

Content homophily in Twitter

Ζώγος Σωτήριος, Τριανταλή Δήμητρα

Πανεπιστήμιο Ιωαννίνων
Τμήμα Μηχανικών Η/Υ και Πληροφορικής

<https://sotiriszogog.github.io/Online-Social-Networks-and-Media/>

Περίληψη- Τα μέσα κοινωνικής δικτύωσης όπως είναι το Twitter, το Facebook και άλλα, τα τελευταία χρόνια είναι στο επίκεντρο της προσοχής. Τα συγκεκριμένα κοινωνικά δίκτυα χρησιμοποιούνται από εκατομμύρια χρήστες καθημερινά οι οποίοι τα χρησιμοποιούν κυρίως για επικοινωνία, διαφήμιση και ψυχαγωγία. Πλέον βάσει της χρήσης που τυγχάνουν καθημερινά διαθέτουν μεγάλο όγκο δεδομένων που είναι διαθέσιμα προς όλους τους χρήστες. Στα πλαίσια του Μεταπτυχιακού Προγράμματος με τίτλο ‘Μηχανική Δεδομένων και Υπολογιστικών Συστημάτων’ και συγκεκριμένα στο μάθημα ‘OnlineSocialNetworksandMedia’ επιλέξαμε να μελετήσουμε την ομοιότητα που εμφανίζεται μεταξύ διαφορετικών χρηστών στο Twitter. Επιλέξαμε 4 κατηγορίες από προφίλ (ποδόσφαιρο, πολιτική, ειδήσεις και μουσική) και για την κάθε κατηγορία διαλέξαμε 2 αντιπροσώπους. Στην πρώτη και τέταρτη κατηγορία επιλέξαμε προφίλ που έχουν κοινά χαρακτηριστικά (ίδια ομάδα και ίδιο είδος μουσικής αντίστοιχα), ενώ στη δεύτερη και τρίτη κατηγορία επιλέξαμε αντίπαλα προφίλ (πολιτικοί και ειδησεογραφικοί αντίπαλοι αντίστοιχα). Για καθένα από τα 8 προφίλ δημιουργήσαμε 3 αρχεία. Στο πρώτο αποθηκεύαμε τα ονόματα των ακολούθων του, στο δεύτερο τα hashtags που χρησιμοποίησαν και στο τρίτο τα διαφορετικά hashtags. Στη συνέχεια, συγκρίναμε την ομοιότητα τόσο μεταξύ προφίλ που ανήκουν στην ίδια κατηγορία, όσο και μεταξύ διαφορετικών κατηγοριών. Τα αποτελέσματα έδειξαν ότι συνήθως τα προφίλ που ανήκουν στην ίδια κατηγορία εμφανίζουν το μεγαλύτερο jaccard_similarity.

I. ΕΙΣΑΓΩΓΗ

Τα τελευταία 10 χρόνια η δραστηριότητα των χρηστών στα κοινωνικά δίκτυα έχει εκτοξευθεί σε υψηλά επίπεδα, με τον όγκο των δεδομένων να γίνεται τεράστιος. Οι ενεργοί χρήστες σε κάθε χώρα αυξάνονται συνεχώς, ενώ παράλληλα αυξάνεται και η συχνότητα με την οποία χρησιμοποιούν τις εφαρμογές κοινωνικής δικτύωσης, καθώς και ο χρόνος που παραμένουν συνδεδεμένοι σε αυτές. Η αυξανόμενη απασχόληση με το φαινόμενο της κοινωνικής δικτύωσης δημιουργεί πρόσφορο έδαφος για τη δημιουργία νέων κοινωνικών δικτύων, καθένα από τα οποία παρέχει πληθώρα εργαλείων έκφρασης και επικοινωνίας στους χρήστες του. Ενδεικτικά, οι χρήστες μπορούν να συμμετέχουν σε online κοινότητες και ομάδες συζητήσεων, να εγγράφονται σε λίστες ενημέρωσης ή ακόμα και να γράφουν οι ίδιοι κείμενα σε προσωπικούς λογαριασμούς και ιστολόγια, με αποτέλεσμα να δημιουργείται ένας μεγάλος όγκος πληροφορίας, που δεν είναι πάντα δυνατό να παρατηρηθεί και να μελετηθεί.

Ένα από τα πιο δημοφιλή σύγχρονα κοινωνικά δίκτυα είναι το Twitter, το οποίο αποτελεί το πιο διαδεδομένο μέσο μικρών αναρτήσεων. Χαρακτηριστικά αναφέρεται ότι, οι εγγεγραμμένοι σε αυτό χρήστες ξεπερνούν το 1 δισεκατομμύριο παγκοσμίως. Αυτό καθιστά το συγκεκριμένο κοινωνικό δίκτυο ένα από τα μεγαλύτερα που υπάρχουν σήμερα και, ίσως, το πιο άμεσο και πιο εύκολα προσβάσιμο δίκτυο επικοινωνίας και ενημέρωσης. Τα χαρακτηριστικά αυτά έχουν επιφέρει μεγάλες αλλαγές στην καθημερινότητα πολλών χρηστών, με αποτέλεσμα να υπάρχει μεγάλη ανάγκη, αλλά και πρόσφορο πεδίο έρευνας στο κομμάτι της αποδοτικής και γρήγορης αξιοποίησης των δεδομένων για νέες υπηρεσίες και γνώση προς το χρήστη. Για τους λόγους αυτούς, η εργασία πραγματεύεται τη συλλογή και διαχείριση των πληροφοριών που παρέχει το Twitter, έχοντας ως αντικειμενικό σκοπό την εξεύρεση γνώσης και πληροφοριών, οι οποίες δεν είναι άμεσα ορατές από έναν απλό χρήστη.

Το παρόν paper αποτελείται από τις εξής ενότητες:

- 1) Περίληψη
- 2) Εισαγωγή
- 3) Ορισμοί εννοιών εργασίας
- 4) Συλλογή δεδομένων
- 5) Ανάλυση δεδομένων
- 6) Συμπεράσματα και προτάσεις
- 7) Βιβλιογραφία

II. ΟΡΙΣΜΟΙ ΕΝΝΟΙΩΝ ΕΡΓΑΣΙΑΣ

2.1. Ορισμός Κοινωνικού Δικτύου

Ο όρος κοινωνικό δίκτυο χρησιμοποιείται για να περιγράψει υπηρεσίες παγκόσμιου ιστού οι οποίες επιτρέπουν στους χρήστες τους να δημιουργήσουν προφίλ, δημόσια ή ημι-δημόσια, ούτως ώστε να επικοινωνούν με άλλους χρήστες του ίδιου δικτύου. [2] Ένα κοινωνικό δίκτυο μπορεί να αναπαρασταθεί με αρκετούς τρόπους. Ένας από τους περισσότερο χρήσιμους είναι μέσω της θεωρίας των γράφων. Στη μαθηματική θεωρία των γράφων, ένας γράφος αποτελείται από κόμβους (ή κορυφές), κάποιιοι από τους οποίους (ενδεχομένως όλοι) συνδέονται μεταξύ τους. Στην περίπτωση που όλοι οι κόμβοι συνδέονται μεταξύ τους έχουμε έναν πλήρη γράφο. Οι συνδέσεις (links) μεταξύ των κόμβων, όταν υπάρχουν, μπορεί να είναι είτε τόξα ή γραμμές, αναλόγως του αν υπάρχει ή δεν υπάρχει κάποια κατεύθυνση στις συνδέσεις μεταξύ των κόμβων του γράφου. Έτσι, σύμφωνα με τη θεωρία των γράφων, ένα κοινωνικό δίκτυο είναι μία κοινωνική δομή που αποτελείται από κόμβους, οι οποίοι συνδέονται με έναν ή περισσότερους τύπους

