

**Κατηγοριοποίηση δεδομένων
ειδησεογραφικού χαρακτήρα σε μέσα
κοινωνικής δικτύωσης με χρήση
αλγοριθμικών τεχνικών ανίχνευσης
δομικών ιδιοτήτων διασποράς, κριτηρίων
και μετρικών σημαντικότητας.**

Μπάρζας Ιωάννης

Διπλωματική Εργασία

Επιβλέπων: Σπυρίδων Κοντογιάννης



**ΤΜΗΜΑ ΜΗΧ. Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF IOANNINA**

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο του προπτυχιακού προγράμματος σπουδών του τμήματος Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής της Πολυτεχνικής Σχολής του Πανεπιστημίου Ιωαννίνων.

Ευχαριστώ τον επιβλέποντα καθηγητή μου, κύριο Σπύρο Κοντογιάννη, για τη καλή συνεργασία μας και τη συμβουλευτική καθοδήγηση.

Ιούλιος 2021, Ιωάννινα

Περίληψη

Τα τελευταία χρόνια η ανάπτυξη της τεχνολογίας υπήρξε ραγδαία ιδιαίτερα στον κλάδο των επικοινωνιών και του διαδικτύου. Αυτή η αλματώδης πρόοδος έχει επιφέρει πολλές αλλαγές στη ζωή μας καθώς οι ευκολίες που προσφέρει έχουν αρχίσει να ριζώνουν για τα καλά στη ζωή και την καθημερινότητα του σύγχρονου ανθρώπου και της κοινωνίας. Πλέον, χάρη στην εξάπλωση του διαδικτύου και των μέσων κοινωνικής δικτύωσης, υπάρχει δυνατότητα ταχύτατης διάδοσης τεράστιου όγκου πληροφοριών σε δισεκατομμύρια ανθρώπους δίχως γεωγραφικά όρια με μόλις λίγα κλικ. Παρόλο που με μια πρώτη ματιά αυτό φαίνεται αρκετά ωφέλιμο για μια πραγματική είδηση, στην πραγματικότητα οι συνέπειες αυτής της ανάπτυξης δεν είναι πάντα προς συμφέρον της κοινής γνώμης. Καθημερινά ένας μεγάλος αριθμός ψευδών ειδήσεων διαδίδεται μέσα από το διαδίκτυο και τα social media, με στόχο την παραπληροφόρηση, τη χειραγώγηση και τον αποπροσανατολισμό της κοινωνίας.

Σε αυτή την εργασία στόχος μας είναι η μελέτη της διάδοσης μιας αναδυόμενης ειδησεογραφικής ιστορίας και η καταγραφή των βασικών δομικών χαρακτηριστικών του υποκείμενου μοντέλου διάδοσης, ώστε να κατασκευάσουμε ένα εργαλείο με το οποίο θα είναι εφικτή η σήμανσή της ως εν δυνάμει πραγματικής ή ψευδούς είδησης σε πραγματικό χρόνο.

Λέξεις Κλειδιά: διάδοση πληροφορίας, ψευδείς ειδήσεις, κοινωνικά δίκτυα, μηχανική μάθηση, κατηγοριοποίηση

Abstract

In recent years the development of technology has been rapid especially in the field of communications and the Internet. This rapid progress has brought many changes in our lives, as the facilities it offers have prevailed in the life of modern man and society. Nowadays, thanks to the proliferation of the Internet and social media, it is possible to quickly spread huge amounts of information to billions of people across geographical boundaries with just a few clicks. Although at first glance this seems quite beneficial for a true news story, in reality the consequences of this development are not always in the public interest. Every day a large number of fake news is spread through the Internet and social media, with the aim of misinforming, manipulating, and disorienting the society.

In this work, our goal is to study the dissemination of information and audit in real-time the main characteristics of the structure in the underlying propagation network for the evolving story, in order to construct a tool with which it will be possible to label the evolving news stories as potentially true or fake, in real time.

Keywords: information propagation, fake news, social networks, early detection, machine learning, classification

Πίνακας περιεχομένων

Κεφάλαιο 1. Εισαγωγή	1
1.1 Αντικείμενο της διπλωματικής.....	1
1.2 Οργάνωση του τόμου	3
Κεφάλαιο 2. Ορισμός του προβλήματος	5
2.1 Στόχος της εργασίας.....	5
2.2 Σχετικές εργασίες και τεχνολογίες	6
2.3 Ανάλυση απαιτήσεων	7
Κεφάλαιο 3. Σχεδίαση & Υλοποίηση	9
3.1 Περιγραφή θέματος	10
3.2 Εξόρυξη και προεπεξεργασία των δεδομένων.....	16
3.2.1 <i>Twitter API</i>	16
3.2.2 <i>Τεχνικές εξόρυξης δεδομένων</i>	18
3.3 Διαθέσιμες βάσεις δεδομένων	20
3.3.1 <i>Twitter Dataset</i>	20
3.3.2 <i>FakeNewNet Dataset</i>	21
3.3.3 <i>Pheme -R</i>	21
3.4 Δομή Ιστοριών και Γραφημάτων διάδοσης της πληροφορίας.....	24
3.5 Προεπεξεργασία και φιλτράρισμα των δεδομένων	26
3.6 Ομαδοποίηση των δεδομένων.....	28
3.7 Αλγόριθμοι επίλυσης του προβλήματος.....	29
3.7.1 <i>CumulativeRank</i>	29
3.7.2 <i>Collective Influence</i>	37
3.7.3 <i>SimRank</i>	40
Κεφάλαιο 4. Πειραματική αξιολόγηση	43
4.1 Μεθοδολογία πειραματισμού	43
4.2 Αναλυτική παρουσίαση αποτελεσμάτων	52
4.2.1 <i>Σύγκριση ανά αλγόριθμο</i>	53
4.2.2 <i>Σύγκριση ανά κριτήριο</i>	61
4.3 Σύνοψη και συμπεράσματα.....	67

4.4	Μελλοντικές επεκτάσεις	70
-----	------------------------------	----

Κεφάλαιο 1. Εισαγωγή

1.1 Αντικείμενο της διπλωματικής

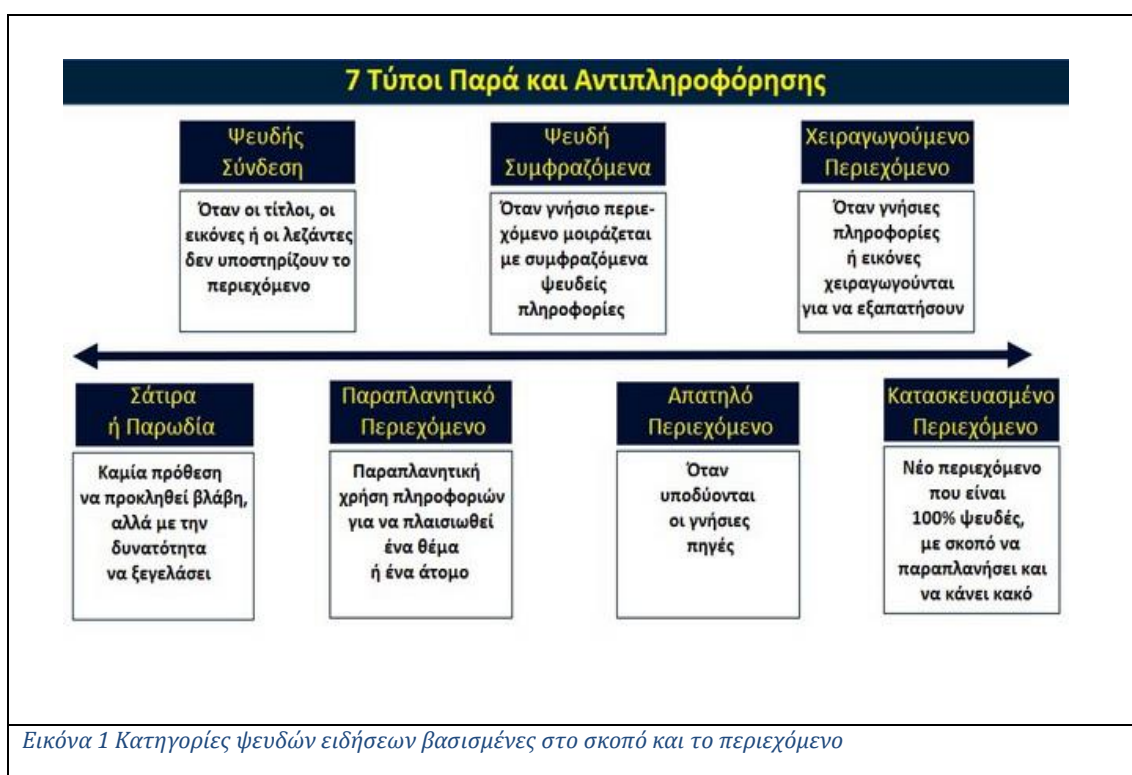
Το φαινόμενο παραπληροφόρησης, επίσης γνωστό ως «fake news», έχει λάβει νέες διαστάσεις στην ψηφιακή εποχή. Παρά το γεγονός ότι το φαινόμενο της διασποράς ψευδών ειδήσεων δεν είναι νέο, είναι πλέον γεγονός ότι τα κοινωνικά δίκτυα στο διαδίκτυο αποτελούν πρωτογενή πηγή πληροφόρησης για τις ειδήσεις αντικαθιστώντας τον κλασσικό τρόπο πληροφόρησης μέσω των ΜΜΕ. Αυτή η νέα πραγματικότητα οδηγεί στη αποδυνάμωση των παραδοσιακών μέτρων και πρακτικών της εγκυρότητας των ΜΜΕ. Πλέον ο κώδικας δεοντολογίας των δημοσιογράφων, η επαγγελματική προσέγγιση, η ανεξάρτητη αρχή ελέγχου, η ευθύνη του εκδότη, παύουν να υφίστανται, αφού η διασπορά των ειδήσεων λαμβάνει χώρα χρησιμοποιώντας πρακτικές που ξεφεύγουν της εφαρμογής των πρακτικών αυτών, καθώς με τη χρήση του διαδικτύου και των κοινωνικών δικτύων μπορούν πλέον όλοι να δημοσιεύουν ειδήσεις. Με άλλα λόγια, δεν χρησιμοποιείται πλέον κανένα φίλτρο αξιοπιστίας, αφήνοντάς τον κάθε χρήστη εκτεθειμένο στη χειραγωγή μέσω της διασποράς ψευδών ειδήσεων καθώς δεν είναι σε θέση να γνωρίζει αν μια είδηση είναι πραγματική ή ψευδής.

Έχουν διατυπωθεί διάφοροι ορισμοί για τις ψευδείς ειδήσεις. Οι ψευδείς ειδήσεις, είναι συχνά εντυπωσιακές, πληροφορίες που διαδίδονται με την μορφή ενός ειδησεογραφικού άρθρου. Παρακάτω παρατίθεται ο ορισμός (Karishma Sharma, 2018) που αποδίδει καλύτερα το νόημά τους στη δική μας προσέγγιση.

Fake News (Sharma et al 2019)

Ειδησεογραφικό άρθρο ή μήνυμα που δημοσιεύεται και μεταδίδεται σε ΜΜΕ ή/και κοινωνικά δίκτυα, το οποίο περιλαμβάνει εσφαλμένη πληροφορία, ανεξάρτητα από τους σκοπούς δημιουργίας του.

Παρακάτω παραθέτουμε έναν πίνακα που κατηγοριοποιεί τις ψευδείς ειδήσεις με βάση το περιεχόμενο και τον σκοπό τους. Η συγκεκριμένη κατηγοριοποίηση προτάθηκε από την Claire Wardle (Claire Wardle, 2017).



Λόγω της σημαντικότητας αλλά και των ζοφερών επιπτώσεων που μπορεί να προκαλέσει το φαινόμενο της διασποράς των ψευδών ειδήσεων, τα τελευταία χρόνια αποτελεί ένα από τα κυριότερα προβλήματα με έναν μεγάλο αριθμό από έρευνες και μελέτες που έχουν διεξαχθεί και έχουν ως κύριο στόχο να δώσουν λύση στο πρόβλημα διάδοσής τους σε πρώιμο στάδιο, έτσι ώστε να ελαχιστοποιηθεί ο αντίκτυπος της παραπληροφόρησης. Το πρόβλημα έχει προσεγγιστεί από πολλές πλευρές και οπτικές. Η λεκτική ανάλυση του συναισθήματος του κειμένου με χρήση Μηχανικής Μάθησης (Machine Learning), ανάλυση των δομικών χαρακτηριστικών του γραφήματος

διάδοσης της πληροφορίας , ανάλυση των χρονικών χαρακτηριστικών του γραφήματος διάδοσης της πληροφορίας είναι κάποιες από τις οπτικές προσεγγίσεις.

Τον Μάρτιο του 2018, οι Vosoughi et al (S. Vosoughi, 2018) μελέτησαν τη διασπορά των ειδήσεων στο Twitter και κατέληξαν ότι οι ψευδείς ειδήσεις έχουν την τάση να διαδίδονται γρηγορότερα, βαθύτερα και ευρύτερα μεταξύ των χρηστών. Συγκεκριμένα απέδειξαν, ότι οι ψευδείς ειδήσεις είναι 70% πιθανότερο να διαδοθούν και ταξιδεύουν 6 φορές πιο γρήγορα σε σχέση με τις αληθείς. Αυτό σημαίνει ότι η δομή του υποκείμενου δικτύου διασποράς των ειδήσεων διαφοροποιείται, ανάλογα με την εγκυρότητα της είδησης.

Γι' αυτό και εμείς με τη σειρά μας προσεγγίζουμε το πρόβλημα ανίχνευσης των δομικών ιδιοτήτων της διασποράς των ειδήσεων στα μέσα κοινωνικής δικτύωσης και πιο συγκεκριμένα στο Twitter, εστιάζοντας στο πλήθος των «σημαντικότερων» χρηστών στους οποίους οφείλεται η διασπορά σε όλο το δίκτυο. Προτείνουμε μία διαδικασία εξέτασης της σημαντικότητας των κόμβων που συμμετέχουν στο δίκτυο διάδοσης των ιστοριών, μέχρι να πετύχουμε όσο το δυνατόν πιο σαφή διάκριση μεταξύ των πραγματικών και των ψευδών ειδήσεων (στο σύνολο των ειδήσεων ή ανά κατηγορία). Κατασκευάζουμε επίσης ένα εργαλείο το οποίο μπορεί να αποτιμά διαρκώς την εγκυρότητα μιας αναδυόμενης είδησης , κατά τις πρώτες ώρες της διάδοσής της.

1.2 Οργάνωση της εργασίας

Η εργασία δομείται σε 5 κεφάλαια. Στο δεύτερο κεφάλαιο παραθέτουμε το θέμα της εργασίας και περιγράφουμε υπάρχουσες προσεγγίσεις και διαθέσιμες τεχνολογίες που αντιμετωπίζουν το πρόβλημα του εντοπισμού ψευδών ειδήσεων.

Στο τρίτο κεφάλαιο αναλύουμε τον τρόπο με τον οποίο προσεγγίσαμε το ζητούμενο. Περιγράφουμε αναλυτικά πώς συλλέξαμε αναδυόμενες ειδησεογραφικές ιστορίες από το Twitter, και πώς σχεδιάσαμε τη δική μας βάση δεδομένων για πειραματισμό. Περιγράφουμε επίσης μερικές από τις σχετικές βάσεις δεδομένων που διατίθενται στο Διαδίκτυο. Παρουσιάζουμε τέλος τις αλγοριθμικές τεχνικές σήμανσης αναδυόμενων ιστοριών, με τις οποίες επιχειρούμε να επιλύσουμε το πρόβλημα.

