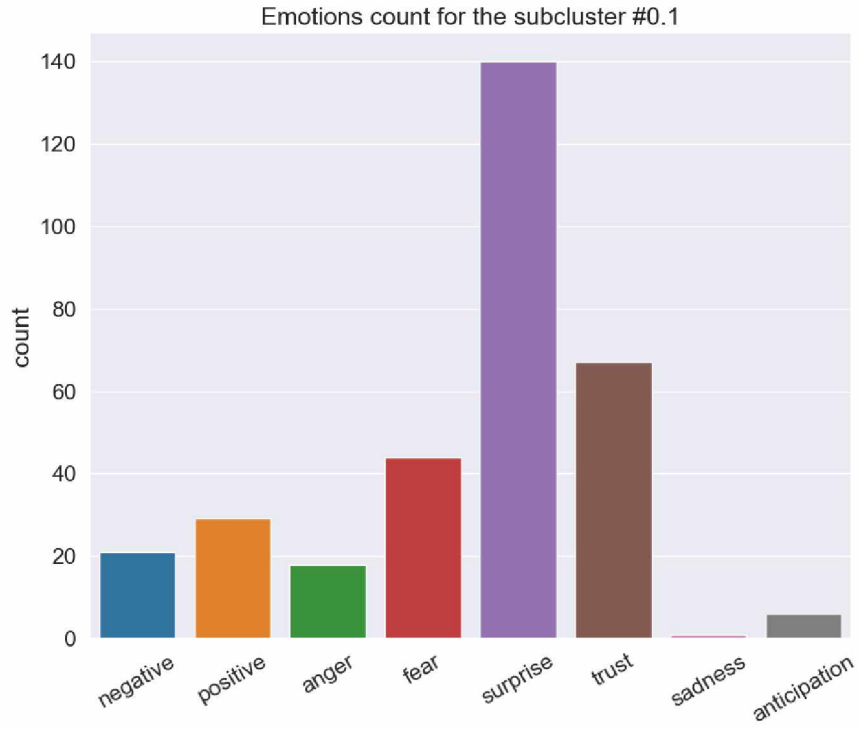
- **Συστάδα #0.0:** 'go', 'campaign', 'know', 'family', 'would', 'state', 'make', 'support', 'new', 'america'
- **Συστάδα #0.1:** 'donald', 'donald trump', 'lil', 'via', 'lil wayne', 'wayne', 'election', 'endorse', 'support', 'take'

- **Συστάδα #0.2:** 'lie', 'son', 'stand', 'like', 'rally', 'tell', 'need', 'come', 'via', 'know'
- **Συστάδα #0.3:** 'hunter', 'hunter biden', 'business', 'email', 'deal', 'amp', 'report', 'greenwald', 'criminal', 'corruption'
- **Συστάδα #0.4:** 'win', 'white', 'biden win', 'house', 'white house', 'make', 'next', 'week', 'secretary', 'want'

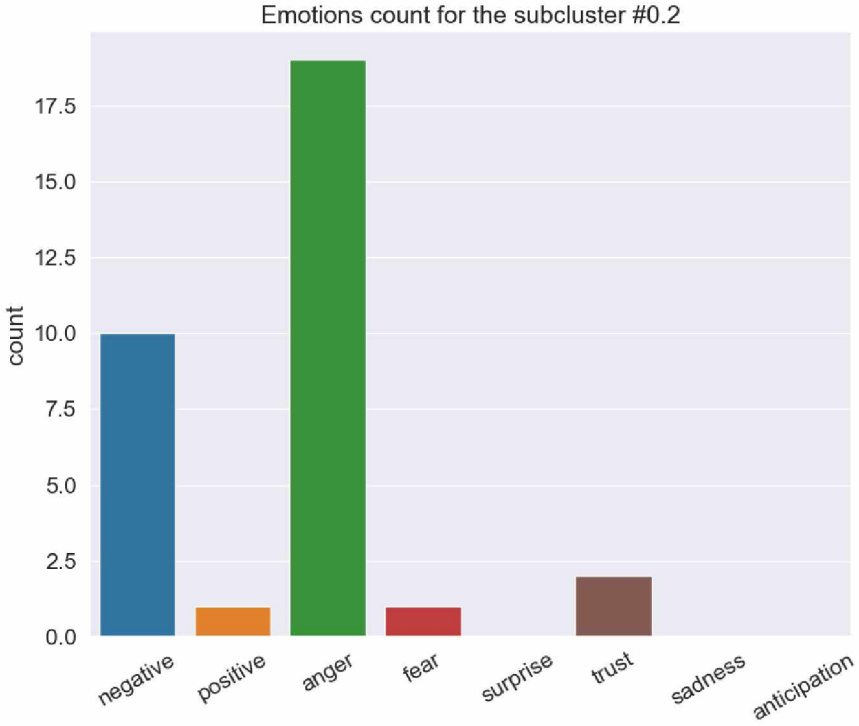
Στα παρακάτω διαγράμματα φαίνονται πως κατανέμονται τα συναισθήματα για την κάθε υπό-συστάδα.



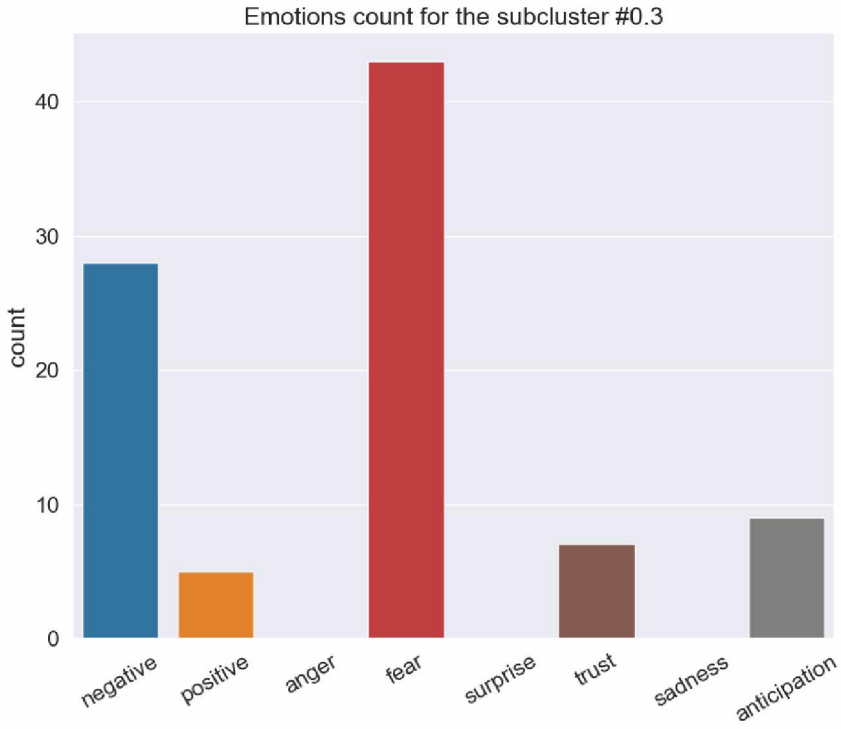
Εικόνα 6.10 Συστάδα 0.0



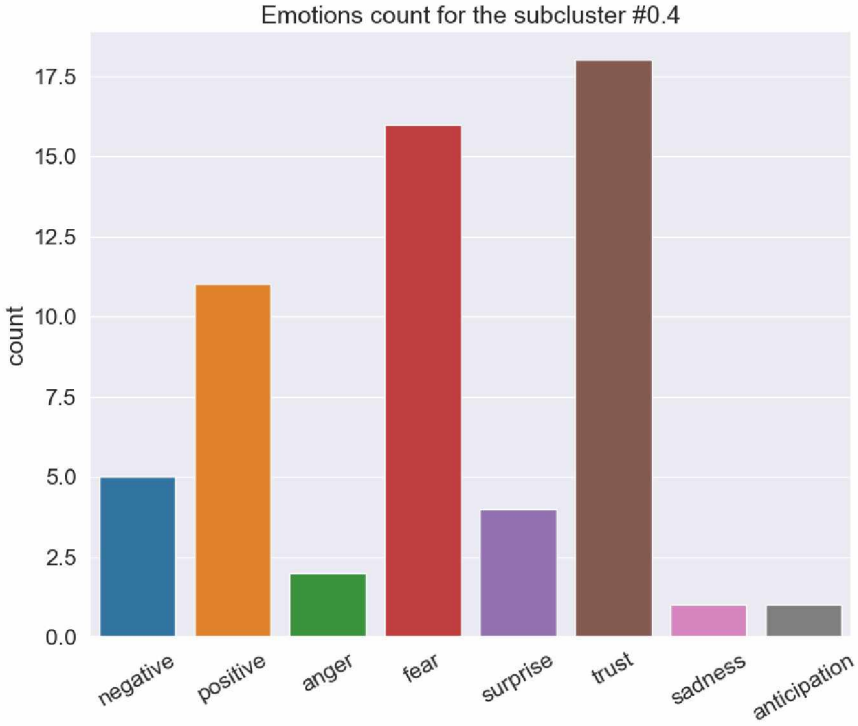
Εικόνα 6.11 Συστάδα 0.1



Εικόνα 6.12 Συστάδα 0.2



Εικόνα 6.13 Συστάδα 0.3



Εικόνα 6.14 Συστάδα 0.4

Από την ανάλυση των υπό-συστάδων της μεγαλύτερης σε μέγεθος συστάδας (συστάδα 0), παρατηρούμε ότι από τις πιο γενικές ισχυρές λέξεις της μπορούμε να εξάγουμε πιο έγκυρα συμπεράσματα. Όπως αναφερθήκαμε και προηγουμένως για την συστάδα 4, η οποία ήταν σχετική με τα δικαιώματα των αφροαμερικάνων, έτσι και τώρα η υπό-συστάδα 0.1 περιέχει κι αυτή tweets από τα οποία ο κόσμος μένει έκπληκτος από την υποστηρικτική στάση του ράπερ Lil Wayne στο πρόσωπο του Ντόναλντ Τραμπ. Ενδιαφέρον παρουσιάζει η υπό-συστάδα 0.4 η οποία χαρακτηρίζεται από σχόλια εμπιστοσύνης προς τον Τζο Μπάιντεν ως τον επόμενο πρόεδρο των Ηνωμένων Πολιτειών. Τέλος, αναμενόμενες είναι οι υπό-συστάδες 0.2 και 0.3 οι οποίες εκφράζουν τον θυμό και τον φόβο αντίστοιχα, όσον αφορά τις μυστικές οικονομικές διαπραγματεύσεις του γιού του Τζο Μπάιντεν.