αλληλεξάρτησης, όπως φιλία, συγγένεια, κοινά ενδιαφέροντα, οικονομικές συναλλαγές, αντιπάθεια, κοινές πεποιθήσεις, γνώση ή κύρος. Οι κόμβοι ανταλλάσσουν μεταξύ τους πόρους, οι οποίοι τους κρατούν συνδεδεμένους. Οι πόροι μπορεί να περιλαμβάνουν πληροφορίες, δεδομένα, κοινωνική υποστήριξη, ή οικονομική υποστήριξη. Κάθε είδος ανταλλασσόμενου πόρου θεωρείται σαν μια σχέση στο κοινωνικό δίκτυο, και τα άτομα που διατηρούν μια τέτοια σχέση λέγεται ότι διατηρούν έναν δεσμό. [4]

2.2. Ορισμός Μέσων Κοινωνικής Δικτύωσης

Ανατρέχοντας στη σχετική βιβλιογραφία, προκύπτει μια πληθώρα ορισμών για τα μέσα κοινωνικής δικτύωσης, γεγονός που καταδεικνύει το τεράστιο ενδιαφέρον της ακαδημαϊκής κοινότητας για το φαινόμενο αυτό. Ένας βασικός ορισμός είναι των Kaplan και Haenlein, σύμφωνα με τον οποίο: “Τα μέσα κοινωνικής δικτύωσης ορίζονται σαν ένα σύνολο από διαδικτυακές εφαρμογές που βασίζονται στα ιδεολογικά και τεχνολογικά θεμέλια του Web 2.0 και επιτρέπουν την δημιουργία και την ανταλλαγή περιεχομένου από τους χρήστες”. [3]

Τα μέσα κοινωνικής δικτύωσης διέπονται από πέντε κοινά βασικά χαρακτηριστικά, τα οποία είναι η συμμετοχή, η διαφάνεια, η συνομιλία, η κοινότητα και η συνεκτικότητα. [1]

Συμμετοχή (Participation): Ένα από τα πιο διακριτικά χαρακτηριστικά των μέσων κοινωνικής δικτύωσης είναι η συμμετοχική τους φύση, που δίνει στα ενδιαφερόμενα μέρη την ευκαιρία να αλληλεπιδράσουν μεταξύ τους. Με την ενθάρρυνση της συμμετοχής και τη δημιουργία σχολίων από όσους ενδιαφέρονται, τα μέσα κοινωνικής δικτύωσης θολώνουν τα όρια μεταξύ των μέσων ενημέρωσης και του κοινού.

Ανοικτότητα (Openness): Τα μέσα κοινωνικής δικτύωσης είναι ανοιχτά σε ανατροφοδότηση και συμμετοχή από τους χρήστες, ενώ υπάρχουν ελάχιστα εμπόδια στην πρόσβαση και στην χρήση του περιεχομένου. Η διαφάνεια ενισχύεται από την ίδια τη φιλοσοφία των μέσων κοινωνικής δικτύωσης και τη διαθεσιμότητα μηχανισμών για τη δημιουργία και την κοινή χρήση περιεχομένου.

Συνομιλία (Conversationality): Σε σύγκριση με τα παραδοσιακά μέσα ενημέρωσης που απλώς μεταδίδουν το περιεχόμενο στο ακροατήριο (συνομιλία μονής κατεύθυνσης), τα μέσα κοινωνικής δικτύωσης επιτρέπουν την αμφίδρομη επικοινωνία (συνομιλία διπλής κατεύθυνσης).

Κοινότητα (Community): Τα μέσα κοινωνικής δικτύωσης επιτρέπουν στα άτομα και τους οργανισμούς που συμμετέχουν σε αυτά να επικοινωνούν μόνο με τους ανθρώπους με τους οποίους θέλουν να συνδέονται. Με άλλα λόγια, επιτρέπουν την εύκολη και άμεση δημιουργία κοινοτήτων που μοιράζονται κοινά ενδιαφέροντα, όπως, για παράδειγμα την αγάπη τους για τη μαγειρική ή τη φωτογραφία. Στο σημείο αυτό πρέπει να διευκρινιστεί ότι, η δημιουργία κοινοτήτων δεν είναι αυτός καθαυτός ο σκοπός των μέσων κοινωνικής δικτύωσης, όσο η επικοινωνία των χρηστών που εμφανίζουν κοινά ενδιαφέροντα για μια δεδομένη χρονική περίοδο. Για παράδειγμα, οι τηλεθεατές που δημοσιεύουν σχόλια σχετικά με μια είδηση ή ένα βίντεο έχουν κάτι κοινό, αλλά δεν είναι απαραίτητο να θεωρηθούν ότι ανήκουν σε μια κοινότητα.

Συνεκτικότητα (Connectedness): Τα περισσότερα μέσα κοινωνικής δικτύωσης αναπτύσσουν την συνεκτικότητα τους κάνοντας χρήση συνδέσεων προς άλλες ιστοσελίδες, πόρους και

ανθρώπους. Η συνεκτικότητα μπορεί να θεωρηθεί και ως ένα πρότυπο διαπροσωπικών σχέσεων που χαρακτηρίζονται από εμπιστοσύνη. Η συνεκτικότητα συνδέεται επίσης με την εγγύτητα και την ταύτιση με τους άλλους. Υπό αυτή την έννοια, τα άτομα με υψηλή συνεκτικότητα αισθάνονται κοντά με τους άλλους και μπορούν να ταυτιστούν και να συνεργαστούν, ενώ οι χρήστες με χαμηλή συνεκτικότητα αισθάνονται απομακρυσμένοι από τους άλλους και δεν είναι κατάλληλοι για τις κοινωνικές σχέσεις.

2.3. Twitter

Το twitter είναι ένα από τα δημοφιλέστερα μέσα κοινωνικής δικτύωσης παγκόσμια. Πιο συγκεκριμένα είναι μια υπηρεσία η οποία δίνει την δυνατότητα στους χρήστες της να κρατούν επαφή, να ενημερώνονται για τις καθημερινές δραστηριότητες των φίλων τους και των ανθρώπων που τους ενδιαφέρει. Επίσης όπως και άλλα κοινωνικά μέσα δικτύωσης δίνει την δυνατότητα σε χρήστες να το χρησιμοποιούν για εμπορικούς σκοπούς.

Στο Twitter κάθε χρήστης δημιουργεί δωρεάν το δικό του προφίλ, το οποίο αποτελείται από εικόνες και διάφορες πληροφορίες, όπως μίνι βιογραφικό, περιοχή διαμονής, username κλπ. Από προεπιλογή, κάθε λογαριασμός δημοσιεύει δημόσια, με οποιονδήποτε χρήστη του διαδικτύου να μπορεί να προβάλει τα μηνύματα του. Βέβαια, υπάρχει και επιλογή για ιδιωτικά tweets, αν και δε συνιστάται για την καλύτερη εμπειρία χρήσης της συγκεκριμένης εφαρμογής.

