ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Μελέτη, υλοποίηση και σύγκριση μεθόδων ανίχνευσης χρηστών με μεγάλη επιρροή στα μέσα κοινωνικής δικτύωσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Θωμά Ε. Λαγού

Επιβλέπων: Δημήτριος Θ. Ασκούνης Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2017

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Μελέτη, υλοποίηση και σύγκριση μεθόδων ανίχνευσης χρηστών με μεγάλη επιρροή στα μέσα κοινωνικής δικτύωσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Θωμά Ε. Λαγού

Επιβλέπων: Δημήτριος Θ. Ασκούνης Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 3η Οκτωβρίου 2017.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτριος Ασκούνης	Ιωάννης Ψαρράς	Χάρης Δούκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2017

(Υπογραφή)

.....

Θωμάς Λαγός

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Ηλεκτρονικών Υπολογιστών Ε.Μ.Π.

Περίληψη

Η διάδοση της πληροφορίας στα μέσα κοινωνικής δικτύωσης, όπως το Twitter, αποτελεί μοναδικό φαινόμενο, καθώς η εικόνα των ανθρώπων που συζητάνε και ανταλλάσσουν ιδέες σε χώρους όπως το πανεπιστήμιο έχει μετασχηματιστεί σε μια πλατφόρμα στην οποία τα άτομα μπορούν να συζητούν με οποιονδήποτε από οπουδήποτε. Στη συγκεκριμένη εργασία γίνεται μια εκτενής περιγραφή για το πως διαχέεται η πληροφορία στα μέσα κοινωνικής δικτύωσης, για το πως διάφοροι αλγόριθμοι δίνουν τις δικές τους λύσεις κατάταξης των ατόμων με μεγάλη επιρροή στα μέσα κοινωνικής δικτύωσης, για το πως μπορούν να γίνουν προβλέψεις για το πόσο πολύ θα διαδοθεί μια πληροφορία και ακόμα και για το πως η διάχυση της πληροφορίας έχει προεκτάσεις σε άλλους τομείς, όπως η κοινωνιολογία. Επιπλέον της θεωρητικής ανάλυσης, σχεδιάστηκε, υλοποιήθηκε και αξιολογήθηκε ένα εργαλείο κατάταξης χρηστών με βάση την επιρροή τους στα μέσα κοινωνικής δικτύωσης. Προκειμένου να γίνει αυτή η κατάταξη δημιουργήθηκαν 3 μεγάλες θεματικές κατηγορίες, έκτακτα νέα, αθλητικά, προσωπική υγεία, με συνολικά 874 άτομα που δημοσίευσαν πάνω από 250,000 tweets σε διάστημα 56 ημερών. Στην συνέχεια, προκειμένου να δημιουργηθεί ένας υπογράφος του Twitter μελετήθηκαν πάνω από 4 εκατομμύρια χρήστες. Σημαντική συνεισφορά της παρούσας εργασίας στην αναγνώριση ατόμων με μεγάλη επιρροή στα μέσα κοινωνικής δικτύωσης είναι η εισαγωγή μιας μεταβλητής που επιβραβεύει άτομα που καταφέρνουν να επηρεάζουν άλλα άτομα σε ώρες που δεν θεωρούνται ώρες αιχμής. Επιπλέον, ο αλγόριθμος που δημιουργήθηκε καταφέρνει να δημιουργήσει μια κατάταξη των ατόμων που ανταποκρίνεται στην πραγματικότητα, αποδεικνύοντας ότι σχέσεις όπως το πλήθος των ακόλουθων δεν είναι ικανές αλλά αναγκαίες προκειμένου να γίνει μια σωστή κατάταξη, εντοπίζοντας ταυτόχρονα χρήστες που παρά το γεγονός ότι έχουν σημαντικά μικρότερο πλήθος ακόλουθων καταφέρνουν να έχουν μεγάλη επιρροή.

Λέξεις κλειδιά: επηρεάζοντας, αλγόριθμος, κατάταξη, διάδοση, πληροφορία

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

The dissemination of information on social media, like Twitter, brings a new era in human communications and interactions, since face-to-face discussions and exchange of ideas are being replaced by online communities where people communicate and discuss with almost anyone from almost everywhere. In this thesis an extensive description is made of how information is disseminated in social media, how different algorithms give their own ranking solutions for people with a strong influence on social media, how to predict how much a specific publication will be disseminated, and even how diffusion of information has extensions to other areas, such as sociology. In addition to the theoretical analysis, a user rating tool was designed, implemented, and evaluated based on the influence of users on social media. In order to make this ranking, 3 major thematic categories were created (extraordinary news, sports, personal health), with a total of 874 users who published over 250,000 tweets in 56 days. Furthermore, in order to create a Twitter subgraph, over 4 million users were studied. An important contribution of this work in the detection of users with strong influence on social media is the introduction of a variable that rewards individuals who manage to influence other people at hours of a day that are not considered as peak hours. Moreover, our algorithm manages to create a ranking that responds to reality, demonstrating that relationships such as the number of followers are not sufficient, but are necessary in order to make a correct ranking. Another significant result of our algorithm is that it manages to distinguish users with a lot of followers from influencers.

Keywords: influencer, algorithm, ranking, dissemination, information

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

Πίνακας Περιεχομένων Εικόνων

1 *Εισαγωγή*

1.1 Αντικείμενο διπλωματικής

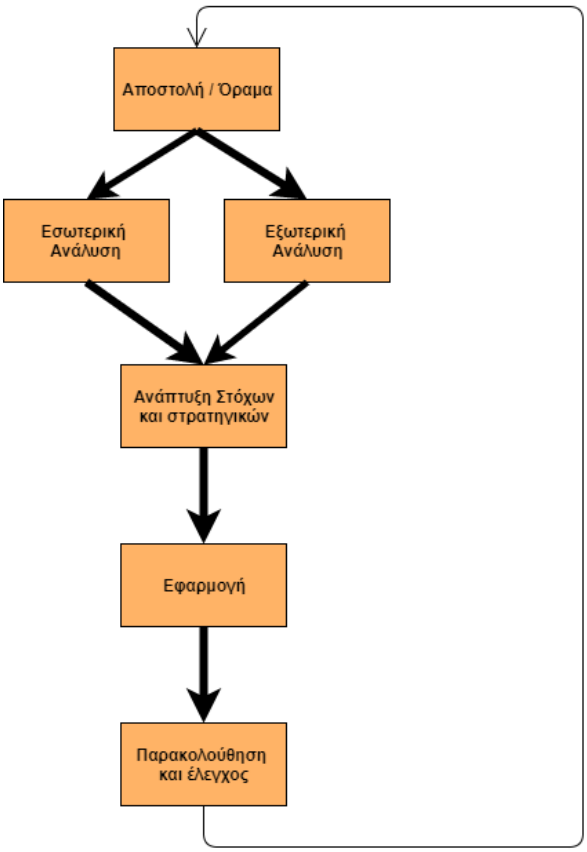
Τα μέσα κοινωνικής δικτύωσης αποτελούν τα τελευταία χρόνια έναν ιδιαιτέρως δημοφιλή τρόπο επικοινωνίας, ενημέρωσης και έκφρασης απόψεων και συναισθημάτων γύρω από πληθώρα θεμάτων, χάρη στην ευκολία στην πρόσβαση και τη χρήση και την ταχύτητα στην επικοινωνία. Ωστόσο ο βαθμός διάχυσης της πληροφορίας εξαρτάται, πέρα από το ίδιο το περιεχόμενό της, και από το άτομο που την παράγει/μοιράζεται, δηλαδή από το βαθμό επιρροής του στο εκάστοτε μέσο και για το εκάστοτε θέμα συζήτησης.

Στόχος της παρούσας διπλωματικής ήταν αρχικά η μελέτη, ανάλυση και σύγκριση μεθόδων ανίχνευσης χρηστών με ιδιαίτερα μεγάλη επιρροή στο μέσο κοινωνικής δικτύωσης που λέγεται Twitter. Στη συνέχεια ήταν η δημιουργία και υλοποίηση ενός αλγορίθμου στα πλαίσια μιας διαδικτυακής εφαρμογής που θα βοηθά εταιρείες να αναγνωρίσουν τους χρήστες του Twitter με τη μεγαλύτερη επιρροή στον τομέα που δραστηριοποιούνται.

1.1.1 Στρατηγικό management

Η συγκεκριμένη εργασία αποτελεί ένα τμήμα της έννοιας του στρατηγικού Management, καθώς η εύρεση ατόμων με επίδραση στο διαδίκτυο βοηθάει μια επιχείρηση πολύπλευρα. Συγκεκριμένα, στο χώρο των εταιρειών, τα κοινωνικά δίκτυα χρησιμοποιούνται ευρέως τόσο ως κανάλια προώθησης, όσο και ως πηγή πληροφοριών

σχετικά με τις επιθυμίες των καταναλωτών, όπως αυτές διαφαίνονται από τις συζητήσεις τους. Έχει συνεπώς ιδιαίτερη σημασία για μια εταιρεία να κατανοήσει ποια από τα εκατομμύρια μηνύματα που ανταλλάσσονται καθημερινά μπορούν να παρέχουν χρήσιμες πληροφορίες, είτε για επιλογή ατόμων που μπορούν να συμβάλουν σε προωθητικές ενέργειες, είτε για να δημιουργήσει πιο ανταγωνιστικά προϊόντα και υπηρεσίες σύμφωνα με την πρόβλεψη των τάσεων της αγοράς. Προκειμένου να γίνει πιο κατανοητή η σύνδεση των παραπάνω με τον Στρατηγικό Management επισημαίνεται ότι το Στρατηγικό Management αποτελεί θεμελιώδη παράγοντα επιτυχίας των επιχειρήσεων, καθώς διαμορφώνει την φυσιογνωμία και εκφράζει τη δυναμική τους. Ένας απλός ορισμός του Στρατηγικού Management είναι «η διαδικασία ευθυγράμμισης ολόκληρης της επιχείρησης με το περιβάλλον της, με κάποιο αποτέλεσμα κατά νου». Το «αποτέλεσμα» υπονοεί συγκεκριμένους επιχειρησιακούς στόχους, αποστολή και γενικά επιδιωκόμενα αποτελέσματα. Το σχήμα που ακολουθεί παρουσιάζει ένα μοντέλο που γενικά περιγράφει τη διαδικασία που ακολουθεί το στρατηγικό Management. Η διαδικασία απαραίτητα ξεκινά με την αποστολή (mission) και το όραμα (vision) της επιχείρησης. Εφόσον αυτές οι δύο έννοιες έχουν ξεκαθαριστεί, πραγματοποιούνται δύο ειδών αναλύσεις (εξωτερική και εσωτερική), που αφορούν αντίστοιχα στο εξωτερικό και εσωτερικό περιβάλλον της επιχείρησης και, βέβαια, καταλήγουν σε μια αναλυτική περιγραφή της παρούσας κατάστασης. Στην συνέχεια, όπως παρουσιάζεται και στο σχήμα, αναπτύσσονται στρατηγικές (δηλαδή τίθενται στόχοι και αναπτύσσονται στρατηγικές και προγράμματα), οι οποίες εφαρμόζονται. Η επιμέρους διαδικασία της εφαρμογής των στρατηγικών παρακολουθείται και ελέγχεται με σκοπό την ανά πάσα στιγμή παρέμβαση για μεταβολές ή για προσαρμογή των στόχων, στρατηγικών ή προγραμμάτων [23].



Εικόνα 1: Η διαδικασία που ακολουθεί το στρατηγικό Management [23]

1.1.2 Συνεισφορά

Οι κυριότερες συνεισφορές της παρούσας διπλωματικής εργασίας είναι:

- Εκτενής συγκριτική ανάλυση διαφορετικών προσεγγίσεων προκειμένου να δοθεί μια ξεκάθαρη εικόνα για τις εξελίξεις στον συγκεκριμένο τομέα.
- Προσέγγιση και εφαρμογή ιδεών από διαφορετικά γνωστικά αντικείμενα, όπως για παράδειγμα από το πως ξεκίνησαν να προτείνονται τα προϊόντα μέσω ηλεκτρονικών μηνυμάτων (e-mail).
- Δημιουργία και υλοποίηση πρωτοποριακού αλγορίθμου, για τον εντοπισμό ατόμων με επίδραση στο διαδίκτυο.
- Παρακολούθηση του Twitter και εξαγωγή δεδομένων από αυτό για εκτεταμένο χρονικό διάστημα (συνολικά 3 μήνες).

- Διεξαγωγή αποτελεσμάτων με συμπεράσματα που επιβεβαιώνονται από εργασίες διαφορετικών ερευνητών.

1.2 Οργάνωση κειμένου

Μετά την εισαγωγή ακολουθεί το δεύτερο κεφάλαιο όπου γίνεται η περιγραφή διαφόρων εργασιών σχετικών με το γνωστικό αντικείμενο της διπλωματικής εργασίας και η συγκριτική ανάλυση αυτών. Στην συνέχεια στο τρίτο κεφάλαιο περιγράφεται ο αλγόριθμος εύρεσης ατόμων με μεγάλη επιρροή στα μέσα κοινωνικής δικτύωσης που υλοποιήθηκε προκειμένου να βρεθούν τα άτομα με υψηλή επίδραση. Ακολουθεί το τέταρτο κεφάλαιο που περιέχει μια αναλυτικότερη περιγραφή των λεπτομερειών της υλοποίησης του αλγορίθμου. Στο πέμπτο κεφάλαιο παρουσιάζονται τα αποτελέσματα της έρευνας και γίνεται ο σχολιασμός αυτών των αποτελεσμάτων. Στο έκτο κεφάλαιο παρουσιάζεται ο επίλογος και στο τέλος παρατίθεται η βιβλιογραφία.

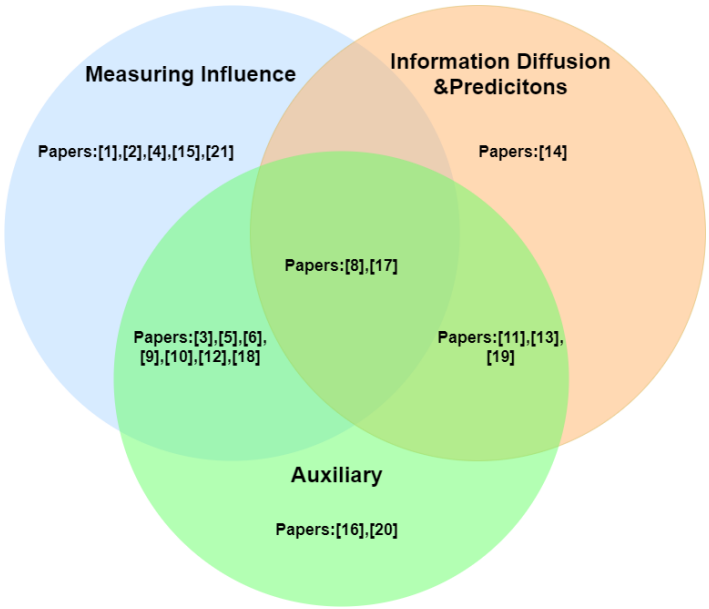
2 *Σχετικές εργασίες*

Η μελέτη του ανθρώπινου παράγοντα στην διάδοση της πληροφορίας στα μέσα κοινωνικής δικτύωσης είναι ένα ζήτημα το οποίο έχει απασχολήσει ιδιαίτερα την επιστημονική κοινότητα τα τελευταία χρόνια. Είναι κοινώς αποδεκτό πλέον ότι για να διαδοθεί ένα οποιοδήποτε κομμάτι πληροφορίας απαιτούνται τα κατάλληλα άτομα για να το προωθήσουν. Η ποσοτικοποίηση της επίδρασης που ασκεί κάποιο άτομο μέσω των κοινωνικών δικτύων που ενδεχομένως συμμετέχει, έχει υποκειμενικό χαρακτήρα και εξαρτάται από πολλές παραμέτρους (π.χ τον ορισμό της επίδρασης). Για την συγγραφή της παρούσας διπλωματικής εργασίας αναζητήθηκαν πληροφορίες από έρευνες σχετικά με:

1. Την **κατάταξη** ατόμων με επίδραση στο διαδίκτυο χρησιμοποιώντας διάφορους αλγορίθμους διάταξης (Measuring Influence). Τα άρθρα που βρίσκονται σε αυτήν την κατηγορία χρησιμοποιούν διαφορετικούς αλγορίθμους συνδυάζοντας α. διαφορετικά χαρακτηριστικά των ατόμων με επίδραση στο διαδίκτυο, β. την δομή του γράφου που ορίζει το μέσο κοινωνικής δικτύωσης και γ. τα δεδομένα που είναι διαθέσιμα για επεξεργασία. Αυτοί οι αλγόριθμοι είναι και αυτοί που δίνουν τις διάφορες «σχετικές» λύσεις στο πρόβλημά μας.
2. Τον τρόπο με τον οποίο **διαδίδεται** η πληροφορία στο διαδίκτυο (Information Diffusion). Πολλές φορές όταν λύνεις ένα πρόβλημα αυτόματα λύνεται και ένα επόμενο. Αυτή ακριβώς η ιδέα εφαρμόζεται και στην συγκεκριμένη περίπτωση. Αν κάποιος γνωρίζει τον τρόπο με τον οποίο διαδίδεται η πληροφορία σε ένα κοινωνικό δίκτυο μπορεί να ρυθμίσει τις παραμέτρους του με τέτοιον τρόπο ώστε να διαδώσει την πληροφορία που θέλει και κατά συνέπεια να μετρήσει αυτές τις μεταβλητές δημιουργώντας μια πιο έγκυρη κατάταξη. Επιπλέον, στην συγκεκριμένη κατηγορία συμπεριλαμβάνονται οι έρευνες σχετικά με το πως κανείς μπορεί να προβλέψει (Predictions) το πόσο πολύ (viral) μπορεί να διαδοθεί ένα κομμάτι πληροφορίας.

3. Παράλληλα με τις δύο κατηγορίες έγινε μια έρευνα σε διάφορες προσεγγίσεις και ιδέες που βοηθάνε στην κατανόηση του προβλήματος προς επίλυση και ταυτόχρονα επιλύουν **υποπροβλήματα** στην συγκεκριμένη περιοχή (Auxiliary). Επιπλέον, εδώ βρίσκονται διάφορες λύσεις σε ποικίλα γνωστικά αντικείμενα που μπορούν να εμπνεύσουν την επίλυση του δικού μας προβλήματος χρησιμοποιώντας αναλογίες.

Για τις παραπάνω κατηγορίες ορίζονται τρία σύνολα προκειμένου να απεικονιστούν οι θεματικές κατηγορίες στις οποίες ανήκουν τα άρθρα που μελετήθηκαν. Για παράδειγμα άρθρα που πραγματεύονται το πως αξιολογούνται τα άτομα με επίδραση στο διαδίκτυο θα βρίσκονται στην πρώτη κατηγορία. Άρθρα τα οποία αξιολογούν τα άτομα με επίδραση στο διαδίκτυο και επιπλέον επιλύουν υποπροβλήματα, όπως ο εντοπισμός αυτοματοποιημένων προγραμμάτων που συμπεριφέρονται σαν χρήστες εξαπατώντας τον αλγόριθμο κατάταξης ατόμων με επίδραση στο διαδίκτυο, θα ανήκουν στις κατηγορίες 1&3 (στην τομή των συνόλων 1 και 3). Η σχηματική αναπαράσταση έχει καθαρά ρόλο κατηγοριοποίησης και δεν επισημαίνει την σημαντικότητα των άρθρων ανάλογα με την θέση την οποία κατέχουν.



Εικόνα 2: Σχηματική αναπαράσταση κατηγοριοποίησης εργασιών

Στη συνέχεια, βάση των παραπάνω συνόλων θα γίνει η παρουσίαση των άρθρων. Έστω ότι τα τρία σύνολα ονομάζονται M, I, A, όπου M={τα άρθρα τα οποία ανήκουν στο σύνολο Measuring Influence}, I={τα άρθρα τα οποία ανήκουν στο σύνολο Information

Diffusion & Predictions}, A={τα άρθρα τα οποία ανήκουν στο σύνολο Auxiliary}. Τα σύνολα τα οποία θα αναλυθούν στην συνέχεια είναι: M-I-A (τα άρθρα που περιέχονται μόνο στο M), I-M-A, A-I-M, M∩A (τα άρθρα που περιέχονται στο M και το A), I∩A, M∩I∩A. Για κάθε ένα από αυτά τα σύνολα θα ακολουθηθεί η εξής δομή:

- Θα δίνονται οι ορισμοί του Επηρεάζοντος (Influencer) για το κάθε άρθρο.
- Θα παρουσιάζονται οι προσεγγίσεις που ακολουθεί το κάθε άρθρο προκειμένου να λύσουν τα προβλήματα που πραγματεύονται.
- Θα παρουσιάζονται τα πλεονεκτήματα της κάθε προσέγγισης.
- Θα παρουσιάζονται τα μειονεκτήματα της κάθε προσέγγισης.

2.1 Κατηγορία: Μέτρο επίδρασης

Η συγκεκριμένη ενότητα αναλύει τα άρθρα που βρίσκονται στο σύνολο M-I-A.

2.1.1 Ορισμός του Επηρεάζοντος

Ένας πρώτος ορισμός των I. Anger και C. Kittl [1], θεωρεί ότι πριν η πληροφορία προκαλέσει κάποια αλλαγή στην συμπεριφορά θα πρέπει πρώτα να έρθει στα χέρια του χρήστη και στην συνέχεια να υποστεί επεξεργασία από τον χρήστη. Οπότε στόχος είναι να βρει κανείς άτομα που προκαλούν αυτό το μοτίβο στο κοινό τους.

Μια διαφορετική προσέγγιση του Daniel Tunkelang [2], υποστηρίζει ότι η δύναμη κάποιου ως προς την επίδραση είναι η ποσότητα της προσοχής που μπορεί να σου δώσει το κοινό σου συν την ποσότητα της προσοχής που μπορεί να σου φέρει το κοινό σου μέσω του δικού του κοινού.

Απλοποιώντας κάθε σχέση ο Thomas Renault [4], θεωρεί απλά ότι φίλοι ατόμων με επίδραση στο διαδίκτυο πάνω σε συγκεκριμένα θέματα, ως φίλοι, θα έχουν και αυτοί επίδραση στο διαδίκτυο και θα μιλάνε για παρόμοια θέματα.

Τέλος, έχοντας στο μυαλό το επιθυμητό αποτέλεσμα που θέλει να πετύχει κανείς με την μελέτη των ατόμων με επίδραση στο διαδίκτυο οι M. Cha, H. Haddadi, F. Benevenuto και K. Gummadi [15], ορίζουν την επίδραση ως την εκδήλωση μιας συμπεριφοράς που μπορεί να προκαλέσει ένα άτομο με την παροχή της πληροφορίας.

Επιπλέον, για την μελέτη τους χρησιμοποιούν την ήδη υπάρχουσα ιδέα του Watt και Dodd [21] σύμφωνα με την οποία η επιτυχία στην διάδοση της πληροφορίας δεν εξαρτάται από το άτομο που ξεκινάει να την διαδίδει, αλλά από το κατά πόσο η συγκεκριμένη πληροφορία θα ήταν ούτως ή άλλως μια πληροφορία η οποία θα διαδίδονταν. Για αυτόν τον λόγο ονομάζουν τα άτομα τα οποία έρχονται πρώτοι σε επαφή με την πληροφορία (early adopters) ή αλλιώς τα άτομα τα οποία εκφράζουν την κοινή γνώμη (opinion leaders), «κατά τύχη » Influencers.

2.1.2 Προσεγγίσεις που ακολουθούνται για τον εντοπισμό ατόμων με επίδραση στο διαδίκτυο

Οι I. Anger και C. Kittl θέλοντας να συμπεριλάβουν το πως ο χρήστης έρχεται σε επαφή με την πληροφορία και στο πως αυτός αλληλοεπιδρά με αυτήν μέσα στο κοινωνικό δίκτυο του Twitter δημιούργησαν τους δείκτες: ri&rRT

I. ri είναι ένας δείκτης αλληλεπίδρασης και χρησιμοποιείται με την έννοια του πλήθους των ατόμων που επαναδημοσιεύουν (retweet) μια δημοσιευμένη πληροφορία ενός χρήστη Α ή αναφέρουν (mention) τον χρήστη Α σε κάποιο σχόλιο και διαιρείται με το σύνολο των ακόλουθων (followers) του χρήστη Α. (συζητηση-ο-κεντρική προσέγγιση)
II. Ο δεύτερος δείκτης, ο rRT είναι ο λόγος των αναδημοσιεύσεων ενός χρήστη προς το πλήθος των αναφορών που έχει αυτός ο χρήστης (Retweet/Mention) (περιεχομενο-κεντρική προσέγγιση).