6.2 Tweets τελευταίου μήνα

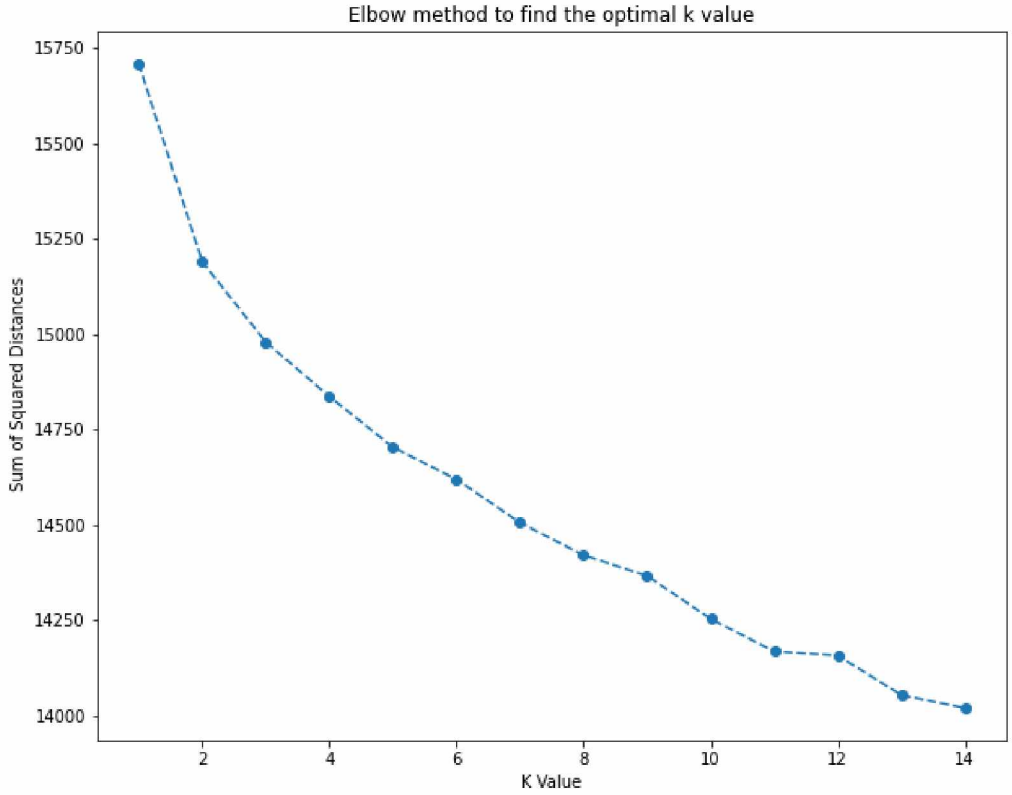
Το δεύτερο σύνολο δεδομένων, που έχουμε στην διάθεση μας προς συσταδοποίηση, συλλέχθηκε σε περίοδο ενός μήνα πριν από τις Αμερικάνικες Προεδρικές εκλογές. Αρχικά, έγινε προεπεξεργασία των δεδομένων, όπως και στο προηγούμενο σύνολο, διότι παρατηρήθηκαν αρκετά ελλιπή δεδομένα καθώς και tweets τα οποία συλλέχθηκαν από χώρες εκτός των Ηνωμένων Πολιτειών και σε διαφορετικές γλώσσες πέρα των αγγλικών. Στη συνέχεια, έγινε προσπάθεια συσταδοποίησης και εξερεύνηση θεματικών ενοτήτων σε ολόκληρο το σύνολο. Η προσπάθεια αυτή απέτυχε, καθώς δεν ήταν δυνατόν να εντοπιστούν θεματικές ενότητες από τις πιο συχνές και σημαντικές λέξεις κάθε συστάδας. Έτσι, αποφασίστηκε να γίνει διαχωρισμός του συνόλου δεδομένων, σύμφωνα με λέξεις κλειδιά που αντιπροσωπεύουν τους δύο υποψήφιους. Το πρώτο σύνολο δημιουργήθηκε με λέξεις κλειδιά σχετικές με τον Ντόναλντ Τραμπ («trump», «donald», «donald trump», «republicans») και περιέχει 16144 tweets, ενώ το δεύτερο με λέξεις σχετικές με τον Τζο Μπάιντεν («joe», «biden», «joe biden», «democrats», «democratics») και αποτελείται από 13359 σχόλια.

6.2.1 Συσταδοποίηση και συναισθηματική ανάλυση

Σε αυτή την υπο-ενότητα θα εφαρμοστεί ο αλγόριθμος K-Means, με τη βοήθεια του κανόνα του αγκώνα για την επιλογή του αριθμού των συστάδων, και για τα δύο σύνολα, όπως αυτά προέκυψαν μετά τον διαχωρισμό.

Tweets σχετικά με τον Ντόναλντ Τραμπ

Στο διάγραμμα της εικόνας 6.15 ο κανόνας του αγκώνα υποδεικνύει ότι η επιλογή k=11, είναι αρκετά καλή. Παρόλα αυτά επιλέγεται k=12, διότι οι συστάδες που



Εικόνα 6.15 Κανόνας του αγκώνα για tweets σχετικά με τον Τραμπ

δημιουργούνται είναι πιο ξεκάθαρες για την εύρεση θεματικών ενοτήτων.

Στη συνέχεια εφαρμόζεται ο αλγόριθμος K-Means για τις 12 συστάδες και παρακάτω παρουσιάζεται η λίστα με τις δέκα πιο συχνές και σημαντικές λέξεις για κάθε μία.

Οι δέκα πιο συχνά εμφανιζόμενες λέξεις κάθε συστάδας:

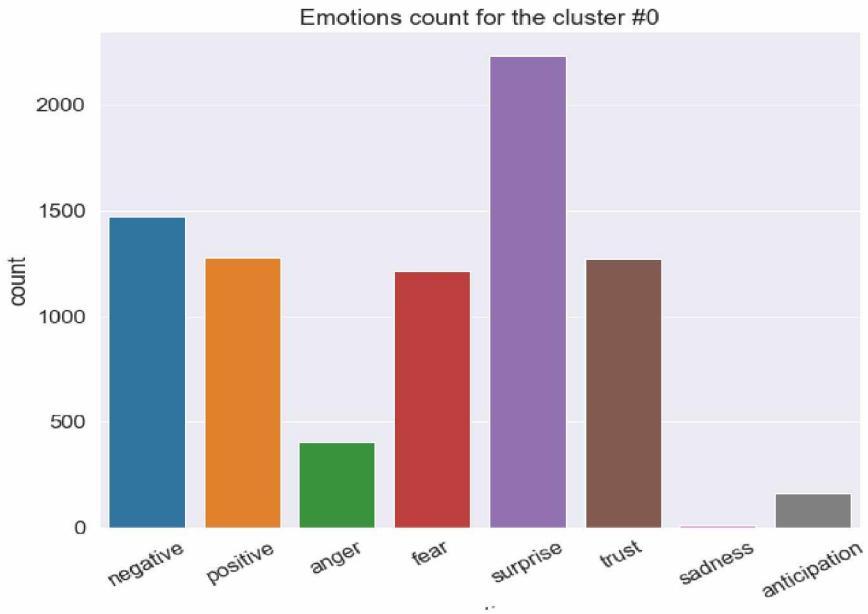
- **Συστάδα #0:** 'amp', 'say', 'get', 'like', 'go', 'people', 'via', 'republicans', 'know', 'lie'
- **Συστάδα #1:** 'york independent', 'via follow', 'trump agonistes', 'three act', 'rise fall', 'follow new', 'american epic', 'agonistes american', 'agonistes', 'fall via'
- **Συστάδα #2:** 'vote', 'vote trump', 'republicans', 'trump vote', 'vote biden', 'make', 'vote vote', 'people', 'get', 'biden'
- **Συστάδα #3:** 'biden', 'joe', 'joe biden', 'donald', 'donald trump', 'debate', 'trump biden', 'biden trump', 'via', 'shirt'
- **Συστάδα #4:** 'hollis', 'medications', 'treatment best', 'multiple fold', 'medications help', 'law pay', 'pay multiple', 'father law', 'hollis father', 'help hollis'
- **Συστάδα #5:** 'win', 'trump win', 'election', 'win election', 'biden', 'vote', 'go', 'say', 'lose', 'republicans'
- **Συστάδα #6:** 'donald', 'donald trump', 'president', 'president donald', 'via', 'click', 'say', 'america', 'great', 'take'
- **Συστάδα #7:** 'wi', 'count', 'mi', 'amp', 'pa', 'able', 'mail', 'stop', 'vote count', 'pa mi'
- **Συστάδα #8:** 'de', 'en', 'el', 'que', 'la', 'por', 'los', 'un', 'las', 'para'
- **Συστάδα #9:** 'trump please', 'please vote', 'anti trump', 'anti', 'please', 'vote', 'lie', 'kill', 'trump lie', 'pandemic'
- **Συστάδα #10:** 'campaign', 'trump campaign', 'end trump', 'trump nightmare', 'nightmare', 'biden', 'end', 'biden campaign', 'donald', 'around'
- **Συστάδα #11:** 'president', 'president trump', 'rally', 'say', 'read', 'campaign', 'lie', 'hold', 'america', 'biden'

Από τις σημαντικότερες λέξεις κάθε συστάδας, καθώς και από μελέτη των tweets που περιέχονται μέσα σε αυτές, εξάγονται οι παρακάτω θεματικές ενότητες:

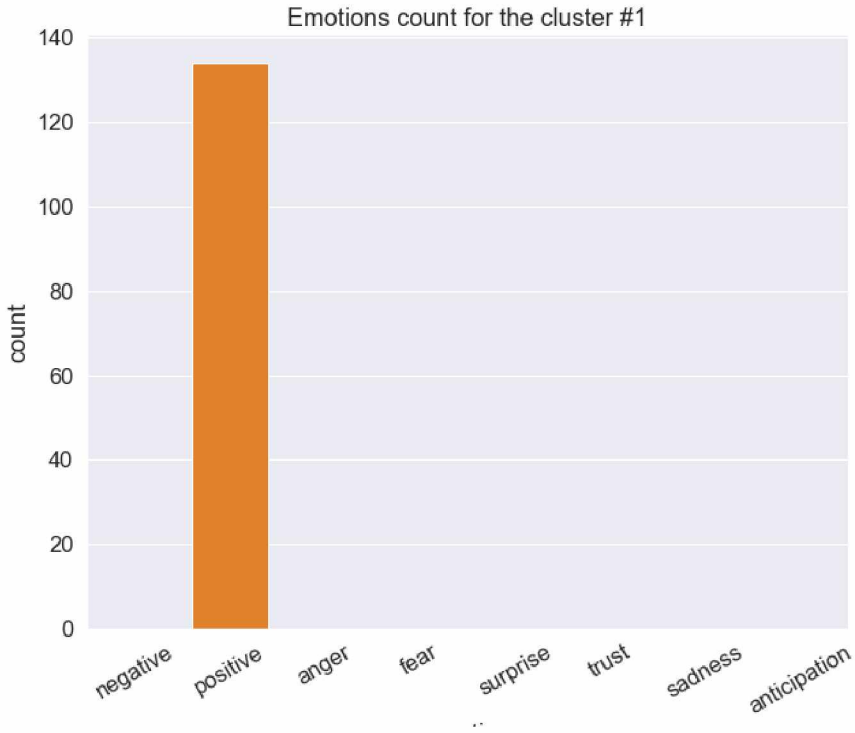
- **Συστάδα #0:** tweets αρνητικής αίσθησης για το Ρεπουμπλικανικό κόμμα
- **Συστάδα #1:** tweets σχετικά με το μυθιστόρημα «Trump agonistes»
- **Συστάδα #2:** -