Στο Twitter δε συναντάει κανείς φιλίες, αλλά ακόλουθους, ή αλλιώς followers. Δεν είναι απαραίτητο για δύο χρήστες να ακολουθεί ο ένας τον άλλον. Κάθε χρήστης ακολουθεί όποιους θέλει και ακολουθείται από οποιονδήποτε. Αυτό σημαίνει πως ο καθένας διαμορφώνει την αρχική του σελίδα με περιεχόμενο το οποίο επιθυμεί.

Οι κυριότερες λειτουργίες του Twitter είναι το retweet το οποίο χρησιμοποιείται για κοινοποίηση ενός tweet, το favorite το οποίο χρησιμοποιείται περισσότερο ως το like του Facebook, και το reply το οποίο χρησιμοποιείται για απάντηση σε ένα tweet. Μία από τις σημαντικότερες λειτουργίες των Κοινωνικών Δικτύων, τα hashtags, καθιερώθηκαν από το Twitter και χρησιμοποιούνται σε μεγάλο βαθμό για συγκέντρωση όλων των tweets γύρω από μια συζήτηση.

2.4. Εξόρυξη Δεδομένων

Σύμφωνα με τους WittenandFrank, η Εξόρυξη Δεδομένων (ΕΔ) (DataMining (DM)) ορίζεται ως η διαδικασία ανακάλυψης προτύπων μέσα από δεδομένα, δίνοντας έτσι έμφαση στη διάσταση της Μηχανικής Μάθησης. [14]

Τα τελευταία χρόνια μεγάλη διάδοση γνωρίζει και η Ανάλυση Κοινωνικών Δικτύων. Η ανάλυση κοινωνικών δικτύων χρησιμοποιεί τη θεωρία δικτύων για να αναλύσει κοινωνικές σχέσεις. Χρησιμοποιείται ευρύτατα στη σύγχρονη κοινωνιολογία, αλλά σχετίζεται και με την ανθρωπολογία, την κοινωνική ψυχολογία, τη βιολογία και άλλους επιστημονικούς κλάδους. Μεγάλη ώθηση στην ανάλυση κοινωνικών δικτύων έδωσε η ραγδαία διάδοση των ιστοτόπων κοινωνικής δικτύωσης, όπως το Facebook.

Παλαιότερα η μελέτη κοινωνικών δικτύων περιοριζόταν σε μικρές ομάδες, γιατί ήταν αδύνατη η συλλογή περισσότερων δεδομένων. Πλέον είναι διαθέσιμος ένας τεράστιος όγκος δεδομένων, που αφορά εκατοντάδες ανθρώπους και τις σχέσεις

τους. Τα δεδομένα αυτά, με τη μορφή γράφων, είναι διαφορετικού τύπου από τα δεδομένα που τηρούνται στις παραδοσιακές σχεσιακές βάσεις δεδομένων, και η επεξεργασία τους συνιστά μια πρόκληση, καθώς μπορούν να αποτελέσουν πηγή χρήσιμης και διαφορετικής πληροφορίας. Τα δεδομένα των κοινωνικών δικτύων αξιοποιούνται από διάφορους φορείς και βεβαίως μπορούν να αξιοποιηθούν από τις επιχειρήσεις.

Η δομή του δικτύου είναι μια πολύτιμη πηγή πληροφοριών, καθώς ο αναλυτής μπορεί να εντοπίσει «κοινότητες» χρηστών. Ο καθορισμός των κοινοτήτων βασίζεται στην ύπαρξη πυκνών δεσμών μεταξύ των μελών της κοινότητας και αραιότερων δεσμών με τα μη μέλη. Πέρα από τη στατική δομή του δικτύου, ενδιαφέρον παρουσιάζει και η δυναμική ροή πληροφοριών μέσα σε αυτό. Οι σχέσεις μεταξύ των κόμβων αποτελούν κανάλια διάδοσης πληροφοριών και αλληλεπίδρασης. Ένα μήνυμα μεταδίδεται μέσα στο δίκτυο και επηρεάζει τα μέλη. Ορισμένοι κόμβοι, λόγω προσωπικής μόρφωσης, ενδιαφερόντων, έντονης δραστηριότητας στο δίκτυο κλπ. επηρεάζουν σε σημαντικότερο βαθμό άλλους κόμβους απ' ότι πιο αδρανείς κόμβοι. Ο εντοπισμός και η επιρροή των δυναμικών κόμβων επιτρέπει την ταχύτερη και αποτελεσματικότερη διάδοση ιδεών στο δίκτυο.

Το Twitter, μπορεί να θεωρηθεί μακράν ως το online κοινωνικό δίκτυο που έχει μελετηθεί και αναλυθεί περισσότερο. [5] Αναλυτικότερα, τα κύρια χαρακτηριστικά που διαφοροποιούν το Twitter, συγκριτικά με άλλα μέσα κοινωνικής δικτύωσης, καθιστώντας το ένα εργαλείο πρόσφορο για ανάλυση είναι:

1. Η ελεύθερη πρόσβαση στα δεδομένα: Το μεγαλύτερο μέρος των δεδομένων του Twitter είναι ελεύθερο στο κοινό και συνεπώς είναι δυνατή η συλλογή ενός επαρκούς αριθμού δεδομένων.
2. Η χρονική αλληλουχία των γεγονότων: Τα tweets περιέχουν χρονοσφραγίδα (timestamp) επιτρέποντας έτσι την παρακολούθηση των γεγονότων με χρονολογική σειρά.
3. Ο περιορισμός στο κείμενο ανάρτησης: Ο κάθε χρήστης μπορεί να δημοσιεύσει μόνο σύντομα μηνύματα, μέχρι 140 χαρακτήρες. Με τον τρόπο αυτό, όλη η πληροφορία περιορίζεται σε λίγες λέξεις και είναι ευκολότερο για μια εφαρμογή να την επεξεργαστεί.
4. Το Κοινωνικό γράφημα (socialgraph): Οι χρήστες του Twitter δημιουργούν ένα κοινωνικό γράφημα με ιδιαίτερη δομή. Όταν λέμε ότι ένας χρήστης «ακολουθεί» (following) έναν άλλον, τότε εννοούμε ότι παρακολουθεί τα tweets του. Αντίθετα αυτοί που παρακολουθούνται από άλλους χρήστες έχουν «οπαδούς» (followers).

2.5. Κοινωνική Επιρροή

Ως κοινωνική επιρροή (socialinfluence) ορίζουμε τη διαδικασία με την οποία ένα άτομο ή μία ομάδα (influencers) προσπαθεί να αλλάξει τον τρόπο σκέψης ή συμπεριφοράς ενός άλλου (ατόμου ή ομάδας), προκειμένου αυτό να προσαρμοστεί στη δική του άποψη ή συμπεριφορά. Παραδοσιακά, το ρόλο των influencers έπαιζαν τα μέσα μαζικής ενημέρωσης, οι πολιτικοί, καθώς και άλλα δημόσια πρόσωπα. Με το πέρασμα των χρόνων όμως και με την εξάπλωση των μέσων κοινωνικής δικτύωσης, είναι πολλοί εκείνοι οι χρήστες που καταφέρνουν να ασκούν επιρροή σε άλλους, χωρίς απαραίτητα να είναι δημόσια πρόσωπα.