Χρησιμοποιώντας μόνο την πληροφορία που περιέχουν αυτοί οι δείκτες και συνδυάζοντας τους με έναν απλό μέσο όρο πολλαπλασιασμένο με 100% αντιστοιχίζεται ένας αριθμός για τον κάθε χρήστη. Ταξινομώντας τους χρήστες με βάση αυτόν τον αριθμό προκύπτει η κατάταξη των χρηστών βάση της επίδρασής τους. Να τονιστεί ότι λόγω του ότι οι αναδημοσιεύσεις ενός χρήστη και οι αναφορές ως προς έναν χρήστη είναι δυνατόν να είναι περισσότερες από το πλήθος των δημοσιεύσεων ενός χρήστη οι αριθμοί που αντιστοιχίζονται στους χρήστες μπορεί να είναι μεγαλύτεροι από 100%. [1]

Σε άρθρο του ο Daniel Tunkelang, παρουσιάζει μια ιδέα ανάλογη με τον Αλγόριθμο του PageRank. Ο PageRank είναι ένας από τους πιο διάσημους αλγόριθμους του 20ου αιώνα και χρησιμοποιείται (με αρκετές τροποποιήσεις πλέον) για την κατάταξη των

ιστοσελίδων του διαδικτύου. Ο νέος αυτός αλγόριθμος που παρουσιάζεται δεν κατατάσσει πλέον τις ιστοσελίδες αλλά τα άτομα με επίδραση στο διαδίκτυο και ονομάζεται TunkRank (T). Η υπολογιστική φόρμουλα που ακολουθεί ο TunkRank είναι η παρακάτω:

$$T(X) = \sum_{Y \in Followers(X)} p^{notice} * \frac{1 + p^{retweet} * T(Y)}{Following(Y)}$$

$$p^{notice} = συνολική\ προσοχή\ που\ ο\ χρήστης\ αφιερώνει\ στο\ Twitter$$

$$p^{retweet} = πιθανότητα\ ο\ χρήστης\ να\ κάνει\ επαναδημοσιεύση\ (retweet)$$

Το δεύτερο αυτό άρθρο συνεχίζει να χρησιμοποιεί την λογική που χρησιμοποιήθηκε στο πρώτο άρθρο, με την έννοια ότι συμπεριλαμβάνει: α. τις επαναδημοσιεύσεις που δέχεται ένας χρήστης σε μια δημοσίευση του, β. το πως οι χρήστες έρχονται σε επαφή με την πληροφορία (p^{notice}) και γ. το πόσα άτομα ακολουθούν τον εκάστοτε χρήστη. Η διαφορά εδώ είναι ότι συνδέονται με τελείως διαφορετικό τρόπο αυτά τα στοιχεία μεταξύ τους, χρησιμοποιώντας έναν αναδρομικό τύπο και συμπεριλαμβάνοντας πληροφορία σε μορφή πιθανοτήτων. [2]

Η προσέγγιση που ακολουθείται από τον Thomas Renault αποτελεί έναν ιδιαίτερα απλό αλγόριθμο βασιζόμενο στην ιδέα του ορισμού του για τον Επηρεάζοντα. Χρησιμοποιώντας μια γενικώς αποδεκτή γνώση ορίζει τα 10 πιο σημαντικά Ιστολόγια (Blogs) πάνω σε ένα συγκεκριμένο θέμα, εξάγει πληροφορίες για τους δημιουργούς των Ιστολογίων και στην συνέχεια τους εισάγει στην λίστα των αποτελεσμάτων του. Έχοντας εξαγάγει την πληροφορία για τους 10 χρήστες που ο ίδιος εισήγαγε στην λίστα βρίσκει τους λογαριασμούς αυτών των χρηστών στο Twitter (ή στο Facebook) και με την βοήθεια του Twitter API βρίσκει τους φίλους αυτών των χρηστών στο Twitter. Στην συνέχεια εισάγει στην λίστα των αποτελεσμάτων τα 20 άτομα με την μεγαλύτερη επικάλυψη κοινών φίλων. Η διαδικασία συνεχίζεται για 3 ακόμα επαναλήψεις. Στην 2η επανάληψη βρίσκει την μεγαλύτερη επικάλυψη των 30 φίλων που βρίσκονται στην λίστα και εισάγει τους 20 πρώτους κ.ο.κ. Το αποτέλεσμα είναι οι 90 πιο επιδρώντες χρήστες του Twitter σχετικά με την θεματική ενότητα στην οποία ανήκουν τα Ιστολόγια που προστέθηκαν στην αρχή. Να τονιστεί ότι αν και επεξεργάζεται ελάχιστη πληροφορία από το διαδίκτυο τα αποτελέσματα που βγαίνουν είναι αρκετά κοντά στα αποτελέσματα που βγάζει ο Αλγόριθμος του Klout (Ο αλγόριθμος του Klout είναι ένας αλγόριθμος που θα αναφέρουμε αργότερα και χρησιμοποιείται από πολλούς ερευνητές για την αξιολόγηση των αποτελεσμάτων τους). Η συγκεκριμένη ιδέα αποτελεί μια

τελείως διαφορετική προσέγγιση από τις δύο παραπάνω που προαναφέρθηκαν και η λειτουργία της απαιτεί την γνώση των πρώτων 10 χρηστών. [4]

Οι M. Cha, H. Haddadi, F. Benevenuto και K. Gummadi [15], προκειμένου να μετρήσουν την επίδραση των χρηστών δημιούργησαν ένα σύνολο από 6 εκατομμύρια χρήστες του Twitter, τον Αύγουστο του 2009. Στην συνέχεια για κάθε έναν από αυτούς τους χρήστες βρήκαν τις 3 μεταβλητές με τις οποίες μετράνε την επίδραση:

1. Πλήθος Ακόλουθων (Indegree Influence)
2. Πλήθος Επαναδημοσιεύσεων (Retweet Influence)
3. Πλήθος Αναφορών (Mention Influence)

Το επόμενο βήμα ήταν να δημιουργήσουν τρεις ταξινομήσεις, σε φθίνουσα σειρά, για τους χρήστες βάση των παραπάνω 3 μεταβλητών. Επιπλέον, για να συσχετίσουν τις μεταβλητές μεταξύ τους χρησιμοποίησαν τον συντελεστή συσχέτισης για κατατάξεις του Spearman, που έχει τύπο συσχέτισης:

$$\rho = 1 - \frac{6\sum(x_i - y_i)^2}{N^3 - N}$$

Ο συγκεκριμένος συντελεστής δίνει τον βαθμό συσχέτισης μεταξύ δύο κατατάξεων, των xi και yi. Εν προκειμένω ο συντελεστής συσχέτισης να δώσει πιο σωστά αποτελέσματα, επικεντρώνονται μόνο στην περιοχή όπου βρίσκονται τα άτομα με επίδραση (χρήση του 10% και 1% βάση του δείκτη «Πλήθος Ακόλουθων»). Για την εύρεση αυτής της περιοχή κάνουν την παραδοχή ότι άτομα τα οποία έχουν επίδραση στο διαδίκτυο θα έχουν ένα αξιόλογο αριθμό ακόλουθων (followers). Μετά από αυτό το κοσκίνισμα των χρηστών, αφού κράτησαν δηλαδή το 10% και 1% των καλύτερων χρηστών, βάση του δείκτη των ακόλουθων, συμπέραναν ότι υπάρχει μεγάλη συσχέτιση μεταξύ των μεταβλητών Πλήθος Επαναδημοσιεύσεων και Πλήθος Αναφορών. Έτσι συνάγουν ότι χρήστες που αναφέρονται συχνά επαναδημοσιεύονται και συχνά. Ωστόσο για το πλήθος των ακόλουθων αναφέρουν ότι δεν συσχετίζεται με τις άλλες δύο μεταβλητές. Συνεπώς άτομα τα οποία έχουν απλά πολλούς ακόλουθους δεν σημαίνει απαραίτητα ότι είναι και άτομα με επίδραση. Ιδιαίτερη μνεία πρέπει να δοθεί στο πείραμα που έκαναν για την κανονικοποίηση των μεταβλητών τους, δηλαδή στην διαίρεση των επαναδημοσίευσεων και αναφορών με το σύνολο των δημοσιεύσεων του χρήστη. Από το συγκεκριμένο πείραμα αναφέρουν ότι προκύπτουν διαφορετικά αποτελέσματα και μάλιστα όχι τόσο κοντά όσο αυτά που θα ήθελαν. Συνεπώς, ένα επιπλέον συμπέρασμα που εξάγουν είναι ότι ακόμα και μικρές αλλαγές στις μεταβλητές

προς μελέτη συνάγουν εντελώς διαφορετικά συμπεράσματα. Στην συνέχεια, εξετάζουν αν οι χρήστες με επίδραση στο διαδίκτυο πάνω σε μια θεματική ενότητα διατηρούν αυτήν τους την επίδραση σε διαφορετικές θεματικές ενότητες, απαντώντας θετικά στον προβληματισμό τους. Δηλαδή αναφέρουν ότι αν κάποιος είναι άτομο με επίδραση στο διαδίκτυο θα έχει επίδραση πάνω σε πολλές θεματικές κατηγορίες. Τέλος, θέλοντας να απαντήσουν στο ερώτημα για το πως κανείς μπορεί να αποκτήσει επίδραση στο διαδίκτυο, δηλώνουν ότι η επίδραση δεν είναι κάτι το οποίο κερδίζεται κατά τύχη αλλά μέσα από μια σειρά προσπαθειών. Για άλλη μια φορά βλέπουμε ότι συνδυάζοντας με απλό τρόπο τις μεταβλητές (πλήθος ακόλουθων, πλήθος επαναδημοσιεύσεων, πλήθος αναφορών) βγαίνουν αποτελέσματα που βασίζονται στην πραγματικότητα. [15]

2.1.3 Πλεονεκτήματα

Οι I. Anger και C. Kittl δημιουργώντας δύο απλούς δείκτες και συνδυάζοντας τους με κατάλληλο τρόπο καταφέρνουν να δημιουργήσουν μια απλή και αποτελεσματική προσέγγιση, εξάγοντας αποτελέσματα παρόμοια με τον αλγόριθμο του Klout. Επιπλέον, λόγω της απλότητας του αλγορίθμου τα αποτελέσματα εξάγονται ιδιαίτερα γρήγορα (γρήγορη σύγκλιση) και η υλοποίηση του αλγορίθμου καθίσταται ιδιαίτερα εύκολη (υπάρχει ήδη διαθέσιμη). Να τονιστεί ότι με τον όρο απλότητα εννοούμε το γεγονός ότι πρόκειται για αλγορίθμους οι οποίοι βρίσκονται στα πρώτα στάδια της εξέλιξής τους και συνεπώς το βάθος του δέντρου που δημιουργείται για την εύρεση λύσης είναι ιδιαίτερα μικρό. [1]

Λόγω της ίδιας απλότητας στην προσέγγισή του ο Thomas Renault καταφέρνει να πετυχαίνει τα ίδια πλεονεκτήματα επιλύοντας με διαφορετικό τρόπο το πρόβλημα εύρεσης ατόμων με επίδραση στο διαδίκτυο. [4]

Ακολουθώντας μια προσέγγιση που χρησιμοποιεί το ίδιο απλούς δείκτες, ο TunkRank καταφέρνει να έχει τα ίδια πλεονεκτήματα λόγω της απλότητας του αλγορίθμου. Επιπλέον όμως, εισάγει δύο επιπρόσθετα απαραίτητα στοιχεία στην ανάλυση δεδομένων μεγάλου όγκου: α. Την πιθανότητα να συμβεί ένα γεγονός, δίνοντας με αυτόν τον τρόπο δυνατότητα προσαρμογής της πιθανότητας με την πάροδο του χρόνου και β. χρησιμοποιεί αναδρομική σχέση, εισάγοντας με αυτόν τον τρόπο μια σειρά αλληλεξαρτήσεων μεταξύ των χρηστών. [2]

Η ίδια απλότητα διέπει και τους 3 δείκτες που χρησιμοποιούν οι M. Cha, H. Haddadi, F. Benevenuto και K. Gummadi δίνοντας τους τα προφανή πλεονεκτήματα. Οι επιπλέον συνεισφορές μέσω της ερευνάς τους είναι το συμπέρασμα ότι: α. χρήστες οι οποίοι αναφέρονται συχνά επαναδημοσιεύονται και συχνά, β. χρήστες οι οποίοι είναι καλοί στο να αναφέρονται συχνά από άλλους χρήστες καταφέρνουν να αναφέρονται συχνά σε πολλές διαφορετικές θεματικές ενότητες, γ. άτομα τα οποία θεωρούνται αυθεντίες στην διάδοση της γνώσης μπορούν να επηρεάζουν το κοινό τους σε πολλές διαφορετικές θεματικές ενότητες, δ. όσον αφορά την οικονομική πολιτική που μπορεί να ακολουθηθεί για την χρήση ατόμων με επίδραση στο διαδίκτυο είναι πιο αποδοτικό να δίνει κανείς τα χρήματά του σε άτομα τα οποία έχουν μεγάλη επίδραση από το να δίνει την ίδια ποσότητα χρημάτων σε πολλά άτομα τα οποία έχουν μικρότερη επίδραση και ε. τέλος, δίνουν εξηγήσεις για το πως οι χρήστες διατηρούν την επίδραση τους κατά την διάρκεια των χρόνων.[15]

2.1.4 Μειονεκτήματα

Η απλότητα της προσέγγισης των αλγορίθμων: των I. Anger και C. Kittl [1], του TunkRank [2] και του Thomas Renault [4], ορίζει τον περιορισμό να μην μπορούν να χρησιμοποιηθούν απευθείας οι αλγόριθμοί τους αλλά ως βάση για παραπέρα έρευνα.

Οι I. Anger και C. Kittl, επιπλέον, δεν λαμβάνουν υπόψιν τους το πλήθος των ατόμων που ακολουθούν τους χρήστες καθώς και την ποιότητα των αλληλεπιδράσεων μεταξύ των χρηστών. Τέλος, αγνοούν την πιθανότητα οι χρήστες να αλληλοεπιδράσουν με το περιεχόμενο του Twitter. [1]

Ο TunkRank, του Daniel Tunkelang, όντας στα πρώτα στάδια της εξέλιξης του δεν καταφέρνει να δώσει ακριβή αποτελέσματα αδυνατώντας να υπολογίσει την δυναμική των επαναδημοσιεύσεων και το ενδεχόμενο αλληλεπίδρασης του χρήστη με τις δημοσιεύσεις (π.χ. με τα σχόλια των χρηστών).

Ο Thomas Renault βασιζόμενος μόνο σε μια βασική ιδέα δεν χρησιμοποιεί πολλά δεδομένα από το διαδίκτυο. Επιπλέον, ο αλγόριθμος του εν προκειμένω να δουλέψει απαιτεί την γνώση της αρχικής λίστας και τα αποτελέσματα εξαρτώνται από την επιλογή της αρχικής λίστας.[4]

Αντιμετωπίζοντας λίγο διαφορετικά προβλήματα οι M. Cha, H. Haddadi, F. Benevenuto και K. Gummadi δεν κατάφεραν να διαχωρίσουν κανονικούς χρήστες από

ηλεκτρονικά ρομπότ (bots) και ως συνέπεια δεν κατάφεραν να αξιοποιήσουν την πληροφορία των επαναδημοσιεύσεων και το πλήθος των ατόμων που ακολουθούν έναν χρήστη. Επιπλέον, κατά την συλλογή των δεδομένων δεν γνώριζαν τον χρόνο δημιουργίας των σχέσεων φιλίας των χρηστών (follow links).

2.2 Κατηγορία: Διάδοση της πληροφορίας – προβλέψεις

Η συγκεκριμένη ενότητα αναλύει τα άρθρα που βρίσκονται στο σύνολο I-M-A.

2.2.1 Προσεγγίσεις που ακολουθούνται για την πρόβλεψη των δημοσιεύσεων

Οι Q. Zhao, M. Erdogdu, H. He, A. Rajaraman, J. Leskovec [14], με το συγκεκριμένο άρθρο τους σε σχέση με τα άρθρα [7],[8], στηρίζουν ότι μπορούν να γίνουν προβλέψεις για το πόσο πολύ θα διαδοθεί ένας σύνδεσμος και κατά συνέπεια ποιοι σύνδεσμοι θα διαδοθούν περισσότερο (κατάταξη σε επίπεδο δημοσιεύσεων), με χρήση ενός στατιστικού μοντέλου. Η διαφοροποίηση με όλα τα προηγούμενα άρθρα που προσπαθούν να κάνουν προβλέψεις έγκειται στο ότι σε αυτήν την εργασία δεν προσεγγίζουν την λύση τους με χρήση τεχνικών μηχανικής μάθησης, που είναι αρκετά δαπανηρή μέθοδος, περιέχοντας επιπλέον αρκετή αβεβαιότητα όσον αφορά την επιλογή των χαρακτηριστικών εξαγωγής για την εκμάθηση των μοντέλων και καταναλώνοντας αρκετούς υπολογιστικούς πόρους. Αναγνωρίζοντας λοιπόν ότι είναι αδύνατον να γίνουν προβλέψεις με τεχνικές μηχανικής μάθησης όπως άλλωστε συμφωνούν και οι [7],[8], ακολουθούν ένα στατιστικό μοντέλο που ονομάζεται «self-exciting point process». Στόχος τους είναι να κάνουν σε πραγματικό χρόνο προβλέψεις για το πόσο πολύ θα διαδοθεί ένας σύνδεσμος. Αρχικά εντοπίζουν ότι κάθε δημοσίευση περνάει από 2 στάδια: α. κρίσιμο (supercritical), b. μη κρίσιμο (subcritical). Όταν οι δημοσιεύσεις βρίσκονται στο κρίσιμο στάδιο τότε δεν μπορούν να γίνουν σωστές προβλέψεις. Όταν βρίσκονται στο μη κρίσιμο στάδιο, τότε μπορούν να γίνουν προβλέψεις. Το κλειδί εδώ είναι ότι οι δημοσιεύσεις βρίσκονται στο κρίσιμο στάδιο για πολύ μικρό χρονικό διάστημα, οπότε τις περισσότερες φορές οι προβλέψεις τους είναι ορθές. Συγκεκριμένα αν τρέξουν τον αλγόριθμό τους για 10 λεπτά δίνουν αποτελέσματα με σφάλμα 25%, ενώ αν τον τρέξουν για 1 ώρα το σφάλμα είναι 15%.

Επιπλέον, προβλέπουν κατά 30% καλύτερα αποτελέσματα από τους καλύτερους μέχρι σήμερα διαθέσιμους αλγορίθμους.

2.2.2 Πλεονεκτήματα

Το μοντέλο τους δεν απαιτεί τεχνικές μηχανικής μάθησης (machine learning) και ακριβές μηχανικές τεχνικές. Επιπλέον, δημιούργησαν μια απλή και αποδοτική φόρμουλα που επιτρέπει να δίνουν απαντήσεις σε πραγματικό χρόνο, σε αντίθεση με τις τεχνικές μηχανικής μάθησης που δεν προσφέρουν αυτήν την δυνατότητα. Τρέχοντας τον αλγόριθμό τους για μια ώρα το σφάλμα πρόβλεψης ανέρχεται μόνο στο 15%. Το στατιστικό τους μοντέλο ξεπερνάει το πρόβλημα της εξαγωγής πολλών δεδομένων από το διαδίκτυο καθώς χρειάζονται μόνο βασικά δεδομένα (την ώρα επαναδημοσίευσης και το βαθμό του κόμβου) για να τρέξει ο αλγόριθμός τους. Αυτό αποτελεί αρκετά σημαντικό γεγονός γιατί η πληροφορία που χρησιμοποιείται για να γίνουν προβλέψεις μπορεί να χρησιμοποιηθεί και για την κατάταξη των ατόμων με επίδραση στο διαδίκτυο.

2.2.3 Μειονεκτήματα

Πρόκειται για μια καινούργια ιδέα, οπότε δεν έχουν συμπεριλάβει δεδομένα που αφορούν το περιεχόμενο των δημοσιεύσεων και επιπλέον δεν έχουν διαθέσιμη την δομή του κοινωνικού δικτύου (network structure). [14]

2.3 Κατηγορία: Επιπλέον υλικό

Η συγκεκριμένη ενότητα αναλύει τα άρθρα που βρίσκονται στο σύνολο A-M-I.

2.3.1 Προσεγγίσεις που ακολουθούνται

Στην κατεύθυνση αξιολόγησης της αποδοτικότητας του Επηρεάζοντος, όσον αφορά την εύρεση της αγοραστικής δύναμης των ατόμων που επηρεάζονται, έρχεται να απαντήσει ο Perry Marshall [16], χρησιμοποιώντας τον κανόνα 80/20. Ο

συγκεκριμένος κανόνας πηγάζει από θεωρίες του μάρκετινγκ και στηρίζει ότι το 20% των πελατών μιας εταιρίας είναι αυτό που δημιουργεί το 80% των πωλήσεών της. Αυτό σημαίνει ότι μια σημαντική μετρική, πέρα του αριθμού των ατόμων που ενδεχομένως να επηρεάζει κάποιος, είναι και τον ποιον επηρεάζει. Το συγκεκριμένο ερώτημα το αξιολογούν και το μετρούν στην συνέχεια οι D. Garcia1, P Mavrodiev, D Casati, F. Schweitzer [18], αλλά στην κατεύθυνση της ποσότητας της επίδρασης που έχουν τα άτομα που επηρεάζονται και όχι στην κατεύθυνση της αγοραστικής τους δύναμης. Πρέπει να τονιστεί ότι έχει πολύ μεγαλύτερη αξία αν κάποιος επηρεάζει λίγα άτομα αλλά αυτά τα λίγα βρίσκονται στο 20% που προηγουμένως αναφέρθηκε. Με την συγκεκριμένη αναφορά, θίγεται μια παράμετρος που είναι σημαντική να ληφθεί υπόψιν για την αξιολόγηση της ποιότητας των ατόμων που επηρεάζει ο Επηρεάζων.

Στην έρευνα τους οι N. Spasojevic, Z. Li, A. Rao, P. Bhattacharyya [20], εντοπίζουν ποιες είναι οι καλύτερες ώρες για να δημοσιεύει ένας χρήστης προκειμένου να βελτιστοποιήσει την πιθανότητα συμμετοχής και αντίδρασης του κοινού του. Η μελέτη τους περιλαμβάνει την εβδομαδιαία μέτρηση των αντιδράσεων των χρηστών σε διάφορες ώρες τις ημέρας για διάστημα 8 εβδομάδων. Αυτό που βρίσκουν είναι ότι οι περισσότερες αντιδράσεις που λαμβάνει μια δημοσίευση συμβαίνουν σε διάστημα 2 ωρών από την στιγμή της δημοσίευσης. Επιπλέον, δείχνουν ότι κάθε κοινωνικό μέσο δικτύωσης έχει τα δικά του χαρακτηριστικά και συμπεριφέρεται με διαφορετικό τρόπο.

2.4 Κατηγορία: Μέτρο επίδρασης & επιπλέον υλικό

Η συγκεκριμένη ενότητα αναλύει τα άρθρα που βρίσκονται στο σύνολο ΜΟΑ.

2.4.1 Ορισμός του Επηρεάζοντος

Η επίδραση ως αλλαγή της συμπεριφοράς διατυπωμένη από τους A. Rao, N. Spasojevic, Z. Li and T. Dsouza [3], εκφράζεται ως η δράση αντίδραση μιας πράξης, με την έννοια ότι αν κάποιος δημοσιεύσει μια πληροφορία και υπάρξει κάποιου είδους αντίδραση από κάποιο άλλο άτομο, τότε λέμε ότι ο πρώτος επηρέασε τον δεύτερο. Επιπλέον, δίνουν ιδιαίτερη έμφαση στην διαφορετική ποσοτικοποίηση αυτών των αντιδράσεων, ανάλογα με τον τρόπο αντίδρασης.

Στην συνέχεια οι N. Booth, J. A. Matic [5], βασιζόμενοι αποκλειστικά στην διαφήμιση μιας εταιρίας και γενικότερα στην εικόνα της προς το ευρύτερο κοινό, αναφέρουν ως επίδραση την δυνατότητα που έχει ένα άτομο να δημιουργεί συζητήσεις για μια εταιρία ως συνάρτηση του αντίκτυπου αυτών των συζητήσεων στο ευρύτερο κοινό.

Παραθέτοντας ένα σύνολο από ορισμούς των ατόμων με επίδραση στο διαδίκτυο και αφήνοντας τον χρήστη να διαλέξει κάθε φορά τον ορισμό που επιθυμεί, οι M. Petychakis κ.α [9], λαμβάνουν υπόψη τα εξής:

1. Πόσο συχνά οι χρήστες αναφέρονται σε συζητήσεις τρίτων.
2. Πόσο συχνά αναφέρονται οι γνώμες κάποιων ατόμων από τρίτους.
3. Πόσο αποδεκτές και αρεστές γίνονται οι γνώμες κάποιων ατόμων από τρίτους.
4. Πόσο γνωστά και αναγνωρισμένα από το ευρύτερο κοινό είναι κάποια άτομα.

Οι J. Weng, E. Lim, J. Jiang, Q. He [10], έχοντας στο μυαλό τους το σύνολο των ατόμων που προσεγγίζει η πληροφορία που διαδίδει ένας χρήστης ενός κοινωνικού δικτύου, δημιουργούν έναν ορισμό για την επίδραση που βασίζεται στην ποσότητα της πληροφορίας που λαμβάνει το κοινό ενός χρήστη σε πρώτο στάδιο και σε δεύτερο στάδιο το άθροισμα της επίδρασης που ασκεί το κοινό του χρήστη στο δικό του κοινό αντίστοιχα.