- **Συστάδα #3:** -
- **Συστάδα #4:** μείωση φόρων και δωρεάν φαρμακευτική περίθαλψη
- **Συστάδα #5:** -
- **Συστάδα #6:** -
- **Συστάδα #7:** -
- **Συστάδα #8:** -
- **Συστάδα #9:** tweets κατά του Τραμπ λόγω της πανδημίας
- **Συστάδα #10:** η «εφιαλτική» καμπάνια του Τραμπ
- **Συστάδα #11:** -

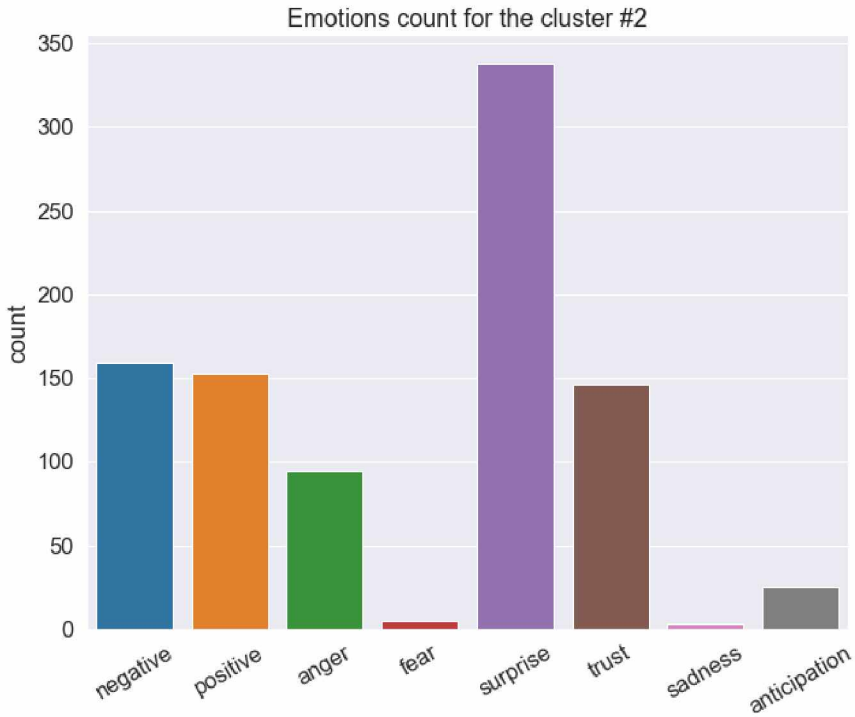
Για τις συστάδες 2, 3, 5, 6 και 11 η διαδικασία εντοπισμού θεματικών ενοτήτων καθίσταται αρκετά δύσκολη, καθώς οι σημαντικότερες λέξεις που τις αποτελούν είναι γενικές. Μία πρόβλεψη η οποία μπορεί να γίνει είναι ότι σε αυτές τις συστάδες γίνεται σύγκριση των δύο υποψηφίων για διάφορα θέματα. Στις συστάδες 7 και 8 παρατηρούνται λέξεις της ισπανικής γλώσσας καθώς κατά την διάρκεια της προεπεξεργασίας ήταν αδύνατο να παραληφθούν. Στα διαγράμματα των εικόνων 6.16 έως 6.27 εφαρμόζεται συναισθηματική ανάλυση για όλες τις συστάδες, από την οποία θα εξαχθούν τα τελικά συμπεράσματα που αφορούν τον Ντόναλντ Τραμπ.



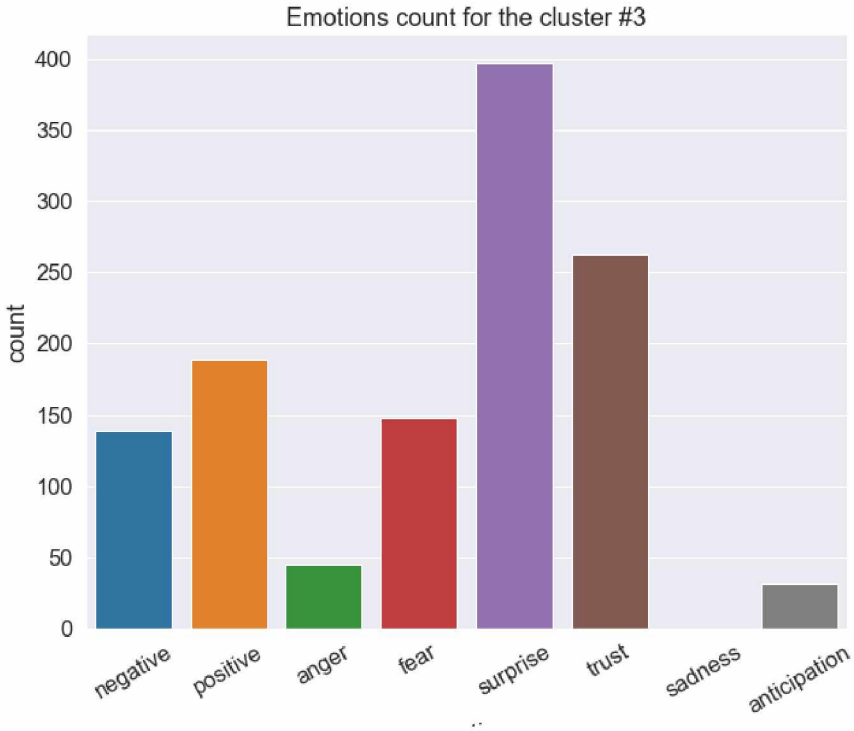
Εικόνα 6.16 Συστάδα 0 (Ρεπουπλικανικόό κόμμα)



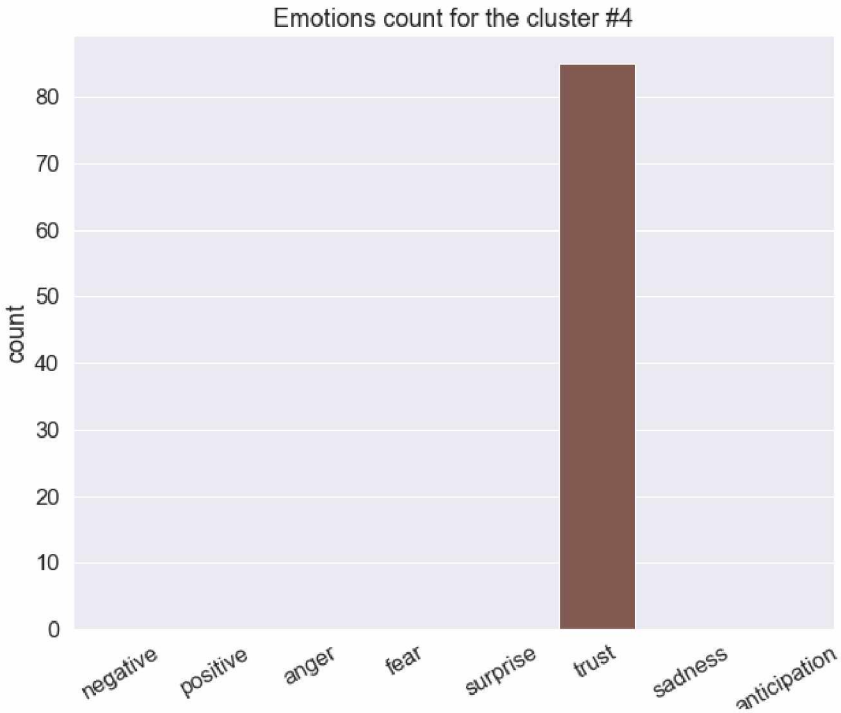
Εικόνα 6.17 Συστάδα 1 (μυθιστόρημα Trump Agonistes)



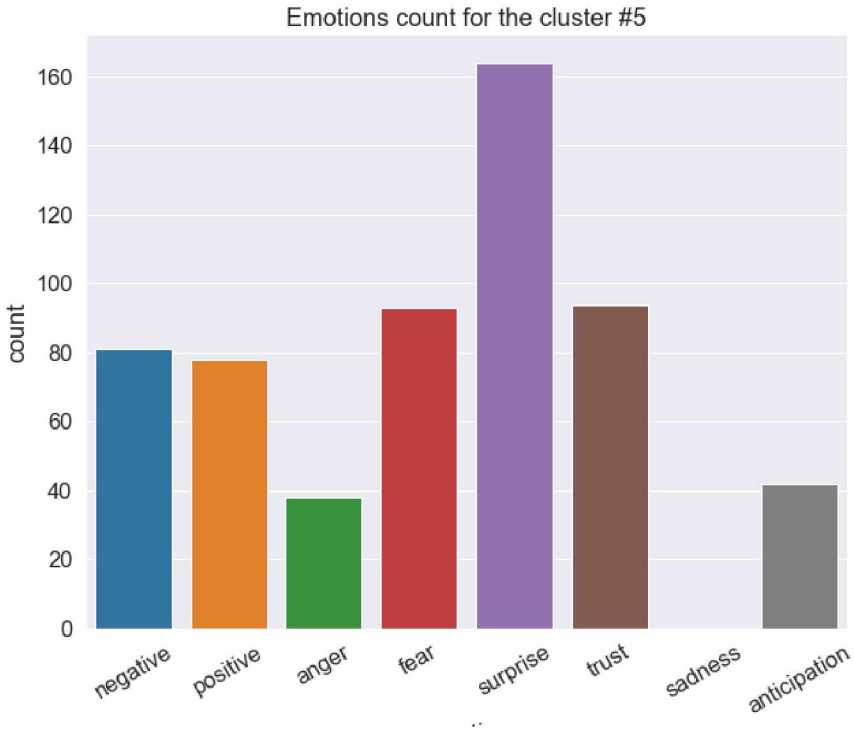
Εικόνα 6.18 Συστάδα 2 (-)



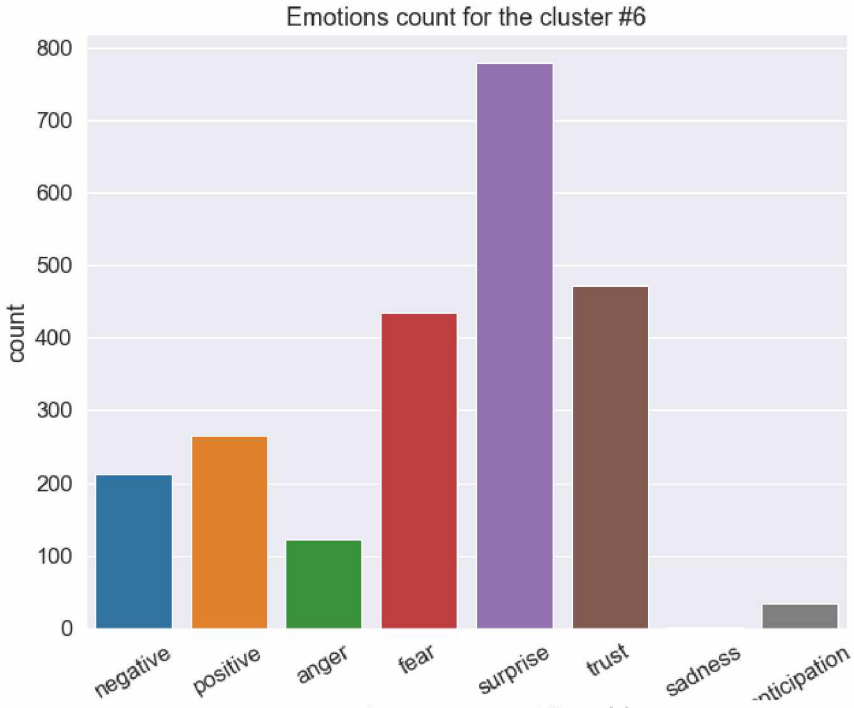
Εικόνα 6.19 Συστάδα 3(-)