Δεδομένου ότι δεν υπάρχει κάποιος κοινός αποδεκτός τρόπος να ποσοτικοποιηθεί και να αναλυθεί η επιρροή ενός ατόμου, υπάρχει πλούσια ερευνητική δραστηριότητα στον συγκεκριμένο

τομέα. Το Twitter ειδικότερα προσφέρεται ως μέσο μελέτης της επιρροής των χρηστών, καθώς παρέχει αρκετά στοιχεία αναφορικά με το δίκτυο του κάθε χρήστη (ποιοι τον ακολουθούν και ποιους ακολουθεί), καθώς επίσης και με τη δραστηριότητά του (tweets, retweets και mentions).

2.6. API

Η Διεπαφή Προγραμματισμού Εφαρμογών, εν συντομία API (ApplicationProgrammingInterface) είναι ένα εργαλείο το οποίο διευκολύνει την αλληλεπίδραση με προγράμματα ηλεκτρονικών υπολογιστών και υπηρεσίες web. Πολλές υπηρεσίες web παρέχουν APIs στους προγραμματιστές προκειμένου να αλληλεπιδρούν με τις υπηρεσίες τους και να έχουν πρόσβαση σε δεδομένα με ένα προγραμματιστικό τρόπο. ΤοTwitter έχει δώσει ιδιαίτερη σημασία στη δημιουργία ενός κομψού και απλού API, το οποίο είναι διαισθητικό και εύκολο στη χρήση.

Με το RESTAPI του Twitter, οι προγραμματιστές αναζητούν ή τραβούν μόνο ένα συγκεκριμένο αριθμό από tweets που έχουν ήδη δημοσιευθεί, ο οποίος περιορίζεται από τα όρια ταχύτητας (RateLimits) του Twitter. Για έναν μεμονωμένο χρήστη, ο μέγιστος αριθμός των tweets που μπορεί να λάβει είναι τα τελευταία 3.200 tweets, ανεξάρτητα από τα κριτήρια αναζήτησης. Επιπλέον, υπάρχει περαιτέρω περιορισμός στον αριθμό των αιτήσεων που μπορούν να γίνουν σε ένα ορισμένο χρονικό διάστημα (180 αιτήσεις σε διάστημα 15 λεπτών).

Το StreamingAPI του Twitter επιτρέπει την αναζήτηση δεδομένων, σε σχεδόν πραγματικό χρόνο, από tweets τα οποία μόλις έχουν δημοσιευθεί, χρησιμοποιώντας είτε βασικά είτε OAuth πρωτόκολλα ασφαλείας. Με το StreamingAPI, οι χρήστες καταχωρούν συγκεκριμένα κριτήρια (hashtags, ονόματα χρηστών, τοποθεσίες, θέσεις κλπ.) και όσο δημοσιεύονται tweets που ταιριάζουν με τα κριτήρια αυτά, ωθούνται απευθείας στο χρήστη μαζί με πληροφορίες για το συγγραφέα του εκάστοτε tweet. Είναι περισσότερο ώθηση δεδομένων από το Twitter, παρά τράβηγμα των δεδομένων από τον τελικό χρήστη.

Το StreamingAPI έχει το μειονέκτημα πως το πραγματικό ποσοστό του συνόλου των tweets, που λαμβάνουν οι χρήστες, ποικίλλει σε μεγάλο βαθμό με βάση τα κριτήρια αναζήτησης και τις τρέχουσες κυκλοφοριακές συνθήκες. Παρ' όλα αυτά τα οφέλη της ύπαρξης μίας πραγματικού χρόνου ροής δεδομένων Twitter, καθιστούν ιδιαίτερα σημαντική την ενσωμάτωση του StreamingAPI σε διάφορους τύπους εφαρμογών.

2.7. OAuth

Το OAuth είναι ένα πρωτόκολλο εξουσιοδότησης που επιτρέπει την αποστολή ασφαλών εξουσιοδοτημένων αιτήσεων προς το TwitterAPI. Οι χρήστες δεν χρειάζεται να μοιράζονται τους κωδικούς τους με άλλες εφαρμογές, αυξάνοντας έτσι την ασφάλεια του λογαριασμού τους. Επιπλέον, υπάρχουν πολλές συμβατές βιβλιοθήκες με την υλοποίηση του OAuth στο Twitter, που το καθιστούν πρότυπο πρωτόκολλο ασφαλείας.

Η εξουσιοδότηση μόνο της εφαρμογής (application-onlyauthentication) είναι μια μορφή ταυτοποίησης που επιτρέπει σε μία εφαρμογή να εκτελέσει αιτήσεις API για δικό της λογαριασμό, χωρίς κάποιο πλαίσιο χρήστη. Οι API κλήσεις εξακολουθούν να τελούν υπό περιορισμούς, οι οποίοι όμως υπόκεινται σε ολόκληρη την εφαρμογή σε σύνολο και όχι ανά χρήστη. Οι API μέθοδοι που υποστηρίζουν αυτή τη μορφή εξουσιοδότησης περιέχουν δύο όρια ταχύτητας στην τεκμηρίωση

τους, ένα ανά χρήστη (application-userauthentication) και ένα

ανά εφαρμογή (application-onlyauthentication).

III. ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ

3.1. Προετοιμασία

Το πρώτο βήμα για την υλοποίηση της εργασίας ήταν η δημιουργία ενός λογαριασμού στο Twitter. Μέσω του λογαριασμού ζητήσαμε άδεια προγραμματιστή για να αποκτήσουμε πρόσβαση στα δεδομένα του Twitter. Ταυτόχρονα κάναμε αίτηση και για δημιουργία εφαρμογής στη σελίδα <https://developer.twitter.com/en/apps/create>. Αφού εγκρίθηκαν και οι δύο αιτήσεις μας ξεκινήσαμε την υλοποίηση της εργασίας.

3.2. Επιλογή Λογαριασμών

Επιλέξαμε 4 κατηγορίες λογαριασμών.

Η πρώτη κατηγορία περιλαμβάνει 2 ποδοσφαιριστές που παίζουν στην ίδια ομάδα, τον Νείμάρ και τον Εμπαπέ.

Ο Νείμάρ ντα Σίλβα Σάντος Τζούνιορ, ή απλώς Νείμάρ, είναι βραζιλιάνος ποδοσφαιριστής. Αγωνίζεται με την ομάδα της Παρί Σεν Ζερμέν και η θέση του είναι στην κορυφή της επίθεσης ή και ως πλάγιος επιθετικός. Γεννήθηκε στο Σάο Πάολο στις 5 Φεβρουαρίου 1992. Ξεκίνησε την ποδοσφαιρική του καριέρα στα 11 του, από την Σάντος το 2003. [6]

Ο Κιλιάν Εμπαπέ Λοτέν, είναι Γάλλος επαγγελματίας ποδοσφαιριστής ο οποίος αγωνίζεται στην θέση του επιθετικού για λογαριασμό της Παρί Σεν Ζερμέν. Γεννήθηκε στο Παρίσι στις 20 Δεκεμβρίου 1998. Ο Εμπαπέ πραγματοποίησε το ντεμπούτο του στη Λιγκ 1 στις 2 Δεκεμβρίου 2015 όταν εισήλθε στον αγωνιστικό χώρο ως αλλαγή στην θέση του Φάμπιο Κοεντράο στο 88ο λεπτό της αναμέτρησης. [7]

Η δεύτερη κατηγορία περιλαμβάνει 2 αντίπαλους πολιτικούς που ήταν πρόσφατα στην επικαιρότητα λόγω των βρετανικών εκλογών, τον Τζόνσον και τον Κόρμπιν .