Αντίθετα, χρησιμοποιώντας ιδέες από το γνωστικό αντικείμενο της κοινωνιολογίας και συγκεκριμένα του πολυβραβευμένου επιστήμονα M. Granovetter (1973) οι Aral Sinan και Dylan Walker [12], ορίζουν 2 μεταβλητές:

I. Ισχυρότητα δεσμών (Tie Strength) που ορίζει την σημαντικότητα της σχέσης μεταξύ των ατόμων και
II. Ενσωμάτωση (Embeddedness), που ορίζει το πλήθος των κοινών φίλων που δύο άτομα μοιράζονται.

Να τονιστεί εδώ ότι η μορφή που έχουν τα κοινωνικά δίκτυα σήμερα, με την έννοια της μορφής του γράφου που έχει σχηματιστεί, μελετάται πολύ πριν την δημιουργία του Twitter.

Τέλος, στηρίζοντας την επίδραση σε δύο έννοιες, την «δημοτικότητα» (popularity) και «φήμη» (reputation), οι D. Garcia, P. Mavrodiev, D. Casati, F. Schweitzer [18], θεωρούν ότι για τον υπολογισμό της επίδρασης είναι απαραίτητη και η μέτρηση του μεγέθους του κοινού που μπορεί να επηρεάσει ένας χρήστης αλλά και το είδος του κοινού που επηρεάζει.

2.4.2 Προσεγγίσεις που ακολουθούνται για τον εντοπισμό ατόμων με επίδραση στο διαδίκτυο

Βασιζόμενοι στον αλγόριθμο του Klout (ο οποίος δεν είναι δημοσιευμένος) και επεκτείνοντάς τον, οι A. Rao, N. Spasojevic, Z. Li and T. Dsouza [3], παρουσιάζουν μια προσέγγιση τελείως διαφορετική από αυτές που έχουν παρουσιαστεί μέχρι τώρα. Για να κατατάξουν τα άτομα με επίδραση στο διαδίκτυο, προτείνουν την σταθμισμένη άθροιση 3600 χαρακτηριστικών, που λαμβάνουν πληροφορία καθημερινά από πολλές διαφορετικές πλατφόρμες κοινωνικών δικτύων όπως και την προσθήκη πληροφορίας από τον έξω κόσμο (Wikipedia & Άρθρα). Η συγκεκριμένη εργασία χρησιμοποιεί τεχνικές μηχανικής μάθησης (machine learning) κατασκευάζοντας το σύστημα με τέτοιον τρόπο, ώστε να μπορούν συνέχεια να προστίθενται και να αφαιρούνται νέα χαρακτηριστικά (επεκτασιμότητα συστήματος). Προκειμένου να μελετήσουν όλο το κοινωνικό δίκτυο, το οποίο θεωρούν απαραίτητη προϋπόθεση για σωστή αξιολόγηση των χρηστών, χωρίζουν το γράφο του κοινωνικού δικτύου σε υπογράφους (batch processing), δημιουργώντας με αυτόν τον τρόπο διαφορετικά χαρακτηριστικά (features) ανά χρήστη. Επιπλέον, χωρίζουν κάποιους δείκτες επίδρασης σε προσωρινούς (1. Σχόλια, 2. Απαντήσεις Σχολίων, 3. Μου αρέσει (Like), 4. Αναφορές, 5. Επαναδημοσιεύσεις, κ.α) και κάποιους άλλους σε μόνιμους (1. Ακόλουθοι, 2. Φίλοι, 3. Θαυμαστές, 4. Συνδρομητές, κ.α). Για την εύρεση των προσωρινών δεικτών απαιτείται μια παρατήρηση που διαρκεί 90 μέρες και προφανώς οι τιμές των αποτελεσμάτων αλλάζουν διαρκώς με την πάροδο του χρόνου. Εστιάζοντας στην ποσοτικοποίηση της επίδρασης παρουσιάζουν την εξής περίπτωση: Το να προκαλέσει μια δημοσίευση 100 αντιδράσεις από 10 άτομα είναι τελείως διαφορετικό από το να προκαλέσει 100 αντιδράσεις από 50 άτομα. Η πρώτη περίπτωση μας λέει ότι αυτός ο χρήστης έχει πολύ μεγάλη επίδραση στους 10 αντιδρώντες, ενώ η δεύτερη περίπτωση μας λέει ότι ο χρήστης έχει μεγαλύτερη απήχηση. Τέλος, δείχνουν ότι ο αλγόριθμός τους είναι επεκτάσιμος σε πολλές θεματικές κατηγορίες. Βλέπουμε για άλλη μια φορά την σημαντικότητα μελέτης της επίδρασης σε διαφορετικές θεματικές ενότητες και επιπλέον την δυναμική που προσφέρουν οι τεχνικές μηχανικής μάθησης.

Από την άλλη πλευρά οι N. Booth, J. A. Matic [5], υποστηρικτές των πιο κλασσικών μεθόδων, χρησιμοποιώντας ένα σύνολο δεικτών για την ποσοτικοποίηση και αξιολόγηση της ποιότητας και της ποσότητας της επίδρασης των bloggers στο διαδίκτυο, δημιουργούν μια προσαρμοσμένη στον χρήστη προσέγγιση. Υπάρχουν

πάρα πολλές παράμετροι με τις οποίους μπορεί κάποιος να φτιάξει μια κατάταξη για το πόσο πολύ μπορεί να επηρεάσει κάποιος το κοινό του. Αντί λοιπόν να παραδίδουν κάποιον αλγόριθμο έτοιμο, δίνουν την δυνατότητα στον χρήστη να σταθμίσει ο ίδιος τις μεταβλητές που θεωρεί πιο σημαντικές μέσα από ένα σύνολο διαθέσιμων μεταβλητών (1. Viewers per Month, 2. Linkages, 3. Post Frequency, 4. Media Citation Score, 5. Industry Score, 6. Social Aggregator Rate, 7. Engagement Index, 8. Subject Topic Related Post, Qualitative Subject/Topic-Related Posts). Το αποτέλεσμα, Index Score, προκύπτει ως σταθμισμένος μέσος όρος των μεταβλητών, με μεταβλητές και βάρη που διάλεξε ο εκάστοτε χρήστης. Στην συνέχεια για να δώσουν ακόμη ένα επίπεδο προσαρμογής στις ανάγκες του χρήστη (second level customization) δημιουργούν τις «Κατηγορίες» (Tiers). Για παράδειγμα οι χρήστες που ανήκουν στην Κατηγορία Α τείνουν να προσανατολίζονται σε δημοσιεύσεις που αφορούν νέα της επικαιρότητας. Οι χρήστες που ανήκουν στην Κατηγορία Β τείνουν να έχουν δημοσιεύσεις που αφοράνε συγκεκριμένα επιστημονικά ζητήματα όπου είναι πολύ δύσκολο να βρεθεί αλλού αυτή η πληροφορία πέρα από την κατηγορία Β. Οι χρήστες που ανήκουν στην κατηγορία C επηρεάζουν ένα μικρότερο κοινό αλλά πολύ πιο έντονα. Αυτές οι κατηγορίες είναι και παράμετροι που συζητήθηκαν και από τους A. Rao, N. Spasojevic, Z. Li and T. Dsouza [3], αλλά δεν οροθετήθηκαν όπως εδώ. Να τονιστεί ότι το αποτέλεσμα του αλγορίθμου τους είναι η επίδραση που έχει κάθε blogger σε διάφορες θεματικές κατηγορίες. Αυτό κάνει την συγκεκριμένη προσέγγιση να επιβεβαιώνει ακόμη μια φορά την πεποίθηση ότι η ανάλυση των ατόμων με επίδραση πρέπει να γίνεται πάνω σε διαφορετικές θεματικές κατηγορίες.

Οι M. Danisch, N. Dugu, A. Perez [6], αντιλαμβανόμενοι το κύριο μειονέκτημα του αλγορίθμου του Klout, την αδυναμία του να ξεχωρίσει πραγματικούς χρήστες από χρήστες που απλά επιζητάνε υψηλά σκορ ή από ηλεκτρονικά ρομπότ (bot) που εκτελούν αυτοματοποιημένες διαδικασίες για αυτόν τον σκοπό, τους λεγόμενους «Κοινωνικούς Καπιταλιστές» (social capitalist), βρίσκουν την πιθανότητα ένας χρήστης να είναι Κοινωνικός Καπιταλιστής και επαναπροσδιορίζουν την κατάταξη του αλγορίθμου του Klout βάση αυτής της πιθανότητας. Ο λόγος που ο αλγόριθμος του Klout όπως και άλλοι γνωστοί αλγόριθμοι που μετράνε την επίδραση στο διαδίκτυο, δεν καταφέρνουν να εντοπίσουν τους χρήστες που δημιουργούν το πρόβλημα, είναι ότι αυξάνουν τα σκορ των χρηστών ανάλογα με τον αριθμό των ακόλουθων που έχουν. Μια λύση που προτείνουν σε αυτήν την κατεύθυνση είναι η μελέτη των συμβόλων #

(hashtags) που περιλαμβάνουν προτάσεις όπως: «#TeamFollowBack», «#instantfollowbackdedicated», «#teamautofollow». Επιπλέον, θέλοντας να συμπεριλάβουν και μια σταθερή παράμετρο στην μέτρησή τους χρησιμοποιούν τον λόγο Φίλοι/Ακόλουθοι. Ο διαχωρισμός των χρηστών και η εύρεση της πιθανότητάς τους να είναι Κοινωνικοί Καπιταλιστές υλοποιείται με τεχνικές μηχανικής μάθησης. Η συγκεκριμένη εργασία είναι ιδιαίτερα χρήσιμη καθώς η πιθανότητα του να είναι κανείς Κοινωνικός Καπιταλιστής μπορεί να προσαρμόσει και να βελτιώσει την κατάταξη οποιουδήποτε αλγορίθμου.

Στην συνέχεια οι M. Petychakis, E. Biliri, A. Arvanitakis κ.α [9], ακολουθούν μια διαφορετική κατεύθυνση όσον αφορά την αξιοποίηση των ατόμων με επίδραση στο διαδίκτυο. Προκειμένου να μειώσουν το πρόβλημα μελέτης του τεράστιου όγκου δεδομένων που είναι διαθέσιμο στο διαδίκτυο, για εντοπισμό προτιμήσεων των καταναλωτών όσον αφορά τα προϊόντα, χρησιμοποιούν τα άτομα με επίδραση στο διαδίκτυο για πιο στοχευμένη επιλογή της πληροφορίας προς επεξεργασία. Για την εύρεση των ατόμων με επίδραση στο διαδίκτυο χρησιμοποιούν μια τροποποιημένη έκδοση του αλγορίθμου του PageRank. Στην συνέχεια για να βρουν τις θεματικές κατηγορίες τις οποίες συζητάει ο κάθε Επηρεάζων, βρίσκουν την συσχέτιση που έχουν οι νέες τάσεις των καταναλωτών με το περιεχόμενο των συζητήσεων που εμπλέκονται. Τέλος, η ποικιλία των ορισμών που παρέχουν για προσδιορισμό ατόμων με επίδραση στο διαδίκτυο δίνει την δυνατότητα στους χρήστες να διαλέγουν ανάμεσα σε πολλούς αλγορίθμους, παραθέτοντας με αυτόν τον τρόπο μια βελτιωμένη προσαρμογή στις ανάγκες των χρηστών (customization) σε σχέση με την πρόταση των N. Booth, J. A. Matic [5].

Οι J. Weng, E. Lim, J. Jiang, Q. He [10], είναι οι πρώτοι που αναφέρουν τον όρο ομοφυλία (homophily), δηλαδή την τάση που έχουν τα άτομα να ακολουθούν άλλα άτομα λόγω των θεματικών ενδιαφερόντων που δημοσιεύουν. Στην συγκεκριμένη εργασία τους προσπαθούν να εντοπίσουν τέτοιους χρήστες ξεχωρίζοντας τους από χρήστες που ακολουθούνε άλλους χρήστες για λόγους αμοιβαιότητας (reciprocity), προσδίδοντας με αυτόν τον τρόπο μια καλύτερη κατανομή επίδρασης. Τονίζουν ότι έχει μεγαλύτερη σημασία να ακολουθείται ένας χρήστης λόγω ενδιαφέροντος του περιεχομένου των δημοσιεύσεων που δημοσιεύει πάρα για λόγους κοινωνικών σχέσεων. Αρχικά αναλύοντας τις δημοσιεύσεις των χρηστών εντοπίζουν τις θεματικές κατηγορίες για τις οποίες ενδιαφέρονται οι χρήστες και κατασκευάζουν σχέσεις

θεματικών ενδιαφερόντων μεταξύ των χρηστών. Στην συνέχεα δημιουργώντας έναν αλγόριθμο εμπνευσμένο από τον PageRank, τον TwitterRank, τον εφαρμόζουν προκειμένου να μετρήσουν την επίδραση των χρηστών προσμετρώντας και την ομοιότητα των χρηστών ως προς τις θεματικές κατηγορίες αλλά και την δομή της σχέσης που συνδέει τους δύο χρήστες (link structure). Ο τρόπος με τον οποίο εμπεριέχουν της θεματικές κατηγορίες για τις οποίες ενδιαφέρονται οι χρήστες, γίνεται προσδίδοντας την πιθανότητα να ενδιαφέρεται για μια συγκεκριμένη θεματική κατηγορία. Τέλος, πέρα από την επίδραση του Επηρεάζοντος πάνω σε μια συγκεκριμένη θεματική κατηγορία βρίσκουν και την γενική επίδραση που έχει ο Επηρεάζων στο Κοινωνικό δίκτυο.

Οι Aral Sinan και Dylan Walker [12] θέλοντας να καταλάβουν πως ένας χρήστης μπορεί να επηρεάσει κάποιον άλλον χρήστη (peer influence) στην υιοθέτηση προϊόντων, μελέτησαν τις δομικές συνθήκες κάτω από τις οποίες η επίδραση μεγιστοποιείται. Έτσι ερευνώντας τις διαπροσωπικές σχέσεις των χρηστών εντοπίζουν κάτω από ποιες συνθήκες μεγιστοποιείται η επίδραση στο διαδίκτυο μεταξύ των χρηστών. Μέσα από τυχαία επεξεργασία των μηνυμάτων που έχουν σταλεί από 1.3 εκατομμύρια χρήστες του Facebook προσδιορίζουν τις μεταβλητές της Ισχυρότητας των Δεσμών (Tie Strength) και της Ενσωμάτωσης (Embeddedness). Επεκτείνοντας τον ήδη γνωστό ορισμό της Ισχυρότητας των δεσμών μελετάνε: α. πώς οι χρήστες γνωρίστηκαν μεταξύ τους, β. πόσο πρόσφατη είναι η γνωριμία τους, γ. το πλήθος των κοινών ενδιαφερόντων που έχουν και δ. το πόσο συχνά επικοινωνούν μεταξύ τους. Η συνεισφορά τους είναι στο να δώσουν ιδέες για καλύτερες στρατηγικές μάρκετινγκ και γενικότερα καλύτερη δημόσια πολιτική (public policy) [12].

Οι D. Garcia1, P Mavrodiev, D Casati, F. Schweitzer [18], μελετούν το κοινωνικό δίκτυο του Twitter για 7 χρόνια, προσπαθώντας να προσδιορίσουν δύο κύριες μεταβλητές. Η πρώτη αφορά το πόσο γνωστός (Popularity) είναι ένας χρήστης στο κοινωνικό δίκτυο και βρίσκουν ότι αυτή η μεταβλητή προσδιορίζει το εύρος της επίδρασης που έχει ένας χρήστης. Η δεύτερη μεταβλητή ποσοτικοποιεί την φήμη (Reputation), μέσω της οποίας βρίσκουν ότι χρήστες με υψηλή φήμη επηρεάζουν άτομα με υψηλή φήμη, το οποίο αποτέλεσμα συμβαδίζει με την πρόταση του Thomas Renault [4]. Συγκρίνοντας αυτές τις δύο μεταβλητές βρίσκουν ότι η «φήμη» παρουσιάζει καλύτερες προβλεπτικές ιδιότητες όσον αφορά την αδράνεια των χρηστών (inactivity) από την «δημοτικότητα». Βλέποντας πώς η μια μεταβλητή επηρεάζει την

άλλη, βρίσκουν ότι η δημοτικότητα αυξάνει με την αύξηση της φήμης μόνο σε περιπτώσεις όπου οι χρήστες έχουν ήδη υψηλό δείκτη δημοτικότητας. Στην συνέχεια συνδέοντας τις δύο μεταβλητές με την επίδραση δείχνουν ότι η επίδραση αυξάνεται υπογραμμικά με την αύξηση της δημοτικότητας και πως η «φήμη» αποτελεί μια καλύτερη μεταβλητή πρόβλεψης της επίδρασης από το πλήθος των ακόλουθων. Συνδυάζοντας λοιπόν τους δύο δείκτες πετυχαίνουν μια καλύτερη εκτίμηση για την επίδραση που ασκούν οι χρήστες στο διαδίκτυο.

2.4.3 Πλεονεκτήματα

Ο αλγόριθμος του Klout είναι ένας από τους πρώτους αλγορίθμους για εύρεση επίδρασης στο διαδίκτυο και χρησιμοποιείται από πολλούς ερευνητές ως μέτρο αξιολόγησης των αποτελεσμάτων τους. Πρωτοεμφανίστηκε το 2008 και ως εκ τούτου διαθέτει μεγάλο χρόνο επεξεργασίας και βελτιώσεων. Σημαντικός είναι ο παράγοντας ότι ξεκίνησε ως Νεοφυής Επιχείρηση και ότι έως τώρα έχουν επενδυθεί πολλά χρήματα στην εταιρία (41.5 εκατομμύρια), Μια σημαντική διαφοροποίηση του συγκεκριμένου αλγορίθμου σε σχέση με άλλους αλγορίθμους είναι το γεγονός ότι προσπαθεί να εκμεταλλευτεί κάθε δυνατή πληροφορία του διαδικτύου συνδυάζοντας μακροχρόνια με βραχυχρόνια χαρακτηριστικά επίδρασης και ταυτόχρονα συνδυάζει αυτήν την πληροφορία με δεδομένα από τον πραγματικό κόσμο (Wikipedia). Τέλος, ο τρόπος με τον οποίο έχει κατασκευαστεί το σύστημα στο οποίο τρέχει ο αλγόριθμός τους του προσδίδει ιδιότητες επεκτασιμότητας, δημιουργώντας ένα ιδιαίτερα ευέλικτο σύστημα. [3]

Οι N. Booth, J. A. Matic [5], παρουσιάζουν μια προσαρμοσμένη στον χρήστη προσέγγιση δίνοντας την δυνατότητα στις εταιρίες να διαφημίζουν τα δυνατά τους σημεία δίνοντας έμφαση στην ποσοτική και ποιοτική ανάλυση των αποτελεσμάτων τους. [5]

Η εργασία των M. Danisch, N. Dugu, A. Perez [6], βελτιώνει και επεκτείνει έναν υπάρχοντα αλγόριθμο για εντοπισμό των Κοινωνικών Καπιταλιστών μελετώντας πέρα από τα τοπολογικά στοιχεία των δημοσιεύσεων και το περιεχόμενο τους. Η δημιουργία του ισορροπημένου Klout σκορ μειώνει το σκορ των Κοινωνικών Καπιταλιστών (Social Capitalist), αφού γίνεται αναπροσαρμογή των αποτελεσμάτων κατάταξης.

Οι M. Petychakis κ.α [9], με την προσέγγιση τους δίνουν την δυνατότητα προσαρμογής του ορισμού των ατόμων με επίδραση στο διαδίκτυο και ως εκ τούτου την επιλογή αλγορίθμων προσαρμοσμένων στις ανάγκες των χρηστών.

Ο TwitterRank αποδίδει καλύτερα από αυτόν τον αλγόριθμο που χρησιμοποιεί το Twitter, από τον PageRank και από την παραλλαγή του PageRank προσαρμοσμένο σε θεματικές κατηγορίες (topic sensitive PageRank). Λόγω της φύσης του αλγορίθμου εντοπίζει τις θεματικές ενότητες για τις οποίες ενδιαφέρονται οι χρήστες διαθέτοντας δύο αποτελέσματα: κατατάξεις επίδρασης των χρηστών με βάση τις θεματικές κατηγορίες και κατατάξεις συνολικής επίδρασης στο κοινωνικό δίκτυο.

Οι Aral Sinan and Dylan Walker [12], βρίσκουν ότι επηρεαζόμαστε περισσότερο από άτομα που ζουν στην ίδια πόλη με εμάς, αλλά οι προτιμήσεις έχουν μεγαλύτερη συσχέτιση με άτομα που βρίσκουν από την ίδια πόλη που μεγαλώσαμε. Μελετώντας τους κοινούς φίλους που έχουν δύο φίλοι, δείχνουν ότι για κάθε ένα άτομο που μοιράζονται δύο φίλοι, εξασκούν 0,6% περισσότερη επίδραση στους κοινούς τους φίλους. Τέλος, τα άτομα που πήγαν στο ίδιο πανεπιστήμιο ασκούν 13,55 φορές περισσότερη επίδραση ο ένας στον άλλον.

2.4.4 Μειονεκτήματα

Δυστυχώς ο αλγόριθμος του Klout δεν είναι δημοσιευμένος, συνεπώς το μόνο που μπορούμε να χρησιμοποιήσουμε είναι οι ιδέες που παρουσιάζονται. Ένα άλλο σημαντικό μειονέκτημα είναι ότι πολλές φορές μπερδεύει τις επαναδημοσιεύσεις με επίδραση. Επιπλέον, δεν συμπεριλαμβάνει την εφαρμογή Pinterest, που είναι μια από τις πιο διαδεδομένες εφαρμογές πλέον. Τέλος, αδυνατεί να αναγνωρίσει τα ηλεκτρονικά ρομπότ (bot) δίνοντας τους την δυνατότητα να κερδίζουν υψηλό Klout σκορ.[3]

Η δυνατότητα της προσαρμογής που προσφέρουν οι N. Booth, J. A. Matic [5], δεν επιτρέπει την εξαγωγή γενικών σκορ, στηριζόμενοι στην πεποίθηση ότι η σύγκριση έχει νόημα κάτω από τις ίδιες προϋποθέσεις (μεταβλητές). [5]

Τα προβλήματα της έρευνας των M. Danisch, N. Dugu, A. Perez [6], είναι το μικρό πλήθος δεδομένων εισόδου και το γεγονός ότι δεν δημιουργούν ένα καινούργιο σκορ από την αρχή αλλά στηρίζονται σε ένα υπάρχον. Τέλος, τα αποτελέσματα τους δεν

γενικεύονται καθώς στηρίζονται σε παρατηρήσεις που υφίστανται στο Κοινωνικό δίκτυο του Twitter.

Οι M. Petychakis κ.α., αντιμετωπίζουν παρομοίως το πρόβλημα του μικρού αριθμού των δεδομένων εισόδου. Επιπλέον, προκειμένου να βγουν ακόμα καλύτερα αποτελέσματα απαιτούνται περισσότερα κοινωνικά δίκτυα και περισσότεροι αλγόριθμοι.

Ο αλγόριθμος του TwitterRank αφορά μια νέα προσέγγιση, συνεπώς έχει αρκετά περιθώρια βελτίωσης. Πέρα από τις θεματικές κατηγορίες που ενδιαφέρονται τα άτομα ο αλγόριθμος οφείλει να συμπεριλάβει και στοιχεία για τους χρήστες (π.χ. συχνότητα σχολίων σε δημοσιεύσεις άλλων χρηστών). Επιπλέον, λαμβάνει υπόψιν τις το πλήθος των ακόλουθων συνεπώς είναι ευάλωτος από επιτηδείους. Τέλος, ο διαχωρισμός βάση των θεματικών κατηγοριών γίνεται σε ένα στιγμιότυπο του Twitter καθώς η συλλογή των δεδομένων γίνεται μια φορά.

Τα μειονεκτήματα που παρουσιάζονται στην εργασία των D. Garcia1, P Mavrodiev, D Casati, F. Schweitzer [18], είναι πως η χρησιμότητα ενός κοινωνικού διαδικτύου μειώνεται με την δημοσιότητα, γιατί ο χρήστης δεν μπορεί να επεξεργαστεί την ποσότητα της πληροφορίας που δέχεται και η γενίκευση των αποτελεσμάτων τους δεν είναι εφικτή λόγω του ότι χρησιμοποιήθηκε τεράστια ποσότητα πληροφορίας από το Twitter.[18]

2.5 Κατηγορία: Διάδοση της πληροφορίας – προβλέψεις & επιπλέον υλικό

Η συγκεκριμένη ενότητα αναλύει τα άρθρα που βρίσκονται στο σύνολο I∩A.