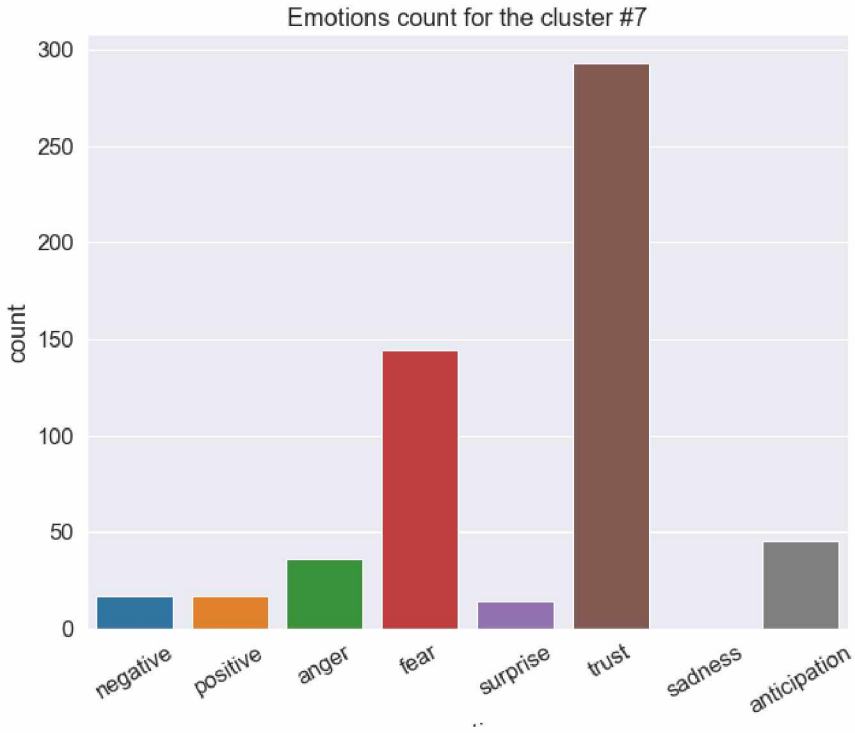
Εικόνα 6.20 Συστάδα 4 (φόροι και φάρμακα)



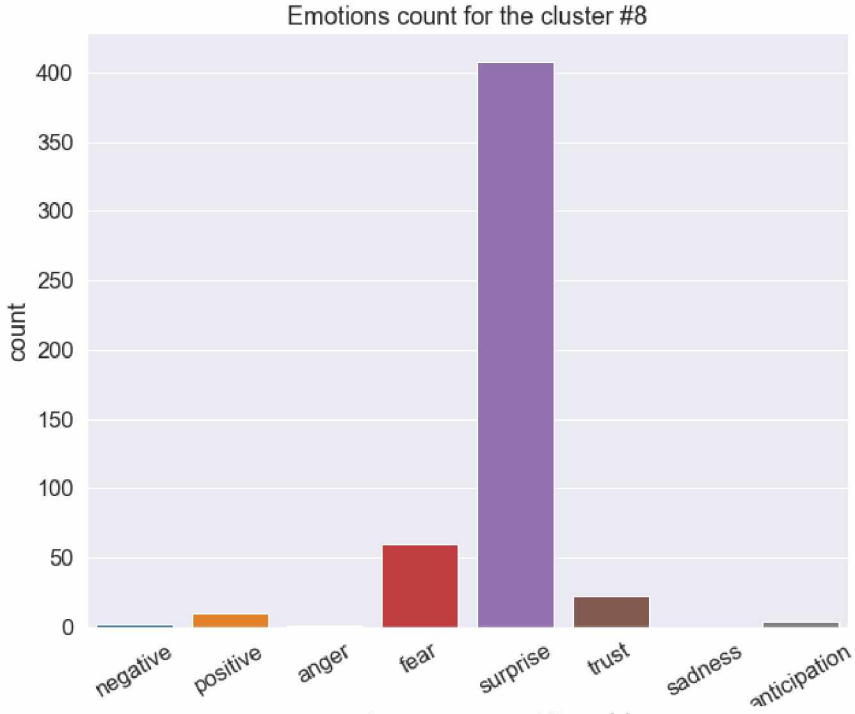
Εικόνα 6.21 Συστάδα 5 (-)



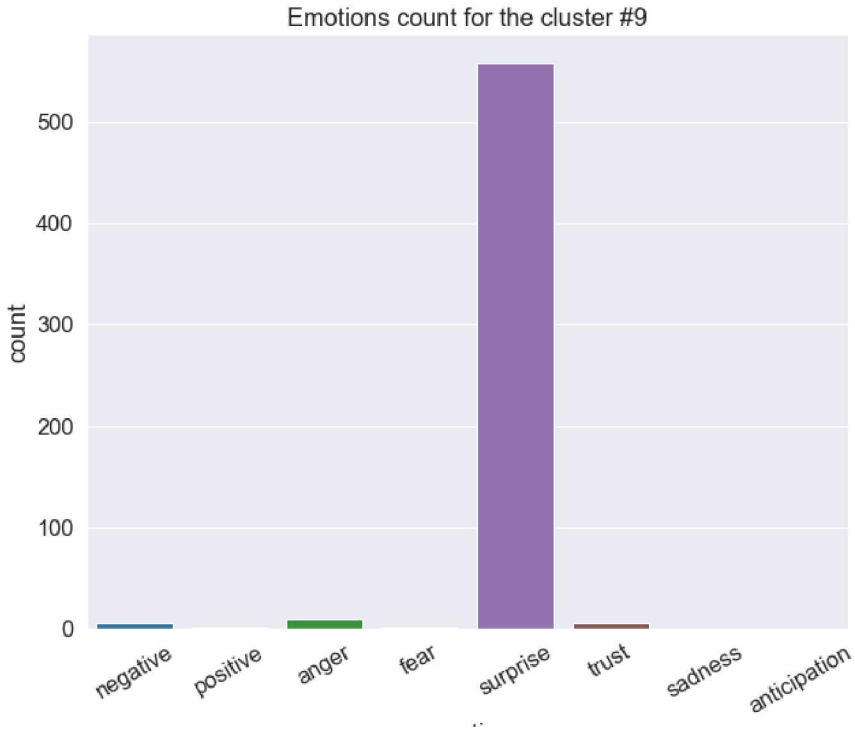
Εικόνα 6.22 Συστάδα 6 (-)



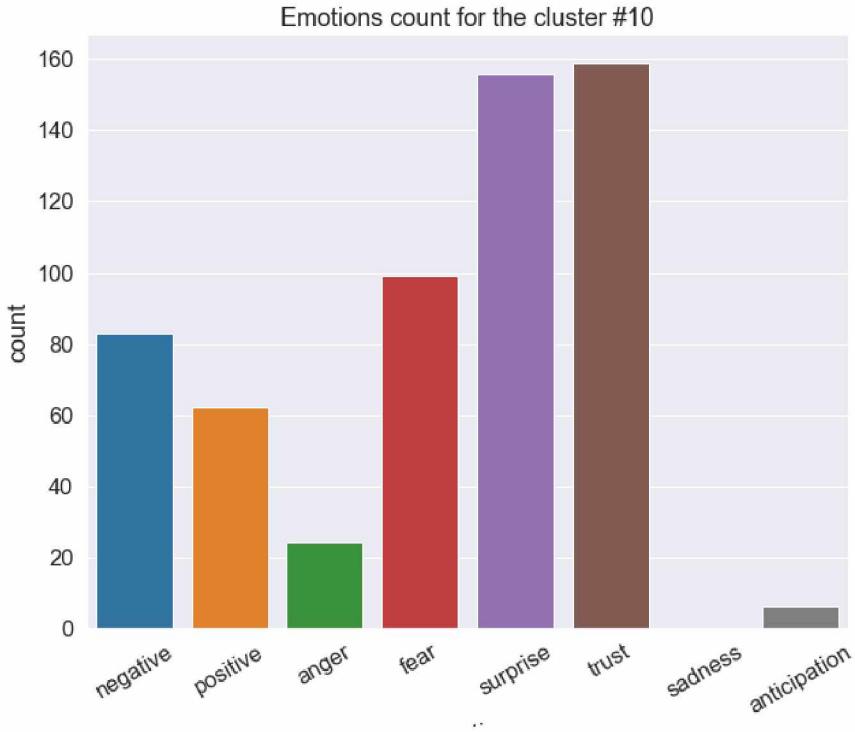
Εικόνα 6.23 Συστάδα 7 (-)



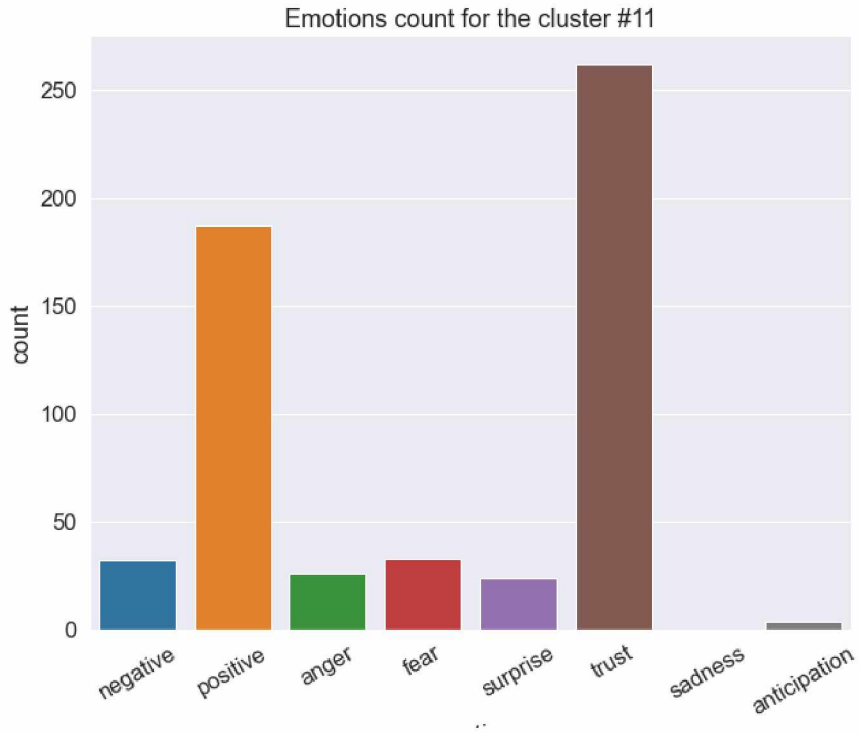
Εικόνα 6.24 Συστάδα 8 (-)



Εικόνα 6.26 Συστάδα 9 (αντί-Τραμπ)



Εικόνα 6.25 Συστάδα 10 (εφιαλτική καμπάνια Τραμπ)



Εικόνα 6.27 Συστάδα 11 (-)