Ο Αλεξάντερ Μπόρις ντε Φέφελ Τζόνσον είναι Βρετανός πολιτικός, συγγραφέας, πρώην δημοσιογράφος και σημερινός Πρωθυπουργός του Ηνωμένου Βασιλείου και ηγέτης του Συντηρητικού Κόμματος από τον Ιούνιο 2019. Γεννήθηκε στη Νέα Υόρκη στις 19 Ιουνίου 1964 από μια πλούσια βρετανική οικογένεια της υψηλής τάξης. Ο Τζόνσον αποτελεί μια αμφιλεγόμενη μορφή στην πολιτική και στην δημοσιογραφία της Βρετανίας. [8]

Ο Τζέρεμι Κόρμπιν είναι Βρετανός πολιτικός και επικεφαλής του Εργατικού Κόμματος του Ηνωμένου Βασιλείου. Γεννήθηκε στις 26 Μαΐου 1949 στο Τσίπεναμ, Ουίλτσερ του Ηνωμένου Βασιλείου. Ο Τζέρεμι Κόρμπιν είναι υποστηρικτής σοσιαλιστικών πολιτικών στην οικονομία. [9]

Η τρίτη κατηγορία περιλαμβάνει 2 από τα πιο γνωστά Μέσα Ενημέρωσης παγκόσμια, το BBC και το CNN.

Η British Broadcasting Corporation, γνωστότερη απλά ως BBC, είναι βρετανικός δημόσιος ραδιοτηλεοπτικός φορέας και το μεγαλύτερο δίκτυο οπτικοακουστικών και διαδικτυακών μεταδόσεων ανά τον κόσμο. Ιδρύθηκε το 1922 ως σύμπραξη έξι μεγάλων εταιριών μεταδόσεων και συστάθηκε με βασικό διάταγμα, ενώ λειτουργεί βάσει της συμφωνίας με τον υπουργό πολιτισμού, μέσωσ ενημέρωσης και αθλητισμού της χώρας . Η έδρα της βρίσκεται στο Ουέστμινστερ στο κεντρικό Λονδίνο, είναι ο παλαιότερος εθνικός ραδιοτηλεοπτικός οργανισμός στον κόσμο και ο μεγαλύτερος ανά αριθμό εργαζομένων.[10]

Το Cable News Network, γνωστότερο ως CNN, είναι ένα αμερικανικό καλωδιακό και δορυφορικό ενημερωτικό

τηλεοπτικό δίκτυο που ανήκει στην Turner Broadcasting System, τμήμα της WarnerMedia. Ιδρύθηκε το 1980 από τον Αμερικανό ιδιοκτήτη μέσωσ μαζικής ενημέρωσης Τεντ Τέρνερ ως 24ωρο ειδησεογραφικό τηλεοπτικό κανάλι. Από το ξεκίνημά του, ήταν το πρώτο που παρείχε 24ωρη ειδησεογραφική κάλυψη στις ΗΠΑ. Η έδρα του είναι στο κτήριο του CNN Center στην Ατλάντα.[11]

Η τέταρτη κατηγορία περιλαμβάνει 2 ροκ συγκροτήματα, τους Muse και τους Placebo.

Οι Muse είναι αγγλικό alternative rock συγκρότημα από το Τέινμπαουθ του Ντέβον, που δημιουργήθηκε το 1994, από τους Matthew Bellamy (φωνητικά, κιθάρα, πιάνο), Christopher Wolstenholme (μπάσο, φωνητικά, πλήκτρα, κιθάρα, φουσαρμόνικα) και τον Dominic Howard (ντραμς, συνθεσάιζερ). Τα μέλη των Muse έπαιζαν σε διαφορετικές σχολικές μπάντες κατά τη φοίτησή τους στο Teignmouth Community College στις αρχές της δεκαετίας του '90, αλλά ο σχηματισμός των Muse ξεκίνησε όταν ο Bellamy πέρασε από οντισιόν για τη θέση του κιθαρίστα, στην μπάντα του Dominic Howard, τους Carnage Mayhem. [12]

Οι Placebo είναι βρετανικό ροκ μουσικό συγκρότημα που ιδρύθηκε το 1994 από τους Brian Molko (φωνή και κιθάρα) και Stefan Olsdal (μπάσο). Ένας σταθμός του λονδρέζικου μετρό αποτέλεσε την αιτία για τη δημιουργία των Placebo. Οι ιδρυτές μπορεί να πήγαιναν επί επτά χρόνια στο ίδιο σχολείο, το American International School of Luxembourg, δίχως να έχουν μιλήσει σχεδόν καθόλου. Ωστόσο, το 1994, συναντήθηκαν τυχαία στον σταθμό του South Kensington και το ίδιο βράδυ αποφάσισαν να φτιάξουν το γκρουπ. Αν και αρχικά είχαν το όνομα Ashtray Heart, το συγκρότημα σύντομα κατέληξε στο όνομα Placebo.[13]

3.3. Υλοποίηση Κώδικα και Συλλογή Δεδομένων

Ως γλώσσα προγραμματισμού για την εργασία επιλέξαμε την Python. Θέσαμε 2 περιορισμούς στα προγράμματά μας. Πρώτον, ένα προφίλ στο Twitter θα θεωρείται ενεργό και κατ' επέκταση θα μπορούμε να το συμπεριλάβουμε στη μελέτη μας, όταν έχει κάνει τουλάχιστον 40 posts και δεν έχει τεθεί ως ιδιωτικό. Δεύτερον, η αναζήτηση των hashtags θα γίνεται στα posts που έχουν γίνει το τελευταίο εξάμηνο. Ως μέγιστο όριο πλήθους ακολούθων θεωρήσαμε το 200. Η επιλογή των 200 ακολούθων και του εξαμήνου έγινε ώστε να είναι μεγάλο το πλήθος των διαφορετικών hashtags και να εξάγουμε έγκυρα αποτελέσματα. Όταν οι αριθμοί ήταν μικρότεροι, τα hashtags μας έδιναν σχεδόν μηδενική ομοιότητα στα προφίλ και δεν μπορούσαμε να βγάλουμε συμπεράσματα.