2.5.1 Προσεγγίσεις που ακολουθούνται

Οι E. Bakshy, I. Rosenn, C. Marlow, L. Adamic [11], μελετάνε το πως το «Τοίχος Ροής Ειδήσεων» (News Feed - NF) του Facebook, όπου εμφανίζονται οι δημοσιεύσεις των χρηστών με τους οποίους έχει συνδεθεί κάποιος χρήστης, μέσω φιλίας, επηρεάζει την διάδοση της πληροφορίας. Ένα κρίσιμο ερώτημα που καλούνται να απαντήσουν είναι

τι θα συνέβαινε αν κάποιες αλληλεπιδράσεις μεταξύ των χρηστών δεν πραγματοποιούνταν ποτέ! Για την μελέτη του NF στην διάδοση της πληροφορίας πραγματοποιούν το εξής πείραμα: κάποιοι χρήστες εκτίθενται σε πληροφορία μέσω του NF και κάποιοι άλλοι όχι (εδώ η πληροφορία μπορεί να μπει στο Facebook μόνο με εξωστρεφείς πόρους). Συγκρίνοντας την δραστηριότητα των δημοσιεύσεων των χρηστών ανάμεσα σε αυτές τις δύο καταστάσεις προσδιορίζουν την σημαντικότητα του NF στην διάδοση της πληροφορίας και συμπεραίνουν ότι τα άτομα τα οποία εκτίθενται στην πληροφορία του NF είναι πολύ πιο πιθανό να δημοσιεύσουν μια πληροφορία και μάλιστα αυτή η δημοσίευση συμβαίνει πολύ νωρίτερα από ότι θα συνέβαινε αν δεν πραγματοποιούνταν η έκθεση της πληροφορίας στο NF. Τι συμβαίνει όμως όταν δύο άτομα σχεδόν ταυτόχρονα δημοσιεύουν έναν σύνδεσμο; Τα αποτελέσματα σε αυτό το ερώτημα μέσω της έκθεσης ατόμων στο NF και όχι, δίνουν τις εξής περιπτώσεις:

- Το άτομο δημοσίευσε τον σύνδεσμο, εξαιτίας του ότι ενημερώθηκε από το NF, μέσω μιας δημοσίευσης ενός φίλου του στο Facebook.
- Η δημοσίευση συνέβη γιατί το άτομο επισκέπτεται τις ίδιες ιστοσελίδες με τον φίλο και τα δύο γεγονότα συνέβησαν ανεξάρτητα μεταξύ τους, πληροφορία την οποία δεν λαμβάνει υπόψη το [7].
- Η δημοσίευση προέκυψε γιατί το άτομο ενημερώθηκε από μια πηγή έξω από το Facebook, όπου ο φίλος του ατόμου είχε δημοσιεύσει τον σύνδεσμο και εξωτερικά αλλά και εσωτερικά του Facebook.

Επιπλέον, τονίζονται οι αιτιατές σχέσεις που οδηγούν σε συμπεριφορές δημοσιεύσεων (sharing activities). Στην συνέχεια μελετάνε την Ισχυρότητα των Δεσμών των φίλων (Tie Strength) με: α. την συχνότητα επικοινωνίας των χρηστών μέσω προσωπικών μηνυμάτων, β. την δημόσια επικοινωνία μέσω σχολιασμών (comments) σε δημοσιεύσεις (posts) χρηστών, γ. το πλήθος των πραγματικών γεγονότων που συνδέουν τους φίλους όπως το να βρίσκονται στην ίδια φωτογραφία και δ. το πλήθος των κοινών δημοσιεύσεων που δύο φίλοι σχολιάζουν. Βλέπουμε λοιπόν ότι πέρα από την πρώτη παράμετρο όλες οι υπόλοιπες είναι διαφορετικές, για την μέτρηση της Ισχυρότητας των Δεσμών, σε σχέση με τις παραμέτρους που χρησιμοποιούν στην εργασία τους οι Aral Sinan και Dylan Walker [12].

Διεξάγοντας μια πολύ μεγάλη έρευνα οι J. Leskovec, L. A. Adamic, and B. A. Huberman [13], μελετάνε το πως διαδίδεται η πληροφορία αναλύοντας την υιοθέτηση προϊόντων μέσω του διαδικτύου (viral marketing), χρησιμοποιώντας «από άτομο σε

άτομο» συστάσεις (person to person recommendation). Έτσι εντοπίζουν ομάδες ατόμων, κατηγορίες προϊόντων και τιμολογιακές πολιτικές για τα οποία το μάρκετινγκ μέσω διαδικτύου είναι πιο αποδοτικό. Ως συμπέρασμα της παραπάνω έρευνας βρίσκουν ότι ανάλογα με την κατηγορία στην οποία ανήκουν τα προϊόντα απαιτείται διαφορετικός τρόπος προσέγγισης του πελάτη. Για παράδειγμα αν κάποιος θέλει να δει μια ταινία, ο αριθμός των ηλεκτρονικών μηνυμάτων (e-mail) που πρέπει να δεχτεί κάποιος προτείνοντάς του ταινίες πρέπει να είναι σαφώς μεγαλύτερος από τον αριθμό των ηλεκτρονικών μηνυμάτων (e-mail) που πρέπει να δεχτεί για προτάσεις βιβλίων. Αυτό οφείλεται στο ότι απαιτείται τελείως διαφορετική επένδυση χρόνου στο να διαβάσει κάποιος ένα βιβλίο από το να δει μια ταινία. Συνεπώς διαφορετική κατηγορία προϊόντων απαιτεί διαφορετική στρατηγική προώθησης. Με τον ίδιο τρόπο λοιπόν οι Επιρεάζοντες θα πρέπει να χρησιμοποιούν διαφορετικές στρατηγικές ανάλογα με την κατηγορία των προϊόντων που προωθούν. Ένα άλλο συμπέρασμα της έρευνάς τους είναι ότι όσο περισσότερα προϊόντα προτείνει ένα άτομο τόσο λιγότερο καταφέρνει να επηρεάσει. Συνεπώς μια παράμετρος που θα πρέπει να ληφθεί υπόψιν στην κατάταξη των ατόμων είναι το κοντινό ιστορικό προτεινόμενων προϊόντων των Επηρεάζοντων.

Μια νέα προσέγγιση για το πως διαδίδεται η πληροφορία στο Twitter ακολουθείται από τους R. Pfitzner, A. Garas, F. Schweitzer [19], όπου προσπαθούν να βρουν ποιες δημοσιεύσεις στο Twitter έχουν περισσότερες πιθανότητες επαναδημοσίευσης βασιζόμενοι στην συναισθηματική απόκλιση (emotional divergence) που περιέχεται μέσα σε μια δημοσίευση. Αρχικά βρίσκουν αν μια δημοσίευση έχει θετικό, αρνητικό ή ουδέτερο περιεχόμενο (emotional polarity) και στην συνέχεια εισάγουν μια μεταβλητή που απεικονίζει την «συναισθηματική απόκλιση» (d). Για παράδειγμα η πρόταση «Μισώ το γεγονός ότι σε αγαπώ» έχει μεγάλη συναισθηματική απόκλιση. Αναλύοντας τα δεδομένα βάση του συναισθηματικού περιεχομένου και αποδίδοντας τιμές στην μεταβλητή της συναισθηματικής απόκλισης βρίσκουν ότι υπάρχει ένα κατώφλι συναισθηματικής απόκλισης, όπου για αυτήν την τιμή και επάνω οι δημοσιεύσεις έχουν πέντε φορές μεγαλύτερη πιθανότητα να επαναδημοσιευτούν. Η συγκεκριμένη ιδέα μπορεί να βοηθήσει να επεκτείνουμε την εργασία [14], ή ακόμα και να εισάγουμε μια μεταβλητή που να προσδιορίζει την συναισθηματική απόκλιση των Επηρεάζοντων και μέσω αυτής της μεταβλητής να γίνεται μια αναπροσαρμογή στην επίδρασή τους.

2.5.2 Πλεονεκτήματα

Σύμφωνα με τους E. Bakshy, I. Rosenn, C. Marlow, L. Adamic [11], τα άτομα τα οποία εκτίθενται στην πληροφορία του NF, έχουν πολύ μεγαλύτερη πιθανότητα να επαναδημοσιεύσουν και μάλιστα η επαναδημοσίευση πραγματοποιείται πολύ πιο γρήγορα. Επιπλέον, οι χρήστες που δημοσιεύουν τους ίδιους συνδέσμους με τους φίλους τους δημοσιεύουν περίπου την ίδια ώρα με αυτούς ακόμα και αν δεν υπάρχει προηγουμένως έκθεση της πληροφορίας στο Facebook. Όσον αφορά την πιθανότητα της επαναδημοσίευσης, στηρίζουν ότι αυξάνει με το πλήθος των φίλων που επαναδημοσιεύουν. Ως συμπέρασμα της Ισχυρότητας των Δεσμών συμπεραίνουν ότι δίνει καλύτερες προβλέψεις για δραστηριότητα εκτός Facebook και με αυτόν τον τρόπο έρχονται σε αντίφαση με το άρθρο [12]. Τέλος, παρατηρούν ότι οι Ασθενείς Δεσμοί βοηθάνε στην διάδοση πρωτότυπης πληροφορίας, γεγονός που έρχεται σε συμφωνία με το άρθρο [17] όπου στηρίζουν ότι η πληροφορία δεν ακολουθεί την κλασική από πάνω προς τα κάτω προσέγγιση.

Οι R. Pfitzner, A. Garas, F. Schweitzer [19], ακολουθώντας μια συγκεκριμένη παράμετρο ξεπερνούν τον πρότυπο τύπο στατιστικής πρόβλεψης και βαθμολογούν το περιεχόμενο των συναισθημάτων που ενέχει μια δημοσίευση. Το συναίσθημα που περιέχεται σε ένα μήνυμα είναι ένα μέτρο που μπορεί να διαχωρίσει ποιες δημοσιεύσεις θα επεκταθούν περισσότερο.

2.5.3 Μειονεκτήματα

Σύμφωνα με τους E. Bakshy, I. Rosenn, C. Marlow, L. Adamic [11], είναι αδύνατον να πει κανείς με σιγουριά μέσα από ανάλυση διαθέσιμων στοιχείων στο ίντερνετ (observational data) αν μια συγκεκριμένη συμπεριφορά οφείλεται στο ότι κάποιος την προκάλεσε ή στο ότι τα άτομα που συμμετέχουν στην εκδήλωση αυτής της συμπεριφοράς έχουν κοινά ενδιαφέροντα.

Σύμφωνα με το πείραμα που διεξάγουν οι J. Leskovec, L. A. Adamic, and B. A. Huberman [13], ο μόνος τρόπος για να πουν ότι κάποιος επηρέασε κάποιον άλλον είναι όταν ένα άτομο αγοράσει από έναν πωλητή και στην συνέχεια στείλει ένα ηλεκτρονικό μήνυμα για αυτήν του την αγορά σε κάποιον φίλο του και αυτό το μήνυμα οδηγήσει τον φίλο του να αγοράσει ένα προϊόν από τον ίδιο πωλητή. Τα ηλεκτρονικά μηνύματα μπορεί να θεωρηθούν ως ανεπιθύμητη αλληλογραφία (spam) πριν καν διαβαστούν από τους χρήστες. Υπάρχει αμφιβολία από τους αποδέκτες των ηλεκτρονικών προτάσεων

αν όντως δέχονται το ηλεκτρονικό μήνυμα γιατί όντως το προτείνει ο φίλος τους ή απλά γιατί ο φίλος τους θέλει να κερδίσει κάποιου είδους έκπτωση. Τέλος, μόνο το ένα τρίτο αυτών που αγοράζουν μπαίνουν στην διαδικασία προώθησης.

Τα αποτελέσματα της έρευνας των R. Pfitzner, A. Garas, F. Schweitzer [19], δεν μπορούν να επεκταθούν για εφαρμογές διαφορετικές από το Twitter καθώς η συναισθηματική απόκλιση μετριέται σε δημοσιεύσεις με περιορισμένο αριθμό κειμένου. Επιπλέον, η έρευνα τους αντιμετωπίζει όλους τους χρήστες ως ισάξιους αγνοώντας ενδεχομένως χρήστες με μεγαλύτερη επίδραση.

2.6 Κατηγορία: Μέτρο επίδρασης & διάδοση της πληροφορίας – προβλέψεις & επιπλέον υλικό

2.6.1 Ορισμός του Επηρεάζοντος

Οι E. Bakshy, J. Hofman, W. Mason, D. Watts [7], θέλοντας να αποφύγουν την αμφισημία που υφίσταται στον ορισμό του ο Επηρεάζων, αποφεύγουν την θεωρία ότι κάποιοι χρήστες επηρεάζουν το κοινό τους, ενώ κάποιοι άλλοι όχι και αντιμετωπίζουν όλους τους χρήστες ισάξια. Έτσι μελετάνε τους χρήστες που «σπέρνουν» (seed) την πληροφορία στο διαδίκτυο, δηλαδή άτομα τα οποία δημοσιεύουν πρώτοι έναν σύνδεσμο σε σχέση με τους φίλους τους. Ως συνέπεια των παραπάνω ορίζεται ως επίδραση η δυνατότητα να μπορεί κάποιος να προωθήσει την ζητούμενη πληροφορία στο κοινό του.

Οι D. M. Romero, W. Galuba, S. Asur, B. A. Huberman [8], αντιλαμβανόμενοι την αδράνεια των χρηστών (user passivity), που εκφράζει την δυσκολία ενός χρηστή να επηρεαστεί από άλλους χρήστες, συμπέρασμα που δημιουργείται από την προηγούμενη έρευνα [7], θεωρούν ότι για να μπορέσει κάποιος να είναι Επηρεάζων θα πρέπει να ξεπεράσει αυτό το πρόβλημα. Συνεπώς ορίζουν ως επίδραση το πλήθος των επισκέψεων που δέχεται ένας δημοσιευμένος σύνδεσμος (URL). Τονίζοντας ότι η επαναδημοσίευση της πληροφορίας είναι καθοριστικός παράγοντας στην διάχυση της πληροφορίας.[8],[1],[2],[15]

Οι M. Cha, F. Benevenuto, H. Haddadi, K. Gummadi [17], θεωρώντας ότι όλοι οι χρήστες του twitter είναι ικανοί να διαδώσουν την πληροφορία χωρίζουν τους χρήστες σε 3 μεγάλες κατηγορίες και ορίζουν ως επίδραση την αποδοτικότητα με την οποία οι 3 αυτές κατηγορίες διαδίδουν την πληροφορία. Η αποδοτικότητα μεταφράζεται ως το μέγεθος του κοινού που καταφέρνουν να ενημερώσουν ή να επηρεάσουν με ένα συγκεκριμένο κομμάτι πληροφορίας που προσπαθούν να διαδώσουν. Στην συγκεκριμένη έρευνα σε σχέση με την εργασία [7] μελετάνε ξεχωριστά το ποιος είναι ο πρώτος που «σπέρνει» το κομμάτι πληροφορίας στο διαδίκτυο και ποιος βοηθάει περισσότερο να διαδοθεί.

2.6.2 Καινοτομικά Στοιχεία

Προκειμένου οι E. Bakshy, J. Hofman, W. Mason, D. Watts [7], να μετρήσουν την επίδραση που έχουν οι χρήστες εξετάζουν πόσο μακριά διαδίδεται (βάθος του δέντρου διάχυσης της πληροφορίας) ένας σύνδεσμος που εισάγει ένας χρήστης μέσω της δημοσίευσής του. Δηλαδή, εντοπίζουν ποιος δημοσίευσε πρώτος έναν σύνδεσμο, ακολουθώντας τους φίλους του ατόμου που δημοσίευσε τον συγκεκριμένο σύνδεσμο. Αν το άτομο που δημοσίευσε τον σύνδεσμο είναι ο πρώτος ανάμεσα στους φίλους του, τότε παίρνει όλους του πόντους αυτός. Αν δεν είναι ο πρώτος που δημοσίευσε τον σύνδεσμο, τότε βρίσκουν τον πρώτο που δημοσίευσε αυτόν τον σύνδεσμο ακολουθώντας τους φίλους αυτού του ατόμου. Η διαδικασία επαναλαμβάνεται μέχρι να βρεθεί άτομο που δεν έχει φίλο που να δημοσίευσε τον συγκεκριμένο σύνδεσμο. Πρέπει να τονιστεί εδώ ότι δεν μελετούν τις επαναδημοσιεύσεις, αλλά το πλήθος των δημοσιεύσεων για έναν συγκεκριμένο σύνδεσμο μέσω ενός κλειστού συνόλου που απαρτίζεται από φίλους και ακόλουθους που κάνουν δυνατή την μεταφορά της πληροφορίας. Αυτή ακριβώς είναι και η ειδοποιός διαφορά σε σχέση με τα υπόλοιπα άρθρα. Ως γενικό συμπέρασμα καταλήγουν στο ότι τα άτομα που καταφέρνουν να έχουν μεγάλο βάθος διάδοσης είναι ελάχιστα και αυτοί που εν τέλει το καταφέρνουν είναι άτομα που ήταν Επηρεάζοντες στο παρελθόν ή άτομα τα οποία έχουν μεγάλο πλήθος ακόλουθων. Στην συνέχεια, προκειμένου να υπολογίσουν τα άτομα με επίδραση υπολόγισαν για κάθε χρήστη όλες τις δημοσιεύσεις τις οποίες δημοσίευσε πρώτος ένας χρήστης και βρήκαν τον λογάριθμο της μέσης τιμής του βάθους διάδοσης των δημοσιεύσεων αυτών. Φτιάχνοντας έναν αλγόριθμο για να προβλέψουν ποιος

χρήστης ή σύνδεσμος θα δημιουργήσει μεγαλύτερο βάθος διάδοσης, συμπεραίνουν ότι τα αποτελέσματα που εξάγονται είναι αναξιόπιστα (δεν μπορούν να γίνουν προβλέψεις). Το οποίο συμπέρασμα και αντιτίθεται με την εργασία των Q. Zhao κ.α [14] που κάνουν προβλέψεις χρησιμοποιώντας ένα καθαρά μαθηματικό μοντέλο σε αντίθεση με τις τεχνικές μηχανικής μάθησης που χρησιμοποιούνται στην παρούσα εργασία. Στην συνέχεια, προκειμένου να βελτιώσουν τον αλγόριθμο πρόβλεψης προσπαθούν να συμπεριλάβουν το περιεχόμενο (content) της πληροφορίας που προσπαθεί να διαδοθεί και καταλήγουν ότι η συγκεκριμένη παράμετρος δεν επηρεάζει τον αλγόριθμο πρόβλεψης. Τέλος, ερευνώντας την βέλτιστη οικονομική πολιτική για πληρωμή ατόμων με επίδραση στο διαδίκτυο καταλήγουν στο προφανές ότι τα άτομα με υψηλή επίδραση είναι και τα πιο αποτελεσματικά, αλλά τονίζουν ότι κάτω από κάποιες προϋποθέσεις η βέλτιστη επιλογή αφορά άτομα με μεσαία ή ακόμα και λιγότερη επίδραση. Δηλαδή, είναι προτιμότερο κάποιες φορές να κατανείμει κανείς τους χρηματικούς πόρους σε περισσότερα άτομα με λιγότερη επίδραση. [7]

Οι D. M. Romero, W. Galuba, S. Asur, B. A. Huberman [8], διαφοροποιούνται από τις προηγούμενες εργασίες συμπεριλαμβάνοντας στον αλγόριθμο τους, πέρα από τα στατιστικά στοιχειά των χρηστών (πλήθος ακόλουθων κ.λ.π.), την αδράνεια των χρηστών (user passivity) όσον αφορά την προώθηση της πληροφορίας (π.χ. μέσω επαναδημοσιεύσεων). Αναφέρουν ότι κάποιος μπορεί να βρει την επίδραση που έχει ο Α στον Β όταν μετρήσει πόσες φορές ο Β επαναδημοσιεύει τον Α, [1], [2], [15], αλλά στηρίζουν ότι μια τέτοια μετρική δεν εντοπίζει την επίδραση που έχει κάποιος στο συνολικό δίκτυο. Για αυτό τον λόγο βρίσκουν την σχετική επίδραση (relative influence) του κάθε χρήστη ως προς το συνολικό δίκτυο και την συνδυάζουν με την αδράνεια των χρηστών (user passivity). Με αυτούς τους δύο δείκτες καταφέρνουν να μετρήσουν ποσοτικά και ποιοτικά την επίδραση. Ο αλγόριθμος τους ονομάζεται IP και βασίζεται στον αλγόριθμο του HITS, που βρίσκει εγκεκριμένες ιστοσελίδες και κεντρικούς ιστοτόπους που οδηγούν σε αυτές τις σελίδες. Στην συνέχεια στηρίζουν ότι ένας καλός αλγόριθμος κατάταξης των χρηστών θα έχει και καλές προβλεπτικές ιδιότητες. Με την έννοια ότι από την στιγμή που έχουν αξιολογηθεί οι χρήστες, αναμένουν παρόμοια αποτελέσματα και στο μέλλον για τους χρήστες (αυτός είναι και ένας από τους λόγους που αξιολογούνται οι χρήστες). Διαφοροποιούμενοι από το [7] και αντιλαμβανόμενοι ότι μια ακριβή πρόβλεψη για το πόσες φορές θα πατηθεί ένας σύνδεσμος από τους χρήστες είναι αδύνατον να υπάρξει, προβλέπουν το άνω όριο του

πλήθους των κλικ που μπορεί να δεχτεί ένας σύνδεσμος μέσω του IP αλγορίθμου τους. Τέλος, για την αξιολόγηση του αλγορίθμου τους, IP-Influence score, συγκρίνουν μεθόδους που χρησιμοποιούνται σε διαφορετικούς τομείς ως μετρικές κατάταξης (H-Index, PageRank, Number of retweets, Number of Followers) [8].

Εν συνεχεία, οι M. Cha, F. Benevenuto, H. Haddadi, K. Gummadi [17] επιχειρούν να μελετήσουν πως οι διάφορες ομάδες χρηστών συμβάλουν στην διάδοση της πληροφορίας. Η διαφοροποίηση σε σχέση με τις υπόλοιπες εργασίες έγκειται στο ότι προσπαθούν να εξηγήσουν γιατί κάποιες δημοσιεύσεις γίνονται πολύ γνωστές μέσω του Twitter. Για τον λόγο αυτό χωρίζουν τους χρήστες σε 3 μεγάλες κατηγορίες: 1) «Πηγές μαζικών μέσων ενημέρωσης» (άτομα με > 100,000 ακόλουθους), 2) «Απλός Λαός» (με < 300 ακόλουθους) και 3) οι ενδιάμεσοι που ονομάζονται «Ευαγγελιστές». Τα χαρακτηριστικά που θεωρούν ότι βοηθούν στην διάδοση της πληροφορίας είναι: α. ο λόγος Φίλοι/Ακόλουθοι (Outdegree/Indegree), όπου εντοπίζουν ότι όσο λιγότερο δημοφιλής είναι ένας χρήστης τόσο πιο ενεργά ακολουθεί άλλους χρήστες, β. η αμοιβαιότητα (reciprocity) δηλαδή το κατά πόσο ένας χρήστης Α ακολουθεί τον Β επειδή ο Β πρώτος ακολούθησε τον Α, μέσω αυτού του δείκτη επιβεβαιώνουν την θεωρία τους ότι οι «Πηγές μαζικών μέσων ενημέρωσης» χρησιμοποιούν το twitter καθαρά για διάδοση της πληροφορίας, ενώ οι υπόλοιπες ομάδες ως μέσο κοινωνικής δικτύωσης, γ. το πλήθος δημοσιεύσεων (number of tweets), όπου βρίσκουν ότι η πρώτη κατηγορία έχει τις περισσότερες δημοσιεύσεις, η 2η κατηγορία έχει τις μισές δημοσιεύσεις και η 3η κατηγορία ελάχιστες. Στην συνέχεια χρησιμοποιώντας 6 αρκετά διαδεδομένες δημοσιεύσεις του 2009 προσπαθούν να ερμηνεύσουν πως οι 3 ομάδες χρηστών συνέβαλλαν στην διάδοση αυτών των δημοσιεύσεων. Τα αποτελέσματα που παίρνουν είναι τα εξής:

1. Η 2η και 3η κατηγορία περιέχει τα άτομα με τους περισσότερους διασκορπιστές της πληροφορίας και τις περισσότερες δημοσιεύσεις.
2. Η 1η κατηγορία περιέχει τους χρήστες που είναι απαραίτητοι αλλά και επαρκείς για να φτάσει η πληροφορία στο αναμενόμενο κοινό.
3. Η 2η κατηγορία βοηθάει αρκετά την 1η στο να φτάσει η πληροφορία σε άτομα που χωρίς αυτούς δεν θα έφτανε.

Συνεχίζοντας την μελέτη τους προσπαθούν αυτήν την φορά να μελετήσουν πως οι 3 κατηγορίες χρηστών συμβάλλουν στην διάδοση όχι τόσο γνωστών δημοσιεύσεων και συμπεραίνουν ότι τα μοτίβα διάδοσης παραμένουν τα ίδια με αυτά των αρκετά

διαδεδομένων δημοσιεύσεων. Τέλος, θέλοντας να μελετήσουν το «ποιος επηρεάζει ποιον» βρήκαν ότι η πληροφορία δεν ακολουθεί την κλασική από Πάνω προς τα Κάτω λογική, δηλαδή από τους πιο διάσημους στους λιγότερο διάσημους, καθώς μεγάλο μέρος της πληροφορίας πρωτοδημοσιεύεται από την 2η και 3η κατηγορία και στην συνέχεια υιοθετείται από την 1η κατηγορία χρηστών.

2.6.3 Πλεονεκτήματα

Οι E. Bakshy, J. Hofman, W. Mason, D. Watts [7], είναι οι πρώτοι που χρησιμοποιούν το βάθος του δέντρου διάδοσης της πληροφορίας ως μετρική, συμπεραίνοντας ότι οι περισσότεροι χρήστες δεν διαδίδουν την πληροφορία. Εν συνεχεία, εντοπίζουν ότι σύνδεσμοι που αξιολογήθηκαν ως πιο ενδιαφέροντες ή άφηναν θετικά συναισθήματα στον χρήστη ήταν πιο πιθανό να διαδοθούν, συμπέρασμα που υποστηρίζεται και από το άρθρο [19].