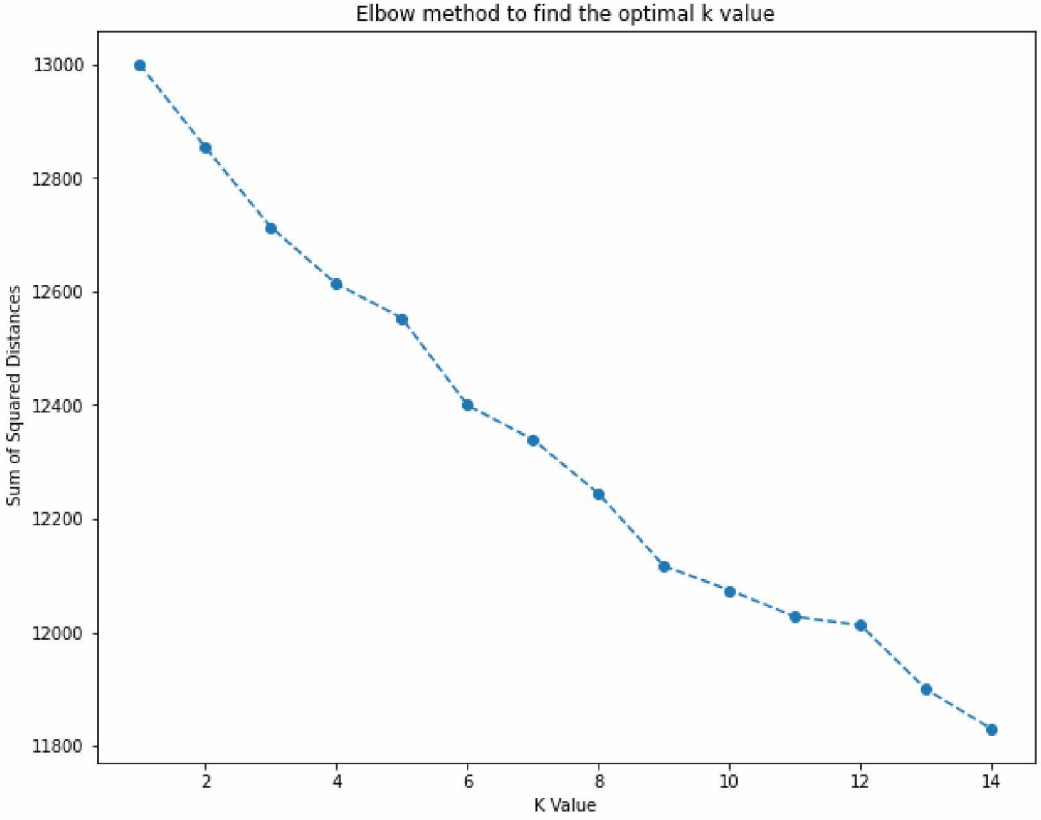
Επομένως, αφού ολοκληρώθηκε η συναισθηματική ανάλυση των συστάδων, μπορούμε να δούμε ότι για τις συστάδες που δεν καταφέραμε να εξάγουμε κάποια θεματική ενότητα, τα συναισθήματα είναι κυρίως έκπληξης με δυσάρεστη χροιά. Αυτό μπορούμε να το διαπιστώσουμε παρατηρώντας τα παραπάνω διαγράμματα, στα οποία υπερισχύει το συναίσθημα της έκπληξης ακολουθούμενο από κάποιο άλλο αρνητικό συναίσθημα.

Ενδιαφέρον παρουσιάζει η συστάδα 1, η οποία αποτελείται εξ ολοκλήρου από το θετικό συναίσθημα. Μία γρήγορη έρευνα στα δεδομένα μπορεί να μας το επιβεβαιώσει, καθώς το μυθιστόρημα «Trump Agonistes» [31], το οποίο δημοσιεύτηκε στις 3 Μαΐου και περιγράφει την δυσλειτουργική προεδρία του Τραμπ, βρίσκει σύμφωνους όλους τους χρήστες. Ακόμη, όπως φαίνεται στο διάγραμμα της συστάδας 4, το οποίο αποτελείται από tweets τα οποία είναι σχετικά με την μείωση των φόρων και την δωρεάν φαρμακευτική περίθαλψη (πιθανώς λόγω της

πανδημίας), κατακλύζεται από σχόλια εμπιστοσύνης προς το πρόσωπο του Ντόναλντ Τραμπ.

Tweets σχετικά με τον Τζο Μπάιντεν

Από το διάγραμμα της εικόνας 6.28, παρατηρείται ότι η επιλογή k=9 για την εφαρμογή του αλγορίθμου K-means είναι η καταλληλότερη.



Εικόνα 6.28 Κανόνας του αγκώνα για tweets σχετικά με τον Μπάιντεν

Μετά την εφαρμογή του αλγορίθμου K-means για τις 9 συστάδες παρουσιάζεται η λίστα με τις πιο σημαντικές λέξεις της κάθε συστάδας:

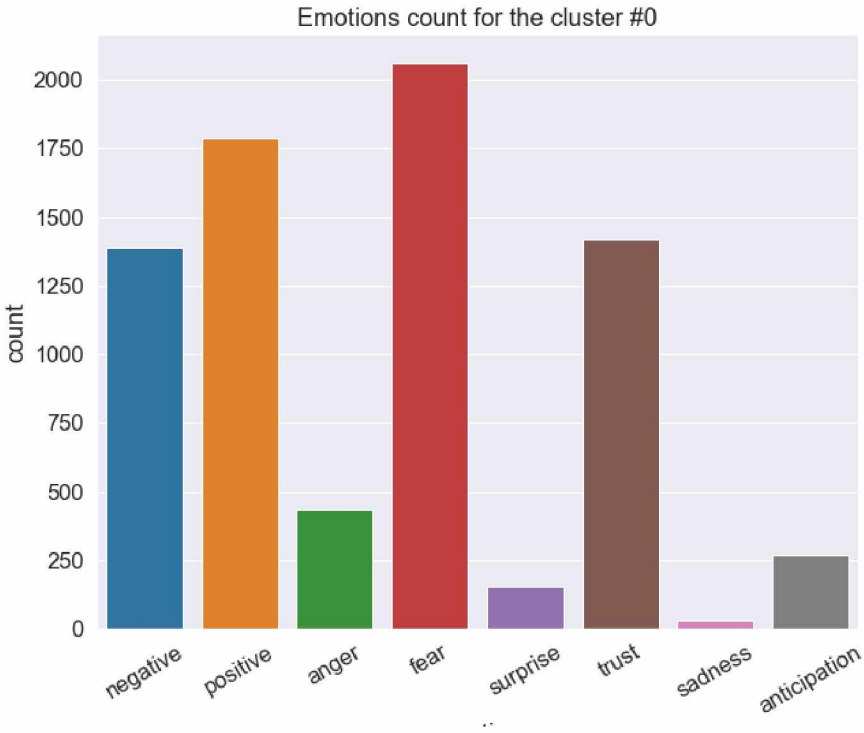
- **Συστάδα #0:** 'amp', 'go', 'via', 'get', 'president', 'democrats', 'win', 'campaign', 'know', 'like'
- **Συστάδα #1:** 'vote', 'vote joe', 'vote biden', 'reason vote', 'reason', 'trump', 'biden vote', 'want', 'vote trump', 'vote vote'
- **Συστάδα #2:** 'hunter', 'hunter biden', 'email', 'business', 'laptop', 'deal', 'via', 'post', 'partner', 'family'

- **Συστάδα #3:** 'sick parkinson', 'parkinson disease', 'biden sick', 'parkinson', 'disease', 'sick', 'watch share', 'share joe', 'please watch', 'share'
- **Συστάδα #4:** 'say', 'biden say', 'trump', 'say joe', 'say biden', 'amp', 'people', 'would', 'like', 'get'
- **Συστάδα #5:** 'click', 'click joes', 'joes picture', 'joes', 'page', 'picture', 'retweet', 'memes click', 'picture take', 'page find'
- **Συστάδα #6:** 'wolfman', 'wolfman joe', 'show day', 'joe show', 'watch wolfman', 'day', 'show', 'watch', 'facts', 'angry'
- **Συστάδα #7:** 'trump', 'donald trump', 'donald', 'president', 'debate', 'trump biden', 'biden trump', 'president trump', 'via', 'election'
- **Συστάδα #8:** 'harris', 'biden harris', 'kamala harris', 'kamala', 'biden kamala', 'vote', 'campaign', 'via', 'vote joe', 'bus'

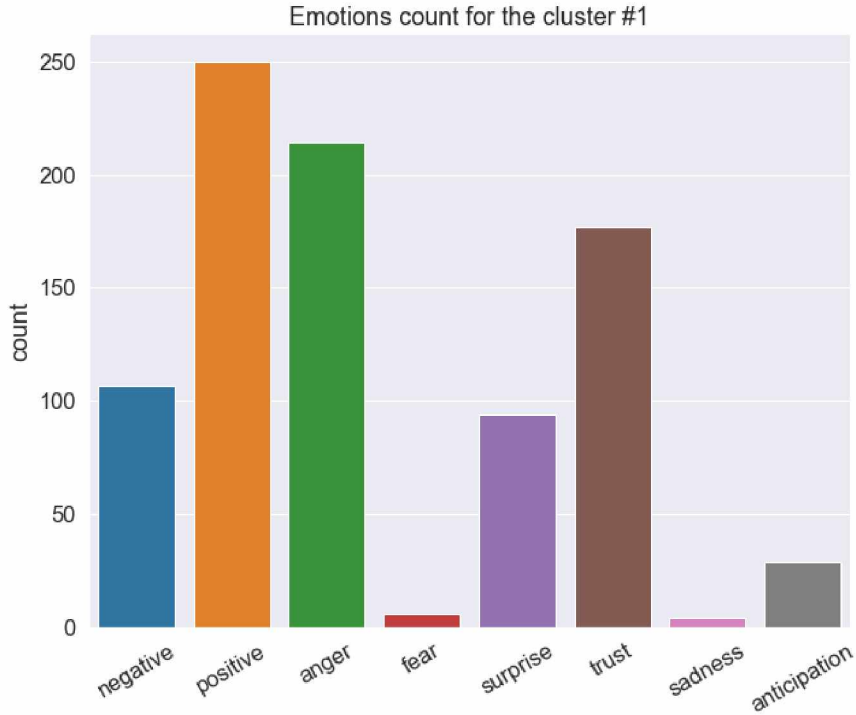
Από τις σημαντικότερες λέξεις κάθε συστάδας, καθώς και από μελέτη των tweets που περιέχονται μέσα σε αυτές, εξάγονται οι παρακάτω θεματικές ενότητες:

- **Συστάδα #0:** Γενικά tweets για πιθανή νίκη του δημοκρατικού κόμματος
- **Συστάδα #1:** Λόγοι να ψηφιστεί ο Τζο Μπάιντεν
- **Συστάδα #2:** tweets που αφορούν τον Hunter Biden, γιο του Joe Biden
- **Συστάδα #3:** Φήμες πως ο Μπάιντεν υποφέρει από την ασθένεια του Parkinson
- **Συστάδα #4:** -
- **Συστάδα #5:** -
- **Συστάδα #6:** Tweets σχετικά με ένα προφίλ στο Twitter με πολιτικό περιεχόμενο
- **Συστάδα #7:** Πολιτική διαμάχη των δύο αντιπάλων
- **Συστάδα #8:** tweets που αφορούν την προεκλογική καμπάνια της αντιπροέδρου των ΗΠΑ, Kamala Harris