Στο τέταρτο κεφάλαιο παρουσιάζουμε την αξιολόγηση των προτεινόμενων αλγοριθμικών τεχνικών σήμανσης. Αναλύουμε τον σκοπό όλων των πειραμάτων και το πλαίσιο στο οποίο διεξήχθησαν τα πειράματα αυτά. Συγκρίνουμε τα αποτελέσματα μεταξύ τριών διαφορετικών αλγορίθμων σήμανσης και μετρικών ποιότητας που χρησιμοποιήσαμε και περιγράφουμε τα αποτελέσματα που προέκυψαν.

Στο πέμπτο κεφάλαιο συνοψίζουμε τη συνεισφορά της παρούσας διπλωματικής. Αναλύουμε τα συμπεράσματά μας και παραθέτουμε προτάσεις βελτίωσης της προσέγγισής μας.

Κεφάλαιο 2. Ορισμός του προβλήματος

2.1 Στόχος της εργασίας

Στόχος της εργασίας είναι η μελέτη της διασποράς των ειδήσεων σε μέσα κοινωνικής δικτύωσης καθώς και ο καθορισμός εκείνων των δομικών χαρακτηριστικών διάδοσης που καθορίζουν το είδος της είδησης.

Η ιδέα πίσω από τα κοινωνικά δίκτυα είναι ότι ένας χρήστης μπορεί να δημοσιεύσει ένα σύντομο μήνυμα δημοσίευση ή να αλληλεπιδράσει με μια δημοσίευση άλλου χρήστη γνωστοποιώντας τη στη δική του διαδικτυακή κοινότητα, συμβάλλοντας με αυτόν τον τρόπο στη διάδοση της. Ως διαδικτυακή κοινότητα ορίζεται μια ομάδα χρηστών που είτε είναι φίλοι με τον χρήστη είτε τον ακολουθούν ανάλογα με τις δυνατότητες που προσφέρει το κοινωνικό δίκτυο.

Μέσα στο κοινωνικό δίκτυο οποιοσδήποτε χρήστης μπορεί να δημοσιεύσει μια είδηση κάνοντας μια δημοσίευση, χωρίς να πραγματοποιείται κανένας έλεγχος εγκυρότητας και αξιοπιστίας. Σαν αποτέλεσμα αυτού, τα κοινωνικά δίκτυα αποτελούν ένα μέσο διάδοσης πολλών ψευδών ειδήσεων που ως στόχο έχουν τα χειραγωγήσουν και να αποπροσανατολίσουν στοχευμένα τους χρήστες. Ένα τέτοιο παράδειγμα πλατφόρμας κοινωνικής δικτύωσης (ανοιχτού κώδικα) αποτελεί το Twitter, το οποίο ενδείκνυται για την μελέτη του φαινομένου παρέχοντας το twitter api για να μπορούμε να αποκτήσουμε πρόσβαση στα δεδομένα και να τα μελετήσουμε βγάζοντας τα δικά μας αποτελέσματα και συμπεράσματα.

Στο πλαίσιο της παρούσας εργασίας, παρακολουθήσαμε και μελετήσαμε πώς διαδίδονται οι ειδήσεις στο Twitter. Αναπαραστήσαμε το μοντέλο διάδοσης σε κατευθυνόμενους γράφους, των οποίων οι κόμβοι αφορούν κάποια αντίδραση σε μια αναδυόμενη ιστορία. Χρησιμοποιήσαμε τρεις διαφορετικές τεχνικές για να ανιχνεύσουμε τις δομικές διαφορές μεταξύ των υπογραφημάτων του Twitter γραφήματος που ενεργοποιούνται για να διαδώσουν τις ιστορίες, ώστε να είμαστε σε

θέση να παρέχουμε μια έγκαιρη σήμανση των αναδυόμενων ιστοριών ως δυνητικά ψευδείς ειδήσεις ή ως δυνητικά αληθείς.

2.2 Σχετικές εργασίες και τεχνολογίες ανίχνευσης ψευδών ειδήσεων

Ο εντοπισμός ψευδών ειδήσεων είναι μια δύσκολη διαδικασία που δεν απαιτεί μόνο ο αναγνώστης/ακροατής να είναι ενημερωμένος σχετικά με τα θέματα επικαιρότητας, αλλά και να έχει ορθή κρίση και την ικανότητα να διασταυρώνει τις πηγές του και να βρίσκει αποδείξεις ή αντιφάσεις. Σε αυτή την κατεύθυνση δημοσιογράφοι και έγκυρες ιστοσελίδες (fact checking sites) αναλαμβάνουν να επικυρώσουν ή να αμφισβητήσουν όποια σημαντική νέα είδηση κυκλοφορεί, παραθέτοντας προηγούμενες επιβεβαιωμένες δημοσιεύσεις ως αποδεικτικά στοιχεία του πορίσματός τους. Κάποιες από αυτές είναι το Snopes, το οποίο διερευνά φήμες και αστικούς μύθους από το 1994, το Politifact, το οποίο επιπλέον συνεργάζεται με το Facebook, σε μία γενικευμένη προσπάθεια καταπολέμησης του φαινομένου στις πλατφόρμες κοινωνικής δικτύωσης, το factcheck.afp, το factcheck και το truthorfiction τα οποία επίσης κινούνται σε αυτή την κατεύθυνση ταυτοποίησης των ειδήσεων, τόσο σαν ένα άρθρο σε κάποιο vlog όσο και σαν δημοσιεύσεις στα κοινωνικά δίκτυα.

Όπως αναφέραμε, οι χρήστες – αναγνώστες αδυνατούν να ξεχωρίσουν μια ψευδή είδηση από μια πραγματική καθώς οι ψευδείς ειδήσεις εξελίσσονται διαρκώς σε επίπεδο κειμένου για να μιμούνται ως προς τα ποιοτικά χαρακτηριστικά τις πραγματικές ειδήσεις. Γι' αυτόν τον λόγο και πολλοί ερευνητές τα τελευταία χρόνια εκμεταλλεύονται την δομή του Twitter το οποίο παρέχει δυνατότητα ανάκτησης δεδομένων της αλληλεπίδρασης των χρηστών και της πορείας διάδοσης των ειδήσεων με σκοπό την προσέγγιση του προβλήματος από πολλές οπτικές.

Μία πρώτη προσέγγιση αναλύει το περιεχόμενο του κειμένου της είδησης. Οι Qazvinian et al (Vahed Qazvinian, 2011) εστίασαν σε λεκτικά μοτίβα και στη μέτρηση της συχνότητας διαφορετικών μερών του λόγου (POS tags) που διαφοροποιούν τις ψευδείς από τις πραγματικές ειδήσεις. Όσον αφορά στο περιεχόμενο του κειμένου, πολλές έρευνες έχουν εστιάσει στη συσχέτιση κειμένου και συνοδευτικού υλικού. Ο

Silverman (Craig Silverman, 2015) μελέτησε τη συσχέτιση τίτλου και κειμένου και παρατήρησε ότι σε ένα δείγμα 1600 ειδήσεων το 13% δεν ανταποκρινόταν στους τίτλους τους. Τεχνικές επίλυσης που εστιάζουν σε σαφώς καθορισμένα χαρακτηριστικά του λόγου, της γλώσσας και του συναισθήματος, υλοποιούνται με κατηγοριοποιητές όπως νευρωνικά δίκτυα, δέντρα απόφασης και Support Vector Machines (SVM). Αν και τέτοιες τεχνικές έχουν αποδειχθεί αξιόπιστες, το περιεχόμενο των κειμένων μπορεί εύκολα να αλλοιωθεί για να οδηγήσει σε λανθασμένα συμπεράσματα.

Οι Pomerleau και Rao το 2017 (Rao, 2017) οργάνωσαν το πρώτο Fake News Challenge: FCN-15, με σκοπό την ανάπτυξη εργαλείων τεχνητής νοημοσύνης (AI) που να κατηγοριοποιούν ειδησεογραφικά άρθρα σε 4 κλάσεις {agree, disagree, discuss, unrelated} βάσει της συσχέτισης τους με τους τίτλους τους. Στον διαγωνισμό συμμετείχαν 50 ομάδες εκ των οποίων νικήτρια αναδείχθηκε η SWEN με το εργαλείο SOLAT, επιτυγχάνοντας ποσοστό 82.02%. Ο κατηγοριοποιητής SOLAT συνδυάζει ένα MLP με 3 κρυφών επιπέδων και δέντρα απόφασης με ενίσχυση κλίσης (gradient-boosted decision trees). Δεύτερη ήρθε η Athene, η οποία παρουσίασε ένα εργαλείο 5 πολυεπίπεδων MLP, 6 κρυμμένων επιπέδων, με ποσοστό 81.97%, ενώ η UCLMR κατέκτησε την τρίτη θέση με ποσοστό 81.72% και ένα MLP ενός κρυφού επιπέδου.

Άλλες προσπάθειες εντοπισμού των ψευδών ειδήσεων επικεντρώνονται στο προφίλ των χρηστών που δημοσιεύουν τις ειδήσεις. Οι Castillo et al (Carlos Castillo, 2011) έφτιαξαν έναν κατηγοριοποιητή ειδήσεων σύμφωνα με τα κοινωνικά χαρακτηριστικά των χρηστών (πλήθος ακόλουθων, πλήθος χρηστών που ακολουθεί). Επιπλέον, έχει αποδειχθεί ότι αυτοματοποιημένοι λογαριασμοί (bots) επιταχύνουν τη διάδοση ψευδών ειδήσεων. Οι Shu et al (Kai Shu, 2017) κατέληξαν ότι για ένα δείγμα 10,000 χρηστών το 22% από τους χρήστες που διέδιδαν ψευδείς ειδήσεις ήταν bots, ενώ το αντίστοιχο ποσοστό εμπλεκόμενων bots σε πραγματικές ειδήσεις έφτανε το 9%.

Οι Zubiaga et al (Arkaitz Zubiaga M. L., 2016) απέδειξαν ότι οι χρήστες δυσκολεύονται να διακρίνουν τις ψευδείς από τις αληθείς ειδήσεις στο πρώιμο στάδιο διάδοσης. Οι Zhao et al (Zhe Zhao, 2015) επικεντρώθηκαν στη δυναμική εξέλιξη των tweets που σχετίζονται με κάθε είδηση σε πραγματικό χρόνο.

2.3 Ανάλυση απαιτήσεων

Η αλματώδης εξέλιξη της τεχνητής νοημοσύνης εισήγαγε νέες μεθόδους μηχανικής μάθησης (machine learning) και νευρωνικών δικτύων (neural networks), οι οποίες μπορούν να χρησιμοποιηθούν για την κατασκευή κατηγοριοποιητών για τις

ειδήσεις. Παρόλο που η μεγάλη αύξηση της υπολογιστικής ισχύος έθεσε νέα στάνταρ, τέτοιες διαδικασίες είναι αρκετά ακριβές τόσο σε χώρο αφού απαιτούνται μεγάλα σύνολα δεδομένων όσο και σε χρόνο αφού χρειάζεται χρόνο για την εκπαίδευση του δικτύου. Αυτό σημαίνει ότι τέτοιες μέθοδοι δεν ενδείκνυνται για κατηγοριοποίηση σε πρώιμο στάδιο. Για να μπορέσουμε να περιορίσουμε όσο είναι δυνατόν περισσότερο τις συνέπειες διάδοσης μια ψεύτικης είδησης, είναι πολύ σημαντικό να μπορεί να ανιχνευτεί όσο το δυνατόν πιο γρήγορα.

Σε αυτή την κατεύθυνση εκπόνησαν τη διπλωματική τους εργασία ο Αλέξης Πολύζος και Ιωάννης Πολύζος, με τίτλο «Measuring the Propagation Speed of Information Spread in Social Networks for Real-Time Fake News Detection», Πανεπιστήμιο Ιωαννίνων, 2019, (Alexios Polyzos, 2019) εστιάζοντας στην ταχύτητα διάδοσης των ειδήσεων και όχι στον τύπο ή το περιεχόμενο. Κατασκεύασαν ένα εργαλείο εξόρυξης ειδήσεων από το Twitter και συγκέντρωσαν ένα δείγμα 22 αληθών ειδήσεων και 21 ψευδών. Για τις αληθείς ειδήσεις απέδειξαν ότι το 87.71% των ακμών στα δίκτυα διάδοσης των ιστοριών συνδέουν απευθείας τον κόμβο – πηγή της ιστορίας με κάποιον παραλήπτη της, ενώ το αντίστοιχο ποσοστό για τις ψευδείς ήταν 61.36%. Χρησιμοποίησαν επίσης μετρικές ομοιότητας (SimRank) για να συγκρίνουν την ομοιότητα του κόμβου – πηγή με τους υπόλοιπους εμπλεκόμενους κόμβους και κατέληξαν στο συμπέρασμα ότι για τη διάδοση των πραγματικών ειδήσεων σε ένα δίκτυο ευθύνεται κατά κύριο λόγο ο χρήστης - πηγή, ενώ για τη διάδοση των ψευδών ειδήσεων υπεύθυνοι είναι ένα μεγαλύτερο υποσύνολο εμπλεκόμενων χρηστών. Σύμφωνα με τη θεωρία των γράφων, θα λέγαμε ότι το γράφημα διάδοσης μίας πραγματικής είδησης μοιάζει με αστέρι (star graph), ενώ το γράφημα διάδοσης μίας ψευδούς είδησης εμφανίζει στο εσωτερικό του μικρότερα υπό - δέντρα, τα οποία ξεκινάνε από τον κόμβο – πηγή.

Επίσης προς την ίδια κατεύθυνση εκπόνησε τη διπλωματική της εργασία η Κωνσταντίνα Μπουρογιάννη με τίτλο «Αλγοριθμικές τεχνικές ανίχνευσης δομικών ιδιοτήτων της διασποράς δεδομένων ειδησεογραφικού χαρακτήρα σε μέσα κοινωνικής δικτύωσης » Πανεπιστήμιο Ιωαννίνων, 2020, εστιάζοντας αποκλειστικά στο μοτίβο διάδοσης των ειδήσεων. Προς αυτή την κατεύθυνση, υλοποιήθηκε ο αλγόριθμος Collective Influence (Bourogianni, 2020) για την αξιολόγηση της σημαντικότητας των κόμβων που εμπλέκονται στο γράφημα διάδοσης κάθε ιστορίας. Το γενικότερο συμπέρασμα είναι ότι ο κόμβος – πηγή είναι ο σημαντικότερος κόμβος για τις περισσότερες ιστορίες. Για τις υπόλοιπες ο Collective Influence μπορεί να διακρίνει ότι οι ψευδείς ειδήσεις συγκεντρώνονται στις χαμηλότερες τιμές της 2ης -στιγμής στο κανονικοποιημένο διάνυσμα με τις σημαντικότητές κόμβων, ενώ οι πραγματικές

ειδήσεις στις υψηλότερες. Δηλαδή ότι οι πραγματικές ιστορίες τείνουν να έχουν μεγαλύτερη ανομοιομορφία στη διαμοίραση της σημαντικότητας γιατί μεγαλύτερο ποσοστό σημαντικότητας εμφανίζεται στον κόμβο πηγή ενώ στις ψευδείς υπάρχει μεγαλύτερη ομοιομορφία καθώς το ίδιο ποσοστό σημαντικότητας συνήθως διαμοιράζεται σε περισσότερους κόμβους του γραφήματος διάδοσης. Τα αποτελέσματα του Collective Influence έδειξαν ότι μπορεί να διακρίνει όλες τις πραγματικές ειδήσεις σε ένα δείγμα, αλλά με τον κίνδυνο να αξιολογήσει επιπλέον ψευδείς ειδήσεις ως πραγματικές αφού μπορεί να διακρίνει ένα μικρό ποσοστό των ψευδών ειδήσεων. Ο SimRank, από την άλλη, κατάφερε να διακρίνει τις ειδήσεις μόνο στο δεύτερο σύνολο δεδομένων και μόνο στις χαμηλότερες τιμές της 2ης – στιγμής, ενώ για τις υψηλότερες τιμές συσσωρευμένης σημαντικότητας έδειξε ότι το πλήθος σημαντικών κόμβων των πραγματικών ιστοριών ξεπερνά αυτό των ψευδών, χωρίς να πετύχει σαφή διάκριση μεταξύ τους με βάση τη 2η – στιγμή.