Οι D. M. Romero, W. Galuba, S. Asur, B. A. Huberman [8] μέσω του αλγορίθμου τους καταφέρνουν να εντοπίζουν ηλεκτρονικά ρομπότ (bot) και άτομα που μόνος στόχος τους είναι να έχουν πολλούς ακόλουθους (spammers), λύνοντας το πρόβλημα των [3], [15]. Ο τρόπος με τον οποίο έχουν κατασκευάσει τον αλγόριθμό τους, επιτρέπει ως είσοδο και διαφορετικής μορφής γράφων επίδρασης από αυτόν που χρησιμοποιήθηκε στην έρευνά τους. Επιπλέον, ο αλγόριθμός τους παρουσιάζει την καλύτερη προβλεπτική ικανότητα σε σχέση με τους συγκρινόμενους αλγορίθμους (PageRank, Number of followers). Η προβλεπτική αυτή δύναμη μπορεί να χρησιμοποιηθεί και για αξιολόγηση του περιεχομένου των δημοσιεύσεων, βασιζόμενη στο ποιος αναφέρθηκε στην συγκεκριμένη δημοσίευση.

Οι M. Cha, F. Benevenuto, H. Haddadi, K. Gummadi [17], για την έρευνά τους χρησιμοποιούν έναν τεράστιο όγκο δεδομένων, 58 servers, και εντοπίζουν τον ρόλο που παίζουν 3 μεγάλες ομάδες χρηστών στην διάδοση της πληροφορίας.

2.6.4 Μειονεκτήματα

Οι E. Bakshy, J. Hofman, W. Mason, D. Watts [7] χρησιμοποιούν έναν πολύ συγκεκριμένο ορισμό για τον Επηρεάζοντα μελετώντας ένα πολύ συγκεκριμένο χαρακτηριστικό της διάδοσης της πληροφορίας. Ο τρόπος με τον οποίο αποδίδουν το

σκορ περιλαμβάνει περιπτώσεις που στην πραγματικότητα θεωρούνται ανεξάρτητα γεγονότα. Μέσω της μετρικής τους δεν καταφέρνουν να συμπεριλάβουν ως μετρική τις επαναδημοσιεύσεις που θεωρείται απαραίτητη παράμετρος για μέτρηση της επίδρασης. Τέλος, δεν λαμβάνουν υπόψιν το περιεχόμενο της πληροφορίας που διαδίδεται.

Οι D. M. Romero, W. Galuba, S. Asur, B. A. Huberman [8], σε πολλά από τα δεδομένα τους δεν έχουν διαθέσιμες τις επαναδημοσιεύσεις (retweet) των χρηστών και σε πολλές από αυτές που έχουν διαθέσιμες δεν λαμβάνουν υπόψιν τον χρόνο που συνέβησαν. Καμία από τις μετρικές που εξετάζουν δεν μπορεί να προβλέψει τον ακριβή αριθμό των κλικ σε συνδέσμους, συνεπώς υπάρχει η πιθανότητα να χρησιμοποίησαν λάθος αλγορίθμους για την σύγκρισή τους.

Οι M. Cha, F. Benevenuto, H. Haddadi, K. Gummadi [17], δεν λαμβάνουν υπόψιν ουσιαστικές παραμέτρους όπως οι πραγματικές σχέσεις των ανθρώπων, την σημαντικότητα των οποίων παραθέτουν τα άρθρα [11], [12]. Επιπλέον, σε πολλά δεδομένα δεν είχαν διαθέσιμο τον χρόνο δημιουργίας των σχέσεων μεταξύ των φίλων όπως συμβαίνει και με όλα τα υπόλοιπα άρθρα.

3 *Θεωρητικό Υπόβαθρο*

(Αλγόριθμος - Υλοποίηση)

3.1 Εισαγωγή

Καθώς τα μέσα μαζικής ενημέρωσης γίνονται ολοένα και πιο αναπόσπαστο κομμάτι της καθημερινότητάς μας, θα ήταν ιδιαίτερα χρήσιμο να υπάρχει ένας μοναδικός τρόπος μετάφρασης της επίδρασης που ασκούν τα άτομα μεταξύ τους μέσω των μέσων κοινωνικής δικτύωσης. Αυτό ενισχύεται ακόμα περισσότερο αν σκεφτεί κανείς ότι η πλέον διαδεδομένη μηχανή αναζήτησης ιστοσελίδων είναι η Google η οποία χρησιμοποιείται από το 77.43% του συνόλου των χρηστών, εκτελώντας καθημερινά 4.5 δισεκατομμύρια αναζητήσεις [24]. Σε αντίθεση όμως με τις μηχανές αναζήτησης ιστοσελίδων η αναζήτηση των ατόμων με μεγάλη επιρροή στο διαδίκτυο βρίσκεται σε πρώιμα στάδια με συνέπεια να μην υπάρχει μια καθολική απάντηση στο ερώτημα εύρεσης των ατόμων με υψηλή επιρροή. Ο λόγος που συμβαίνει αυτό είναι γιατί ο κάθε αλγόριθμος που χρησιμοποιείται μετράει και αναδεικνύει διαφορετικά χαρακτηριστικά των χρηστών. Συνεπώς, υπάρχουν αρκετές αμφισημίες που πρέπει να απαντηθούν ώστε το πρόβλημα να περάσει στο επόμενο στάδιο. Λαμβάνοντας υπόψιν όλους αυτούς τους περιορισμούς έγινε μια προσπάθεια κατασκευής αλγορίθμου που αξιολογεί απαραίτητες μεταβλητές που πρέπει κανείς να λάβει υπόψιν προκειμένου να μετρήσει την επίδραση. Επιπλέον, τονίζεται ότι η απουσία γενικής αλήθειας που να δίνει μοναδική κατάταξη ατόμων με επίδραση στο διαδίκτυο μετασχημάτισε το πρόβλημα εύρεσης βέλτιστων βαρών, το οποίο είναι απαραίτητο για να δοθεί μοναδική λύση, για ένα σύνολο από μεταβλητές σε πρόβλημα συσχέτισης μεταβλητών. Τέλος, αφού βρέθηκαν οι συσχετίσεις δόθηκε μια κατάταξη των χρηστών με χρήση ενός απλού μέσου όρου των μεταβλητών προκειμένου να αξιολογηθεί η αθροιστική συμπεριφορά των μεταβλητών.

3.2 Περιγραφή υψηλού επιπέδου του Αλγορίθμου

Ορίζοντας την επίδραση ως την δυνατότητα που έχει κάποιος στο να προκαλέσει μια συμπεριφορά, καθίσταται απαραίτητη η χρήση των δεικτών ri&rRT όπως αυτοί περιγράφονται από τους I. Anger και C. Kittl [1]. Στόχος ήταν να εξαλειφθούν τα μειονεκτήματα αυτού του αλγορίθμου προσθέτοντας νέα χαρακτηριστικά. Αρχικά, η

ιδέα φαινόταν αρκετά ελκυστική, αλλά γρήγορα ανακαλύφθηκε ότι περιέχει εγγενή προβλήματα. Αυτό οφείλεται στους διαφορετικούς περιορισμούς που έχει σήμερα το Twitter στην εξαγωγή των δεδομένων. Το Twitter λοιπόν, δεν επιτρέπει την εξαγωγή του πλήθους των επαναδημοσιεύσεων και αναφορών που έχει δεχτεί ένας χρήστης. Αυτό που είναι διαθέσιμο από το κοινωνικό μέσο είναι η μεταβλητή statuses_count που προσδιορίζει τον αριθμό των tweets (συμπεριλαμβανομένων των retweets) που έχει δημοσιεύσει ένας χρήστης συνολικά, ενώ για τα retweets και τις αναφορές των χρηστών δίνονται πληροφορίες μόνο σε επίπεδο tweet και όχι σε μορφή συνολικού αριθμού που απαιτεί ο αλγόριθμος των I. Anger και C. Kittl [1]. Αυτοί οι περιορισμοί οδήγησαν σε μια τροποποίηση της αρχικής ιδέας, διατηρώντας παράλληλα τον στόχο της εξάλειψη των μειονεκτημάτων του Αλγορίθμου των I. Anger και C. Kittl [1]. Ξεκινώντας από αυτήν την βάση κατασκευάστηκε ένας αλγόριθμος που όπως προαναφέρθηκε επιλύει προβλήματα άλλων αλγορίθμων και επιπλέον προσπαθεί να συνδυάσει μακροχρόνια και βραχυχρόνια χαρακτηριστικά. Η κατασκευή του αλγορίθμου αποτελείται από 5 βήματα. Τα βήματα αυτά παρουσιάζονται εδώ συνοπτικά προκειμένου να δοθεί μια συνολική εικόνα του αλγορίθμου και αναλύονται στην συνέχεια εκτενέστερα.

- Το πρώτο στάδιο περιλάμβανε την κατηγοριοποίηση χρηστών σε θεματικές ομάδες. Η κατηγοριοποίηση έγινε κατά την συλλογή των δεδομένων και συνολικά δημιουργήθηκαν τρεις μεγάλες κατηγορίες.
- Το δεύτερο στάδιο περιείχε την δημιουργία των μακροχρόνιων χαρακτηριστικών. Τα μακροχρόνια χαρακτηριστικά αντιπροσωπεύονται από τις μεταβλητές: followers, followers/friends & quality_of_followers.
- Στο τρίτο στάδιο δημιουργήθηκαν βραχυχρόνια χαρακτηριστικά και αφαιρέθηκαν τα άτομα που λειτουργούσαν με αυτοματοποιημένο τρόπο με κύριο στόχο την αύξηση των favourite, followers κ.λ.π. (εντοπισμός και αφαίρεση των bots). Για την αφαίρεση των λεγόμενων bot έγινε χρήση της διαδικασίας που προτείνεται από τους M. Danisch, N. Dugu, A. Perez [6]. Για την δημιουργία των βραχυχρόνιων χαρακτηριστικών απαιτούνταν η παρατήρηση όλων των δημοσιεύσεων (χωρίς τα retweets) για κάθε χρήστη κάθε κατηγορίας για διάστημα 9 εβδομάδων. Το συνολικό χρονικό διάστημα από την στιγμή που έγινε η εξαγωγή των χρηστών μέχρι και το τελευταίο ερώτημα διήρκησε 90 μέρες περίπου, διάστημα που χρησιμοποιούν και οι A. Rao, N. Spasojevic, Z. Li and T. Dsouza [3], για την κατασκευή των βραχυχρόνιων χαρακτηριστικών του αλγορίθμου του Klout.
- Στο τέταρτο στάδιο έγινε η αξιολόγηση της ποιότητας των ακόλουθων κάθε χρήστη. Να τονιστεί ότι αυτή η μεταβλητή χρησιμοποιείται από τους D. Garcia1, P Mavrodiev, D Casati, F. Schweitzer [18], αλλά στην συγκεκριμένη εργασία ορίστηκε με διαφορετικό τρόπο. Συγκεκριμένα αυτή η μεταβλητή χρησιμοποιήθηκε ως βασική παράμετρος αξιολόγησης του αλγορίθμου ποσοτικοποιώντας ταυτόχρονα την ιδιότητα που χαρακτηρίζει τον αλγόριθμο του Thomas Renault [4], ότι τα άτομα με επίδραση ακολουθούνται επίσης από άτομα με επίδραση, όπως και την ιδιότητα του Daniel Tunkelang [2], που ορίζει ως επίδραση την ποσότητα της προσοχής που μπορεί να σου δώσει το κοινό σου συν την ποσότητα της προσοχής που μπορεί να σου φέρει το κοινό σου μέσω του δικού του κοινού. Τέλος, για την δημιουργία της μεταβλητής της

ποιότητας των ακόλουθων χρησιμοποιούνται μόνο οι ακόλουθοι του πρώτου επιπέδου και δεν γίνεται έρευνα σε δεύτερο επίπεδο, γιατί όπως αναφέρουν οι E. Bakshy, J. Hofman, W. Mason, D. Watts [7], το βάθος της διάδοσης της πληροφορίας είναι ιδιαίτερα μικρό.

- Στο πέμπτο στάδιο εισάγεται μια νέα μεταβλητή, η μεταβλητή που προσδιορίζει την ώρες που δημοσιεύει ένας χρήστης. Προκειμένου να δημιουργηθεί αυτή η μεταβλητή παρατηρείται η ώρα που λαμβάνουν χώρα οι δημοσιεύσεις. Η ιδέα πηγάζει από την αντιστροφή της πρότασης που αναφέρει ποιες είναι οι καταλληλότερες ώρες για να δημοσιεύει κάποιος όπως μελετάτε από τους N. Spasojevic, Z. Li, A. Rao, P. Bhattacharyya [20]. Συγκεκριμένα η ανεστραμμένη πρόταση οδηγεί σε μια μεταβλητή που επιβραβεύει τα άτομα που καταφέρνουν να πηγαίνουν καλά με δημοσιεύσεις που δημοσιεύονται σε ώρες που δεν είναι ώρες αιχμής.

Τέλος, είναι απαραίτητο να σχολιαστεί ότι ο συγκεκριμένος αλγόριθμος είναι μια τελείως διαφορετική έκδοση από αυτή των I. Anger και C. Kittl [1], εξαλείφοντας ταυτόχρονα τα 3 βασικά του μειονεκτήματα (αξιολογεί τους ακόλουθους, αξιολογεί την ποιότητα των αλληλεπιδράσεων, λαμβάνει βραχυχρόνια και μακροχρόνια χαρακτηριστικά υπόψιν, αφαιρεί τα bots, και τέλος επιβραβεύει άτομα που δημοσιεύουν σε ώρες που δεν είναι ώρες αιχμής)

3.3 Στάδιο συλλογής και διαλογής των χρηστών

Σε αυτό το στάδιο έγινε η εξαγωγή των χρηστών προς μελέτη και η τοποθέτησή τους σε θεματικές κατηγορίες. Συγκεκριμένα η διαδικασία που περιγράφει τον τρόπο εξαγωγής των χρηστών είναι η εξής:

Με κέντρο το Λονδίνο και ακτίνα 1000 μίλια εκτελέστηκαν ερωτήματα στο Twitter με σκοπό την εύρεση χρηστών που στις δημοσιεύσεις τους περιέχουν συγκεκριμένες λέξεις κλειδιά (προϋπόθεση ήταν η λέξη να περιέχεται στην δημοσίευση και να μην είναι απαραίτητα hashtag[1]). Κάθε ερώτημα επέστρεφε το πολύ 100 τυχαίους χρήστες. Για κάθε χρήστη αποθηκεύονταν το μοναδικό αριθμητικό προσδιοριστικό του (user_id) και τα hashtags που περιέχονταν στην δημοσίευσή του. Οι λέξεις κλειδιά (hashtags) χρησιμοποιήθηκαν στην συνέχεια για την διεξαγωγή των επόμενων ερωτημάτων. Ειδικότερα, έγινε μια εκτίμηση για το ποιες λέξεις κλειδιά εκπροσωπούν καλύτερα την κάθε θεματική κατηγορία συνυπολογίζοντας ταυτόχρονα και την συχνότητα εμφάνισης της κάθε λέξης κλειδί. Συνολικά δημιουργήθηκαν 3 θεματικές κατηγορίες. Για την δημιουργία της κάθε θεματικής κατηγορίας συλλέχθηκαν χρήστες από 4 διαφορετικά ερωτήματα. Συνεπώς, κάθε κατηγορία απαρτίστηκε από 400 περίπου χρήστες. Οι 3 θεματικές κατηγορίες που δημιουργήθηκαν ήταν:

[1] Hashtags: Οι λέξεις που περιέχονται σε μια δημοσίευση και έπονται των συμβόλων #

- **Έκτακτα Νέα**: όπου χρησιμοποιήθηκαν λέξεις κλειδιά για την επίθεση στον Λονδίνο.
  (attack, beatifulCity, london, Staystrong)
- **Αθλητικά**: όπου χρησιμοποιήθηκαν λέξεις κλειδιά από τον τελικό του Champions League, το Final four και πιο γενικές λέξεις.
  (ChampionsLeague, FinalFour, sport, win)
- **Προσωπική Υγεία** (ψυχολογία, αυτοβελτίωση): όπου χρησιμοποιήθηκαν πιο γενικές λέξεις.
  (fitness, health, MentalHealth, psychology)

Ιδιαίτερη μνεία πρέπει να αποδοθεί στο γεγονός ότι την περίοδο συλλογής δεδομένων συνέβησαν και σημαντικά γεγονότα. Οι χρήστες συλλέχτηκαν την ημέρα που έγινε η επίθεση στο Λονδίνο, δηλαδή στις 3 Ιουνίου του 2017. Την ίδια ημέρα διεξάγονταν και ο τελικός Champions League, και στις 21 Μαΐου 2017 είχε διεξαχθεί και ο τελικός του Μπάσκετ (κοντινή ημερομηνία).

Αφού λοιπόν συλλέχθηκαν οι χρήστες ανά κατηγορία και ανά λέξει κλειδί, ενώθηκαν ανά κατηγορία όλα τα αρχεία που είχαν προκύψει λόγω των διαφορετικών ερωτημάτων σε ένα ενιαίο αρχείο, όπου προστέθηκαν μόνο τα αναγνωριστικά ids και αφαιρέθηκαν τα διπλότυπα.

Για κάθε χρήστη της κάθε κατηγορίας βρέθηκαν τα στοιχεία του { id_str, followers, friends, favourites, statuses, profile_url } και οι 400 καλύτεροι ακόλουθοι και αποθηκεύτηκαν σε μια βάση δεδομένων. Με τον όρο καλύτεροι ακόλουθοι, θεωρούμε τους ακόλουθους που έχουν τον μεγαλύτερο λόγο followers/friends. Για την συγκεκριμένη διαδικασία μελετήθηκαν όλοι οι ακόλουθοι των 400 περίπου χρηστών της κάθε κατηγορίας. Το μέγεθος της αναζήτησης ήταν 4 εκατομμύρια περίπου. Τέλος, για πρακτικούς λόγους αφαιρέθηκαν όλοι οι χρήστες που είχαν κλειδωμένο (private) λογαριασμό.

3.4 Στάδιο εξαγωγής μακροχρόνιων χαρακτηριστικών

Η διεξαγωγή των μακροχρόνιων χαρακτηριστικών έγινε κατά το στάδιο κατάταξης των χρηστών σε κατηγορίες. Λόγω του περιορισμού των δεδομένων που μπορούσαν να εξαχθούν, το Twitter διέθετε ελάχιστη πληροφορία προς επεξεργασία, δημιουργήθηκαν τρεις δείκτες για την εκπροσώπηση των μακροχρόνιων χαρακτηριστικών των χρηστών. Συγκεκριμένα αυτοί οι δείκτες είναι:

- Ai = followers_count (αντιπροσωπεύει το σύνολο των ακόλουθων ενός ατόμου)
- Bi = followers_count/ friends_count (σύνολο φίλων προς σύνολο ακόλουθων – πλήθος ατόμων που ακολουθούνται προς πλήθος ατόμων που ακολουθούν)
- Ποιότητα_ακόλουθων = Για κάθε έναν από τους 400 ακόλουθους κάθε χρήστη δημιουργήθηκαν οι δείκτες Ai & Bi (400 γιατί έχουμε κρατήσει ήδη τους 400 καλύτερους ακόλουθους από το σύνολο των ακόλουθων ενός χρήστη βάση του

δείκτη Bi). Στην συνέχεια, ταξινομήθηκαν οι 400 ακόλουθοι βάση του μέσου όρου Ai & Bi και έπειτα δημιουργήθηκε η μεταβλητή της ποιότητας των ακόλουθων για 6 διαφορετικές περιπτώσεις. Ποιότητα10, ποιότητα20, ποιότητα50, ποιότητα100, ποιότητα200, ποιότητα400. Κάθε μια από αυτές τις μεταβλητές είναι ο μέσος όρος των 10, 20, 50, 100, 200, 400 ατόμων, όπου κάθε άτομο αντιπροσωπεύεται από τον μέσο όρο των Ai & Bi.

3.5 Στάδιο εξαγωγής βραχυχρόνιων χαρακτηριστικών των χρηστών

Για την εξαγωγή των βραχυχρόνιων χαρακτηριστικών μελετήθηκαν όλες οι δημοσιεύσεις όλων των χρηστών για διάστημα 9 εβδομάδων. Για κάθε χρήστη μελετήθηκαν τα βραχυχρόνια χαρακτηριστικά βάσει των δημοσιεύσεών του (tweets), χωρίς όμως να ληφθούν υπόψη τα retweets καθώς περιέχουν πληροφορίες από το πρώτο άτομο που είχε δημοσιεύσει το εκάστοτε tweet που στην συνέχεια επαναδημοσιεύονταν. Συγκεκριμένα, η συλλογή των δεδομένων γινόταν σε εβδομαδιαία βάση και περιλάμβανε όλα τα tweets που είχε δημοσιεύσει ο κάθε χρήστης στο διάστημα της μιας εβδομάδας. Συνεπώς, κάθε εβδομάδα αποθηκεύονταν σε ένα αρχείο το πλήθος των favourite που δεχόταν ένας χρήστης στο σύνολο των δημοσιεύσεών του (πόσες φορές έγιναν like τα tweets που δημοσίευε ο χρήστης σε χρονικό διάστημα μιας εβδομάδας), το πλήθος των retweet που δεχόταν τα tweets του χρήστη (πόσες φορές επαναδημοσιεύθηκαν τα tweets του χρήστη στο διάστημα μιας εβδομάδας), το σύνολο των tweets που δημοσίευε ο χρήστης σε μια εβδομάδα και τέλος το χρονικό διάστημα που δημοσιεύονταν τα tweets του χρήστη. Για παράδειγμα, αν ένας χρήστης δημοσίευε 10 tweets σε μια εβδομάδα θα αποθηκεύονταν ποια ώρα της ημέρας έγιναν αυτά τα tweets χωρίς να υπάρχει σύνδεση του πόσο καλά πήγε ένα tweet που δημοσιεύθηκε την τάδε ώρα. Αυτό που αποθηκεύονταν ήταν τα συνολικά αποτελέσματα. Έχοντας εξαγάγει λοιπόν αυτά τα δεδομένα για 9 εβδομάδες, η πρώτη εβδομάδα χρησιμοποιήθηκε για να δημιουργηθεί ένα σημείο εκκίνησης, αθροίστηκαν τα αποτελέσματα των 8 πιο πρόσφατων εβδομάδων σε ένα αρχείο. Στην συνέχεια από το τελικό αρχείο διαβάστηκαν τα δεδομένα και δημιουργήθηκαν οι δείκτες:

- Ci = favourite_count /total_tweets (σύνολο των like που δέχτηκαν τα tweets του κάθε χρήστη στο διάστημα 8 εβδομάδων προς σύνολο των tweet σε αυτό το διάστημα)
- Di = retweet_count/total_tweets (σύνολο των retweet που δέχτηκαν τα tweets του κάθε χρήστη στο διάστημα 8 εβδομάδων προς το σύνολο των δημοσιεύσεων σε αυτό το διάστημα)
- hour_posting = μεταβλητή που αντιπροσωπεύει το τι ώρες δημοσιεύει τα tweets του ο κάθε χρήστης. Στον πίνακα D αποθηκεύτηκαν οι 24 κανονικοποιημένες τιμές που αντιπροσωπεύουν το τι ώρες έγιναν τα tweets του χρήστη στο διάστημα των 8 εβδομάδων. Στην συνέχεια όλες αυτές οι τιμές συνδυάστηκαν με τις τιμές του πίνακα PH, όπου περιέχει τις κανονικοποιημένες τιμές της

πρότυπης συχνότητας δημοσίευσης ανά ώρα στην Ευρώπη. Όλες αυτές οι μεταβλητές συνδυάστηκαν στην σχέση:

$$hour\_posting = \sum_{k=0, hour\_posting=0}^{k=23} hour\_posting + D[k] * (1 - PH[k])$$

Τέλος, να σημειωθεί ότι κατά την εξαγωγή των βραχυχρόνιων δεδομένων ελέγχονταν για κάθε tweet αν περιέχονται συγκεκριμένες λέξεις κλειδιά που χρησιμοποιούν οι χρήστες που λειτουργούν με αυτοματοποιημένο τρόπο (bots). Οι συγκεκριμένες λέξεις είναι οι ίδιες που χρησιμοποιούνταν από τους M. Danisch, N. Dugu, A. Perez [6] (#TeamFollowBack, #instantfollowbackdedicated, #teamautofollow).