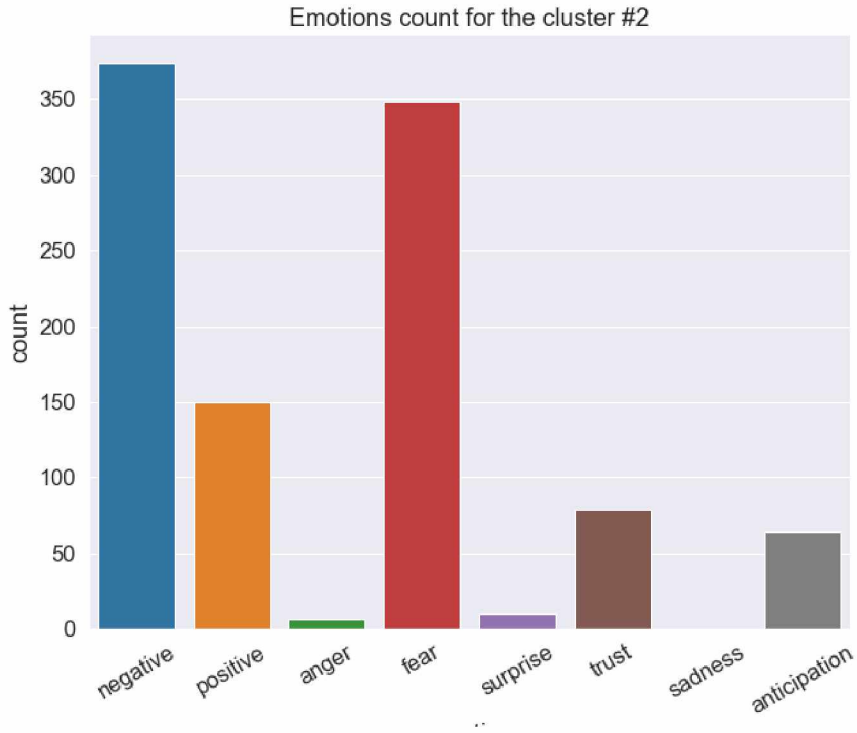
Στα διαγράμματα των εικόνων 6.29 έως 6.37 παρουσιάζεται η συναισθηματική ανάλυση όλων των συστάδων που δημιουργήθηκαν για τα tweets τα οποία είναι σχετικά για τον Τζο Μπάιντεν.



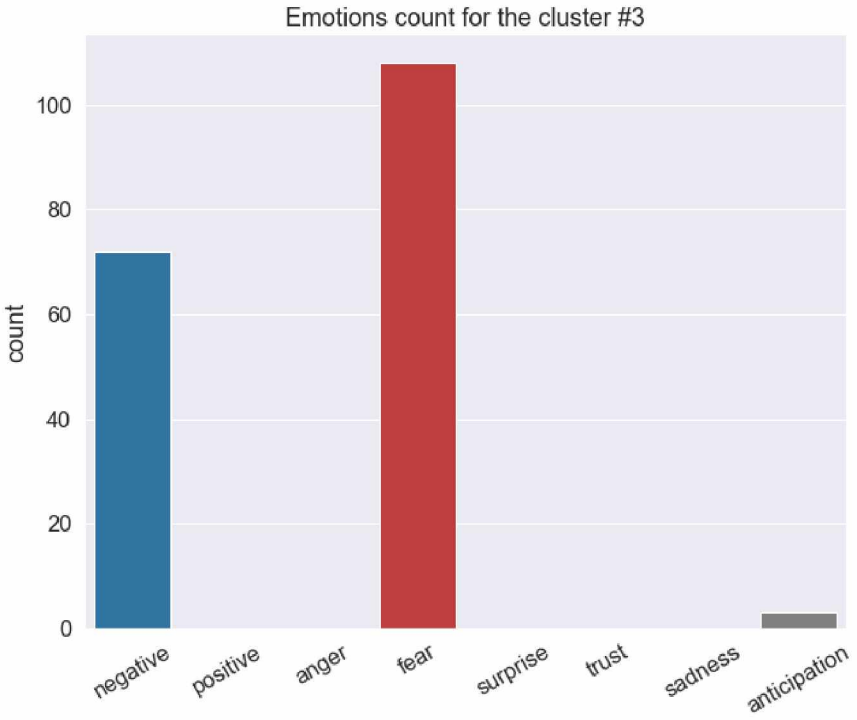
Εικόνα 6.29 Συστάδα 0 (πιθανή νίκη του δημοκρατικού κόμματος)



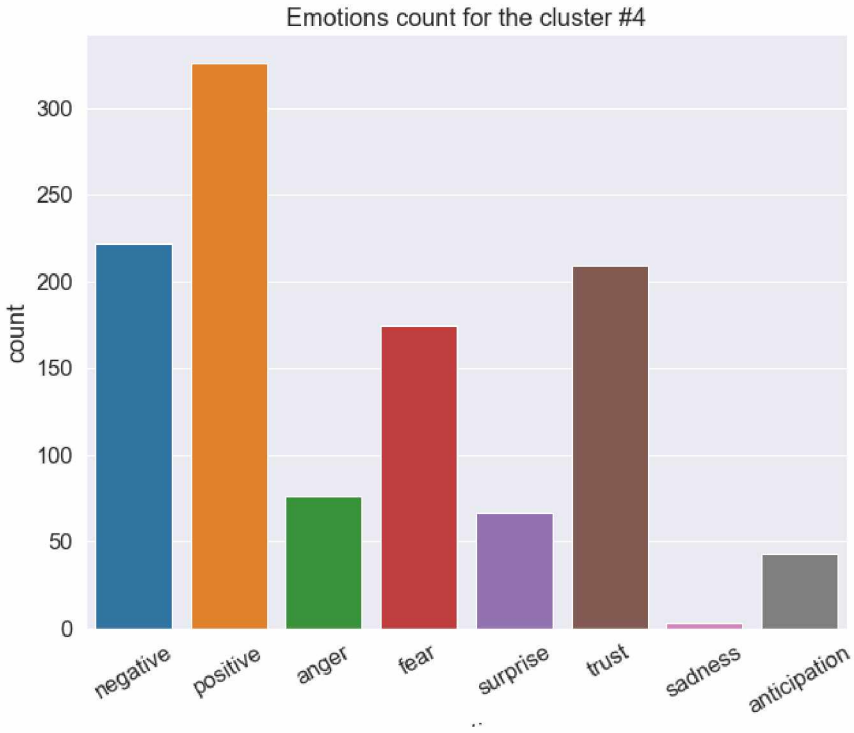
Εικόνα 6.30 Συστάδα 1 (λόγοι ψηφοφοριας για τον Μπάιντεν)



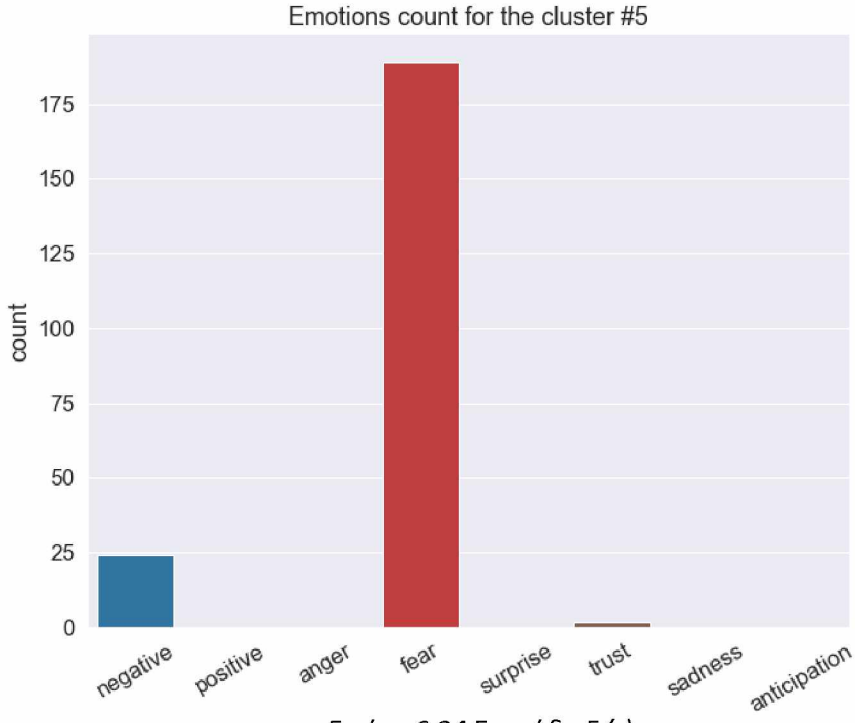
Εικόνα 6.31 Συστάδα 2 (Hunter Biden)



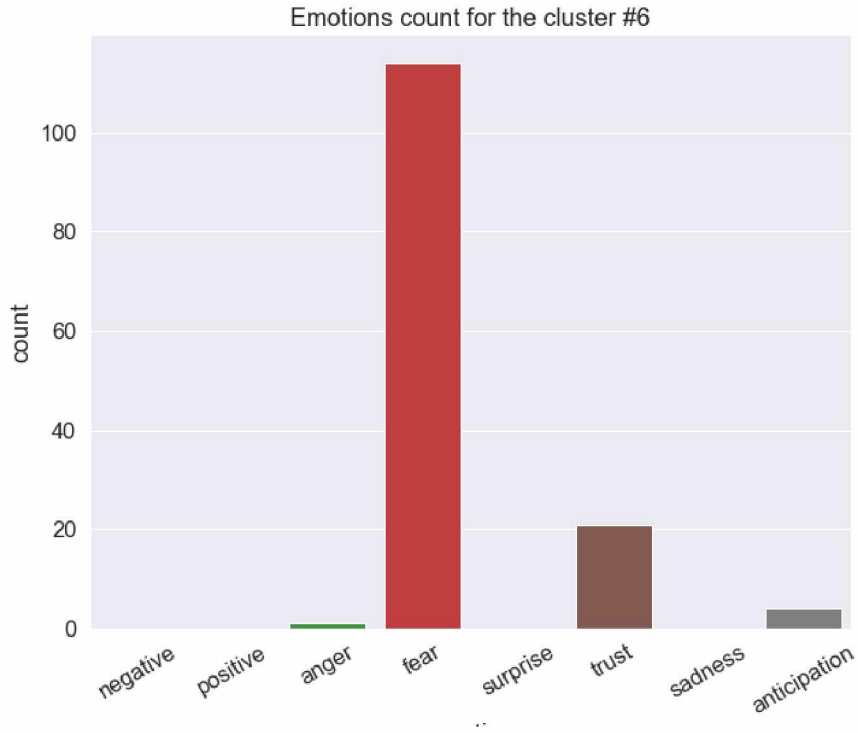
Εικόνα 6.32 Συστάδα 3 (Μπάιντεν και ασθένεια Parkinson)