Η παρούσα εργασία συνεχίζει αυτή τη γραμμή έρευνας, στην κατεύθυνση των network – based προσεγγίσεων και εστιάζουμε αποκλειστικά στο μοτίβο διάδοσης των ειδήσεων. Συγκεκριμένα, για κάθε γκρουπ των ιστοριών που θεωρούμε και για κάθε συγκεκριμένη ιστορία γκρουπ, κατασκευάσαμε το δίκτυο αντιδράσεων των χρηστών στην συγκεκριμένη ιστορία. Στην συνέχεια επεκτείνουμε την παραπάνω έρευνα προτείνοντας επιπλέον μετρικές αξιολόγησης σημαντικότητας των εμπλεκόμενων κόμβων στο δίκτυο διάδοσης. Τέλος, αξιοποιώντας τα πειραματικά μας δεδομένα, προχωρήσαμε σε σύγκριση της προτεινόμενης μεθόδου αξιολόγησης της σημαντικότητας των κόμβων, με τις προηγούμενες.

Αξιοποιώντας ένα σημαντικό τμήμα τυχαία επιλεγμένων ιστοριών σε κάθε γκρουπ για μάθηση, προσδιορίσαμε τις τυπικές τιμές ως προς τη συσσωρευμένη σημαντικότητα των κόμβων για πραγματικές και ψευδείς ειδήσεις χωριστά, χρησιμοποιώντας τρεις διαφορετικές μετρικές αξιολόγησης της σημαντικότητας. Οι υπόλοιπες ιστορίες του συγκεκριμένου γκρουπ κατηγοριοποιήθηκαν με βάση τη δική τους τιμή συσσωρευμένης σημαντικότητας και τις κλειδωμένες πλέον τιμές συσσωρευμένης σημαντικότητας για πραγματικές και ψευδείς ειδήσεις.

Στόχος μας είναι να κατασκευάσουμε ένα εργαλείο για τη σήμανση των αναδυόμενων ειδήσεων, το οποίο να μπορεί να αξιολογήσει την εγκυρότητα της είδησης στις πρώτες ώρες της διάδοσης ως (δυνητικά) πραγματικές ή ψευδείς.

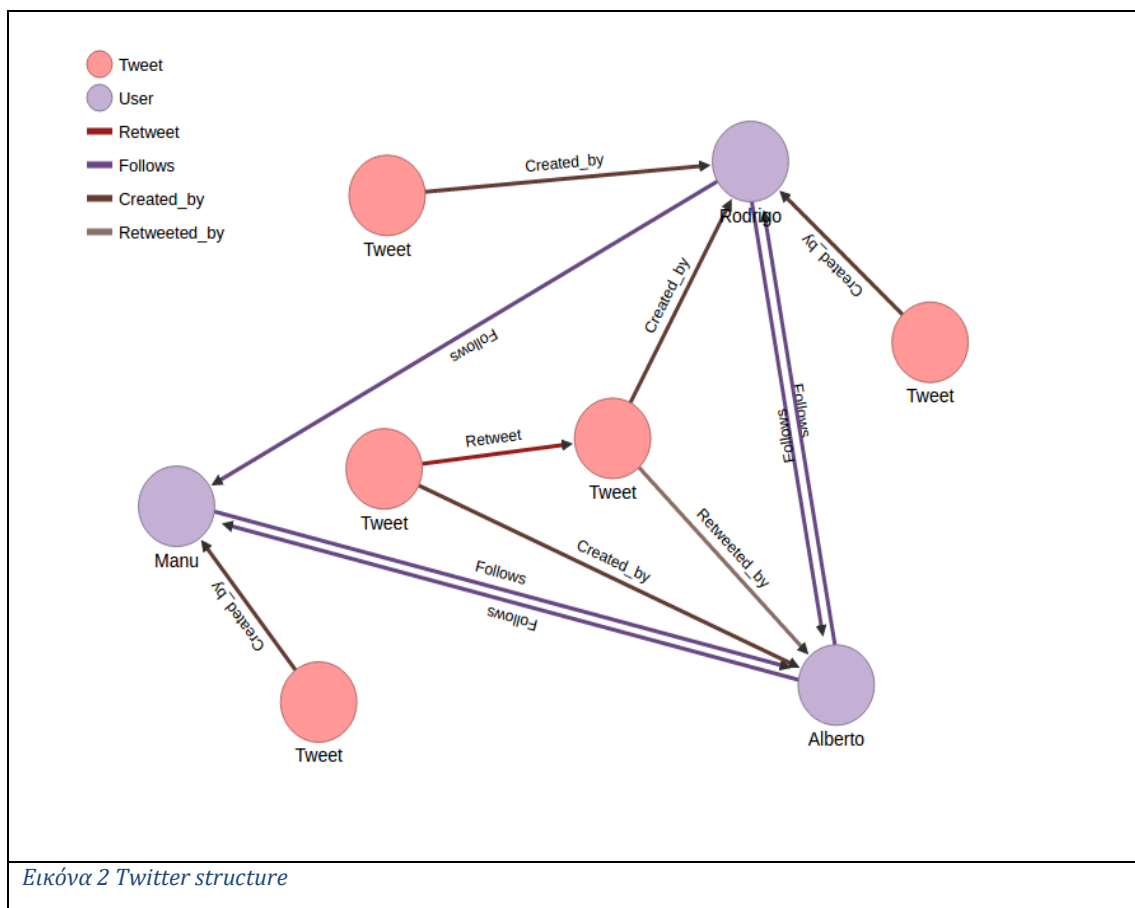
Κεφάλαιο 3. Σχεδίαση & Υλοποίηση

3.1 Περιγραφή Θέματος

Στην παρούσα διπλωματική μελετήθηκαν τα δομικά χαρακτηριστικά του γραφήματος διάδοσης της πληροφορίας στα κοινωνικά δίκτυα και πιο συγκεκριμένα στο Twitter. Η εργασία χωρίζεται σε τέσσερις φάσεις:

- Στη συλλογή και επεξεργασία του συνόλου δεδομένων από το twitter, ώστε να δημιουργηθούν τα γραφήματα διάδοσής των διαφορετικών ιστοριών. Επιπρόσθετα, οι ιστορίες χαρακτηρίστηκαν με βάση το περιεχόμενό τους σύμφωνα με το TF-IDF score, έτσι ώστε για κάθε ιστορία να είναι δυνατός ο προσδιορισμός μιας ομάδας περισσότερο σχετικών με αυτήν ιστοριών.
- Στην υλοποίηση του αλγορίθμου CumulativeRank (Dayong Zhang, 2019) για την αξιολόγηση της σημαντικότητας των κόμβων που εμπλέκονται στο δίκτυο διάδοσης μιας ιστορίας.
- Στην πειραματική εξέταση και σύγκριση των τριών αλγορίθμων CumulativeRank, Collective Influence και SimRank για τον προσδιορισμό των διανυσμάτων σημαντικότητας ανά ιστορία.
- Στην κατασκευή εργαλείου κατηγοριοποίησης των ιστοριών, το οποίο:
 - Μαθαίνει τις τυπικές τιμές συσσωρευμένης σημαντικότητας για ψευδείς ειδήσεις, αλλά και για πραγματικές ειδήσεις, τόσο στο σύνολο των ιστοριών που χρησιμοποιούνται για μάθηση, όσο και στα υποσύνολα ιστοριών μάθησης ανά γκρουπ σχετιζόμενων ιστοριών.
 - Φιλτράρει το σύνολο των ιστοριών που δε συμμετείχαν στη φάση της μάθησης, σύμφωνα με τις τυπικές τιμές συσσωρευμένης σημαντικότητας για πραγματικές και ψευδείς ειδήσεις.

Το Twitter αποτελεί ένα δημοφιλές κοινωνικό δίκτυο μικροϊστολογίων ανοιχτού λογισμικού (open source). Εσωτερικά του το μεγαλύτερο ποσοστό των προφίλ των χρηστών είναι δημόσια προφίλ, επιτρέποντας στις ειδήσεις – tweets να διαδοθούν ευκολότερα και ταχύτερα αφού είναι προσβάσιμα από όλους του χρήστες. Πρακτικά η διάδοση της πληροφορίας επιτυγχάνεται μέσω των αντιδράσεων των χρηστών με τις σχέσεις follows, retweeted_by, quoted_by, reply_to και favorite, στο αρχικό δημοσίευμα tweet, γνωστοποιώντας το στο προσωπικό του δίκτυο ακολούθων.



Με τον όρο tweet αναφερόμαστε στη δημοσίευση ενός χρήστη. Το tweet έχει πολλά χαρακτηριστικά. Τα πιο σημαντικά είναι τα id ,user.id που είναι το id του χρήστη που το έγραψε, created_at δηλώνει τον χρόνο που δημιουργήθηκε και το text του. Ολόκληρη η δομή του φαίνεται παρακάτω:

```
{
  "created_at": "Fri Dec 11 18:01:06 +0000 2020",
  "id": 1337457397002948608,
  "id_str": "1337457397002948608",
  "text": "sumpah bingung bgt ditawarin mama hp opp o ren o 4 tapi gua maunya samsung s20+ tp\u2026 https://t.co/qxfy96eRi9",
  "source": "<a href='\"http://twitter.com/download/android\"' rel='\"nofollow\"'>Twitter for Android</a>",
  "truncated": true,
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 960562577280413696,
    "id_str": "960562577280413696",
    "name": "cucubidah",
    "screen_name": "cucudidadah",
    "location": "cucucucudidadah",
    "url": "http://www.ca/pa.com",
    "description": "cubidacubida\n\n[fkik uin] [cyber acc]",
    "translator_type": "none",
    "protected": false,
    "verified": false,
    "followers_count": 123,
    "friends_count": 207,
    "listed_count": 0,
    "favourites_count": 1613,
    "statuses_count": 6058,
    "created_at": "Mon Feb 05 17:15:39 +0000 2018"
  },
  "favorited": false,
  "retweeted": false,
  "filter_level": "low",
  "lang": "in",
  "timestamp_ms": "1607709666859"
}
```

Εικόνα 3 Json μορφή ενός αρχικού tweet

Με τον όρο retweet αναφερόμαστε στην κοινοποίηση ενός tweet αυτούσιου κάποιου χρήστη από έναν δεύτερο χρήστη. Εκτός από τα id, user.id, created_at πρέπει να αναφερθεί και το πεδίο retweeted_status το οποίο δείχνει πάντα στο αρχικό tweet ακόμα και αν το retweet έγινε σε ενδιάμεσο retweet. Η δομή ενός retweet είναι η εξής:

```
{
  "created_at": "Wed Jul 07 16:39:35 +0000 2021",
  "id": 1412813559202009000,
  "id_str": "1412813559202009000",
  "text": "RT @nunidior: I\u0019m young and learning a lot still but one thing \u0026",
  "source": "\u003c\u003ca href=\u0022http://twitter.com/download/iphone\u0022 rel=\u0022nofollow\u0022\u003eTwitter for iPhone\u003c/a\u003e",
  "truncated": false,
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 1408908734483607560,
    "id_str": "1408908734483607560",
    "name": "\u00442\u003c3\u0043c\u0043c\u0043c\u003b9\u00454",
    "screen_name": "tamyjailashe2", "location": "618\u002708\u00fe0f615", "url": "https://linktr.ee/Tamyjailashe", "description": "TSU \u00d83c\u00df93 \u00d8",
    "followers_count": 135,
    "friends_count": 145,
    "listed_count": 0,
    "favourites_count": 116,
    "statuses_count": 304,
    "created_at": "Sat Jun 26 22:03:31 +0000 2021",
    //...
  },
  "geo": null,
  "coordinates": null,
  "place": null,
  "contributors": null,
  "retweeted_status": {
    "created_at": "Wed Jul 07 05:34:25 +0000 2021",
    "id": 1412646164051210241,
    "id_str": "1412646164051210241",
    "text": "I\u002019m young and learning a lot still but one thing I\u002019ve learned so far is to go with my gut\u00d83d\u00dcaf do what you want and fee",
    "source": "\u003ca href=\u0022http://twitter.com/download/iphone\u0022 rel=\u0022nofollow\u0022\u003eTwitter for iPhone\u003c/a\u003e",
    "truncated": true, "in_reply_to_status_id": null, "in_reply_to_status_id_str": null, "in_reply_to_user_id": null, "in_reply_to_user_id_str": null
  }
}
```

Εικόνα 4 Json μορφή ενός retweet

Με τον όρο quote αναφερόμαστε στην κοινοποίηση ενός tweet κάποιου χρήστη από έναν δεύτερο χρήστη, με τη δυνατότητα προσθήκης επιπλέον σχολιασμού. Και εδώ σημαντικά πεδία αποτελούν το id της αντίδρασης (quote) , user.id είναι το id του χρήστη που έκανε το quote , το created_at είναι η στιγμή δημιουργίας και το quoted_status_id δείχνει στο αρχικό tweet. Η δομή ενός quote είναι η εξής:

```
{
  "created_at": "Sun Dec 20 21:00:23 +0000 2020",
  "id": 1340764005338497024,
  "id_str": "1340764005338497024",
  "text": "And Remember the Biden Crime Family!\ud83d\ude0a\ud83c\uddfa\ud83c\uddfa8",
  "source": "<a href='\"http://twitter.com/download/iphone\"' rel='\"nofollow\"'>Twitter for iPhone</a>",
  "truncated": false,
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 28429313,
    "id_str": "28429313",
    "name": "rlm",
    "screen_name": "rlm6789",
    "location": "Oro Valley, AZ",
    "url": null,
    "description": "Business Consulting IPO's/ StartUps. Retired Business Exec. US Navy Beach Jumper Unit One Team 11 Vet.",
    "translator_type": "none",
    "protected": false,
    "verified": false,
    "followers_count": 1437,
    "friends_count": 4889,
    "listed_count": 72,
    "favourites_count": 35947,
    "statuses_count": 34877,
    "created_at": "Thu Apr 02 21:54:13 +0000 2009",
    "utc_offset": null,
    "time_zone": null,
    //...
  },
  "geo": null,
  "coordinates": null,
  "place": null,
  "contributors": null,
  "quoted_status_id": 1340660607142473737,
  "quoted_status_id_str": "1340660607142473737",
  "quoted_status": {
    "created_at": "Sun Dec 20 14:09:31 +0000 2020",
    "id": 1340660607142473737,
    "id_str": "1340660607142473737",
    "text": "REVEALED: 'Simple Math' Shows Biden Claims 13 MILLION More Votes Than There Were Eligible Voters Who Voted in 2020\ud2026 https://t.c",
    "source": "<a href='\"https://mobile.twitter.com\"' rel='\"nofollow\"'>Twitter Web App</a>",
    "truncated": true,
    "in_reply_to_status_id": null,
    "in_reply_to_status_id_str": null,
    "in_reply_to_user_id": null,
    "in_reply_to_user_id_str": null,
    "in_reply_to_screen_name": null,
    "user": {
      "id": 19211550,
      "id_str": "19211550",
      "name": "Jim Hoft",
      "screen_name": "gatewaypundit",
      "location": "St. Louis, Missouri",
      "url": "