3.6 Στάδιο κανονικοποίησης των μεταβλητών

Προκειμένου να μπορούν όλες οι μεταβλητές, μακροχρόνιες και βραχυχρόνιες, να δουλέψουν ταυτόχρονα, κανονικοποιήθηκαν. Για την κανονικοποίηση τους χρησιμοποιήθηκε η σχέση $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$. Απαραίτητη προϋπόθεση για να δουλέψει η σχέση είναι η εύρεση του μέγιστου και του ελάχιστου της κάθε μεταβλητής στην κάθε κατηγορία. Για την κανονικοποίηση της μεταβλητής της ποιότητας των ακόλουθων πρώτα κανονικοποιήθηκαν οι μεταβλητές Ai & Bi, στην συνέχεια βρέθηκε ο μέσος όρος τους και τέλος βρέθηκαν οι μέσοι όροι του συνόλου των ακόλουθων κάθε χρήστη, δηλαδή η κανονικοποίηση ήταν το πρώτο στάδιο και όχι το τελικό. Όσον αφορά την κανονικοποίηση της μεταβλητής που εκπροσωπεί την ώρα δημοσίευσης, δεν χρησιμοποιήθηκε η παραπάνω σχέση, αλλά για κάθε χρήστη διαιρέθηκε το σύνολο των δημοσιεύσεων για κάθε ώρα της ημέρας για το διάστημα των 8 εβδομάδων με το σύνολο των δημοσιεύσεων όλων των ωρών για το διάστημα των 8 εβδομάδων (η ίδια διαδικασία ακολουθήθηκε για τον πίνακα PH, που αρχικά περιείχε το σύνολο των δημοσιεύσεων ανά ώρα για 4.8εκ. δημοσιεύσεις για 10.000 χρήστες, σύμφωνα με έρευνα που αναφέρει τι ώρα να δημοσιεύει κανείς στην Ευρώπη), οπότε δημιουργήθηκαν οι συχνότητες δημοσίευσης του χρήστη για τις διάφορες ώρες του 24ώρου.

4 *Τεχνικές λεπτομέρειες*

4.1 Λεπτομέρειες υλοποίησης

Το πρώτο βήμα που χρειάστηκε προκείμενου να γίνει η υλοποίηση της διπλωματικής ήταν η εξοικείωση με τις εξής γλώσσες προγραμματισμού:

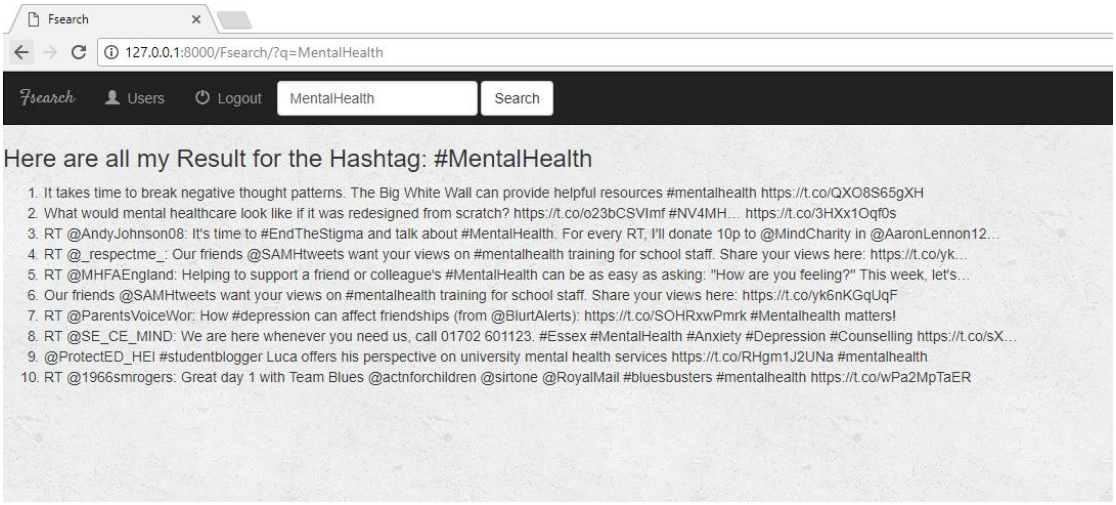
- Django
- Python
- JavaScript
- Html
- Css

Επιπλέον, απαραίτητη προϋπόθεση ήταν η λεπτομερής ανάγνωση της βιβλιογραφίας που περιγράφει πως λειτουργούν τα ερωτήματα στο Twitter (εξοικείωση με το Twitter API) και κατά επέκταση κατανόηση της διαθέσιμης πληροφορίας προς εξαγωγή και ακόμη περισσότερο των περιορισμών που ορίζει το μέσο κοινωνικής δικτύωσης. Τέλος, να τονιστεί ότι η ανάπτυξη του κώδικα έγινε στην πλατφόρμα του Pycharm.

4.1.1 Εξαγωγή χρηστών

Προκειμένου να επιτευχθούν τα παραπάνω κατασκευάστηκε μια διαδικτυακή (web) εφαρμογή που με κέντρο το Λονδίνο (51.500152, -0.126236) και ακτίνα 1000 μίλια εκτελούσε ερωτήματα στο Twitter βάση μιας λέξης κλειδί, και έβρισκε σε αυτό τις πιο πρόσφατες δημοσιεύσεις που περιείχαν την συγκεκριμένη λέξη κλειδί. Στην συνέχεια επιστρέφονταν το πολύ 100 δημοσιεύσεις που περιείχαν αυτήν την λέξη αποθηκεύοντας τα αποτελέσματα σε ένα αρχείο .txt που έπαιρνε ως όνομα την λέξη του ερωτήματος. Η πληροφορίες που αποθηκεύονταν ήταν το αναγνωριστικό κάθε χρήστη (user_id) και μια λίστα με τις λέξεις της δημοσίευσης που βρίσκονταν μετά το hashtag. Τέλος, εμφανίζονταν στην οθόνη του χρήστη τα 10 πιο πρόσφατα αποτελέσματα της αναζήτησης. Στο σχήμα που ακολουθεί αποτυπώνεται το

αποτέλεσμα ενός ερωτήματος που είχε γίνει για την κατηγορία προσωπική υγεία με λέξη αναζήτησης «MentalHealth».



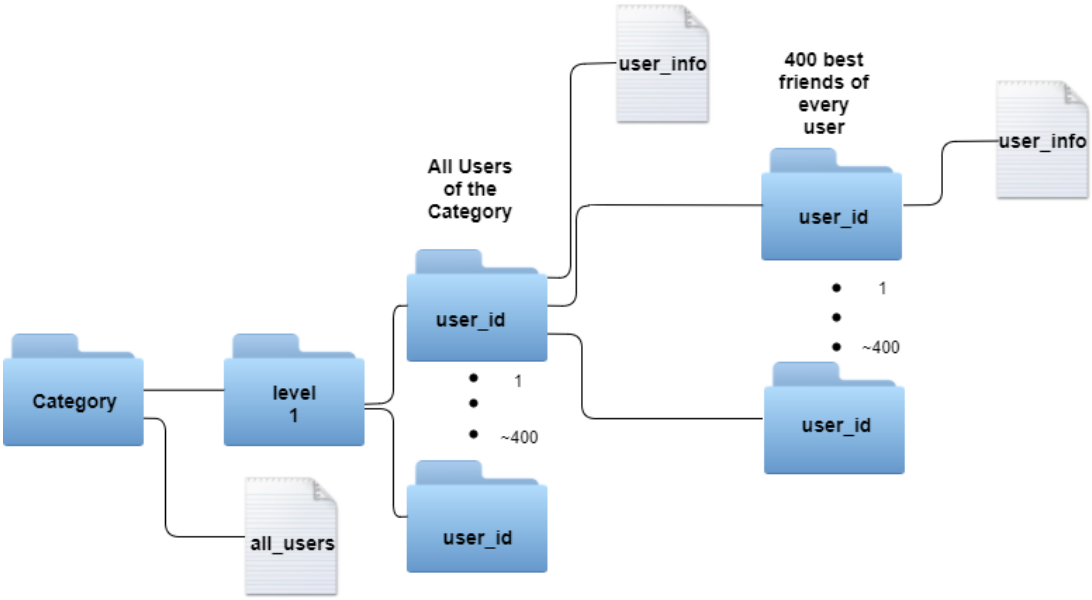
Εικόνα 3: Διαδικτυακή εφαρμογή αναζήτησης δημοσιεύσεων

Με την παραπάνω διαδικασία συλλέγονταν το πολύ 100 χρήστες που περιείχαν στις δημοσιεύσεις τους την εκάστοτε λέξη που χρησιμοποιούνταν για την εκτέλεση του ερωτήματος. Στην συνέχεια ανοίγονταν το κάθε αρχείο, εντοπίζοντας τις λέξεις με την πιο μεγάλη συχνότητα εμφάνισης και με διαισθητικούς κυρίως λόγους διαλέγονταν η επόμενη λέξη κλειδί, από αυτές με την μεγαλύτερη συχνότητα εμφάνισης, που θα συλλέξει τους επόμενους χρήστες. Η διαδικασία επαναλαμβάνονταν μέχρι να συμπληρωθούν οι 4 λέξεις κλειδιά (πρώτη λέξη κλειδί επιλέχθηκε αυθαίρετα). Με αυτόν τον τρόπο έγινε η εξαγωγή και κατηγοριοποίηση των χρηστών στις τρεις μεγάλες κατηγορίες (Προσωπική υγεία, Γενικές ειδήσεις, Αθλητισμός). Τέλος, τα 4 αρχεία της κάθε κατηγορίας ενοποιήθηκαν σε ένα ενιαίο αρχείο που περιείχε όλα τα αναγνωριστικά (user_ids) των χρηστών αφαιρώντας τα διπλότυπα.

4.1.2 Βάση δεδομένων

Αφού λοιπόν βρέθηκαν οι χρήστες δημιουργήθηκε μια βάση δεδομένων όπου αποθηκεύονταν οι πληροφορίες τους. Η διαδικασία που ακολουθήθηκε ήταν αρχικά να διαβαστεί κάθε ένα από τα ενοποιημένα αρχεία της κάθε κατηγορίας και για κάθε έναν χρήστη που περιέχονταν εκεί μέσα διεξάγονταν ένα ερώτημα στο Twitter προκειμένου να βρεθούν οι πληροφορίες που τον χαρακτηρίζουν (id_str, followers_count, friends_count, favourite_count, statuses_count, profile_image_url). Το κάθε .txt αρχείο

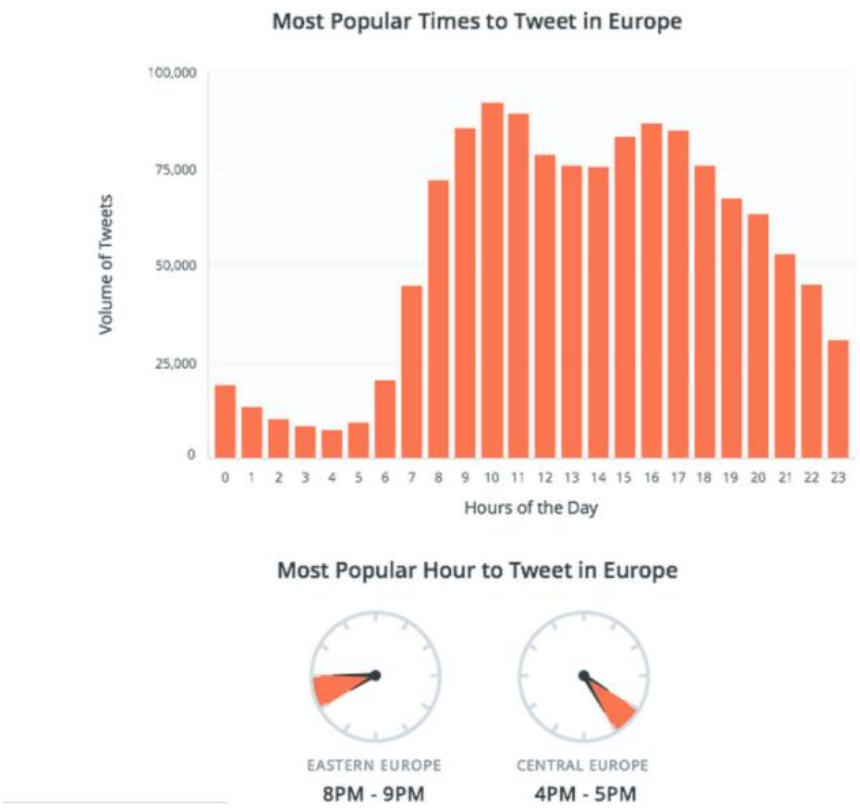
στο οποίο αποθηκεύονταν οι πληροφορίες του κάθε χρήστη και είχε ως όνομα το user_id του εκάστοτε χρήστη αποθηκεύονταν μέσα σε έναν φάκελο που είχε και αυτό ως όνομα το user_id του χρήστη. Ακολουθώντας την ίδια λογική, αποθηκεύονταν στην συνέχεια οι πληροφορίες για τους 400 καλύτερους ακόλουθους του κάθε χρήστη. Στην συνέχεια δίνεται η σχηματική αναπαράσταση της βάσης δεδομένων.



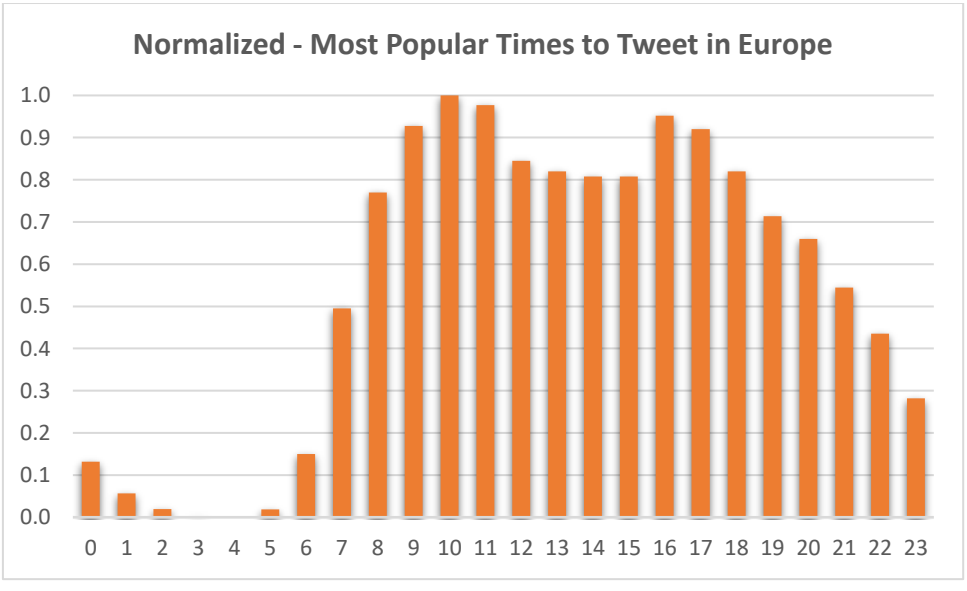
Εικόνα 4: Εικονικό σχήμα αναπαράστασης βάσης δεδομένων

4.1.3 Ώρα δημοσίευσης

Μια ιδιαίτερα σημαντική παράμετρος που διαφοροποιεί την συγκεκριμένη εργασία σε σχέση με τις προηγούμενες είναι η δημιουργία της μεταβλητής που επιβραβεύει τα άτομα που καταφέρνουν να διαδίδουν την πληροφορία σε ώρες της ημέρας όπου η διάδοση της πληροφορίας είναι πιο δύσκολη. Για τον λόγο αυτό από το παρακάτω διάγραμμα που είναι το αποτέλεσμα μιας έρευνας που είχε διεξαχθεί σε 10,000 χρήστες για 4,8 εκατομμύρια tweets, με τελευταία ενημέρωση δεδομένων στις 5/4/2016, προσδιορίζεται μια γενική αλήθεια για το πότε δημοσιεύουν οι χρήστες στην Ευρώπη. Προκειμένου να χρησιμοποιηθεί το παραπάνω γράφημα βρέθηκαν προσεγγιστικά οι τιμές που αντιπροσωπεύει η κάθε στήλη του γραφήματος. Στην συνέχεια αυτές οι τιμές κανονικοποιήθηκαν, αφού διαιρέθηκαν με την μέγιστη τιμή και αποτέλεσαν τις πρότυπες ώρες για να δημοσιεύει κάποιος. Στην συνέχεια παρουσιάζεται το αρχικό και το κανονικοποιημένο πρότυπο διάγραμμα.



Εικόνα 5: Σχήμα αναπαράστασης πρότυπης ώρας δημοσίευσης [22]



Εικόνα 6: Κανονικοποιημένο σχήμα αναπαράστασης πρότυπης ώρας δημοσίευσης [22]

5 *Αξιολόγηση των αποτελεσμάτων*

5.1 Σχολιασμός των αποτελεσμάτων του αλγορίθμου

Το πρώτο στάδιο διεξαγωγής των αποτελεσμάτων περιλάμβανε την κατάταξη των χρηστών κάθε κατηγορίας με τους εξής τρόπους: α. Βάση κάθε μιας από τις 6 μεταβλητές ξεχωριστά (followers, followers/friends, favourites/total_tweets, retweets/total_tweets, hour_index, quality_of_friends_10), β. βάση του μέσου όρου των έξη μεταβλητών και γ. κατάταξη των χρηστών πρώτα ως προς το πλήθος των followers, στην συνέχεια εξαγωγή των 21 πρώτων χρηστών και τέλος κατάταξη αυτών των 21 χρηστών κάθε κατηγορίας βάση του μέσου όρου των υπόλοιπων 5 μεταβλητών. Με τον τελευταίο τρόπο εξασφαλίζονταν να μην μπούνε άτομα στην λίστα των 21 που δεν έχουνε επαρκή αριθμό ακόλουθων.

Επιτυχία δειγματοληψίας

Στην κατηγορία Personal Health (self-improvement) από τους 21 πρώτους χρήστες βάση του πλήθους των followers οι 8 μιλάνε αποκλειστικά για ζητήματα που αφορούν την προσωπική υγεία (8/21). Το αποτέλεσμα αυτό αντιστοιχεί στο 38% των πιθανών επηρεάζοντων. Ο λόγος που στην συγκεκριμένη κατηγορία η συσχέτιση δεν είναι τόσο υψηλή είναι κυρίως συγκυριακός, καθώς οφείλεται εν μέρη στον τρόπο με τον οποίο έγινε η δειγματοληψία (μπορεί να χρειαζόταν παραπάνω λέξεις κλειδιά), αλλά ταυτόχρονα και στο γεγονός ότι εκείνη την περίοδο ενδεχομένως να υπήρχε μια τάση για τέτοιου είδους δημοσιεύσεις από τους χρήστες λόγω της επίθεσης που είχε γίνει στο Λονδίνο. Στις επόμενες δύο κατηγορίες News Oriented και About Sport η επιτυχία της δειγματοληψίας ήταν υψηλότερη, καθώς 10/21 και 14/21 αντίστοιχα είναι τα άτομα που μιλάνε αποκλειστικά για θέματα της θεματικής κατηγορίας στην οποία ανήκουν.

About Sport

followers	followers/ friends	favourites/ total_tweets	retweets/ total_tweets	hour_index	Quality_of_ followers_10	All six metrics	Two Level
metric0	metric1	metric2	metric3	metric4	metric5	SameWeight	SameWeight
123355219	436141986	123355219	123355219	513985476	22903166	123355219	123355219
22903166	2447460852	185179480	185179480	200882064	436141986	436141986	436141986
436141986	22903166	1104776366	137648408	2160126689	244422106	22903166	22903166
137648408	386308403	262340419	837168378	325914120	1102291885	2447460852	2447460852
185179480	65309424	890794543	161729102	21265939	876350444	513985476	92235951
244422106	29203487	137648408	242760367	245172975	137648408	200882064	137648408
2447460852	85788844	293594004	193689356	326354621	185179480	2160126689	185179480
92235951	890794543	193689356	259896187	92235951	390042363	92235951	1102291885
390042363	2371543241	161729102	890794543	245530412	92235951	325914120	1368601795
1368601795	137648408	242760367	339365676	449447183	161729102	21265939	890794543
42376866	365432110	259896187	313827158	40885653	24900238	326354621	2371543241
1102291885	517607733	535411311	22903166	815519982	42376866	245172975	242760367
2371543241	310314587	41584848	436141986	1969659680	517607733	386308403	390042363
890794543	1191500480	288924330	535411311	128521031	103693118	29203487	517607733
41584848	98981265	614046195	41584848	425408030	301527968	245530412	41584848
517607733	242760367	876350444	155260536	2447460852	1368601795	137648408	161729102
24900238	193689356	143390882	876350444	308417103	41584848	815519982	244422106
876350444	262340419	837168378	1455353010	512069020	462819202	185179480	876350444
161729102	1969659680	313827158	448947013	29203487	890794543	1969659680	42376866
242760367	535411311	1265168803	234837692	562419999	111087583	449447183	24900238
1191500480	293594004	144920233	1368601795	2449716486	242760367	40885653	1191500480

Εικόνα 7: Αποτέλεσμα ταξινόμησης των χρηστών της κατηγορίας αθλητικά

Από τον παραπάνω πίνακα διεξάγονται οι εξής παρατηρήσεις:

- Λαμβάνοντας υπόψιν ότι το πλήθος των ακόλουθων που έχει κάποιος είναι απαραίτητη προϋπόθεση για να μπορεί να χαρακτηριστεί ως επηρεάζων, παρατηρείται ότι την μεγαλύτερη σύγκλιση με αυτήν την μετρική παρουσιάζουν η μεταβλητή της ποιότητας των ακόλουθων και η μεταβλητή που προσδιορίζει το πόσο πολύ επαναδημοσιεύεται ένας χρήστης. Ιδιαίτερα σημαντικό είναι το γεγονός ότι και οι υπόλοιπες μετρικές παρουσιάζουν αρκετά ικανοποιητική σύγκλιση. Επιπλέον, η μεταβλητή που προσδιορίζει το τι ώρες δημοσιεύει ένας χρήστης δεν περιλαμβάνει πολλούς από τους 21 πρώτους χρήστες βάση του πλήθους των ακόλουθων (μόνο 2 στους 21), αφού αυτοί ως υποψήφιοι επηρεάζοντες είναι λογικό να δημοσιεύουν σε ώρες αιχμής (γίνεται μια πρώτη επιβεβαίωση της θεωρίας ότι οι επηρεάζοντες δημοσιεύουν σε ώρες αιχμής). Ο λόγος όμως που υπάρχει αυτή η μεταβλητή είναι για να εντοπίζει και να ενισχύει χρήστες που πηγαίνουν καλά παρότι δεν δημοσιεύουν σε ώρες αιχμής.

- Ο απλός μέσος όρος παρουσιάζει σημαντικές αδυναμίες, αφού προστίθενται χρήστες που σε καμία περίπτωση δεν μπορούν να θεωρηθούν επηρεάζοντες, όπως για παράδειγμα ο 513985476, που είναι ένας απλός χρήστης με 6 ακόλουθους που ενδεχομένως να ανέβηκε ψηλά στην λίστα λόγω του γεγονότος ότι δημοσιεύει σε ώρες που δεν είναι ώρες αιχμής. Αυτό το πρόβλημα λοιπόν επιλύεται με την εισαγωγή του μέσου όρου δύο επιπέδων.

- Ο μέσος όρος δύο επιπέδων καταφέρνει και βελτιώνει την σειρά κατάταξης αφού λαμβάνει υπόψιν τις υπόλοιπες μεταβλητές υπό την προϋπόθεση ότι τα άτομα έχουν αρκετούς followers. Αυτό αποδεικνύεται με παρατήρηση των προφίλ των χρηστών που κατάφεραν να αλλάξουν θέση. Για παράδειγμα ο

χρήστης 436141986 παρά το γεγονός ότι έχει 144K ακόλουθους καταφέρνει να ξεπεράσει τον χρήστη 22903166 που έχει 213K ακόλουθους. Αυτό οφείλεται κυρίως στο γεγονός ότι ο πρώτος χρήστης έχει καλύτερο λόγο followers/friends και ταυτόχρονα καταφέρνει να αποσπά περισσότερες επαναδημοσιεύσεις ανά δημοσίευση.