Δημιουργήσαμε συνολικά 3 βασικά προγράμματα. Από το πρώτο πρόγραμμα δημιουργήσαμε 8 παραλλαγές προκειμένου να γίνει παράλληλα η συλλογή και η αποθήκευση δεδομένων από τους 8 λογαριασμούς που επιλέξαμε. Αρχικά στον κώδικα καταχωρήσαμε τους κωδικούς που λάβαμε για τη δημιουργία εφαρμογής, ώστε να αποκτήσουμε πρόσβαση στο API του Twitter, αφού πρώτα επικαιροποιήθηκε η αυθεντικότητά τους. Στη συνέχεια, δημιουργήσαμε ένα αρχείο txt στο οποίο αποθηκεύονται τα ονόματα των ακολούθων που επιλέχθηκαν για

το συγκεκριμένο προφίλ. Για κάθε έναν από τους ακολούθους ελέγξαμε τα posts του τελευταίου εξαμήνου και εφόσον δεν είναι κενά αποθηκεύσαμε σε μια λίστα όλα τα hashtags που χρησιμοποιήσε. Στη λίστα αυτή διαγράψαμε τα διπλότυπα και δημιουργήσαμε μια δεύτερη λίστα με τα μοναδικά hashtags. Τις δύο τελευταίες λίστες τις αποθηκεύσαμε αντίστοιχα σε δύο αρχεία txt με τα ονόματα all και clean για επόμενη χρήση. Στην επόμενη εικόνα φαίνεται ο κώδικας του πρώτου προγράμματος που χρησιμοποιήθηκε για τη συλλογή και αποθήκευση των δεδομένων από τα προφίλ. Για κάθε ένα από τα προφίλ αντικαθιστούσαμε την τιμή “Name” με το όνομα του χρήστη, καθώς επίσης προσθέταμε και το όνομά του στα 3 .txt αρχεία που μας επέστρεφε το πρόγραμμα.

```
1 from tweepy import OAuthHandler
2 from tweepy import API
3 from tweepy import Cursor
4 from datetime import datetime, date, time, timedelta
5 from collections import Counter
6 import sys
7 import tweepy
8
9 auth = tweepy.OAuthHandler("#####")
10 auth.set_access_token("#####", "#####")
11 api = Tweepy.API(auth, wait_on_rate_limit=True)
12 c = 0
13 list1 = []
14 hashtags = []
15 hashtagsall = []
16
17 with open('followerslist.txt', 'w', encoding="utf-8") as f, open('cleanhashtags.txt', 'w', encoding="utf-8") as g, open('allhashtags.txt', 'w', encoding="utf-8") as h:
18     for user in Cursor(api.get_followers, screen_name="Name").items():
19         if (user.protected == True):
20             continue
21         if (user.statuses_count < 40):
22             continue
23         temp = []
24         temp1 = []
25         list1.append(user)
26         f.write("%s\n" % user.screen_name)
27         c += 1
28         if (c > 100):
29             break
30     end_data = datetime.utcnow() - timedelta(days=180)
31     print(c)
32     for status in Cursor(api.user_timeline, screen_name=user.screen_name).items():
33         temp1 = []
34         temp2 = []
35         if (status.status_type == "retweet"):
36             entities = status.entities
37             if "hashtags" in entities:
38                 for ent in entities["hashtags"]:
39                     if ent.is_not_name():
40                         if "text" in ent:
41                             hashtag = ent["text"]
42                             if hashtag is not None:
43                                 temp1.append(hashtag)
44                                 hashtagsall.append(hashtag)
45                                 g.write("%s\n" % hashtag)
46         if status.created_at == end_data:
47             break
48         temp1 = set(temp1)
49         temp2 = list(temp1)
50         for hashtag in temp2:
51             hashtags.append(hashtag)
52             g.write("%s\n" % hashtag)
53     print(len(hashtags))
54     print(len(hashtagsall))
55
56 for user in list1:
57     print(user.screen_name)
58
59 print(len(hashtags))
```

Το επόμενο πρόγραμμα που δημιουργήσαμε μετατρέπει μία λίστα σε ένα σετ (δεν υπάρχουν διπλότυπα αρχεία). Η εισαγωγή των δεδομένων στη λίστα γίνεται από ένα αρχείο .txt. Στην επόμενη εικόνα φαίνεται η μετατροπή αυτή.

```
1 hashtags = []
2
3 with open('hashtags.txt', 'r', encoding="utf8") as filehandle:
4     for line in filehandle:
5         currentPlace = line[:1]
6         hashtags.append(currentPlace)
7
8 hashtags = list(set(hashtags))
9 print(len(hashtags))
```

Το τελευταίο πρόγραμμα που δημιουργήσαμε παίρνει ως είσοδο τις λίστες που περιλαμβάνουν τα δεδομένα που αποθηκεύτηκαν στα αρχεία με τα hashtags και υπολογίζει την ομοιότητα τους με τη χρήση της Jaccard ομοιότητας. Για τον υπολογισμό της Jaccardsimilarity ορίσαμε τη δική μας συνάρτηση jaccard_similarity, καθώς η buildinfunctionjaccard_score της βιβλιοθήκης sklearn.metrics δεν μας έδινε τα επιθυμητά αποτελέσματα.

```
1 hashtagsForName1 = []
2 hashtagsForName2 = []
3
4 def jaccard_similarity(list1, list2):
5     s1 = set(list1)
6     s2 = set(list2)
7     return float(len(s1.intersection(s2))) / len(s1.union(s2))
8
9 with open('name1allhashtags.txt', 'r', encoding="utf8") as filehandle:
10     for line in filehandle:
11         currentPlace = line[:1]
12         hashtagsForName1.append(currentPlace)
13
14 with open('name2allhashtags.txt', 'r', encoding="utf8") as filehandle:
15     for line in filehandle:
16         currentPlace = line[:1]
17         hashtagsForName2.append(currentPlace)
18
19 k = jaccard_similarity(hashtagsForName1, hashtagsForName2)
20 print("Jaccard similarity is: ", k)
```

Τέλος, δημιουργήσαμε ένα ακόμη πρόγραμμα το οποίο εισάγει όλα τα txt αρχεία με τα hashtags που έχουν προκύψει από τα 8 επιλεγμένα προφίλ και υπολογίσαμε την JaccardSimilarity (ανά δύο) για όλους τους δυνατούς συνδιασμούς των παραπάνω προφίλ.

```
1 def jaccard_similarity(list1, list2):
2     s1 = set(list1)
3     s2 = set(list2)
4     return float(len(s1.intersection(s2))) / len(s1.union(s2))
5
6 list1 = []
7 list2 = []
8 list3 = []
9 list4 = []
10 list5 = []
11 list6 = []
12 list7 = []
13 list8 = []
14 listAll = []
15
16 with open('neymarjr5cleanhashtags.txt', 'r', encoding="utf8") as filehandle:
17     for line in filehandle:
18         currentPlace = line[:1]
19         list1.append(currentPlace)
20
21 listAll.append(list1)
22
23 with open('mblist3cleanhashtags.txt', 'r', encoding="utf8") as filehandle:
24     for line in filehandle:
25         currentPlace = line[:1]
26         list2.append(currentPlace)
27
28 listAll.append(list2)
29
30 with open('borisjohnson5cleanhashtags.txt', 'r', encoding="utf8") as filehandle:
31     for line in filehandle:
32         currentPlace = line[:1]
33         list3.append(currentPlace)
34
35 listAll.append(list3)
36
37 with open('jeremycorbyn5cleanhashtags.txt', 'r', encoding="utf8") as filehandle:
38     for line in filehandle:
39         currentPlace = line[:1]
40         list4.append(currentPlace)
41
42 listAll.append(list4)
43
44 with open('BBCWorld5cleanhashtags.txt', 'r', encoding="utf8") as filehandle:
45     for line in filehandle:
46         currentPlace = line[:1]
47         list5.append(currentPlace)
48
49 listAll.append(list5)
50
51 with open('cmilist3cleanhashtags.txt', 'r', encoding="utf8") as filehandle:
52     for line in filehandle:
53         currentPlace = line[:1]
54         list6.append(currentPlace)
55
56 listAll.append(list6)
57
58 with open('muse5cleanhashtags.txt', 'r', encoding="utf8") as filehandle:
59     for line in filehandle:
60         currentPlace = line[:1]
61         list7.append(currentPlace)
62
63 listAll.append(list7)
64
65 with open('PLACEBOWORLD5cleanhashtags.txt', 'r', encoding="utf8") as filehandle:
66     for line in filehandle:
67         currentPlace = line[:1]
68         list8.append(currentPlace)
69
70 listAll.append(list8)
71
72 finalarr = []
73
74 for l in listAll:
75     new = []
76     for m in listAll:
77         k = jaccard_similarity(l,m)
78         new.append(k)
79     finalarr.append(new)
80
81 print("CLEAN")
82 print("    neymar mbappe boris corbyn bbc    cnn muse placebo")
83 names = ["neymar", "mbappe", "boris", "corbyn", "bbc", "cnn", "muse", "placebo"]
84 print(names[5])
85
86 for i in range(0, len(listAll)):
87     t = ""
88     for j in range(0, len(listAll)):
89         t += str(round(finalarr[i][j],3))
90     print(t)
91
92 print(t)
```