```

Εικόνα 5 Json μορφή ενός quote

Με τον όρο reply αναφερόμαστε στην απάντηση σε ένα tweet. Και εδώ σημαντικά πεδία αποτελούν το id της αντίδρασης (reply) , user.id είναι το id του χρήστη που έκανε το reply, το created_at είναι η στιγμή δημιουργίας και το in_reply_to_status_id δείχνει στο αρχικό tweet. Η δομή ενός reply είναι η εξής:

Εικόνα 6 *Json* μορφή ενός *reply*

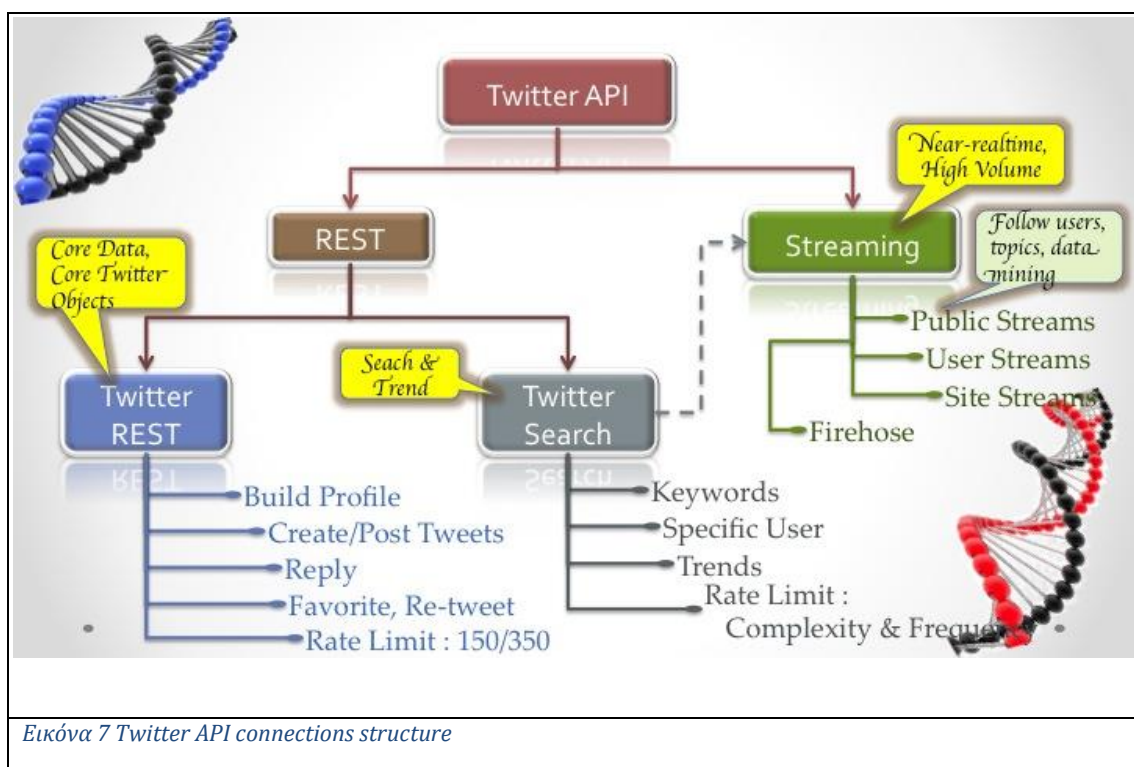
3.2 Εξόρυξη και προεπεξεργασία των δεδομένων

3.2.1 Twitter API

Όπως ήδη αναφέραμε το Twitter έχει δημιουργήσει τη δική του προγραμματιστική διεπαφή Twitter api προκειμένου να είναι εφικτή η συλλογή δεδομένων δραστηριότητας των χρηστών στο κοινωνικό δίκτυο Twitter και η περεταίρω ανάλυση τους με στόχο την εξαγωγή συμπερασμάτων και τη δημιουργία εργαλείων για την ανίχνευση των ψευδών ειδήσεων.

Η διεπαφή του Twitter χωρίζεται σε δύο μέρη, στο REST API και στο Streaming API. Η σύνδεση με το API είναι HTTP σύνδεση. Η REST (Representational State Transfer) API προσφέρει δεδομένα που απαντάνε σε συγκεκριμένα αιτήματα που κάνει ο χρήστης.

Ενώ αντίθετα, η Streaming API επιστρέφει μία ροή δεδομένων για κάθε νέα δραστηριότητα σε σχεδόν πραγματικό χρόνο μέχρι ο χρήστης να διακόψει τη σύνδεση.



Για να μπορέσουμε να συνδεθούμε στη διεπαφή του Twitter, απαιτείται λογαριασμός στο Twitter και στην συνέχεια δημιουργία Twitter Developer Account. Κάθε Developer Account δικαιούται έως 10 εφαρμογές (apps), καθεμία από τις οποίες παρέχει 5 κλειδιά (API consumer key, API consumer secret key, bearer token, access token, access token secret) και η σύνδεση γίνεται με την βιβλιοθήκη tweepy.

Η σύνδεση στη διεπαφή Streaming API φαίνεται παρακάτω:

```
streamListener=StreamListener()  
auth=OAuthHandler(credentials.CONSUMER_KEY,credentials.CONSUMER_SECRET)  
auth.set_access_token(credentials.ACCESS_TOKEN,credentials.ACCESS_TOKEN_SECRET)  
stream=Stream(auth,streamListener)  
stream.filter(track=searchTerms,stall warnings=True)
```

Εικόνα 8 Σύνδεση στη διεπαφή Streaming API του Twitter

Η σύνδεση στη διεπαφή Rest API φαίνεται παρακάτω:

```
### set the connection with twitter api  
  
auth=OAuthHandler(credentials.CONSUMER_KEY,credentials.CONSUMER_SECRET)  
auth.set_access_token(credentials.ACCESS_TOKEN,credentials.ACCESS_TOKEN_SECRET)  
api = tweepy.API(auth,wait_on_rate_limit=True)#,wait_on_rate_limit=True
```

Εικόνα 9 Σύνδεση στη διεπαφή Rest API του Twitter

Τώρα είμαστε έτοιμοι να συνδεθούμε είτε στη διεπαφή Streaming API για την πρόσβαση και συλλογή δεδομένων σε πραγματικό χρόνο ή στη διεπαφή Rest API για την πραγματοποίηση αιτημάτων όπως αιτήματα ανάκτησης ενός tweet ή ενός χρήστη με ένα συγκεκριμένο id , αιτήματα ενημέρωσης για τις σχέσεις μεταξύ των χρηστών (follow), κλπ.

3.2.2 Τεχνικές εξόρυξης δεδομένων

Για την εξόρυξη των δεδομένων υπάρχουν δύο κύριες στρατηγικές, η top – down και η bottom – up . Κατά την top – down, ψάχνουμε δεδομένα που σχετίζονται με ήδη επιβεβαιωμένες ειδήσεις. Αυτό μπορεί να γίνει είτε παρακολουθώντας έγκυρες πηγές, είτε παρακολουθώντας πηγές που αποδεδειγμένα αναγεννούν ψευδείς ειδήσεις. Σε αυτή την περίπτωση το φιλτράρισμα της βάσης είναι σχετικά εύκολο. Κατά την bottom – up εξόρυξη, κατά την οποία συλλέγουμε δεδομένα με συγκεκριμένα φίλτρα, αλλά συγκεντρώνουμε τυχαία μεγάλο όγκο δεδομένων, πχ ειδήσεις της επικαιρότητας που κάνουν τον κύκλο του Διαδικτύου.

Για τη συλλογή δεδομένων από το twitter API, χρειάζεται καθημερινή αναζήτηση στα fact checking sites πολλές φορές μέσα στην μέρα για τον εντοπισμό ψευδών ιστοριών που αφορούν κάποιο tweet. Ο κύριος παράγοντας που καθιστά πολύ απαιτητική και χρονοβόρα την διαδικασία συλλογής βάσης δεδομένων από το twitter API είναι η πολιτική του twitter API. Πιο συγκεκριμένα, για την standard (δωρεάν) έκδοση λογαριασμού, δεν επιτρέπει αναζήτηση πάνω από 200 αντιδράσεις πριν από το τρέχον (re)tweet. Συνεπώς, για να βρούμε μια ιστορία και να την καταγράψουμε από την αρχή της πρέπει να την προλάβουμε πριν αποκτήσει 200 αντιδράσεις ($|likes + retweets + quotes + replies| \leq 200$) πράγμα πολύ δύσκολο αν σκεφτούμε ότι στα fact checking sites ο έλεγχος γίνεται από ανθρώπινο παράγοντα και σχεδόν πάντα η διαδικασία διασταύρωσης για να χαρακτηριστεί είτε πραγματική είτε ψευδής γίνεται αρκετά αργά. Έτσι δεν επιμείναμε αρκετά έτσι ώστε να εμβαθύνουμε σε άλλες ενότητες όπως η κατασκευή σύνθετου εργαλείου κατηγοριοποίησης.

Παρόλα αυτά, συλλέξαμε κάποιες ιστορίες με την top – down τεχνική εξόρυξης. Οι ιστορίες δεν είναι ολόκληρες και η ένωσή τους έγινε κατ' ευθείαν με το αρχικό tweet. Κατά τη συλλογή τους χρησιμοποιήθηκαν φίλτρα και έτσι παρακολουθήσαμε είτε ιστορίες που έχουν κάποιες συγκεκριμένες λέξεις κλειδιά (keywords) ή συνδυασμό τους ή κάποια συγκεκριμένα id ιστοριών με στόχο να περιορίζουμε τον όγκο δεδομένων.

```

class StreamListener(StreamListener):
    def on_data(self, data):
        story=json.loads(data)
        for keyword in searchTerms:
            #quote
            if ('quoted_status_id' in story):# or story['id_str']==ID:
                if(str(story['quoted_status_id'])+'.xlsx' in onlyfiles or (str(story['quoted_status_id']) in ID)):
                    book = load_workbook('stories/'+str(story['quoted_status_id'])+'.xlsx')
                    writer = pd.ExcelWriter('stories/'+str(story['quoted_status_id'])+'.xlsx', engine='openpyxl')
                    writer.book = book
                    writer.sheets = dict((ws.title, ws) for ws in book.worksheets)
                    reader = pd.read_excel(r, 'stories/'+str(story['quoted_status_id'])+'.xlsx')
                    df = pd.DataFrame({"created_at": [story['created_at']], "id": [story['id_str']], "user_id": [story['user']['id_str']]})
                    df.to_excel(writer,index=False,header=False,startrow=len(reader)+1)
                    writer.close()

            #retweet
            elif 'retweeted_status' in story:
                if(str(story['retweeted_status']['id_str'])+'.xlsx' in onlyfiles or (str(story['retweeted_status']['id_str']) in ID)):
                    book = load_workbook('stories/'+str(story['retweeted_status']['id_str'])+'.xlsx')
                    writer = pd.ExcelWriter('stories/'+str(story['retweeted_status']['id_str'])+'.xlsx', engine='openpyxl')
                    writer.book = book
                    writer.sheets = dict((ws.title, ws) for ws in book.worksheets)
                    reader = pd.read_excel('stories/'+str(story['retweeted_status']['id_str'])+'.xlsx')
                    df = pd.DataFrame({"created_at": [story['created_at']], "id": [story['id_str']], "user_id": [story['user']['id_str']]})
                    df.to_excel(writer,index=False,header=False,startrow=len(reader)+1)
                    writer.close()

            #reply
            elif story['in_reply_to_status_id'] is not None:
                if(str(story['in_reply_to_status_id_str'])+'.xlsx' in onlyfiles or (str(story['in_reply_to_status_id_str']) in ID)):
                    book = load_workbook('stories/'+str(story['in_reply_to_status_id_str'])+'.xlsx')
                    writer = pd.ExcelWriter('stories/'+str(story['in_reply_to_status_id_str'])+'.xlsx', engine='openpyxl')
                    writer.book = book
                    writer.sheets = dict((ws.title, ws) for ws in book.worksheets)
                    reader = pd.read_excel('stories/'+str(story['in_reply_to_status_id_str'])+'.xlsx')
                    df = pd.DataFrame({"created_at": [story['created_at']], "id": [story['id_str']], "user_id": [story['user']['id_str']]})
                    df.to_excel(writer,index=False,header=False,startrow=len(reader)+1)
                    writer.close()

            #tweet
            else:
                if(story['id_str'] in ID or story['user']['screen_name'] in ReleableAuthors ):

                    df = pd.DataFrame({"created_at": story['created_at'], "id": story['id_str'], "user_id": story['user']['id_str'], index=[0]})
                    df.to_excel('stories/'+str(story['id'])+'.xlsx', index=False)
                    onlyfiles.append(str(story['id'])+'.xlsx')

```

Εικόνα 10 Κώδικας ανάκτησης ιστοριών που περιέχουν συγκεκριμένες λέξεις κλειδιά ή ιστοριών με συγκεκριμένα ID

3.3 Διαθέσιμες βάσεις δεδομένων

Από προηγούμενες έρευνες διατίθενται ελεύθερα οι βάσεις δεδομένων που συλλέχθηκαν.

3.3.1 Twitter Dataset

Οι Ma et al το 2017 δημοσίευσαν μία βάση δεδομένων (Ma et al, 2017) που συλλέχθηκε από το Twitter. Απαρτίζεται από την Twitter15 των Liu et al για ειδήσεις του 2015 και από την Twitter16 των Ma et al για ειδήσεις του 2016.

3.3.1.1 Twitter15

Η αναζήτηση και η συλλογή των επιβεβαιωμένων ειδήσεων για την βάση, έγινε από τις πηγές Snopes και Emergent. Η βάση περιέχει το γράφημα σε δομή δέντρου για κάθε tweet, ένα αρχείο label, το οποίο περιέχει την κατηγορία κάθε tweet και ένα αρχείο source_tweets, που περιέχει το περιεχόμενο κάθε tweet. Τα δεδομένα της κατηγοριοποιούνται σε 4 κλάσεις – true rumor, false rumor, non-rumor, unverified, όπως φαίνεται παρακάτω:

Twitter15	Statistics
True	372
False	370
Non-Rumor	374
Unverified	374
Source Tweets	1490

Πίνακας 1 Twitter15

Η τεχνική τους υλοποιείται με SVM κατηγοριοποιητές και μπορεί να επιτύχει 78% ακρίβεια ήδη στα 5 πρώτα tweets, 79% ακρίβεια στην πρώτη ώρα, ενώ μπορεί να αποφανθεί για ποσοστό 75% των ειδήσεων πριν από τα μέσα ενημέρωσης με ακρίβεια 80%.