- Σε αυτήν την κατηγορία τα άτομα με περισσότερους από 100K ακόλουθους είναι 3 στους 211 δηλαδή το 1,42% το οποίο είναι μικρότερο από 2%

News oriented

followers	followers/ friends	favourites/ total_tweets	retweets/ total_tweets	hour_index	Quality_of_ followers_10	All six metrics	Two Level
metric0	metric1	metric2	metric3	metric4	metric5	SameWeight	SameWeight
38142380	38142380	493985399	493985399	214811227	38142380	38142380	38142380
1217272591	1217272591	555370896	555370896	562251683	1217272591	493985399	493985399
555370896	1191500480	8.27964E+17	8.27964E+17	2371894197	493985399	1217272591	1217272591
166555111	166555111	1217272591	1217272591	7.51237E+17	474910079	555370896	555370896
493985399	585569365	621120527	474910079	3662879419	125054097	8.27964E+17	8.27964E+17
1029442188	3250799330	474910079	579693969	1691309497	555370896	214811227	474910079
474910079	125054097	1402965540	38142380	322289077	11344112	562251683	302600028
125054097	4689622424	579693969	7.53898E+17	4616402067	329176433	7.51237E+17	125054097
212991450	555370896	26795883	125054097	8.51567E+17	368261129	2371894197	81104863
1191500480	1405339620	463906342	19158571	294913295	176449857	3662879419	25079047
302600028	166314889	3339352228	1514639766	155155595	166314889	1691309497	1460919698
210020861	388764797	38142380	81104863	812627438	1029442188	322289077	2388189796
2388189796	1460919698	125054097	1254901999	3052617434	81104863	4616402067	11344112
11344112	474910079	317332081	176449857	7.41891E+17	1191500480	7.41891E+17	166555111
176449857	971807755	2369713616	569197631	762639852	25079047	294913295	585569365
585569365	981463632	1254901999	621120527	164776163	212991450	579693969	368261129
368261129	329176433	4363463301	153295243	2997741610	196515407	8.51567E+17	212991450
25079047	1546566788	81104863	3376728137	438427119	4698167466	812627438	176449857
81104863	1295315881	7.51237E+17	7.51237E+17	8.01328E+17	302600028	2997741610	210020861
1460919698	2786263893	2786263893	3339352228	625037061	317332081	155155595	1029442188
8.27964E+17	7.53898E+17	3238460239	2997741610	1217272591	1685110555	474910079	1191500480

Εικόνα 8: Αποτέλεσμα ταξινόμησης των χρηστών της κατηγορίας έκτακτα νέα

Από τον παραπάνω πίνακα διεξάγονται οι εξής παρατηρήσεις:

- Και σε αυτήν την κατηγορία την μεγαλύτερη σύγκληση παρουσιάζουν οι μεταβλητές followers και Quality_of_followers_10 με αρκετά ικανοποιητική σύγκληση των άλλων μετρικών πέρα του hour_index.
- Όπως μπορεί κανείς εύκολα να διαπιστώσει ανατρέχοντας στο προφίλ του χρήστη 493985399, ο συγκεκριμένος χρήστης καταφέρνει να αποσπάσει μεγάλο αριθμό επαναδημοσιεύσεων και favourite. Ενδεικτικά αναφέρεται ότι για μια δημοσίευση στο διάστημα 6-Σεπτ-2017 μέχρι το διάστημα 18-Σεπτ-2017 καταφέρνει να συλλέξει 7,4K επαναδημοσιεύσεις και 3,8K favourites. Αυτή του η ιδιότητα τον κάνει να ανεβαίνει στις πρώτες θέσεις και να ξεπερνάει ακόμα και χρήστες με 128K ακόλουθους, έχοντας μόνο 21.9K ακόλουθους.
- Σε αυτήν την κατηγορία οι χρήστες με περισσότερους από 100K ακόλουθους είναι 2 στους 324 χρήστες δηλαδή το 0,62%.

Personal Health

followers	followers/ friends	favourites/ total_tweets	retweets/ total_tweets	hour_index	Quality_of_ followers_10	All six metrics	Two Level
metric0	metric1	metric2	metric3	metric4	metric5	SameWeight	SameWeight
18774246	497145453	18774246	315421768	864760548	18774246	18774246	18774246
2568808680	18774246	315421768	18774246	161943713	19594318	315421768	315421768
437500980	1527289544	267156262	927662736	513514024	437500980	497145453	927662736
14933526	14933526	927662736	437500980	2732461070	315421768	927662736	437500980
777150946912722000	39244173	7.73857E+17	7.73857E+17	7.65085E+17	2732584356	864760548	2568808680
2732584356	315421768	437500980	144299276	389491428	14933526	161943713	14933526
94984926	7.84125E+17	8.16562E+17	8.16562E+17	2260605911	2734087717	513514024	777150946912722000
315421768	7.73857E+17	2296668656	1212055152	2726908617	22824243	2732461070	3338877533
927662736	16000065	144299276	8.60123E+17	3067808998	522421143	437500980	2732584356
39244173	437500980	1487224332	8.59693E+17	79835407	325769081	7.65085E+17	1452143478
22824243	7.84142E+17	494455001	7.95956E+17	8.70789E+17	1254424548	389491428	522421143
3338877533	3026493605	1552769161	2297842659	8.38098E+17	39244173	2726908617	22824243
2734087717	2540108371	237932374	793819092	130133175	2568808680	2260605911	626051508
23751066	15496973	509430719	1454200969	8.65984E+17	927662736	3067808998	94984926
522421143	793819092	4237350017	4237350017	1195827348	330602313	79835407	2233761132
1452143478	1004941927	463219708	4119573141	89779210	2606940668	8.70789E+17	3647285597
1254424548	8.37237E+17	7.95956E+17	8.11706E+17	1602311922	23751066	2568808680	39244173
2233761132	3388596436	223136763	8.34692E+17	28881775	94984926	8.38098E+17	104767916
626051508	1142176501	1242530948	237932374	3180990446	45899012	509430719	1254424548
104767916	97434534	8.11706E+17	7.7179E+17	509430719	2233761132	130133175	23751066
3647285597	4237350017	8.60123E+17	73498916	784365722	130693058	7.73857E+17	2734087717

Εικόνα 9: Αποτέλεσμα ταξινόμησης των χρηστών της κατηγορίας προσωπική υγεία

Από τον παραπάνω πίνακα διεξάγονται οι εξής παρατηρήσεις:

- Για ακόμη μια φορά έχουμε πολύ μεγάλη συσχέτιση των μεταβλητών followers και Quality_of_followers_10. Όπως και των μεταβλητών favourites και retweets.
- Το πλήθος των ατόμων με περισσότερους από 100K ακόλουθους είναι 2 στους 339 χρήστες, δηλαδή 0,59%
- Τέλος, ιδιαίτερα θετικό είναι το γεγονός ότι τα υπόλοιπα συμπεράσματα αυτής της κατηγορίας είναι τα ίδια με αυτά των προηγούμενων δύο κατηγοριών.

Γενικά σχόλια

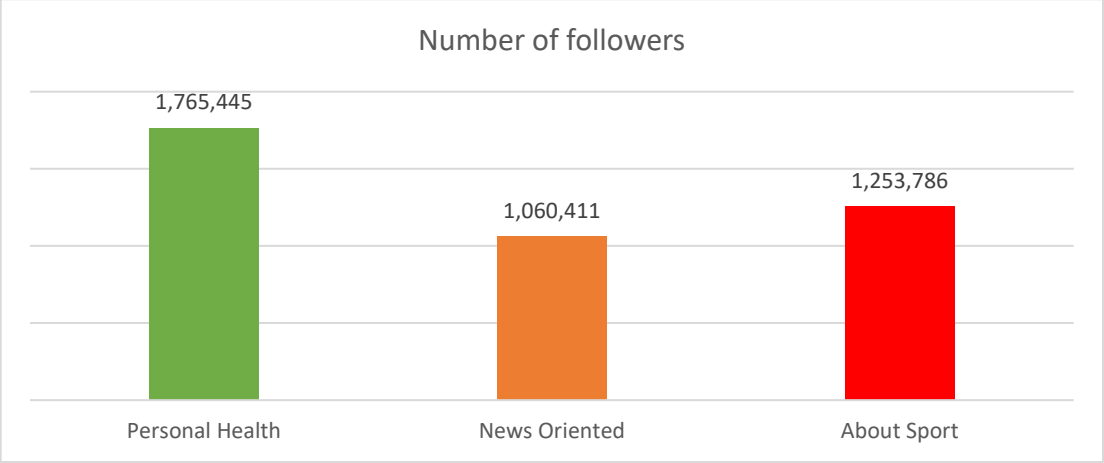
- Το γεγονός ότι η στήλη των followers παρουσιάζει μεγάλη ομοιότητα με την στήλη των Quatlity_of_followers_10 αποδεικνύει την θεωρία του Thomas Renault [4], ότι οι χρήστες με επίδραση ακολουθούνται από χρήστες με επίδραση και ταυτόχρονα αναδεικνύει την επιτυχή λειτουργία αυτής της νέας μεταβλητής.
- Η ομοιότητα της στήλης των followers με την στήλη followers/friends επιβεβαιώνει την θεωρία ότι όσο λιγότερο δημοφιλής είναι ένας χρήστης τόσο πιο ενεργά ακολουθεί άλλους χρήστες, όπως ακριβώς αναφέρουν και οι M. Cha, F. Benevenuto, H. Haddadi, K. Gummadi [17]
- Σε κάθε κατηγορία, ενώ έγινε τυχαία δειγματοληψία επιβεβαιώνεται ο ισχυρισμός των M. Cha, F. Benevenuto, H. Haddadi, K. Gummadi [17], ότι οι επηρεάζοντες αποτελούν περίπου το 2% του συνόλου των χρηστών.
- Και στις 3 κατηγορίες υπάρχει μεγάλη συσχέτιση μεταξύ των μεταβλητών favourite/total_tweets και retweet/total_tweet. Αυτό σημαίνει ότι οι δημοσιεύσεις που καταφέρνουν να αποσπάσουν αρκετές επαναδημοσιεύσεις καταφέρουν να αποσπάσουν και πολλά favourites και αντίστροφα. Επιπλέον σημαίνει ότι χρήστες που αποσπούν πολλές επαναδημοσιεύσεις retweets κατά πάσα πιθανότητα αποσπούν και πολλά favourites.

- Η λίστα με το πλήθος των followers δεν έχει μεγάλη συσχέτιση με τις λίστες των favourite και retweet. Αυτό σημαίνει ότι τα άτομα με τους περισσότερους ακόλουθους δεν σημαίνει απαραίτητα είναι τα άτομα που έχουν και την μεγαλύτερη επίδραση.
- Η αιτία που η λίστα των All_six_metrics παρουσιάζει πολύ μεγάλη ομοιότητα με την λίστα hour_index (και στις 3 κατηγορίες) είναι διότι οι χρήστες σε αυτόν τον δείκτη έχουν περίπου την ίδια τιμή λόγω της ομοιότητας που υπάρχει στις ώρες δημοσιεύσεων (για αυτό εξάλλου ανέβηκαν και ψηλά στην λίστα) και ταυτόχρονα οι πρώτοι χρήστες βάση του δείκτη followers έχουν πολύ μεγάλη απόσταση με τους υπόλοιπους χρήστες (λόγω της μεγάλης διαφοράς στο πλήθος των ακόλουθων) με αποτέλεσμα να δημιουργούν κατά την κανονικοποίηση πολύ μικρούς αριθμούς στους υπόλοιπους χρήστες. Αυτός είναι και ο λόγος που δεν δουλεύει σωστά ο απλός μέσος όρος.

5.2 Σχολιασμός των βραχυχρόνιων αποτελεσμάτων

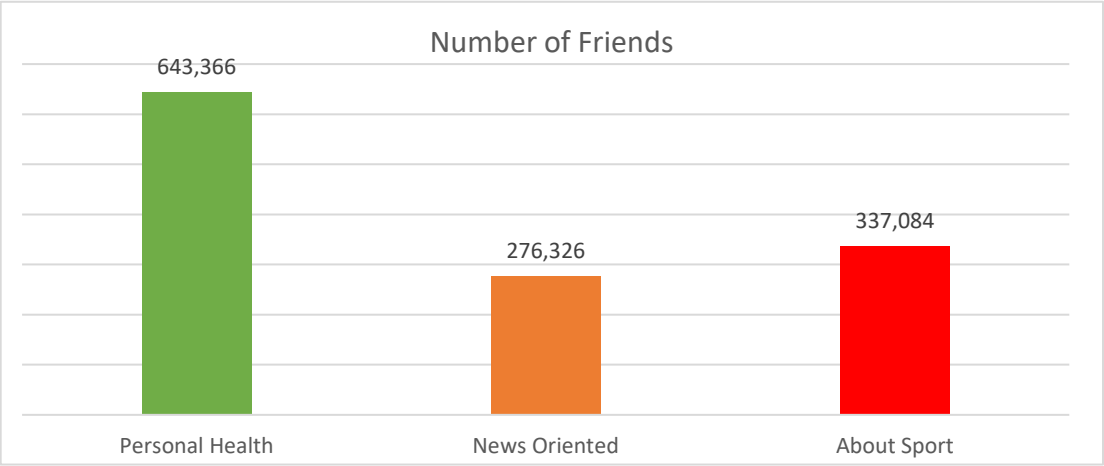
Στην συνέχεια παρουσιάζονται και σχολιάζονται τα στατιστικά αποτελέσματα των αποτελεσμάτων της συλλογής των δεδομένων.

Πλήθος ακόλουθων



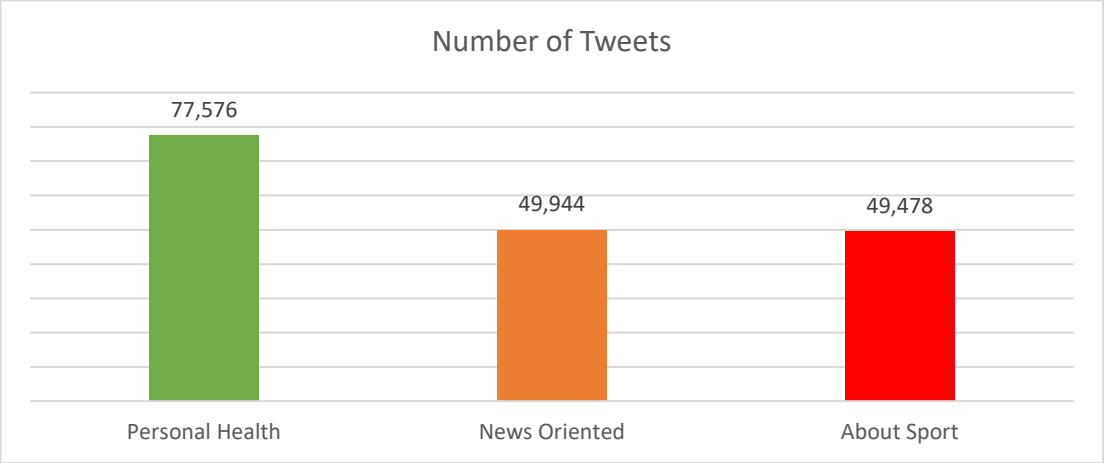
Εικόνα 10: Πλήθος ακόλουθων σε κάθε κατηγορία

Πλήθος φίλων



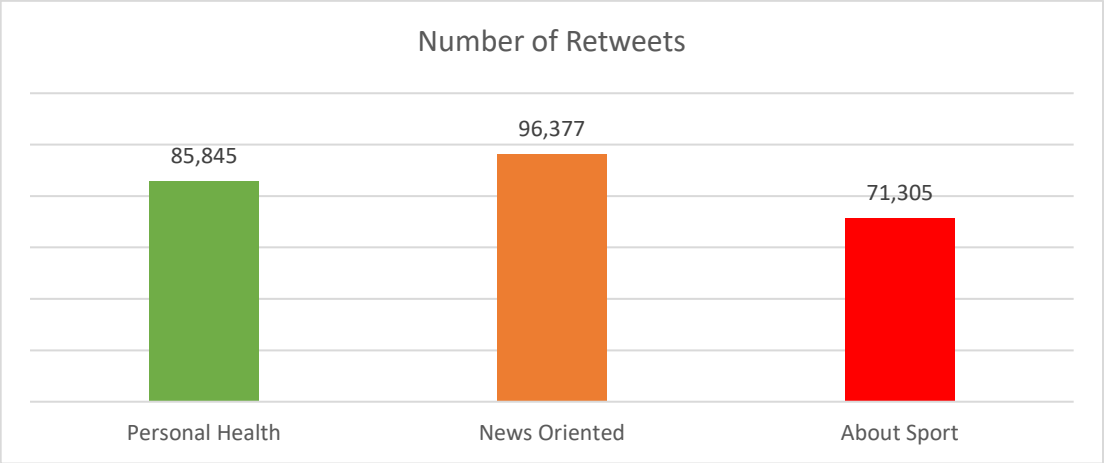
Εικόνα 11: Πλήθος φίλων σε κάθε κατηγορία

Σύνολο δημοσιεύσεων



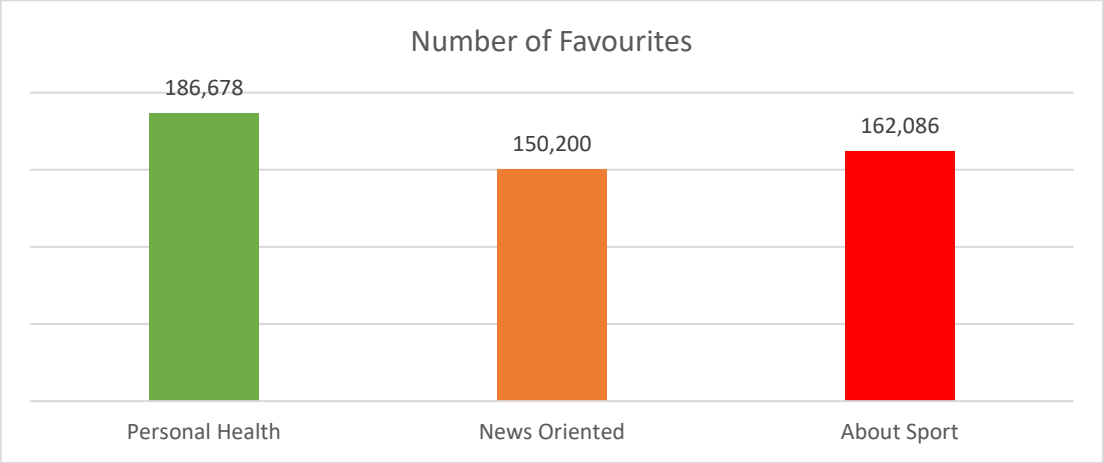
Εικόνα 12: Πλήθος δημοσιεύσεων σε κάθε κατηγορία

Πλήθος επαναδημοσιεύσεων



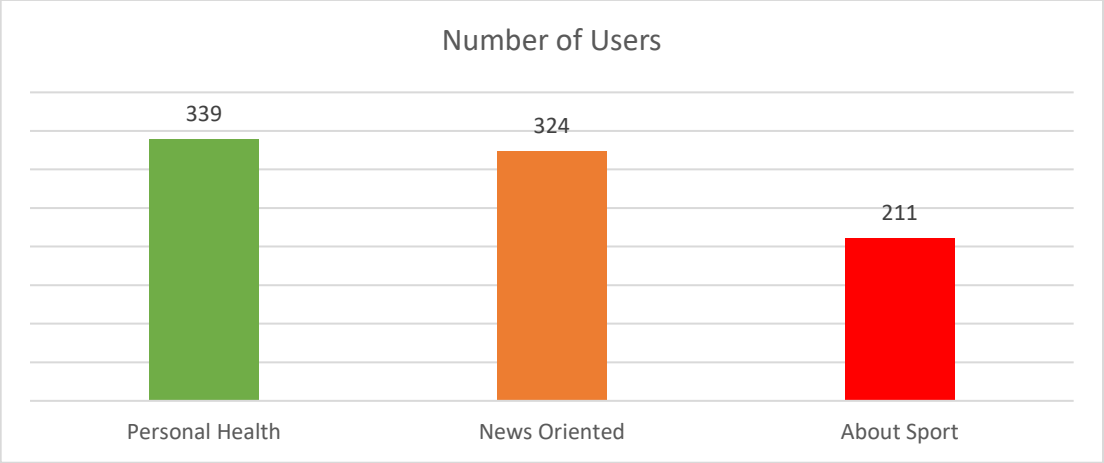
Εικόνα 13: Πλήθος επαναδημοσιεύσεων σε κάθε κατηγορία

Πλήθος των favourite



Εικόνα 14: Πλήθος favourite σε κάθε κατηγορία

Πλήθος χρηστών προς μελέτη σε κάθε κατηγορία



Εικόνα 15: Πλήθος χρηστών σε κάθε κατηγορία

Τα συγκεκριμένα γραφήματα τονίζουν την αναγκαιότητα ύπαρξης και άλλων μετρικών πέραν του πλήθους των ακόλουθων. Σύμφωνα λοιπόν με τα αποτελέσματα συλλογής των δεδομένων:

- Η κατηγορία των αθλητικών έχοντας τις λιγότερες δημοσιεύσεις (περίπου τις μισές σε σχέση με την κατηγορία προσωπική υγεία) και αρκετά λιγότερους ακόλουθους (το 70% περίπου) και ταυτόχρονα περίπου τους μισούς φίλους σε σχέση με την κατηγορία που μιλάει για προσωπική υγεία, καταφέρνει και αποσπά περίπου το ίδιο πλήθος favourite στο ίδιο χρονικό διάστημα. Αυτό

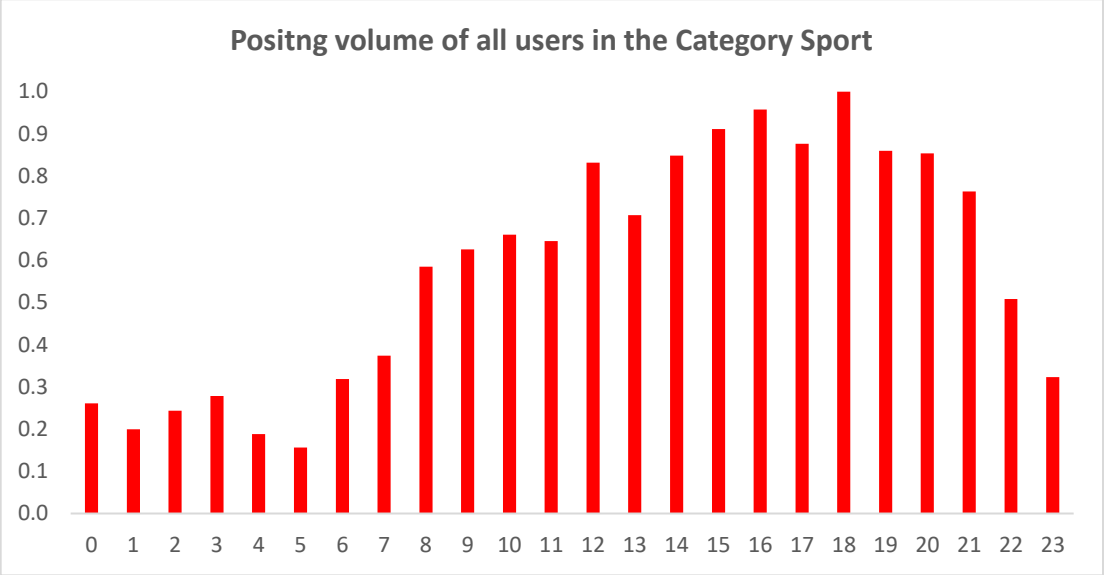
αναδεικνύει την δύναμη που έχουν οι συγκεκριμένοι χρήστες στο να διαδίδουν την πληροφορία και να επηρεάζουν το κοινό τους μέσω των favourite.

- Η κατηγορία των χρηστών που μιλάει γενικότερα για τα νέα έχοντας περίπου και αυτή τις μισές δημοσιεύσεις από την κατηγορία προσωπική υγεία και ταυτόχρονα τους λιγότερους ακόλουθους και φίλους καταφέρνει και αποσπάει τις περισσότερες επαναδημοσιεύσεις. Ιδιαίτερα σημαντικό είναι το γεγονός ότι, ενώ η συγκεκριμένη κατηγορία είναι πολύ καλή στο να κερδίζει τις επαναδημοσιεύσεις των χρηστών κερδίζει ταυτόχρονα και τα λιγότερα favourites σε σχέση με τις υπόλοιπες κατηγορίες.

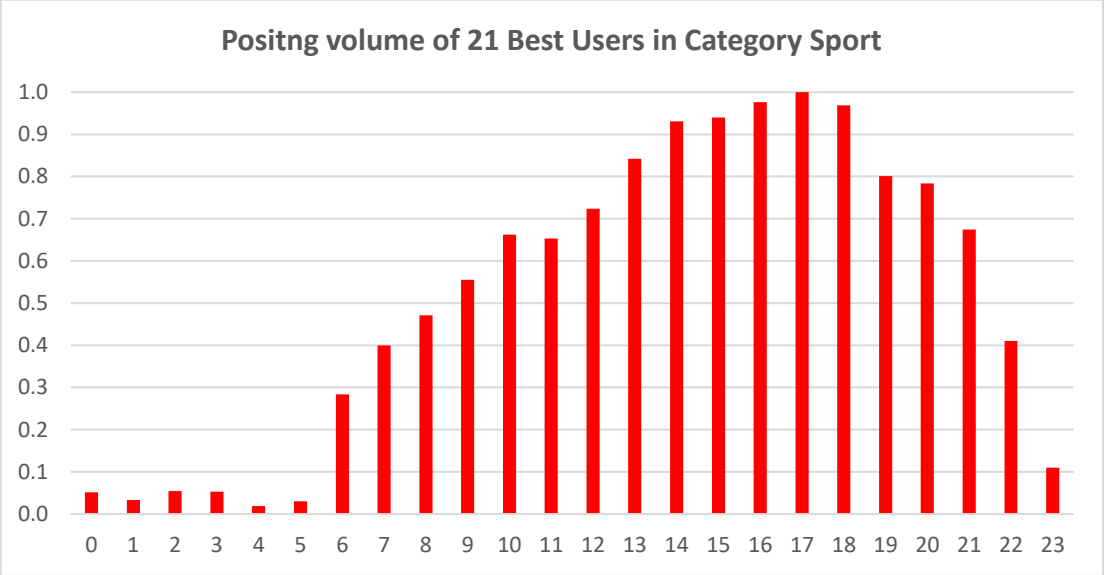
5.3 Σχολιασμός των αποτελεσμάτων της ώρας δημοσίευσης

Στην συνέχεια, παρατίθενται και σχολιάζονται τα αποτελέσματα που αφορούν το τι ώρα δημοσιεύουν οι χρήστες. Προκειμένου να δοθεί καλύτερη εικόνα δημιουργήθηκαν δύο ειδών γραφήματα ανά κατηγορία χρηστών. Στην πρώτη κατηγορία ανήκουν τα γραφήματα που περιέχουν πληροφορίες για τις ώρες που δημοσιεύουν οι 21 πρώτοι χρήστες βάσει του πλήθους των ακόλουθων και στην δεύτερη κατηγορία ανήκουν τα γραφήματα που περιέχουν πληροφορίες για τις ώρες που δημοσιεύουν όλοι οι χρήστες της κατηγορίας. Τα γραφήματα είναι όλα κανονικοποιημένα, δηλαδή αρχικά βρέθηκε το σύνολο των δημοσιεύσεων στην κάθε ώρα της ημέρας, στην συνέχεια βρέθηκε η μέγιστη τιμή και τέλος, κάθε τιμή της ημέρας διαιρέθηκε με την μέγιστη τιμή.