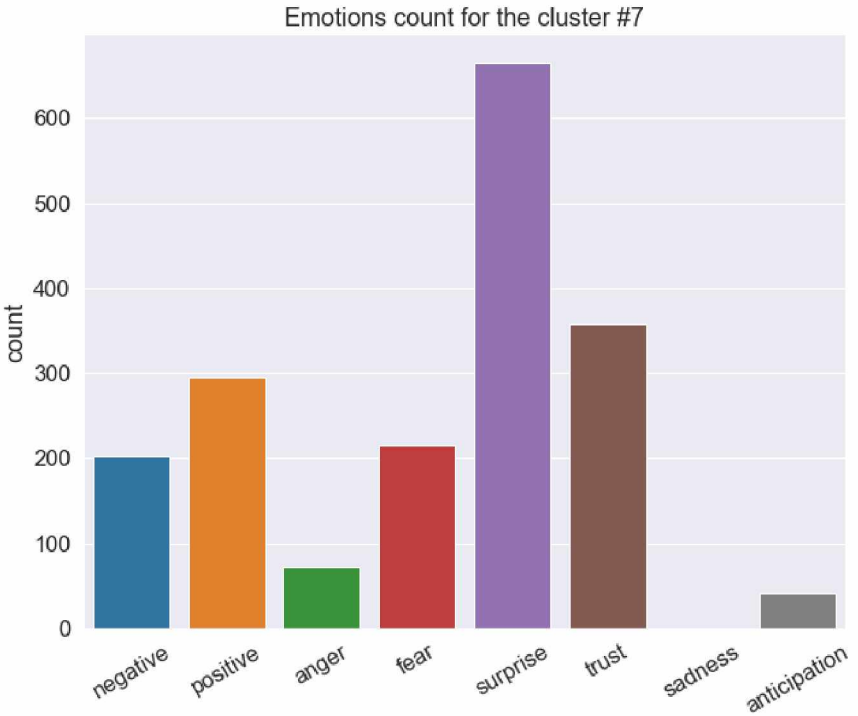
Εικόνα 6.33 Συστάδα 4 (-)



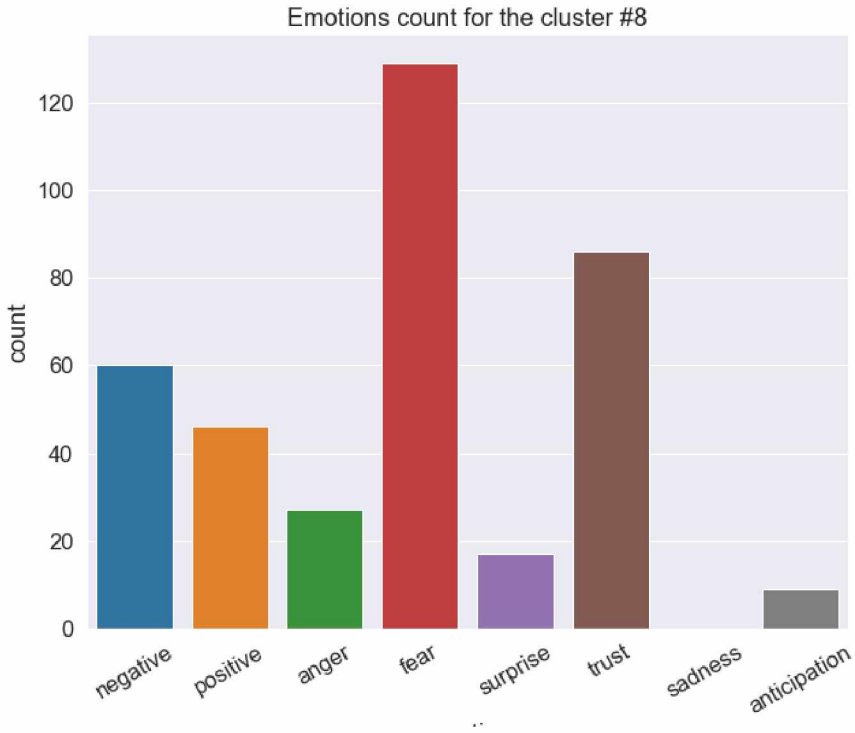
Εικόνα 6.34 Συστάδα 5 (-)



Εικόνα 6.35 Συστάδα 6 (προφίλ με πολιτικό περιεχόμενο)



Εικόνα 6.36 Συστάδα 7 (πολιτική διαμάχη Τραμπ - Μπάιντεν



Εικόνα 6.37 Συστάδα 8 (Kamala Harris)

Η συστάδα 0 η οποία περιέχει τα tweets των χρηστών που σχολιάζουν μία πιθανή νίκη του Τζο Μπάιντεν, χαρακτηρίζεται κυρίως από το συναίσθημα του φόβου, ενώ υπάρχει και μεγάλος αριθμός από θετικά tweets. Η συστάδα 1 περιλαμβάνει σχόλια με τα οποία οι χρήστες παρουσιάζουν κάποιους λόγους για τους οποίους ο Μπάιντεν αξίζει να ψηφιστεί και χαρακτηρίζεται από το θετικό συναίσθημα, αλλά και τον θυμό. Στην συστάδα 3 υπάρχουν φήμες πως ο Μπάιντεν υποφέρει από την νόσο Parkinson. Ο κόσμος δείχνει να είναι φοβισμένος, καθώς εκείνη την περίοδο τίποτα δεν ήταν επιβεβαιωμένο για την κατάσταση υγείας του Μπάιντεν. Όσον αφορά τις δημοσιεύσεις του προφίλ με πολιτικό περιεχόμενο, της συστάδας 6, δείχνει να επικρατεί φόβος, ενώ ο κόσμος μένει έκπληκτος με την πολιτική διαμάχη των δύο αντιπάλων. Τέλος, οι συστάδες 2 (Χάντερ Μπάιντεν) και 8 (Καμάλα Χάρις) χαρακτηρίζονται από tweets αρνητικά και φόβου αντίστοιχα, όπως ακριβώς είδαμε και στην ανάλυση των παρόμοιων συστάδων στα σχόλια της τελευταίας εβδομάδας.

Στην συγκεκριμένη ανάλυση, για τον εντοπισμό θεματικών ενοτήτων, εφαρμόστηκε, και, ο αλγόριθμος Latent Dirichlet Allocation (LDA), ο οποίος στο σύνολο δεδομένων που περιέχει tweets της τελευταίας εβδομάδας των εκλογών λειτούργησε με παρόμοιο τρόπο με τον αλγόριθμο των K-Means. Στο δεύτερο σύνολο δεδομένων, το οποίο περιέχει tweets του τελευταίου μήνα, οι λέξεις των συστάδων που επέδειξε ήταν εντελώς διαφορετικές μεταξύ τους, με αποτέλεσμα να καταφέρουμε να αναδείξουμε ελάχιστες θεματικές ενότητες.

Επομένως, σε αυτή την εργασία, αποφασίστηκε να χρησιμοποιηθεί ο αλγόριθμος των K-Means. Η επιλογή του αριθμού των συστάδων εξετάστηκε με δύο τρόπους. Ο πρώτος ήταν η χρήση του κανόνα του αγκώνα, ο οποίος τελικά επιλέχθηκε για την ανάλυση μας. Η δεύτερη μέθοδος ήταν η βαθμολογία Silhouette, η οποία χρησιμοποιείται για την αξιολόγηση των συστάδων που δημιουργούνται. Ο συγκεκριμένος τρόπος φάνηκε να μην λειτουργεί αποδοτικά, καθώς ανέδειξε μεγάλο αριθμό συστάδων με λέξεις που επαναλαμβάνονταν πολλές φορές σε διαφορετικές συστάδες.

ΚΕΦΑΛΑΙΟ 7

ΣΥΜΠΕΡΑΣΜΑΤΑ

Σκοπός της συγκεκριμένης εργασίας ήταν μέσω της εφαρμογής της συναισθηματικής ανάλυσης στα δεδομένα του Twitter να γίνει πρόβλεψη του εκλογικού αποτελέσματος των τελευταίων εκλογών που έλαβαν χώρα στις Ηνωμένες Πολιτείες Αμερικής. Πέρα από την πρόβλεψη, έγινε σύγκριση των αποτελεσμάτων την ανάλυσης με τα πραγματικά αποτελέσματα, ώστε να διαπιστώσουμε κατά πόσο τα μέσα κοινωνικής δικτύωσης παρέχουν αξιόπιστες και έγκυρες πληροφορίες για παρόμοια πολιτικά γεγονότα.

Αρχικά, στο κεφάλαιο 4, γίνεται παρουσίαση των δεδομένων που έχουμε στην διάθεση μας καθώς και προεπεξεργασία φυσικής γλώσσας των δεδομένων κειμένου. Στη συνέχεια, εφαρμόζεται συναισθηματική ανάλυση με την χρήση του λεξικού VADER της βιβλιοθήκης NLTK της Python και μέσω διαγραμμάτων παρουσιάζεται η δημοτικότητα των δύο κύριων υποψηφίων, Ντόναλντ Τραμπ και Τζο Μπάιντεν. Τέλος, με την χρήση μιας βελτιωμένης μαθηματικής φόρμουλας γίνεται η πρόβλεψη του εκλογικού αποτελέσματος, το οποίο παρόλο που δεν συνάδει με το πραγματικό, καταφέρνει να προσεγγίσει με σημαντική ακρίβεια την διαφορά μεταξύ των υποψηφίων, όπως αυτή καθορίστηκε με το πέρας των εκλογών. Για τον έλεγχο της αξιοπιστίας της μαθηματικής φόρμουλας, έγινε εφαρμογή της σε δεδομένα που συλλέχθηκαν για τον ίδιο σκοπό στις αντίστοιχες εκλογές του 2016, μεταξύ Χίλαρι Κλίντον και Μπαράκ Ομπάμα. Σε αυτή την περίπτωση, παρατηρήθηκε ότι το ποσοστό που αποδόθηκε στην Χίλαρι Κλίντον ήταν σχεδόν ίσο με αυτό που κατάφερε να αποκομίσει στην πραγματικότητα. Είναι αξιοσημείωτο ότι και για τις δύο εκλογικές αναμετρήσεις ο νικητής αναδείχθηκε σωστά.

Στο κεφάλαιο 5 γίνεται εφαρμογή αλγορίθμων κατηγοριοποίησης της μηχανικής μάθησης με στόχο να εντοπιστεί το μοντέλο που κατηγοριοποιεί τα διαθέσιμα δεδομένα με τον καλύτερο τρόπο στο αντίστοιχο συναίσθημα. Στη συνέχεια, γίνεται βελτίωση των υπέρ-παραμέτρων των δύο επικρατέστερων αλγορίθμων, βάσει του

μέτρου F1-score, όπου τελικά διαπιστώθηκε ότι ο αλγόριθμος Μηχανών Διανυσματικής Υποστήριξης καταφέρνει να προβλέψει σωστά το 91% των δεδομένων κειμένου. Τέλος, με την χρήση του αποδοτικότερου μοντέλου γίνεται πρόβλεψη των συναισθημάτων σε νέα δεδομένα, ενώ παράλληλα εφαρμόζεται και συναισθηματική ανάλυση με τη χρήση του VADER λεξικού. Αυτές οι δύο διαφορετικές τεχνικές πετυχαίνουν σχεδόν ίδια αποτελέσματα, πράγμα που σημαίνει ότι η εκπαίδευση του μοντέλου μας λειτούργησε στον βέλτιστο βαθμό.