3.3.1.2 Twitter16

Η Twitter16 συλλέχθηκε από τους Ma et al με τον ίδιο τρόπο που κατασκευάστηκε και η Twitter15. Η αναζήτηση επιβεβαιωμένων ειδήσεων έγινε από το Snopes. Τα δεδομένα της κατηγοριοποιούνται σε 4 κλάσεις – true rumor, false rumor, non-rumor, unverified, όπως φαίνεται παρακάτω.

Twitter16	Statistics
True	205
False	205
Non-Rumor	205
Unverified	203
Source Tweets	818

Πίνακας 2 Twitter16

Το εργαλείο που κατασκεύασαν εκπαιδεύει RNN μοντέλα αξιοποιώντας τη δυναμική εξέλιξη των δημοσιεύσεων για κάθε γεγονός και συγκλίνει γρήγορα υποδεικνύοντας έγκαιρη ανίχνευση δεδομένων πραγματικού χρόνου.

3.3.2 FakeNewNet Dataset

Το FakeNewsNet (Shu et al, 2018) δημοσιεύθηκε το 2018 από τους Shu et al. Συλλέχθηκε από δύο ιστότοπους ελέγχου γεγονότων: το GossipCop και το PolitiFact που περιέχουν περιεχόμενο ειδήσεων με ετικέτες σχολιασμένες από επαγγελματίες δημοσιογράφους και ειδικούς, καθώς και πληροφορίες για το κοινωνικό πλαίσιο. Λόγω της πολιτικής του Twitter δεν διατίθεται online η βάση δεδομένων, αλλά διατίθεται ο κώδικας για την ανάκτηση του.

3.3.3 Pheme -R

Το Pheme δημοσιεύθηκε το 2018 από τους Zubiaga et al και όπως και οι υπόλοιπες συλλέχθηκε από το Twitter. Χωρίζεται σε τρία σύνολα δεδομένων, το PHEME dataset of rumors and non-rumors, το PHEME dataset for Rumor Detection and Veracity

Classification που αποτελεί επέκταση του πρώτου και το PHEME Rumor scheme dataset: journalism use case.

3.3.3.1 PHEME dataset of rumors and non-rumors

Αυτό το σύνολο δεδομένων (Arkaitz Zubiaga Geraldine Wong Sak Hoi, 2016) δημοσιεύθηκε το 2016 και περιέχει μια συλλογή από φήμες και μη-φήμες στο Twitter που δημοσιεύτηκαν κατά τη διάρκεια έκτακτων ειδήσεων. Τα πέντε νέα στα οποία εστίασε το συγκεκριμένο σύνολο δεδομένων έχουν ως εξής:

PHEME	Rumour	None Rumour
Charlie Hebdo	458 (22.0%)	1,621 (78.0%)
Ferguson	284 (24.8%)	859 (75.2%)
Germanwings Crash	238 (50.7%)	231 (49.3%)
Ottawa Shooting	470 (52.8%)	420 (47.2%)
Sydney Siege	522 (42.8%)	699 (57.2%)

Πίνακας 3 PHEME

Τα δεδομένα είναι δομημένα ως εξής. Κάθε συμβάν έχει έναν φάκελο, με δύο υποφακέλους, φήμες και μη φήμες. Αυτοί οι δύο φάκελοι έχουν φακέλους με όνομα ID tweet. Το ίδιο το tweet μπορεί να βρεθεί στον κατάλογο «source-tweet» του εν λόγω tweet και ο κατάλογος «αντιδράσεις» έχει το σύνολο των tweets που ανταποκρίνονται σε αυτό το tweet πηγής.

3.3.3.2 PHEME dataset for Rumour Detection and Veracity Classification

Αποτελεί επέκταση του PHEME που δημοσιεύθηκε το 2016 (Elena Kochkina, 2018) και περιέχει φήμες που σχετίζονται με 9 συμβάντα και κάθε φήμη συνοδεύεται με μια ετικέτα σχετικά με την αληθοτιμή της, η οποία μπορεί να είναι, είτε True, είτε False είτε Unverified. Είναι στην ίδια δομή με το προηγούμενο σύνολο δεδομένων.

3.3.3.3 PHEME Rumour scheme dataset: journalism use case

Πρόκειται για ένα σύνολο δεδομένων που συλλέγεται και σχολιάζεται στην περίπτωση χρήσης δημοσιογραφίας του έργου PHEME FP7 (Arkaitz Zubiaga M. L., 2016). Αυτές οι φήμες σχετίζονται με 9 διαφορετικά νέα. Δημιουργήθηκε για την ανάλυση των φημών στα μέσα κοινωνικής δικτύωσης και περιέχει συνομιλίες στο Twitter που ξεκινούν από ένα tweet που ασχολείται με μια συγκεκριμένη φήμη. Οι συνομιλίες περιλαμβάνουν tweets που ανταποκρίνονται σε αυτά τα tweets. Αυτά τα tweets αξιολογήθηκαν ως προς το περιεχόμενό τους για το αν παρέχουν υποστήριξη, βεβαιότητα και αποδεικτικότητα ως προς τη συγκεκριμένη φήμη.

Το σύνολο δεδομένων περιέχει 330 συνομιλίες (297 στα Αγγλικά και 33 στα Γερμανικά), με έναν φάκελο για κάθε νήμα.

3.4 Δομή Ιστοριών και Γραφημάτων διάδοσης της πληροφορίας

Για τη διπλωματική μας είναι απαραίτητη η κατασκευή ενός γραφήματος που φανερώνει τη διάδοση της πληροφορίας στο twitter και αποτελεί την είσοδο των αλγορίθμων αξιολόγησης σημαντικότητας που εφαρμόζουμε. Πιο συγκεκριμένα το γράφημα διάδοσης μπορεί να κατασκευαστεί με διάφορα κριτήρια. Οι προηγούμενες διπλωματικές που προσέγγισαν το ζήτημα, χρησιμοποίησαν την δομή του έτοιμου γραφήματος που παρέχεται στις παραπάνω διαθέσιμες βάσεις και είναι σε δομή δέντρου. Εμείς, χρησιμοποιήσαμε γράφημα διάδοσης της πληροφορίας σε μορφή DAG, καθώς ο ίδιος χρήστης ενδέχεται να έλαβε γνώση μιας συγκεκριμένης αναδυόμενης ιστορίας από περισσότερους του ενός «φίλους», πριν τελικά αποφασίσει να αντιδράσει.

Πριν προχωρήσουμε στην κατασκευή ενός γραφήματος διάδοσης της πληροφορίας, είναι αναγκαίο να περιγράψουμε την έννοια της ιστορίας. Με τον όρο ιστορία εννοούμε ένα αρχείο με τρεις στήλες που περιέχει πρώτα το αρχικό tweet της είδησης και εν συνεχεία όλες τις αντιδράσεις οι οποίες μπορεί να είναι retweet, quote, reply ή like, και αφορούν το αρχικό tweet. Πιο συγκεκριμένα η δομή μιας ιστορίας σε κάθε γραμμή περιέχει το id της αντίδρασης (id) , το id του χρήστη (user_id) που έκανε την αντίδραση καθώς και τον χρόνο που έγινε η αντίδραση (created_at).

Έχοντας ορίσει την έννοια της ιστορίας, μπορούμε τώρα να περιγράψουμε την κατασκευή του υποκείμενου γραφήματος διάδοσης. Οι κόμβοι του γραφήματος αντιπροσωπεύουν τους χρήστες που αντιδράσανε στον αρχικό tweet, ή τον χρήστη του αρχικού tweet. Οι κατευθυνόμενες ακμές ανάμεσα σε δύο κόμβους αντιπροσωπεύουν τη ροή της πληροφορίας, δηλαδή μια ακμή από τον χρήστη1 στον χρήστη2 σημαίνει ότι:

1. ο χρήστης2 είδε την πληροφορία - ενημερώθηκε από τον χρήστη1 ο οποίος αντέδρασε (και άρα, είδε την ιστορία) πριν από αυτόν, και
2. ο χρήστης2 ακολουθεί τον χρήστη1.

Για την κατασκευή ενός γραφήματος διάδοσης της πληροφορίας σε μορφή δέντρου, αρχικά διατάσσουμε τις αντιδράσεις σε αύξουσα χρονολογική σειρά, αυτό

σημαίνει ότι όσο κατεβαίνουμε πιο κάτω στο αρχείο τόσο πιο πρόσφατες είναι οι αντιδράσεις. Στην συνέχεια, για κάθε αντίδραση (α1), βρίσκουμε την τελευταία αντίδραση (α2) που έγινε πριν από την (α1) και ο χρήστης της (α1) ακολουθεί τον χρήστη της αντίδρασης (α2). Στην περίπτωση που διατρέξουμε όλες τις γραμμές πάνω από την αντίδραση (α1) και δεν βρούμε αντίδραση (α2) που να πληροί τα κριτήρια, θεωρούμε ως (α2) το αρχικό tweet. Τέλος πριν προχωρήσουμε στην επόμενη αντίδραση, προσθέτουμε μια ακμή από την (α2) στην (α1). Η διαδικασία επαναλαμβάνεται για όλες τις αντιδράσεις μέχρι το τέλος της ιστορίας.

Η κατασκευή ενός γραφήματος διάδοσης της πληροφορίας σε μορφή DAG γίνεται πάλι μέσω της κατάλληλης επεξεργασίας της δομής μια ιστορίας. Αρχικά και πάλι διατάσσουμε τις αντιδράσεις σε αύξουσα χρονολογική σειρά, αυτό σημαίνει ότι όσο κατεβαίνουμε πιο κάτω στο αρχείο τόσο πιο πρόσφατες είναι οι αντιδράσεις. Στην συνέχεια, για κάθε αντίδραση (α1), βρίσκουμε όλες τις αντιδράσεις (α2) που έγιναν πριν από την (α1) και ο χρήστης της (α1) ακολουθεί τον χρήστη της αντίδρασης (α2). Στην περίπτωση που διατρέξουμε όλες τις γραμμές πάνω από την αντίδραση (α1) και δεν βρούμε καμία αντίδραση (α2) που να πληροί τα κριτήρια, θεωρούμε (α2) το αρχικό tweet. Τέλος, πριν προχωρήσουμε στην επόμενη αντίδραση, προσθέτουμε μια ακμή από την (α2) στην (α1) για όλες τις αντιδράσεις (α2) που πληρούσαν τα κριτήρια. Η διαδικασία επαναλαμβάνεται για όλες τις αντιδράσεις μέχρι το τέλος της ιστορίας. Η διαφορά με την δεντρική δομή του γράφου είναι ότι περιέχει πιο ολοκληρωμένη την πληροφορία της διάδοσης. Δηλαδή, αν ένας κόμβος έστω χρήστης1 έχει δύο παιδιά τον χρήστη2 και τον χρήστη3 και έστω ότι ο χρήστη2 αντέδρασε πριν από τον χρήστη3 και ο χρήστης3 ακολουθεί και τον χρήστη2, στη μορφή DAG θα προστεθεί ακμή από τον χρήστη2 στον χρήστη3 σχηματίζοντας τρίγωνο. Αυτή η ακμή εισάγεται επειδή δεν είναι ξεκάθαρο αν ο χρήστης3 ενημερώθηκε από τον χρήστη1 ή από τον χρήστη2 ή και από τον χρήστη1 και από τον χρήστη2.

Άρα είναι πλέον ξεκάθαρο ότι αυτή η δομή περιέχει όλα τα μονοπάτια εξάπλωσης της πληροφορίας του αρχικού tweet. Τέλος πρέπει να τονιστεί ότι λόγω της πολιτικής αιτημάτων του Twitter API με το ημερήσιο όριο αιτημάτων η διαδικασία εξόρυξης της πληροφορίας των σχέσεων ακολουθίας των χρηστών είναι πολύ αργή με αποτέλεσμα να καθυστερεί η διαδικασία δημιουργίας του γραφήματος σε μορφή DAG.

3.5 Προεπεξεργασία και φιλτράρισμα των δεδομένων

Για την παρούσα διπλωματική, χρησιμοποιήθηκε κυρίως η έτοιμη βάση δεδομένων twitter15. Η επιλογή αυτή έγινε διότι παρέχεται η σήμανση κάθε ιστορίας με βάση αν είναι πραγματική ή ψευδής είδηση, καθώς επίσης και το κείμενο της είδησης του αρχικού tweet. Επιπρόσθετα, δίνει την χρονική πληροφορία των αντιδράσεων έτσι ώστε να μπορούμε να επεξεργαστούμε τα δεδομένα και να τα μετατρέψουμε στην δομή DAG. Όπως αναφέραμε και πιο πάνω όταν περιγράψαμε την συγκεκριμένη βάση, η twitter15 δίνει το έτοιμο γράφημα διάδοσης σε δομή δέντρου. Επιπλέον πρέπει να αναφερθεί ότι στα αρχικά δεδομένα της βάσης υπάρχουν ασυνέπειες όπως διπλότυπα, αρνητικοί χρόνοι καθυστέρησης και άλλες ασάφειες.

Για την προετοιμασία των δεδομένων, το πρώτο βήμα είναι το φιλτράρισμα των δεδομένων χρησιμοποιώντας μόνο τα δεδομένα από τα δέντρα εκείνων των ιστοριών που έχουν τουλάχιστον 200 αντιδράσεις και έχουν χαρακτηριστεί «true» ή «false» απορρίπτοντας αυτές που χαρακτηρίστηκαν «unverified» ή «non-rumor» αφού δεν μπορούμε να εξάγουμε κάποια χρήσιμη πληροφορία για τον αντικειμενικό στόχο μας. Το φιλτράρισμα με βάση τον αριθμό αντιδράσεων έγινε διότι για μικρότερο μέγεθος ιστοριών δεν υπάρχει «πλούσια» πληροφορία για την εξαγωγή των συμπερασμάτων που μας ενδιαφέρουν. Άρα ξεχωρίζουμε τις ιστορίες εκείνες που το άθροισμα των $|\#retweets + \#quotes + \#replies + \#likes| \geq 200$ ή ισοδύναμα αφού κάθε κόμβος στο γράφημα είναι ένας χρήστης που αντέδρασε, χρησιμοποιούμε τα δέντρα ιστοριών με αριθμό διακεκριμένων κόμβων του δέντρου μεγαλύτερο ή ίσο του 200.

Στη συνέχεια, για τα δέντρα ιστοριών με τουλάχιστον 200 αντιδράσεις γίνεται η μετατροπή των δέντρων σε ιστορίες. Η δομή των δέντρων των ιστοριών είναι της μορφής:

- parent node -> child node με δομή κόμβου της μορφής:

['user id', 'tweet ID', 'post time delay (in minutes)']

Η λογική στη μετατροπή των δέντρων των ιστοριών στην αρχική ιστορία είναι η συλλογή όλων των κόμβων σε μια λίστα. Εν συνεχεία η απαλοιφή των διπλότυπων κάνοντας την λίστα αποθήκευσης σύνολο (set) και τέλος για κάθε κόμβο του συνόλου η εξαγωγή των τριών χαρακτηριστικών id χρήστη, id αντίδρασης, χρόνος δημιουργίας και η αποθήκευση τους σε αρχείο ταξινομημένο σε αύξουσα σειρά με βάση το ακέραιο που αντιπροσωπεύει τον χρόνο δημιουργίας. Πρέπει να τονιστεί σε αυτό το σημείο ότι δεν χρειαζόμαστε την αυτούσια ημερομηνία κάθε αντίδρασης αλλά την πληροφορία της ακριβούς σειράς που έγιναν οι αντιδράσεις.