About Sport

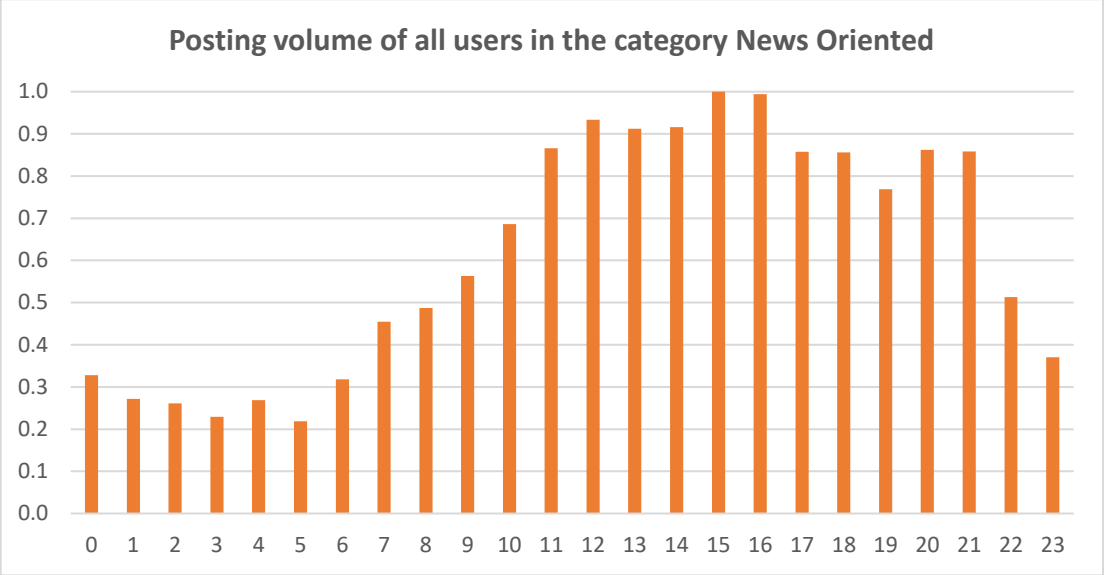


Εικόνα 16: Κανονικοποιημένο σχήμα αναπαράστασης ώρας δημοσίευσης όλων των χρηστών της κατηγορίας αθλητικά

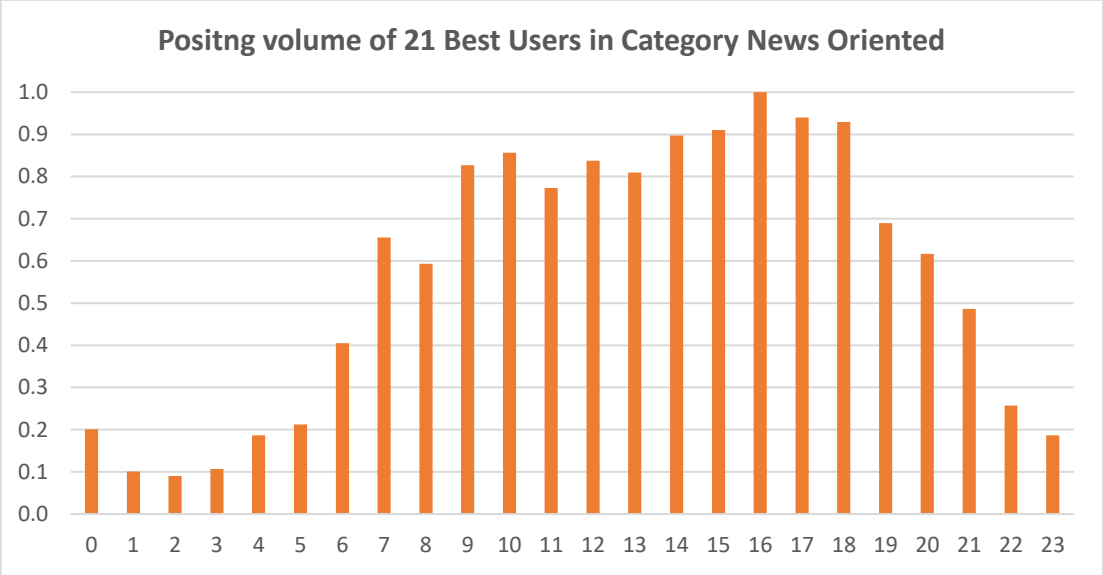


Εικόνα 17: Κανονικοποιημένο σχήμα αναπαράστασης ώρας δημοσίευσης των 21 χρηστών της κατηγορίας αθλητικά

News Oriented

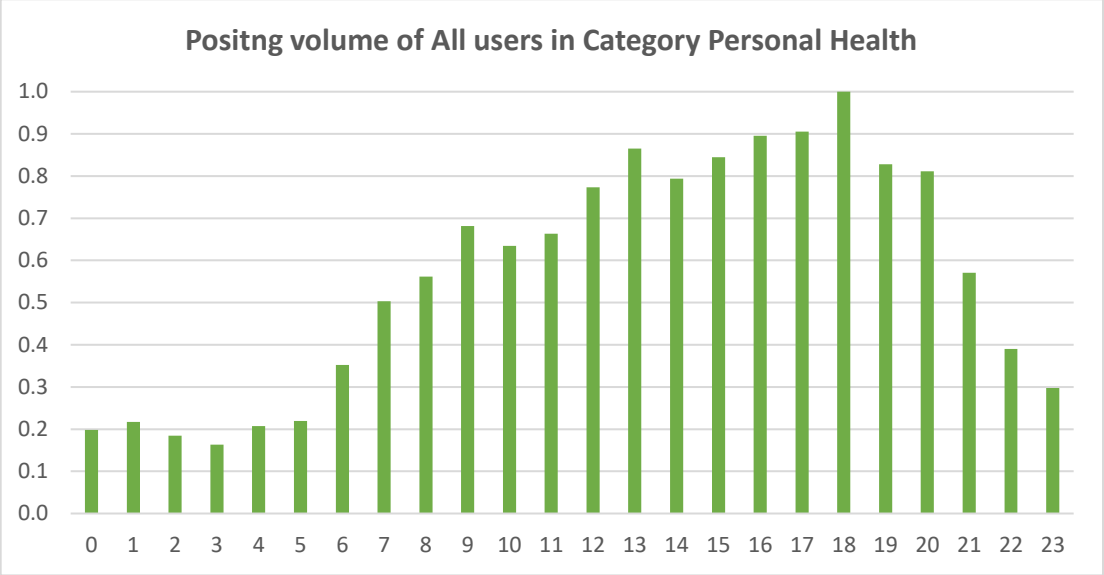


Εικόνα 18: Κανονικοποιημένο σχήμα αναπαράστασης ώρας δημοσίευσης όλων των χρηστών της κατηγορίας έκτακτα νέα

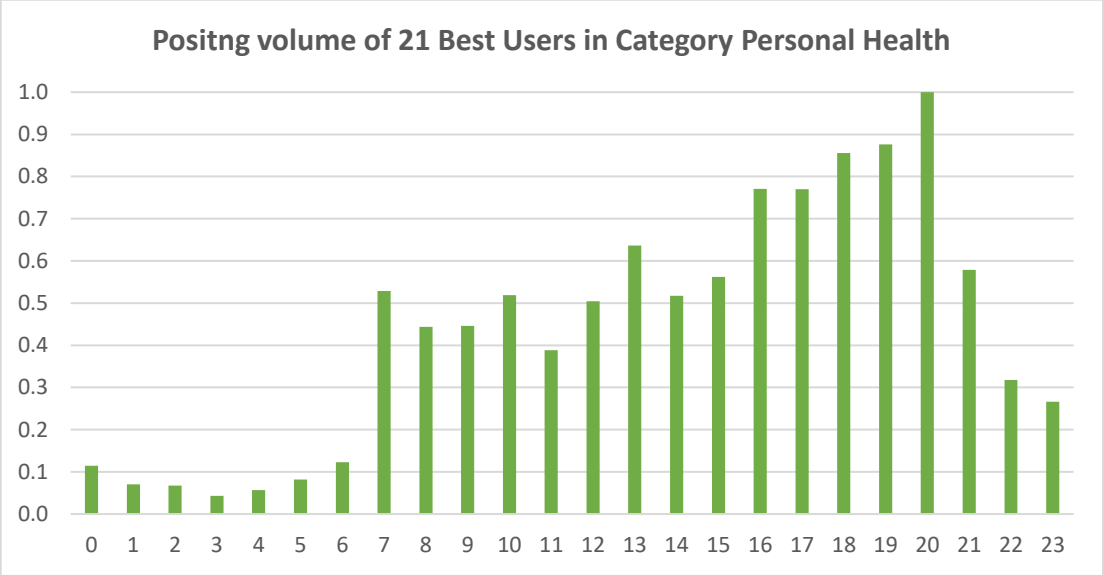


Εικόνα 19: Κανονικοποιημένο σχήμα αναπαράστασης ώρας δημοσίευσης των 21 χρηστών της κατηγορίας έκτακτα νέα

Personal Health



Εικόνα 20: Κανονικοποιημένο σχήμα αναπαράστασης ώρας δημοσίευσης όλων των χρηστών της κατηγορίας προσωπική υγεία



Εικόνα 21: Κανονικοποιημένο σχήμα αναπαράστασης ώρας δημοσίευσης των 21 χρηστών της κατηγορίας προσωπική υγεία

Συνολικά σχόλια

Όπως ακριβώς ήταν αναμενόμενο και στις 3 κατηγορίες οι 21 πρώτοι χρήστες βάση του πλήθους των ακόλουθων έχουν λιγότερες δημοσιεύσεις τις ώρες που δεν θεωρούνται ώρες αιχμής, ενώ στα συνολικά αποτελέσματα ο αριθμός των δημοσιεύσεων σε ώρες που δεν είναι ώρες αιχμής αυξάνεται. Επιπλέον, οι κατανομές των δημοσιεύσεων προσεγγίζουν λογικά τον πρότυπο πίνακα δημοσιεύσεων που παρατίθεται στο κεφάλαιο 4 παρουσιάζοντας όμως τις αναμενόμενες διαφορές. Για παράδειγμα στην κατηγορία της προσωπικής υγείας η μέγιστη τιμή λαμβάνεται στις 18:00, στην κατηγορία των γενικών νέων στις 15:00, στην κατηγορία για τα αθλητικά πάλι στις 18:00 ενώ σύμφωνα με το πρότυπο γράφημα η μέγιστη τιμή λαμβάνεται στις 10:00, ενώ στις 18:00 σύμφωνα με την πρότυπη τιμή έχουμε την 3η καλύτερη ώρα δημοσίευσης της ημέρας.

6 *Επίλογος*

6.1 Σύνοψη και συμπεράσματα

Σήμερα υπάρχουν δισεκατομμύρια χρήστες με λογαριασμούς στα μέσα κοινωνικής δικτύωσης ξεπερνώντας κατά πολύ τον αριθμό των διαθέσιμων ιστοσελίδων του διαδικτύου. Η κατάταξη των ιστοσελίδων βάση της σημαντικότητας των συνδέσμων που περιέχονταν μέσα σε κάθε ιστοσελίδα, της ποσότητας των κλικ που δεχόταν ένας σύνδεσμος, και της γενικότερης αίσθησης που άφηναν οι ιστοσελίδες στους χρήστες, οδήγησε σε πολύτιμες εφαρμογές όπως για παράδειγμα οι μηχανές αναζήτησης. Η εφαρμογή αποδοτικών αλγορίθμων για την ταξινόμηση των χρηστών που συμμετέχουν σε κάποιο μέσο κοινωνικής δικτύωσης έχει τις ίδιες προοπτικές επιτυχίας και με την σειρά της, μπορεί να οδηγήσει σε νέες χρήσιμες εφαρμογές.

Η συγκεκριμένη εργασία αποτελεί μια προσπάθεια δημιουργίας ενός αλγορίθμου που επιλύει τα προβλήματα προηγούμενων εργασιών στο χώρο. Συγκεκριμένα λαμβάνει υπόψη μακροχρόνια και βραχυχρόνια χαρακτηριστικά χρηστών, τα οποία συλλέχθηκαν για διάστημα 3 μηνών, αφαιρώντας αυτοματοποιημένους χρήστες και προσθέτοντας επιπλέον μια καινούργια μεταβλητή, την μεταβλητή που προσδιορίζει τις ώρες που δημοσιεύει ένας χρήστης, με την έννοια ότι οι χρήστες που καταφέρνουν να αποσπάσουν αρκετή προσοχή σε ώρες που δεν αποτελούν ώρες αιχμής οφείλεται να επιβραβεύονται παραπάνω. Η συγκεκριμένη διπλωματική, όσο μπορούμε να γνωρίζουμε, είναι η πρώτη εργασία που λαμβάνει υπόψιν αυτήν την μεταβλητή για σκοπούς μέτρησης της επίδρασης που ασκούν τα άτομα μεταξύ τους. Τέλος, για τον καλύτερο προσδιορισμό των ατόμων με επίδραση κατά το στάδιο της συλλογής των χρηστών έγινε η κατάταξη τους σε τρεις θεματικές κατηγορίες.

Τα αποτελέσματα της εργασίας είναι πολύ ικανοποιητικά καθώς και τα δύο στάδια της διαδικασίας υλοποίησης εξήγαγαν αποτελέσματα και συμπεράσματα που επιβεβαιώνονται λογικά και ταυτόχρονα επαληθεύονται από προηγούμενες εργασίες. Στο πρώτο στάδιο, στο στάδιο της εξαγωγής των δεδομένων, δείξαμε ότι μια θεματική

κατηγορία μπορεί να είναι καλή στο να αποσπά τις περισσότερες επαναδημοσιεύσεις και ταυτόχρονα να αποσπά τα λιγότερα favourite, καταλήγοντας στο συμπέρασμα ότι κάθε θεματική κατηγορία περιέχει χρήστες που αντιδρούν με διαφορετικό τρόπο με την πληροφορία. Επίσης δείξαμε ότι το πλήθος των ακόλουθων και το πλήθος των δημοσιεύσεων μιας θεματικής κατηγορίας είναι ανεξάρτητο από το πλήθος των favourite και το πλήθος των επαναδημοσιεύσεων που η συγκεκριμένη θεματική κατηγορία θα αποσπάσει, συμπεραίνοντας ότι η ποσότητα των δημοσιεύσεων δεν ερμηνεύεται σε καμία περίπτωση ως ποιότητα δημοσιεύσεων. Τέλος, δείξαμε ότι η κατανομή της ώρας των δημοσιεύσεων των ατόμων με επίδραση έρχεται σε συμφωνία με έρευνες που μελετούν πότε πρέπει κανείς να επαναδημοσιεύει, αποδεικνύοντας ταυτόχρονα ότι τα άτομα με επίδραση δημοσιεύουν κατά βάση σε ώρες αιχμής.

Στο δεύτερο στάδιο, το στάδιο συμπερασμάτων του αλγορίθμου, καταλήξαμε στο ότι το πλήθος των ακόλουθων δεν είναι επαρκής μετρική προκειμένου να θεωρηθεί μια κατάταξη πλήρης, αλλά θεωρήθηκε ότι αποτελεί απαραίτητη μετρική για την κατάταξη. Στην συνέχεια, δείξαμε ότι το πλήθος των ακόλουθων που έχει ένας χρήστης παρουσιάζει πολύ μεγάλη συσχέτιση με την ποιότητα των ακόλουθων που ο ίδιος έχει, επιβεβαιώνοντας την πρόταση ότι άτομα με επίδραση ακολουθούνται από άτομα με επίδραση. Επιπλέον, δείξαμε ότι το πλήθος των ακόλουθων δεν έχει πολύ ισχυρή συσχέτιση με τις μεταβλητές των favourite και επαναδημοσιεύσεων, ενώ οι μεταβλητές των favourite και επαναδημοσιεύσεων παρουσιάζουν πολύ μεγάλη συσχέτιση μεταξύ τους. Τα συγκεκριμένα αποτελέσματα οδήγησαν στην πρόταση ότι το πλήθος των ακόλουθων δεν είναι ικανό για να χαρακτηρίσει κάποιο άτομο ως άτομο με μεγάλη επιρροή και ταυτόχρονα ότι τα άτομα που γίνονται favourite είναι πολύ πιθανό να επαναδημοσιεύονται και αντίστροφα. Τέλος, η δειγματοληψία της εξαγωγής των χρηστών σε συγκεκριμένες θεματικές κατηγορίες είχε εύρος επιτυχίας από 38% έως 67% διάστημα ιδιαίτερα ικανοποιητικό αν ληφθεί υπόψιν ότι κάθε θεματική κατηγορία αντιπροσωπεύονταν μόνο από τέσσερις λέξεις κλειδιά.

6.2 Μελλοντικές επεκτάσεις

Η συγκεκριμένη διπλωματική αποτελεί μια πολύ συγκεκριμένη προσέγγιση προσαρμοσμένη στους περιορισμούς του Twitter. Συνεπώς, ένας ιδιαίτερα σημαντικός περιορισμός που επιβλήθηκε από το συγκεκριμένο κοινωνικό μέσο ήταν η έλλειψη της πληροφορίας που προσδιορίζει πόσο συχνά αναφέρεται ένας χρήστης. Προκειμένου να υπολογιστεί αυτός ο δείκτης ήταν απαραίτητη η μελέτη όλων των δημοσιεύσεων του Twitter σε μια συγκεκριμένη χρονική περίοδο, καθώς τα ερωτήματα μέσω Twitter API μπορούν να απαντήσουν μόνο στο ερώτημα που λέει ποιοι χρήστες αναφέρονται σε μια συγκεκριμένη δημοσίευση. Όπως είναι λογικό λοιπόν, στα πλαίσια της διπλωματικής εργασίας δεν ήταν δυνατόν να μελετηθούν όλες οι δημοσιεύσεις που συνέβησαν σε ένα κλειστό σύνολο του Twitter. Αυτό γιατί α. ο χρόνος που θα απαιτούνταν για την μελέτη όλων των δημοσιεύσεων θα ήταν τεράστιος, β. η διαθέσιμη υπολογιστική δύναμη θα ήταν ανεπαρκής και γ. το πλήθος των ερωτημάτων που μπορούσαν να γίνουν στο Twitter ήταν περιορισμένο. Μια βελτίωση που σκοπεύουμε να κάνουμε είναι να εξάγουμε τα tweets των χρηστών που ανήκουν στις τρεις θεματικές κατηγορίες για διάστημα 2 εβδομάδων, κατηγοριοποιώντας τα σε θεματικές ενότητες και στην συνέχεια καταγράφοντας την απόδοση αυτών των tweet ανά θεματική ενότητα. Με αυτόν τον τρόπο θα μπορέσουμε να απαντήσουμε στο ερώτημα που αναφέρει αν ένας Επηρεάζων μιας θεματικής κατηγορίας μπορεί να είναι και Επηρεάζων σε διαφορετική θεματική κατηγορία. Ιδιαίτερα σημαντικό είναι το γεγονός ότι στην συγκεκριμένη προσέγγιση δεν λαμβάνονται καθόλου υπόψιν η ισχυρότητα των δεσμών (tie strength) μεταξύ των φίλων. Αυτή η μεταβλητή θα μπορούσε να λαμβάνει υπόψιν πως και πότε οι χρήστες γνωρίστηκαν μεταξύ τους, πόσο συχνά εμφανίζονται σε κοινές φωτογραφίες, πόσο συχνή είναι η επικοινωνία στο μέσο κοινωνικής δικτύωσης, πόσο συχνά σχολιάζει ο ένας χρήστης έναν άλλον, αν τα άτομα μεγάλωσαν στην ίδια πόλη, αν πήγαν στο ίδιο πανεπιστήμιο και αρκετούς ακόμα παράγοντες. Τέλος, ένα προφανές μειονέκτημα είναι ότι ο διαχωρισμός των χρηστών σε κατηγορίες γίνεται σε ένα στιγμιότυπο του Twitter. Η ιδανική προσέγγιση θα ήταν η συλλογή όλων των βραχυχρόνιων χαρακτηριστικών όπως και η διαλογή των χρηστών να συμβαίνει συνεχόμενα, διατηρώντας ταυτόχρονα το διάστημα παρακολούθησης των τριών μηνών.

7 *Βιβλιογραφία*

[1] Anger, I., & Kittl, C. (2011, September). Measuring influence on Twitter. In Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (p. 31). ACM.

[2] Thenoisychannel. “A Twitter Analog to PageRank”, http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank

https://seo-hacker.com/measure-twitter-influence/

https://github.com/ealdent/tunkrank/blob/master/README.markdown, Available online (Last visited: 28/9/2017)

[3] Rao, A., Spasojevic, N., Li, Z., & DSouza, T. (2015, October). Klout score: Measuring influence across multiple social networks. In Big Data (Big Data), 2015 IEEE International Conference on (pp. 2282-2289). IEEE.

[4] Captaindatascience. “Top fashion influencers on Instagram and Twitter”, http://www.captaindatascience.com/top-fashion-influencers-instagram-twitter-an-algorithmic-approach-8, (Last visited: 23/6/2017)

[5] Booth, N., & Matic, J. A. (2011). Mapping and leveraging influencers in social media to shape corporate brand perceptions. Corporate Communications: An International Journal, 16(3), 184-191.

[6] Danisch, M., Dugué, N., & Perez, A. (2014, October). On the importance of considering social capitalism when measuring influence on Twitter. In BESC 2014-International Conference on Behavioral, Economic, and Socio-Cultural Computing (pp. 1-7). IEEE.

[7] Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011, February). Everyone's an influencer: quantifying influence on twitter.

In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 65-74). ACM.

[8] Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011, March). Influence and passivity in social media. In Proceedings of the 20th international conference companion on World wide web (pp. 113-114). ACM.

[9] Petychakis, M., Biliri, E., Arvanitakis, A., Michalitsi-Psarrou, A., Kokkinakos, P., Lampathaki, F., & Askounis, D. (2016, October). Detecting Influencing Behaviour for Product-Service Design Through Big Data Intelligence in Manufacturing. In Working Conference on Virtual Enterprises (pp. 361-369). Springer International Publishing.

[10] Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010, February). Twitterrank: finding topic-sensitive influential twitterers. In Proceedings of the third ACM international conference on Web search and data mining (pp. 261-270). ACM.

[11] Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012, April). The role of social networks in information diffusion. In Proceedings of the 21st international conference on World Wide Web (pp. 519-528). ACM.

[12] Aral, S., & Walker, D. (2014). Tie strength, embeddedness, and social influence: A large-scale networked experiment. Management Science, 60(6), 1352-1370.

[13] Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007). The dynamics of viral marketing. ACM Transactions on the Web (TWEB), 1(1), 5.

[14] Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., & Leskovec, J. (2015, August). Seismic: A self-exciting point process model for predicting tweet popularity. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1513-1522). ACM.

[15] Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. Icwsm, 10(10-17), 30.

[16] Entrepreneur. “The 80/20 Rule of Sales: How to find your best customers”, https://www.entrepreneur.com/article/229294, (Last visited: 28/9/2017)

[17] Cha, M., Benevenuto, F., Haddadi, H., & Gummadi, K. (2012). The world of connections and information flow in twitter. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 42(4), 991-998.

[18] Garcia, D., Mavrodiev, P., Casati, D., & Schweitzer, F. (2017). Understanding popularity, reputation, and social influence in the twitter society. Policy & Internet.

[19] Pfitzner, R., Garas, A., & Schweitzer, F. (2012). Emotional Divergence Influences Information Spreading in Twitter. ICWSM, 12, 2-5.

[20] Spasojevic, N., Li, Z., Rao, A., & Bhattacharyya, P. (2015, August). When-to-post on social networks. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2127-2136). ACM.

[21] Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. Journal of consumer research, 34(4), 441-458.

[22] BufferApp. “The Biggest Social Media Science Study: What 4.8 Million Tweets Say About the Best Time to Tweet” https://blog.bufferapp.com/best-time-to-tweet-research Available online (Last visited: 28/9/2017)

[23] Academics. “Βιβλίο μαθήματος (Εγχειρίδιο)”, http://academics.epu.ntua.gr/LinkClick.aspx?fileticket=EIlRdVCBgwE%3d&tabid=382&mid=2277, (Last visited: 28/9/2017)

[24] SmartInsights. “Search engine statistics 2017”, http://www.smartinsights.com/search-engine-marketing/search-engine-statistics/, (Last available: 28/9/2017)