IV. ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

4.1. Jaccard Similarity

Στην εργασία επιλέξαμε ως μετρική για τον έλεγχο της ομοιότητας μεταξύ δύο ομάδων την JaccardSimilarity. Έστω δύο κείμενα A,B, που δεν έχουν απαραίτητα το ίδιο μέγεθος. Η Jaccardορίζεται ως: $jaccard(A,B) = |A \cap B| / |A \cup B|$. Η τιμή αυτή κυμαίνεται ανάμεσα στο 0, όταν τα A,B δεν έχουν κοινά στοιχεία, και στο 1, όταν ταυτίζονται.

4.2. Ανάλυση Δεδομένων

Αρχικά μετρήσαμε για καθένα από τους 8 λογαριασμούς το πλήθος των ακολούθων που επιλέξαμε, καθώς και το πλήθος όλων των hashtags (all) που χρησιμοποιήθηκαν, αλλά και των μεμονωμένων (clean). Ως set(clean) ορίσαμε μια νέα λίστα αρχείων που περιλαμβάνει όλα τα μη διπλότυπα hashtags των μεμονωμένων hashtags. Οι αριθμοί που προέκυψαν φαίνονται στον επόμενο πίνακα:

	Neymar	Mbappe	Johnson	Corbyn	BBC	CNN	Muse	Placebo
All	8571	14990	14304	16319	20743	13761	18594	15798
Clean	2920	5086	4184	3695	3666	2465	3900	3979
Set(clean)	2734	4347	3851	3445	3430	2303	3636	3676
People	200	200	200	200	200	200	200	200

Ακόμη, επιλέξαμε να συγκρίνουμε την ομοιότητα και μεταξύ λογαριασμών που εντάσσονται σε διαφορετικές κατηγορίες. Τα αποτελέσματα της Jaccard_similarity φαίνονται στον ακόλουθο πίνακα:

	Neymar	Mbappe	Johnson	Corbyn	BBC	CNN	Muse	Placebo
Neymar	1	0.058	0.029	0.019	0.035	0.037	0.028	0.027
Mbappe	0.058	1	0.029	0.025	0.031	0.031	0.03	0.038
Johnson	0.029	0.029	1	0.069	0.033	0.026	0.027	0.036
Corbyn	0.019	0.025	0.069	1	0.033	0.028	0.032	0.042
BBC	0.035	0.031	0.033	0.033	1	0.033	0.029	0.031
CNN	0.037	0.031	0.026	0.028	0.033	1	0.033	0.029
Muse	0.028	0.03	0.027	0.032	0.029	0.033	1	0.041
Placebo	0.027	0.038	0.036	0.042	0.031	0.029	0.041	1

V. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΡΟΤΑΣΕΙΣ

5.1. Συμπεράσματα

Από τα αποτελέσματα που παρουσιάστηκαν στον παραπάνω πίνακα προέκυψαν οι παρακάτω παρατηρήσεις. Θα σχολιάσουμε τις συγκρίσεις που έγιναν ανάμεσα σε διαφορετικά προφίλ. Για το προφίλ του Νεϊμάρ διαπιστώσαμε ότι το μεγαλύτερο jaccard_similarity των ακολούθων του είναι 0,058 και προέκυψε με την σύγκρισή του με τους ακολούθους του Εμπαπέ. Αντίθετα, για το ίδιο προφίλ, το μικρότερο jaccard_similarity των ακολούθων του είναι 0,019 και προέκυψε με την σύγκρισή του με τους ακολούθους του Κόρμπιν.

Για το προφίλ του Εμπαπέ διαπιστώσαμε ότι το μεγαλύτερο jaccard_similarity των ακολούθων του είναι 0,058 και προέκυψε με την σύγκρισή του με τους ακολούθους του Νεϊμάρ. Αντίθετα, για το ίδιο προφίλ, το μικρότερο jaccard_similarity των ακολούθων του είναι 0,025 και προέκυψε με την σύγκρισή του με τους ακολούθους του Κόρμπιν.

Για το προφίλ του Μπόρις διαπιστώσαμε ότι το μεγαλύτερο jaccard_similarity των ακολούθων του είναι 0,069 και προέκυψε με την σύγκρισή του με τους ακολούθους του Κόρμπιν. Αντίθετα, για το ίδιο προφίλ, το μικρότερο jaccard_similarity των ακολούθων του είναι 0,026 και προέκυψε με την σύγκρισή του με τους ακολούθους του CNN.

Για το προφίλ του Κόρμπιν διαπιστώσαμε ότι το μεγαλύτερο jaccard_similarity των ακολούθων του είναι 0,069 και προέκυψε

με την σύγκρισή του με τους ακολούθους του Μπόρις. Αντίθετα, για το ίδιο προφίλ, το μικρότερο jaccard_similarity των ακολούθων του είναι 0,019 και προέκυψε με την σύγκρισή του με τους ακολούθους του Νεϊμάρ.

Για το προφίλ του BBC διαπιστώσαμε ότι το μεγαλύτερο jaccard_similarity των ακολούθων του είναι 0,035 και προέκυψε με την σύγκρισή του με τους ακολούθους του Νεϊμάρ. Αντίθετα, για το ίδιο προφίλ, το μικρότερο jaccard_similarity των ακολούθων του είναι 0,029 και προέκυψε με την σύγκρισή του με τους ακολούθους των Muse.

Για το προφίλ του CNN διαπιστώσαμε ότι το μεγαλύτερο jaccard_similarity των ακολούθων του είναι 0,037 και προέκυψε με την σύγκρισή του με τους ακολούθους του Νεϊμάρ. Αντίθετα, για το ίδιο προφίλ, το μικρότερο jaccard_similarity των ακολούθων του είναι 0,026 και προέκυψε με την σύγκρισή του με τους ακολούθους του Μπόρις.