Το τελευταίο βήμα για την προετοιμασία των δεδομένων είναι η μετατροπή των ιστοριών σε γραφήματα διάδοσης των ειδήσεων σε μορφή DAG. Η διαδικασία αυτή περιγράφηκε εκτενώς στην ενότητα 3.4 Εντούτοις υπάρχουν δυο ζητήματα που πρέπει να αναφέρουμε:

1. Κατά τη διάρκεια της διαδικασίας δημιουργίας του DAG πραγματοποιείται rest σύνδεση στο twitter api γιατί η διαδικασία απαιτεί την πληροφορία των σχέσεων των χρηστών, δηλαδή τις σχέσεις ακολουθίας τους. Το πρόβλημα σε αυτή τη διαδικασία είναι ότι οι χρήστες που εμπλέκονται ενεργά στην διάδοση ψευδών ειδήσεων συχνά διαγράφονται. Επίσης υπάρχουν χρήστες που έχουν διαγράψει το προφίλ τους ή δεν έχουν δημόσιο προφίλ. Όλες αυτές οι περιπτώσεις αποτελούν σημαντικό εμπόδιο στην συλλογή της πληροφορίας που χρειαζόμαστε για την κατασκευή του γραφήματος. Έτσι για να μπορέσουμε να ξεπεράσουμε αυτό το πρόβλημα όσο αυτό ήταν δυνατό, φορτώσαμε το γράφημα μορφής δέντρου που δίνεται έτοιμο για κάθε ιστορία και αντλήσαμε από τις ακμές του την πληροφορία ακολουθούμενου-ακολουθού, αφού για μια ακμή $X \rightarrow Y$ σημαίνει ότι ο Y ακολουθεί τον X . Αυτή η λύση δεν αντιμετωπίζει τελείως το πρόβλημα, αλλά πάντως το μετριάζει. Αυτό φαίνεται αν υποθέσουμε ότι για έναν κόμβο X που έχει έστω δύο παιδιά Y, Z σε μορφή δέντρου. Παρατηρούμε ότι δεν αντλούμε πληροφορία για το ζεύγος $Y-Z$ πάρα μόνο για τα ζεύγη $X-Y$ και $X-Z$.
2. Κατά την τελική αποθήκευση του γραφήματος, ελέγχουμε η κάθε ακμή να αποθηκεύεται μια φορά και ακμές της μορφής $X \rightarrow X$ να μην αποθηκεύονται γιατί δεν είναι έγκυρες. Ο λόγος ύπαρξης αυτών των περιπτώσεων είναι επειδή υπάρχουν εξ αρχής διπλότυπα στα δεδομένα της βάσης.

3.6 Ομαδοποίηση των δεδομένων

Για τον σκοπό της μελέτης των δομικών διαφορών της διασποράς των πραγματικών και των ψευδών ιστοριών, καθώς και της βέλτιστης κατηγοριοποίησής τους, ομαδοποιήσαμε τις ιστορίες που φιλτράραμε σε γκρουπ με βάση το περιεχόμενο τους. Για την ομαδοποίησή τους σε πρώτη φάση, περνούμε κάθε κείμενο των φιλτραρισμένων ιστοριών και πραγματοποιούμε προεπεξεργασία. Πιο συγκεκριμένα πραγματοποιούμε απαλοιφή από ειδικούς χαρακτήρες, κεφαλαία, συχνές άσκοπες λέξεις (stop words) και εφαρμογή lemmatizing για επαγωγή των λέξεων στις αρχικές του ρίζες με σκοπό τη χρήση κοινής λέξης για τις διαφορετικές μορφές μιας λέξης. Στην συνέχεια, πραγματοποιούμε ανάλυση TF-IDF των επεξεργασμένων κειμένων μέχρι φράσεις μεγέθους δύο και χρησιμοποιούμε cosine similarity στα διανύσματα των TF-IDF σκορ για να συγκρίνουμε τα κείμενα για όλα τα πιθανά ζεύγη ιστοριών. Τέλος, για κάθε ιστορία εξάγουμε τις δέκα πιο κοντινές ιστορίες που έχουν το ψηλότερο cosine similarity σκορ εξασφαλίζοντας ότι σε κάθε γκρουπ μεγέθους έντεκα ιστοριών ότι υπάρχουν τουλάχιστον τρεις ιστορίες από κάθε είδος, πραγματικές και ψευδείς, για να μπορούμε να εξετάσουμε τις διαφορές και να εξάγουμε την πληροφορία που θέλουμε.

Λόγω του τρόπου κατασκευής των γκρουπ υπάρχει επικάλυψη στα γκρουπ γιατί πολλές ιστορίες περιέχονται σε παραπάνω από ένα γκρουπ δημιουργώντας μοτίβα θεματολογίας που θα χρησιμοποιηθούν στο στάδιο του κατηγοριοποιητή. Σε κάποιες περιπτώσεις, για κάποια γκρουπ, αυτή η επικάλυψή φτάνει το 100%, δηλαδή υπάρχουν κάποια γκρουπ που περιλαμβάνουν ακριβώς τις ίδιες ιστορίες, κάτι το οποίο δεν είναι ωφέλιμο. Έτσι φιλτράραμε τα γκρουπ, εξασφαλίζοντας την μοναδικότητά τους ως προς το σύνολο των ιστοριών που περιέχουν.

Παρά τη λεπτομερή διαδικασία που ακολουθούμε κατά την κατασκευή των γκρουπ, υπολογίσαμε τον μέσο όρο ομοιότητας όλων των πιθανών ζευγών του κάθε γκρουπ σαν ένα μέτρο του πόσο συμπαγή είναι τα παραγόμενα γκρουπ. Στη πλειοψηφία των τιμών του μέσου όρου ομοιότητας για κάθε γκρουπ, συγκεντρώνονται στο διάστημα $[0.03, 0.25]$, κάτι που φανερώνει ότι τα γκρουπ δεν είναι αρκετά συμπαγή. Αυτό το αποτέλεσμα καθορίζεται από πολλούς παράγοντες.

Ένας παράγοντας είναι το γεγονός ότι για κάθε ιστορία δεν υπάρχουν πολλές αρκετά όμοιες ιστορίες μέσα στην βάση, και έτσι κατά την δημιουργία των γκρουπ μεγέθους έντεκα (μια ιστορία + τοπ δέκα όμοιες για αυτήν), εσωτερικά στα γκρουπ εισέρχονται και άσχετες ιστορίες για να συμπληρωθεί ο αριθμός. Επίσης, ένας άλλος παράγοντας είναι το γεγονός ότι εξαναγκάζουμε την ύπαρξη τουλάχιστον τριών ιστοριών από κάθε είδος, έτσι υπάρχουν περιπτώσεις που δεν εισάγονται τα κορυφαία

δέκα όμοια επειδή δεν έχουν τρία από κάθε είδους. Κατά συνέπεια, για να εξασφαλιστεί η συνθήκη αναγκαζόμαστε να εισάγουμε κάποιες (≤ 3) ιστορίες που είναι χαμηλότερα στην κατάταξη ομοιότητας.

Εν κατακλείδι, το πρόβλημα ομαδοποίησης των γκρουπ, αποτελεί ένα εντελώς ξεχωριστό και δύσκολο πρόβλημα το οποίο δεν είναι ο στόχος της παρούσας διπλωματικής.

3.7 Αλγόριθμοι επίλυσης του προβλήματος

Στην προσπάθεια αντιμετώπισης του προβλήματος της διάδοσης των ψευδών ειδήσεων στα μέσα κοινωνικής δικτύωσης, καθίσταται αναγκαία η ανίχνευση των δομικών χαρακτηριστικών της διασποράς των ειδήσεων. Εστιάζουμε στην εύρεση των πιο σημαντικών κόμβων του δικτύου, εκείνων δηλαδή που παίζουν τον σημαντικότερο ρόλο στη διάδοση της πληροφορίας. Στόχος μας είναι η εύρεση δομικών διαφορών ανάμεσα στις ψευδείς και στις πραγματικές ιστορίες, ούτως ώστε να υπάρξει ο διαχωρισμός τους και τελικά ή επιτυχή σήμανση των εν δυνάμει ψευδών ειδήσεων σε πραγματικό χρόνο. Υπάρχουν πολλοί αλγόριθμοι κατάταξης (ranking algorithms) που δίνουν λύση στο πρόβλημα εύρεσης των σημαντικότερων κόμβων σε ένα γράφημα, αλλά οι περισσότεροι αλγόριθμοι ποσοτικοποιούν κάποια συγκεκριμένα χαρακτηριστικά των κόμβων με σκοπό τη μέτρηση της σημαντικότητας τους. Εμείς εστιάζουμε σε τρεις διαφορετικούς αλγόριθμους, οι οποίοι εφαρμόζουν πρωτοποριακές ιδέες για τη μέτρηση της σημαντικότητας και τους παρουσιάζουμε παρακάτω.

3.7.1 CumulativeRank

3.7.1.1 Εισαγωγή CR

Στο πλαίσιο της εργασίας υλοποιήσαμε τον αλγόριθμο CumulativeRank (Dayong Zhang, 2019), που προτάθηκε από τους DayongZhang, YangWang & Zhaoxin Zhang το 2019 και αναλύει την σημαντικότητα των κόμβων τόσο σε τοπικό όσο και σε ευρύτερο επίπεδο μέσα στο δίκτυο. Ο αλγόριθμος δέχεται ως είσοδο το DAG γράφημα διάδοσης μιας ιστορίας και επιστρέφει το διάνυσμα σημαντικότητας των κόμβων του γραφήματος εισόδου.

Σε πρώτο βήμα, σε τοπικό επίπεδο ο αλγόριθμος βασίστηκε στη θεωρία των δομικών οπών του Burt (Burt, 1992), που παρέχει την προοπτική για την κατανόηση της τοπικής απόδοσης των κόμβων. Η θεωρία δομικών οπών του Burt υποστηρίζει ότι η δομική θέση μέσα σε ένα κοινωνικό δίκτυο είναι πιο σημαντική από την αντίστοιχη

εξωτερική σχέση ισχύος. Πιο συγκεκριμένα, μια δομική οπή είναι ένα κενό μεταξύ δύο μη συνδεδεμένων κόμβων. Όταν αυτοί οι δύο μη συνδεδεμένοι κόμβοι συνδέονται με έναν τρίτο κόμβο, ο κόμβος γεφύρωσης συνήθως έχει περισσότερα πλεονεκτήματα πληροφοριών και ελέγχου, δεδομένου ότι λειτουργεί ως μεσολαβητής μεταξύ διαφορετικών κόμβων ή ομάδων κόμβων. Προκειμένου να ποσοτικοποιηθεί το πλεονέκτημα ελέγχου των κόμβων γεφύρωσης, ο Burt εισήγαγε τον συντελεστή περιορισμό δικτυακών περιορισμών (NCC) που ορίζεται ως εξής:

$$NCC_i = \sum_{j \in \Gamma(i)} \left(p_{ij} + \sum_{k=1, k \neq i, j}^n P_{ik} P_{kj} \right)^2$$

όπου:

$$p_{ij} = \frac{z_{ij}}{\sum_{j \in \Gamma(i)} z_{ij}}$$

όπου:

- z_{ij} έχει τιμή 1 αν το μονοπάτι i-j υπάρχει διαφορετικά 0.
- $\Gamma(i)$ δηλώνει το σύνολο των κοντινότερων γειτόνων.
- Η ποσότητα του παρανομαστή μετρά την δύναμη έμμεσης σύνδεσης των i, j.

Παρ' όλα αυτά, πειραματική ερευνά έδειξε ότι η μετρική NCC είναι ελλιπής γιατί αξιοποιεί την πληροφορία συνδεσιμότητας που αφορά μόνο τους γείτονες πρώτου επιπέδου οδηγώντας σε μη ακριβή κατάταξη σημαντικότητας. Αυτό έχει ως αποτέλεσμα διαφορετικοί κόμβοι με παρόμοια χαρακτηριστικά και ίδιο αριθμό γεφύρωσης στη γειτονιά πρώτου επιπέδου αλλά διαφορετικά χαρακτηριστικά στη γειτονιά δευτέρου επιπέδου, πχ η πρόσβαση σε γείτονες δευτέρου επιπέδου που έχουν πολλές συνδέσεις, να μην αξιολογούνται και αρά να τους αποδίδονται ίδιες τιμές NCC.

Για να αντιμετωπιστεί η συγκεκριμένη παθογένεια, σε τοπικό επίπεδο ο αλγόριθμος λαμβάνει υπόψη του την άμεση επιρροή των κόμβων γειτονιάς πρώτου και δευτέρου επιπέδου ορίζοντας τον βελτιωμένο δείκτη σημαντικότητας των κόμβων **αναβαθμισμένου συντελεστή δικτυακού περιορισμού** (improved network constraint coefficient (INCC)).

Στο δεύτερο βήμα, προκειμένου να μελετηθεί ένα ευρύτερο επίπεδο επιρροής μέσα στο δίκτυο ο αλγόριθμος εισάγει την έννοια της **αντοχής** (tenacity), μιας μετρικής για να εκτιμηθεί η εξέχουσα θέση του κόμβου στη διατήρηση της συνδεσιμότητας του δικτύου. Δηλαδή ένας κόμβος αξιολογείται με βάση την επίδραση που έχει στο δίκτυο η αφαίρεσή του. Η συγκεκριμένη ιδέα είναι γνωστή ως βέλτιστη διήθηση (optimal percolation). Στη θεωρία των γράφων, ο όρος διήθηση περιγράφει τη συμπεριφορά του δικτύου καθώς αφαιρούνται τυχαία κόμβοι και ακμές. Η βέλτιστη διήθηση περιγράφει το πρόβλημα εύρεσης του ελάχιστου συνόλου κόμβων του δικτύου, των οποίων η αφαίρεση το κατακερματίζει σε μικρότερες συστάδες.

Στο τρίτο βήμα ο αλγόριθμος συνυπολογίζει τα δυο προηγούμενα σκορ του τοπικού και του ευρύτερου επιπέδου σε ένα ενιαίο σκορ, ώστε να προσδιορίσει τους κόμβους με την μεγαλύτερη επιρροή στο δίκτυο. Ο αλγόριθμος βασίζεται στην θεωρία δομικών οπών του Burt. Βασική φιλοσοφία του αλγορίθμου είναι η επίτευξη της χρυσής τομής αναμεσα ακρίβεια αποτελεσματικότητας και στην απλότητα της υλοποίησής του έτσι ώστε να είναι εφαρμόσιμος σε μεγάλης κλίμακας δίκτυα επιτυγχάνοντας σύντομο χρόνο εκτέλεσης με το καλύτερο δυνατό αποτέλεσμα.

3.7.1.2 Περιγραφή του αλγορίθμου CR

Όπως αναφέρουμε πιο πάνω, αλγόριθμος CumulativeRank αποτελείται από τρία βήματα. Στο πρώτο βήμα για τον υπολογισμό της σημαντικότητας των κόμβων βάσει των τοπικών δομικών χαρακτηριστικών, σε πρώτη φάση πρέπει να υπολογιστούν οι τιμές N_i , Q_i , LR_i για όλους τους κόμβους του δικτύου με την σειρά που αναγράφονται γιατί είναι απαραίτητες για τον υπολογισμό των τιμών INCC.

Συγκεκριμένα, για κάθε κορυφή v :

- Η τιμή N_i είναι το άθροισμα όλων των γειτόνων πρώτου επιπέδου και όλων των γειτόνων δευτέρου επιπέδου, για την κορυφή i .