Στο κεφάλαιο 6 γίνεται εφαρμογή του αλγορίθμου συσταδοποίησης K-means, με σκοπό τον εντοπισμό θεματικών ενοτήτων που έπαιξαν ρόλο στην εξαγωγή του εκλογικού αποτελέσματος. Για τον σκοπό αυτό εξετάζονται δύο διαφορετικά σύνολα δεδομένων. Το πρώτο περιέχει tweets τα οποία συλλέχθηκαν την προηγούμενη βδομάδα των εκλογών, ενώ το δεύτερο περιέχει tweets σε διάρκεια ενός μήνα πριν από την διεξαγωγή των εκλογών. Στη συνέχεια εφαρμόστηκε συναισθηματική ανάλυση του NRC λεξικού. Στο πρώτο σύνολο δεδομένων το συναίσθημα που κυριάρχησε ήταν ο φόβος του κόσμου στην ιδέα μιας πιθανής νίκης του Τζο Μπάιντεν. Από την άλλη, θεματικές ενότητες που αφορούσαν τον Ντόναλντ Τραμπ, χαρακτηρίστηκαν ολοκληρωτικά από το αρνητικό συναίσθημα και το συναίσθημα του φόβου, καθώς εκείνη την περίοδο βρισκόμασταν στην κορύφωση της πανδημίας της COVID-19. Στο δεύτερο σύνολο δεδομένων, το οποίο διαχωρίστηκε σε δύο ξεχωριστά σύνολα (ένα για κάθε υποψήφιο), το πρόσωπο του Ντόναλντ Τραμπ χαρακτηρίστηκε με το συναίσθημα της έκπληξης με αρνητική αίσθηση, ενώ επίσης σε αρκετά tweets γινόταν αναφορά σε ένα μυθιστόρημα το οποίο ανέλυε την δυσλειτουργική προεδρία του Τραμπ. Από την άλλη, για τον Μπάιντεν τα σχόλια ήταν κυρίως φόβου, ένω δεν ήταν λίγα εκείνα που ήταν θετικά προς τον ίδιο. Ενδιαφέρον παρουσίασε η συστάδα, η οποία χαρακτηρίζεται από φόβο, όπου γινόταν συζήτηση για τις φημολογίες πως ο Μπάιντεν υποφέρει από την νόσο Parkinson.

ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Η συγκεκριμένη εργασία αποτελεί μια βασική μεθοδολογία για την πρόβλεψη εκλογικών αποτελεσμάτων σε εκλογές παρόμοιας μορφής με αυτές των Ηνωμένων Πολιτειών της Αμερικής. Ως επεκτάσεις αυτής της εργασίας μπορούν να είναι τα ακόλουθα:

1. Πρόβλεψη αποτελέσματος σε διαφορετικές χώρες και σε άλλες γλώσσες, με βάση την χρήση των αντίστοιχων λεξικών κάθε γλώσσας που είναι διαθέσιμα στο διαδίκτυο.
2. Αν υπήρχαν διαθέσιμα λεξικά της ελληνικής γλώσσας, τα οποία να υποστηρίζονται από τους αλγόριθμους συναισθηματικής ανάλυσης της Python, με μικρή τροποποίηση του κώδικα θα βρισκόμασταν σε θέση να προβλέψουμε τα αποτελέσματα των εκλογών της χώρας μας.
3. Μία πιο γενική επέκταση θα ήταν η συλλογή δεδομένων κειμένου και από άλλα μέσα κοινωνικής δικτύωσης, όπως το Facebook, καθώς και η σύγκριση των απόψεων των χρηστών από τις διαφορετικές, αυτές, πλατφόρμες.
4. Αν τα μέσα κοινωνικής δικτύωσης στο μέλλον δώσουν πρόσβαση σε πληροφορίες όπως η ηλικία, το φύλο ή ο τόπος κατοικίας των συμμετεχόντων θα μπορέσουμε εύκολα, με μία μικρή επέκταση της παρούσας εργασίας, να κάνουμε μια πιο διεξοδική μελέτη των παραγόντων που καθορίζουν το αποτέλεσμα μίας εκλογικής αναμέτρησης.

Βιβλιογραφία

[1] Daniel M. Romero, Brendan Meeder, Jon Kleinberg. '' Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter''. https://dl.acm.org/doi/10.1145/1963405.1963503

[2] John Allen Hendricks, Jr. Denton, Robert E., Jody C. Baumgartner, Jenn Burleson Mackay, Jonathan S. Morris, Eric E. Otenyo, Larry Powell, Melissa M. Smith, Nancy Snow, Frederic I. Solop, Brandon C. Waite. '' Communicator-in-Chief: How Barack Obama Used New Media Technology to Win the White House (Lexington Studies in Political Communication)''. https://rowman.com/ISBN/9780739141052/Communicator-in-Chief-How-Barack-Obama-Used-New-Media-Technology-to-Win-the-White-House

[3] Adam Tsakalidis, Symeon Papadopoulos, Alexandra I. Cristea, Ioannis (Yiannis) Kompatsiaris. '' Predicting Elections for Multiple Countries Using Twitter and Polls''. https://www.researchgate.net/publication/273836232_Predicting_Elections_for_Multiple_Countries_Using_Twitter_and_Polls

[4] Ussama Yaqub, Soon Ae Chun, Vijayalakshmi Atluri, Jaideep Vaidya. '' Sentiment based Analysis of Tweets during the US Presidential Elections''. https://www.researchgate.net/publication/317272393_Sentiment_based_Analysis_of_Tweets_during_the_US_Presidential_Elections

[5] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe. '' Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment''. https://ojs.aaai.org/index.php/ICWSM/article/view/14009/13858

[6] Satish Mahadevan Srinivasan, Raghvinder Sangwan, Colin Neill, Tianhai Zu, '' Power of Predictive Analytics: Using Emotion Classification of Twitter Data for

Predicting 2016 US Presidential Elections''.
https://www.thejsms.org/tsmri/index.php/TSMRI/article/view/477

[7] P.-N. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining. Pearson Addison Wesley, Boston, 2006.

[8] J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco, 2011.

[9] '' Understanding the differences between the two main types of machine learning methods'', https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

[10] '' Learn about text mining, which is the practice of analyzing vast collections of textual materials to capture key concepts, trends and hidden relationships.'' https://www.ibm.com/cloud/learn/text-mining

[11] '' All you need to know about text preprocessing for NLP and Machine Learning''. https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html

[12] Anita Kumari Singh, Mogalla Shashi. '' Vectorization of Text Documents for Identifying Unifiable News Articles''. https://pdfs.semanticscholar.org/caf5/b10072c03fc78b4d4a5c007c8e9e1feaa0d4.pdf

[13] '' Natural Language Processing: Text Data Vectorization''. https://medium.com/@paritosh_30025/natural-language-processing-text-data-vectorization-af2520529cf7

[14] '' Quick Introduction to Bag-of-Words (BoW) and TF-IDF for Creating Features from Text''. https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/

[15] ''Sentiment Analysis on Twitter data regarding 2020 US Elections''. https://towardsdatascience.com/sentiment-analysis-on-twitter-data-regarding-2020-us-elections-1de4bedbe866

[16] Andy Januar Wicaksono, Suyoto and Pranowo, "A proposed method for predicting US presidential election by analyzing sentiment in social media," 2016 2nd International Conference on Science in Information Technology (ICSITech), 2016, pp. 276-280, doi: 10.1109/ICSITech.2016.7852647. https://ieeexplore.ieee.org/document/7852647

[17] Tim Hamling, Ankur Agrawal, '' Sentiment Analysis of Tweets to Gain Insights into the 2016 US Election'', Department of Computer Science, Manhattan College, New York https://journals.library.columbia.edu/index.php/cusj/article/view/6359/3023

[18] M. Trupthi, S. Pabboju and G. Narasimha, "Improved feature extraction and classification — Sentiment analysis," 2016 International Conference on Advances in Human Machine Interaction (HMI), 2016, pp. 1-6, doi: 10.1109/HMI.2016.7449189. https://ieeexplore.ieee.org/abstract/document/7449189

[19] '' Measuring Performance: The Confusion Matrix''. https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/

[20] Aditya Mishra, '' Metrics to Evaluate your Machine Learning Algorithm'' https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

[21] H. Zhang, W. Gan and B. Jiang, "Machine Learning and Lexicon Based Methods for Sentiment Classification: A Survey," 2014 11th Web Information System and Application Conference, 2014, pp. 262-265, doi: https://ieeexplore.ieee.org/abstract/document/7058024

[22] '' US Election 2020 Tweets'' https://www.kaggle.com/manchunhui/us-election-2020-tweets

[23] Matanga, Yves. (2017). Analysis of Control Attainment in Endogenous Electroencephalogram Based Brain Computer Interfaces. 10.13140/RG.2.2.10493.05608.

[24] Venkatesh, Ranjitha K. V, '' Classification and Optimization Scheme for Text Data using Machine Learning Naïve Bayes Classifier'', https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8690536

[25] Kibriya A.M., Frank E., Pfahringer B., Holmes G. (2004) Multinomial Naive Bayes for Text Categorization Revisited. In: Webb G.I., Yu X. (eds) AI 2004: Advances in Artificial Intelligence. AI 2004. Lecture Notes in Computer Science, vol 3339. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30549-1_43

[26] '' Support Vector Machine — Introduction to Machine Learning Algorithms'', https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[27] '' All you need to know about decision trees and how to build and optimize decision tree classifier.'', https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

[28] Leif E. Peterson (2009) K-nearest neighbor. Scholarpedia, 4(2):1883. http://www.scholarpedia.org/w/index.php?title=K-nearest_neighbor&action=cite&rev=137311

[29] '' A Complete Guide to the Random Forest Algorithm'', https://builtin.com/data-science/random-forest-algorithm

[30] '' Random Forest Explained [Random Forest explained simply: An easy Introduction to Training, Classification, and Regression]'', https://towardsdatascience.com/random-forest-explained-7eae084f3ebe

[31] Steven Grant, Yvonne Diane Brooks, '' Trump Agonistes: A dystopian novel of conflict between superpowers after Trump's dysfunctional presidency. Paperback – May 3, 2019'',

[32] "Hillary Clinton and Donald Trump Tweets", https://www.kaggle.com/benhamner/clinton-trump-tweets

[33] J. Ramteke, S. Shah, D. Godhia and A. Shaikh, "Election result prediction using Twitter sentiment analysis," 2016 International Conference on Inventive Computation Technologies (ICICT), 2016, pp. 1-5, doi: 10.1109/INVENTIVE.2016.7823280.

[34] A. Tsakalidis, S. Papadopoulos, A. I. Cristea and Y. Kompatsiaris, "Predicting Elections for Multiple Countries Using Twitter and Polls," in IEEE Intelligent Systems, vol. 30, no. 2, pp. 10-17, Mar.-Apr. 2015, doi: 10.1109/MIS.2015.17.

[35] Hasan A, Moin S, Karim A, Shamshirband S. Machine Learning-Based Sentiment Analysis for Twitter Accounts. Mathematical and Computational Applications. 2018; 23(1):11. https://doi.org/10.3390/mca23010011

Παράρτημα

Οι κώδικες και τα δεδομένα που χρησιμοποιήθηκαν για την υλοποίηση της εργασίας βρίσκονται διαθέσιμα **εδώ**, στον προσωπικό μου λογαριασμό στο GitHub.