Για το προφίλ των Muse διαπιστώσαμε ότι το μεγαλύτερο jaccard_similarity των ακολούθων του είναι 0,041 και προέκυψε με την σύγκρισή του με τους ακολούθους των Placebo. Αντίθετα, για το ίδιο προφίλ, το μικρότερο jaccard_similarity των ακολούθων του είναι 0,027 και προέκυψε με την σύγκρισή του με τους ακολούθους του Μπόρις.

Για το προφίλ των Placebo διαπιστώσαμε ότι το μεγαλύτερο jaccard_similarity των ακολούθων του είναι 0,042 και προέκυψε με την σύγκρισή του με τους ακολούθους του Κόρμπιν. Αντίθετα, για το ίδιο προφίλ, το μικρότερο jaccard_similarity των ακολούθων του είναι 0,027 και προέκυψε με την σύγκρισή του με τους ακολούθους του Νεϊμάρ.

Από την παραπάνω ανάλυση επισημαίνουμε τα ακόλουθα:

- Στις κατηγορίες των Ποδοσφαιριστών και των Πολιτικών τα αποτελέσματα επιβεβαίωσαν την αρχική μας υπόθεση, καθώς οι μεγαλύτερες τιμές της jaccard_similarity παρατηρήθηκαν στις μεταξύ τους συγκρίσεις (0,058 και 0,069 αντίστοιχα).
- Στην κατηγορία των ροκ συγκροτημάτων, ενώ για τους Muse το μεγαλύτερο αποτέλεσμα της jaccard_similarity ήταν στη σύγκριση με τους Placebo, για τους Placebo η σύγκρισή τους με τους Muse έρχεται στη δεύτερη θέση καθώς υπολείπεται 0,001 του 0,042 που έχουν στη σύγκρισή τους με τον Κόρμπιν.
- Στην κατηγορία των ειδησεογραφικών, το BBC και το CNN δεν εμφάνισαν μεταξύ τους τη μεγαλύτερη τιμή της jaccard_similarity. Παραδόξως και τα δύο ειδησεογραφικά πρακτορία εμφάνισαν τη μέγιστη τιμή της jaccard_similarity κατά τη σύγκρισή τους με τον Νεϊμάρ.

5.2. Προτάσεις

Για μια πιο έγκυρη και αξιόπιστη έρευνα, θα μπορούσαμε στο μέλλον να χρησιμοποιήσουμε έναν μεγαλύτερο αριθμό ακολούθων για το κάθε ένα επιλεγμένο προφίλ αντί για τους 200 που επιλέξαμε σ' αυτήν την ανάλυση. Ακόμη, θα μπορούσε να γίνει πιο εξονυχιστικός έλεγχος σε κάθε ένα από τα προφίλ των ακολούθων προκειμένου να υπάρχει μια ομοιομορφία σχετικά με τον αριθμό των posts και τον αριθμό των hashtags του καθενός. Μ' αυτόν τον τρόπο θα αποφευχθεί το γεγονός ότι προφίλ με πολλά posts επηρεάζουν τα αποτελέσματα του πειράματος. Τέλος, θα μπορούσαμε να επεκταθούμε και σε περισσότερες κατηγορίες χρηστών του Twitter.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1]: Chan-Olmsted, S. , Cho , M. , Lee ,S.(2013), “*User Perceptions of Social Media: A Comparative Study of Perceived Characteristics and User Profiles by Social Media*”, Online Journal of Communication and Media Technologies, vol.3, issue 4.
- [2]: Chelmis, Charalampos, and Viktor K. Prasanna (2011), "Social Networking Analysis: A State of the Art and the Effect of Semantics." IEEE Third International Conference on Social Computing (SocialCom), 531- 536.
- [3]: Kaplan, Andreas M., and Michael Haenlein (2010), "Users of the world, unite! The challenges and opportunities of Social Media." Business Horizons (Elsevier) 53, no. 1, 59-68.
- [4]: Marsden, Peter V., and Karen E. Campbell (1984), "Measuring Tie Strength." Social Forces, 482-501.
- [5]: Rogers, Richard (2013), "Debanalizing twitter." 5th Annual ACM Web Science Conference. New York: ACM Press, 356-365.
- [6]: Wikipedia, *Νεϊμάρ*. Διαθέσιμο στο δικτυακό τόπο: <https://el.wikipedia.org/wiki/%CE%9D%CE%B5%CF%8A%CE%BC%CE%AC%CF%81>, (τελευταία πρόσβαση στις 27/01/2020).
- [7]: Wikipedia, *Κιλιάν Εμπαπέ*. Διαθέσιμο στο δικτυακό τόπο: https://el.wikipedia.org/wiki/%CE%9A%CE%B9%CE%BB%CE%B9%CE%AC%CE%BD_%CE%95%CE%BC%CF%80%CE%B1%CF%80%CE%AD, (τελευταία πρόσβαση στις 27/01/2020).
- [8]: Wikipedia, *Μπόρις Τζόνσον*. Διαθέσιμο στο δικτυακό τόπο: https://el.wikipedia.org/wiki/%CE%9C%CF%80%CF%8C%CF%81%CE%B9%CF%82_%CE%A4%CE%B6%CF%8C%CE%BD%CF%83%CE%BF%CE%BD, (τελευταία πρόσβαση στις 27/01/2020).
- [9]: Wikipedia, *Τζέρεμι Κόρμπιν*. Διαθέσιμο στο δικτυακό τόπο: https://el.wikipedia.org/wiki/%CE%A4%CE%B6%CE%AD%CF%81%CE%B5%CE%BC%CE%B9_%CE%9A%CF%8C%CF%81%CE%BC%CF%80%CE%B9%CE%BD, (τελευταία πρόσβαση στις 27/01/2020).
- [10]: Wikipedia, *BBC*. Διαθέσιμο στο δικτυακό τόπο: <https://el.wikipedia.org/wiki/BBC>, (τελευταία πρόσβαση στις 27/01/2020).
- [11]: Wikipedia, *CNN*. Διαθέσιμο στο δικτυακό τόπο: <https://el.wikipedia.org/wiki/CNN>, (τελευταία πρόσβαση στις 27/01/2020).
- [12]: Wikipedia, *Muse*. Διαθέσιμο στο δικτυακό τόπο: <https://el.wikipedia.org/wiki/Muse>, (τελευταία πρόσβαση στις 27/01/2020).
- [13]: Wikipedia, *Placebo*. Διαθέσιμο στο δικτυακό τόπο: [https://el.wikipedia.org/wiki/Placebo_\(%CF%83%CF%85%CE%B3%CE%BA%CF%81%CF%8C%CF%84%CE%B7%CE%BC%CE%B1\)](https://el.wikipedia.org/wiki/Placebo_(%CF%83%CF%85%CE%B3%CE%BA%CF%81%CF%8C%CF%84%CE%B7%CE%BC%CE%B1)), (τελευταία πρόσβαση στις 27/01/2020).
- [14]: Witten, I. H., & Frank, E. (2000), “*Data Mining Practical Machine Learning Tools and Techniques with Java Implementations*”. San Francisco, CA: Morgan Kaufmann Publishers.