- Η τιμή Q_i είναι το άθροισμα των τιμών N_j όλων των κορυφών u που είναι γείτονες πρώτου επιπέδου της i .
- Η τιμή LR_i (localRank) είναι το άθροισμα των τιμών Q_j των γειτόνων πρώτου επιπέδου της i .

Οι τύποι παρουσιάζονται παρακάτω:

$$Q_j = \sum_{w \in \Gamma(j)} N_w$$

$$LR_i = \sum_{j \in \Gamma(i)} Q_j$$

Στην συνέχεια αφού έχουν υπολογιστεί οι τιμές N_i, Q_i, LR_i , υπολογίζουμε τις τιμές INCC βάσει του τύπου παρακάτω και έτσι ολοκληρώνεται το πρώτο βήμα.

$$INCC_i = \sum_{j \in \Gamma(i)} \left(p'_{ij} + \sum_{k=1, k \neq i, j}^n p'_{ik} p'_{kj} \right)^2$$

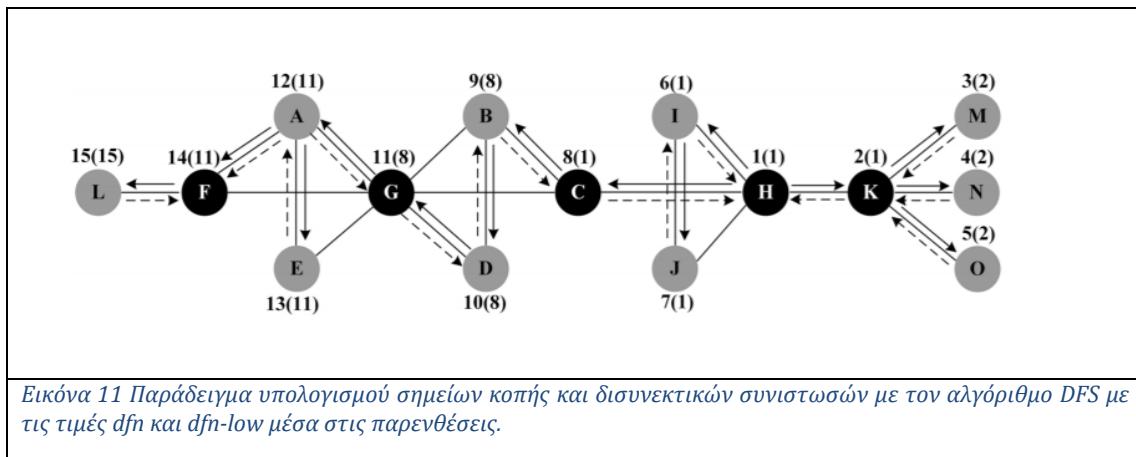
Όπου:

$$p'_{ij} = \frac{Q_j}{LR_i}$$

Στο δεύτερο βήμα για τον υπολογισμό της σημαντικότητας των κόμβων βάσει των ευρέων δομικών χαρακτηριστικών, σε πρώτη φάση και εφόσον έχουμε ορίσει το γράφημα της ιστορίας, είναι αναγκαίος ο υπολογισμός και καθορισμός των σημείων κοπής ή κομβικών σημείων (articulation points) και των δισυνεκτικών συνιστωσών (bi-connected components) του γραφήματος. **Σημεία κοπής** είναι κορυφές που η αφαίρεσή τους προκαλεί την αποσύνδεση του γραφήματος. Οι **δισυνεκτικές**

Για τον υπολογισμό τους εφαρμόζεται ο dfs αλγόριθμος. Πρακτικά ένας κόμβος είναι σημείο κοπής μόνο αν είναι ρίζα του dfs με τουλάχιστον δυο παιδιά ή δεν είναι ρίζα και έχει έναν απόγονο που η ελάχιστη τιμή του min-depth-first-number δεν είναι μικρότερη από την τιμή του κόμβου αυτού, depth-first-number. Ως depth-first-number ενός κόμβου u ορίζουμε τη σειρά με την οποία ο DFS τον επισκέφθηκε και η τιμή min-depth-first-number ή $low(u)$ ενός κόμβου u ορίζεται ως η ελάχιστη depth-first-number τιμή που μπορεί να φτάσουμε από τον κόμβο u ακολουθώντας ένα μονοπάτι απογόνων του ακολουθούμενο από το πολύ μια back-edge ακμή και ορίζεται από τον τύπο παρακάτω:

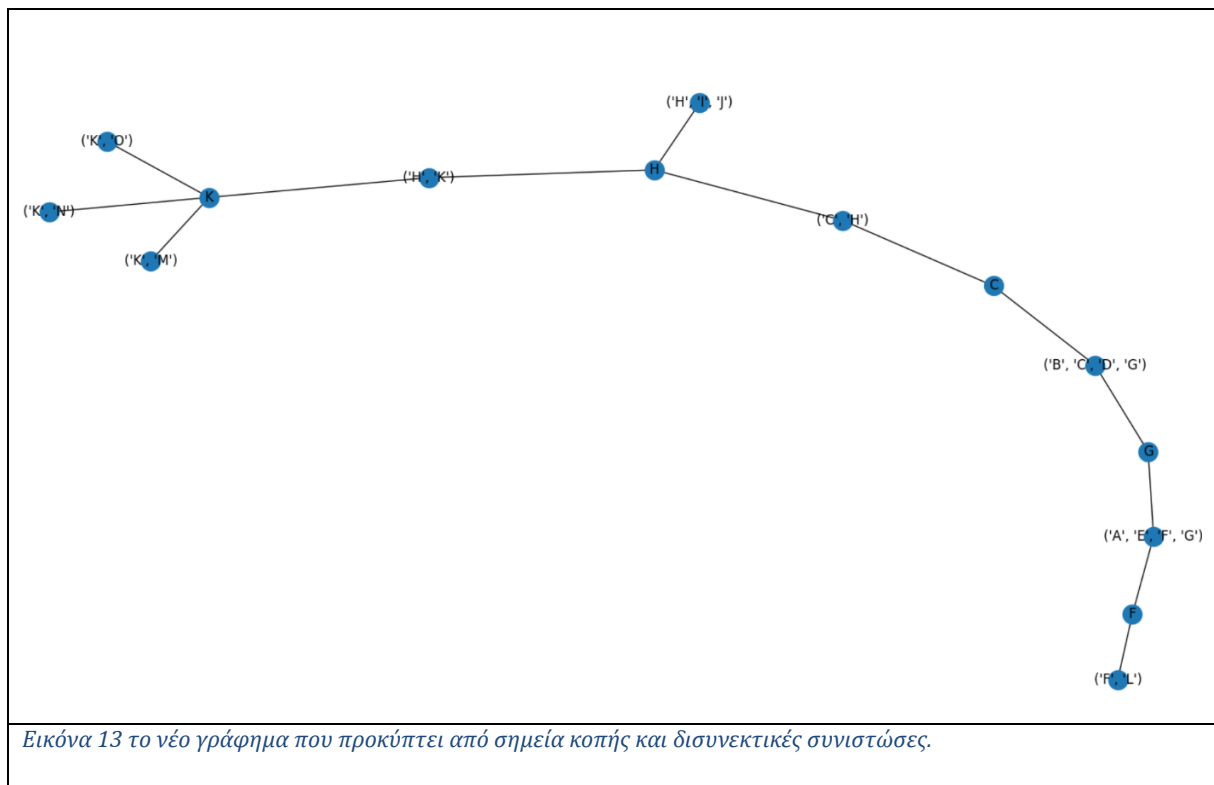
$$low(u) = \min\{dfn(u), \min\{low(w) \mid w \text{ is a child of } u\}, \\ \min\{dfn(w) \mid (u, w) \text{ is a back edge}\}\}$$



Εφόσον έχουν υπολογιστεί τα σημεία κοπής και οι δυσυνεκτικές συνιστώσες, δημιουργούμε ένα νέο γράφημα. Το νέο αυτό γράφημα έχει κόμβους που μπορεί να είναι είτε ένα σημείο κοπής ή να αντιστοιχούν σε μία δυσυνεκτική συνιστώσα. Οι δυσυνεκτικές συνιστώσες περιέχουν σημεία κοπής οπότε οι ακμές του νέου αυτού γραφήματος ορίζονται αναμεσά σε κόμβους που περιέχουν ένα σημείο κοπής τους

οποίους ενώνουμε με κόμβους που περιέχουν μια δισυνεκτική συνιστώσα που περιέχει το συγκεκριμένο σημείο κοπής, έτσι κάνουμε τη σύνδεση και το νέο γράφημα που προκύπτει είναι συνήθως μικρότερο στην περίπτωση που το αρχικό γράφημα είναι κατευθυντικό άκυκλο γράφημα (Directed Acyclic Graph -- DAG), αφού συρρικνώνουμε κάθε δισυνεκτική συνιστώσα του σε έναν κόμβο.

Παρακάτω φαίνεται το νέο γράφημα που προκύπτει από το παράδειγμα της Εικόνα 12. Οι κόμβοι που περιέχουν ένα μόνο κόμβο, όπως οι κόμβοι: K, H, C, G, F είναι κόμβοι που περιέχουν ένα σημείο κοπής (articulation point), ενώ οι υπόλοιποι κόμβοι περιέχουν μια δισυνεκτική συνιστώσα (Bi-component).



Στη συνέχεια, βρίσκουμε για κάθε κορυφή τις τιμές GiantSize και C στο νέο γράφημα. Η τιμή GiantSize για μια κορυφή u ορίζεται ως το μέγεθος της μεγαλύτερης συνιστώσας που προκύπτει στο γράφημα μετά την αφαίρεση της u . Η τιμή C ορίζεται ως το πλήθος των συνιστωσών που προκύπτουν μετά την αφαίρεση της u . Έτσι ο υπολογισμός αυτών των τιμών διαφοροποιείται ανάλογα με το αν η κορυφή είναι κομβικό σημείο ή όχι. Αν δεν είναι, η τιμή GiantSize είναι $|V|-1$ και η τιμή C είναι 1 καθώς

η αφαίρεση της u δεν προκαλεί αποσύνδεση του γραφήματος και αρά υπάρχει μόνο μια συνιστώσα με μέγεθος $|V|-1$ αφού αφαιρέθηκε η κορυφή u .

Στην περίπτωση που η κορυφή u είναι σημείο κοπής τα πράγματα αλλάζουν, αφού η αφαίρεσή της προκαλεί αποσύνδεση του γραφήματος. Έτσι για τον υπολογισμό των τιμών C , για την κορυφή u , η τιμή C θα είναι το πλήθος των κορυφών που αντιστοιχούν σε δισυνεκτικές συνιστώσες και που περιέχουν το κομβικό σημείο u , δηλαδή οι γείτονές της u στο νέο γράφημα. Για την τιμή $GiantSize$ της u , θα χρειαστεί να εφαρμόσουμε DFS σε κάθε υποδένδρο με ρίζα κάθε ένα από τους γείτονες της u και να μετρήσουμε τις τιμές όλων των κόμβων στο υποδένδρο συνολικά. Εφόσον το κάνουμε για όλα τα υποδένδρα επιλέγουμε την μεγαλύτερη τιμή, έτσι έχουμε βρει το μέγεθος της μεγαλύτερης συνιστώσας που προκύπτει μετά την αφαίρεση της u . Οι τιμές κάθε κορυφής ορίζονται ως εξής: Αν η κορυφή αναπαριστά ένα κομβικό σημείο τότε η τιμή της κορυφής ισούται με 1, διαφορετικά αν η κορυφή αναπαριστά δισυνεκτική συνιστώσα, τότε η τιμή της κορυφής ορίζεται ως το πλήθος των κορυφών στους οποίους αντιστοιχεί η κορυφή μείον το πλήθος των κομβικών σημείων που περιέχονται στη κορυφή δισυνεκτικής συνιστώσας.

Αφού υπολογίσαμε τις τιμές $GiantSize$ και C για όλες τις κορυφές μπορούμε να υπολογίσουμε τώρα τις τιμές R σύμφωνα με τον παρακάτω τύπο:

$$R_i = \frac{1 + m(G - i)}{w(G - i)}$$

όπου:

- $m(G-i)$ είναι η $GiantSize$ τιμή για την αφαίρεση της κορυφής i .
- $W(G-i)$ είναι η C τιμή για την αφαίρεση της κορυφής i .

Στο τρίτο βήμα και αφού έχουμε υπολογίσει τις τιμές $INCC$ από το πρώτο βήμα και τις τιμές R από το δεύτερο βήμα, χρειάζεται να κανονικοποιήσουμε τις τιμές R σύμφωνα με τον παρακάτω τύπο και μα βρούμε τις κανονικοποιημένες τιμές TC :

$$TC_i = \frac{R_i - \min_{j=1}^N R_j}{\max_{j=1}^N \{R_j\} - \min_{j=1}^N \{R_j\}}$$

Επιπλέον χρειάζεται να συνοψίσουμε τα δυο σκορ από το τοπικό και το ευρύ επίπεδο υπολογίζοντας τις τιμές CR όπως δηλώνει ο παρακάτω τύπος:

$$CR_i = \frac{INCC_i}{\sum_{j=1}^N INCC_j} + \frac{TC_i}{\sum_{j=1}^N TC_j}$$

Επιπρόσθετα, πρέπει να αναφέρουμε ότι στο επιστημονικό άρθρο του αλγορίθμου CumulativeRank ο τύπος του CR στους παρανομαστές εφαρμόζει ρίζα αλλά κατά την εκτέλεση διαπιστώσαμε ότι στο παράδειγμα υπολογισμού τους, χρησιμοποιούν τον τύπο που δίνουμε και εμείς, δηλαδή χωρίς ρίζες. Τέλος πρέπει να τονίσουμε ότι οι τιμές INCC, R, TC και αρά και οι τιμές CR είναι αντιστρόφως ανάλογες της σημαντικότητας των κορυφών. Αυτό σημαίνει ότι όσο μικρότερες οι τιμές τους, τόσο πιο σημαντική είναι η αντίστοιχη κορυφή. Έτσι για ευκολία υπολογισμού των στιγμών στην πειραματική αξιολόγηση μας, αντί απλά να κανονικοποιήσουμε το διάνυσμα με τις CR τιμές και να χρησιμοποιήσουμε τις κανονικοποιημένες τιμές, αντιστρέψαμε τις τιμές και μετά τις κανονικοποιήσαμε ούτως ώστε οι μεγαλύτερες τιμές να δηλώνουν μεγαλύτερη σημαντικότητα υπολογίζοντας τις τιμές *reversal_CR_normalized* όπως φαίνεται παρακάτω:

$$reversalCRnormalized_i = \frac{\frac{1}{CR_i}}{\sum_{j=1}^N \frac{1}{CR_j}}$$

3.7.2 Collective Influence

3.7.2.1 Εισαγωγή CI

Ο βελτιωμένος αλγόριθμος Collective Influence (Flaviano Morone, 2016) προτάθηκε από τους Morone και Makse το 2015. Αναλύει και αυτός τη μεγιστοποίηση της επιρροής των κόμβων σε σύνθετα δίκτυα, αλλά μόνο μέσω της βέλτιστης διήθησης (optimal percolation). Τα συμπεράσματα στα οποία κατέληξαν μέσα από την ερευνά τους οι συγγραφείς είναι ότι το πλήθος των κόμβων σημαντικής επιρροής είναι αρκετά μικρό, καθώς και ότι κόμβοι με μικρό πλήθος συνδέσεων μπορούν να επηρεάσουν το δίκτυο κάτι τελείως αντίθετο με τις πεποιθήσεις μέχρι τότε. Το 2016 οι Morone et al πρότειναν μία αυστηρή υλοποίηση του Collective Influence σε σχεδόν γραμμικό (quasilinear) χρόνο, με πολυπλοκότητα $O(N \log N)$.

Η βέλτιστη διήθηση ορίζεται ως η διαδικασία κατά την οποία μελετάται η συμπεριφορά του δικτύου καθώς αφαιρούνται τυχαία κόμβοι και ακμές. Αυτή η ιδέα αξιοποιείται και από τον αλγόριθμο CumulativeRank στο δεύτερο βήμα για τον υπολογισμό της αντοχής (Tenacity) τιμών R . Αναλυτικότερα, η βέλτιστη διήθηση αναφέρεται στο πρόβλημα εύρεσης του ελάχιστου συνόλου κόμβων του δικτύου, των οποίων η αφαίρεση το κατακερματίζει σε μικρότερες συνιστώσες. Η βέλτιστη διήθηση βρίσκει εφαρμογή σε πολλά σύγχρονα προβλήματα, όπως προβλήματα ανοσοποίησης δικτύων στην αντιμετώπιση της εξάπλωσης ιών και στη διασπορά πληροφοριών σε κοινωνικά δίκτυα που μελετάμε σε αυτή την διπλωματική. Το πρόβλημα ανήκει στην κλάση των NP – hard προβλημάτων.

Πρακτικά ο αλγόριθμος εντοπίζει το ελάχιστο υποσύνολο κόμβων που επηρεάζει το δίκτυο και η αφαίρεσή των οποίων το αποσυνδέει. Αυτό το σύνολο κόμβων ονομάζεται giant component και οι κόμβοι που την απαρτίζουν influencers.

3.7.2.2 Περιγραφή του αλγορίθμου Collective Influence

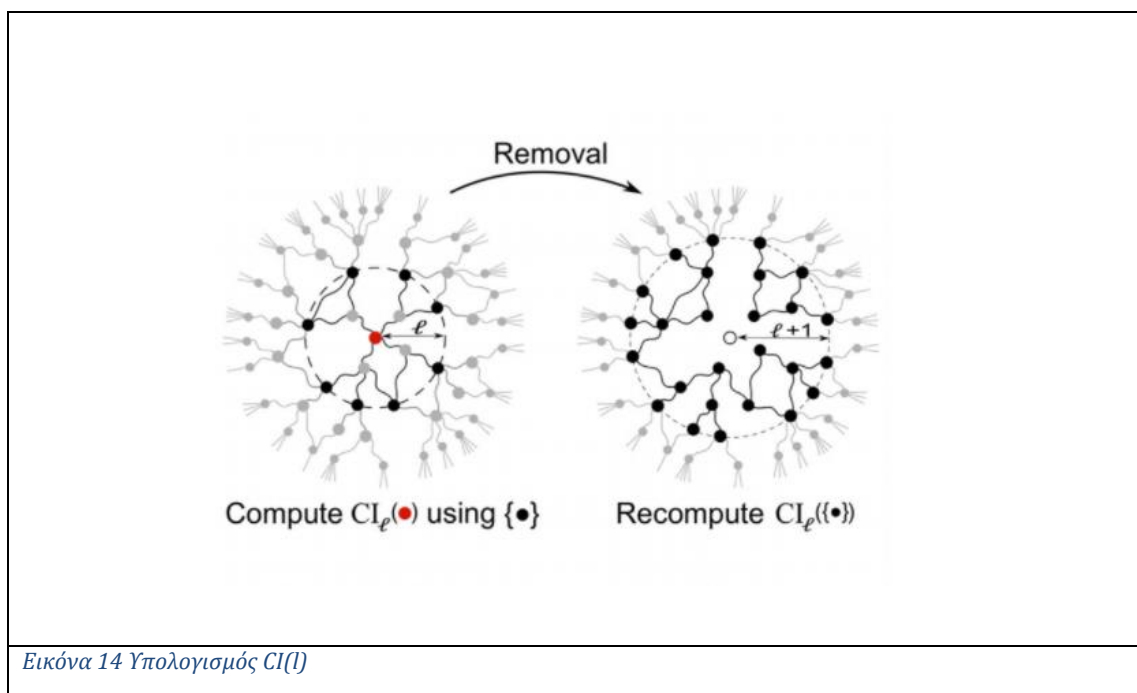
Κάθε κόμβος του δικτύου επηρεάζει τους κόμβους που βρίσκονται στη σφαίρα επιρροής του. Για κάθε κόμβο u ο αλγόριθμος υπολογίζει τη τιμή της μετρικής CI η οποία ορίζεται σύμφωνα με τον παρακάτω τύπο:

$$CI_l(i) = (k_i - 1) \sum_{j \in \partial B(i,l)} (k_j - 1)$$

όπου:

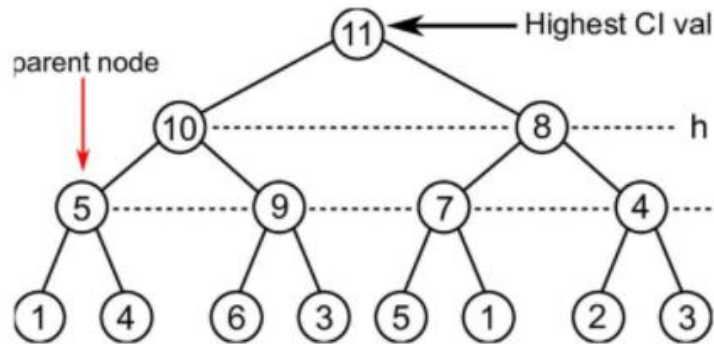
- k_i ο έξω βαθμός του κόμβου i
- k_j ο έξω βαθμός κάθε κόμβου j που ανήκει στη σφαίρα επιρροής $\partial B(i, l)$ με κέντρο τον i και ακτίνα l . Στη σφαίρα $\partial B(i, l)$, ανήκουν όλοι οι κόμβοι που απέχουν απόσταση το πολύ l ακμών από τον κόμβο i .

Σε κάθε επανάληψη, ο αλγόριθμος αφαιρεί τον κόμβο με τη μεγαλύτερη CI τιμή και επικαιροποιεί τους υπολογισμούς μέχρι τελικά να καταστρέψει την giant component.



Εικόνα 14 Υπολογισμός $CI(l)$

Η εκτέλεση του σε σχεδόν γραμμικό χρόνο επιτυγχάνεται χρησιμοποιώντας τις κατάλληλες δομές δεδομένων. Πιο συγκεκριμένα αντί να χρησιμοποιηθούν πίνακες και να ταξινομούνται σε κάθε επανάληψη, χρησιμοποιείται ένας σωρός μεγίστων (max-heap). Πρόκειται για μια δομή δυαδικού δέντρου με το χαρακτηριστικό ότι η τιμή κάθε κόμβου πατέρα είναι πάντα μεγαλύτερη από αυτές των κόμβων παιδιών του. Έτσι, αφού στην ρίζα βρίσκεται η μεγαλύτερη τιμή, μπορούμε να την εξάγουμε κατευθείαν γλιτώνοντας πολυπλοκότητα και αρά χρόνο εκτέλεσης.



Εικόνα 15 Max - heap

Με κάθε αφαίρεση πρέπει να αναδιαταχθούν οι τιμές στον σωρό , ώστε η ρίζα να έχει διαρκώς την εκάστοτε κορυφή με τη μέγιστη CI, μεταξύ των αποθηκευμένων κορυφών στον σωρό. Η αφαίρεση ενός κόμβου διαταράσσει τις τιμές των κόμβων που ανήκουν στη σφαίρα $\partial B(i, l+1)$, δηλαδή των κόμβων που βρίσκονται σε απόσταση $l+1$ από αυτόν που αφαιρείται.

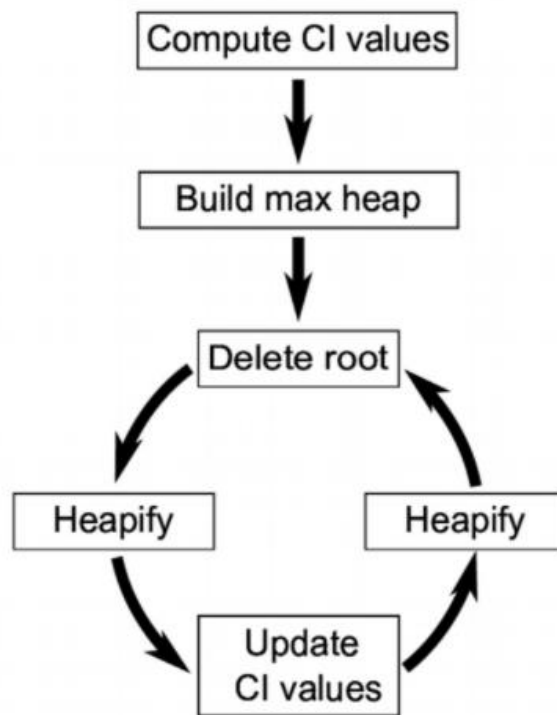
Στη συνέχεια, υπολογίζονται ξανά τις τιμές CI των κόμβων που επηρεάστηκαν από την αφαίρεση δηλαδή ήταν μέσα στην σφαίρα επιρροής του κόμβου που αφαιρέθηκε και αναδιατάσσουμε τις νέες τιμές μέσα στην δομή. Στο τέλος κάθε επανάληψης υπολογίζουμε την ποσότητα $\lambda(l:q)$, που δίνεται από τον τύπο:

$$\lambda(l:q) = \frac{\sum_i CI_l(i)^{\frac{1}{l+1}}}{N < k >}$$

Όπου:

- $<k>$ ο μέσος βαθμός του δικτύου όταν $q = 0$,
- q το πλήθος των κόμβων που αφαιρούνται από το δίκτυο
- N το πλήθος των κόμβων του δικτύου στην τρέχουσα επανάληψη.

Σε κάθε επανάληψη του αλγορίθμου, η τιμή $\lambda(l:q)$ μειώνεται και η εύρεση της giant component με τους σημαντικότερους κόμβους (influencers) ολοκληρώνεται όταν η τιμή $\lambda(l:q)$ πάρει την τιμή 1, οπότε και τερματίζει ο Collective Influence.



Εικόνα 16 Βήματα εκτέλεσης του *Collective Influence*

3.7.3 SimRank

3.7.3.1 Εισαγωγή SR

Ο SimRank είναι ένας αλγόριθμος που παράγει ένα μέτρο ομοιότητας, βασισμένο στην ιδέα τυχαίων περιπάτων του PageRank και μπορεί να εφαρμοστεί σε οποιονδήποτε τομέα με σχέσεις αντικειμένου-προς-αντικείμενο. Μετρά την ομοιότητα των αντικειμένων με βάση τις σχέσεις τους με άλλα αντικείμενα. Έτσι θεωρεί ότι δύο αντικείμενα είναι όμοια αν αναφέρονται - χαρακτηρίζονται από παρόμοια αντικείμενα.

3.7.3.2 Περιγραφή του αλγόριθμου SimRank

Για τον υπολογισμό της ομοιότητας των κόμβων των δικτύων, χρησιμοποιήσαμε την έτοιμη συνάρτηση `SimRank_similarity()` του `NetworkX` της `Python`. Ο αλγόριθμος έχει δυνατότητα επιστροφής της ομοιότητας μεταξύ όλων των κόμβων του δικτύου, της ομοιότητας ενός κόμβου με όλους τους άλλους κόμβους, καθώς και την ομοιότητα μεταξύ δύο συγκεκριμένων κόμβων. Αφού ο στόχος είναι να βρούμε τους κόμβους που παίζουν τον σημαντικότερο ρόλο στην διάδοση της πληροφορίας, μας ενδιαφέρει να βρούμε τους κόμβους εκείνους που μοιάζουν περισσότερο στον κόμβο πηγή της αναδυόμενης ιστορίας, ο οποίος είναι ο σημαντικότερος κόμβος στην διάδοση της πληροφορίας. Αρά, όσο πιο πολύ μοιάζει ένας κόμβος στον κόμβο πηγή, τόσο πιο σημαντικός είναι στην διάδοση της πληροφορίας. Ο αλγόριθμος επιστρέφει ένα λεξικό της `Python` του οποίου τα κλειδιά είναι οι κόμβοι του δικτύου και τιμές, είναι οι τιμές ομοιότητας του καθενός με τον κόμβο – πηγή. Στη συνέχεια διατάσσονται όλοι οι κόμβοι του υποκείμενου γραφήματος διάδοσης της ιστορίας σε φθίνουσα σειρά για να υπολογιστούν οι τιμές των στιγμών.

Το μέτρο ομοιότητας s δύο κόμβων a, b του δικτύου, δίνεται από τον τύπο:

$$s(a, b) = \begin{cases} 1, & a = b \\ \frac{c}{|I(a)||I(b)|} \sum_{j=1}^{|I(b)|} \sum_{i=1}^{|I(a)|} s(I_i(a), I_j(b)) , & I(a), I(b) \neq \emptyset \\ 0, & \text{άλλο} \end{cases}$$

Όπου:

- $I(a), I(b)$ το σύνολο των έσω γειτόνων του a , και του b αντίστοιχα
- $|I(a)|, |I(b)|$ ο πληθάριθμος του $I(a)$ και του $I(b)$ αντίστοιχα
- $I_i(a), I_j(b)$ στοιχείο του $I(a)$ και του $I(b)$ αντίστοιχα
- Ο συντελεστής $c \in [0,1]$ δίνει μεγαλύτερη βαρύτητα στους κοντινούς
- κόμβους των a, b .

Το NetworkX είναι πακέτο της προγραμματιστικής γλώσσας Python για τη δημιουργία, τη διαχείριση και τη μελέτη σύνθετων δικτύων. Διαθέτει πληθώρα έτοιμων συναρτήσεων και μετρικών ανάλυσης δικτύων, όπως ο SimRank που χρησιμοποιήσαμε. Παρακάτω, στο πρώτο σύνδεσμο υπάρχει το εγχειρίδιο της βιβλιοθήκης networkx της Python και στον δεύτερο σύνδεσμο υπάρχουν πληροφορίες για την χρήση της βιβλιοθήκης networkx μέσω της βιβλιοθήκης numpy της Python.

<https://networkx.github.io/>

https://networkx.github.io/documentation/stable/reference/algorithms/generated/networkx.algorithms.similarity.simrank_similarity_numpy.html/

Κεφάλαιο 4. Πειραματική αξιολόγηση

Στην ενότητα αυτή παρουσιάζουμε αναλυτικά πώς διεξήχθη η πειραματική αξιολόγηση των δεδομένων καθώς και η δημιουργία του εργαλείου κατηγοριοποίησης.

4.1 Μεθοδολογία πειραματισμού

Εφόσον έχουμε δημιουργήσει τα γκρουπ ιστοριών από τα δεδομένα και έχουμε δημιουργήσει τα DAG γραφήματα διάδοσης της πληροφορίας των ιστοριών, για κάθε γκρουπ ξεχωριστά φορτώνουμε τις ακμές των γραφημάτων των ιστοριών του. Εν συνεχεία, δημιουργούμε το γράφημα τόσο με τη βιβλιοθήκη `collections.defaultdict` εσωτερικά σ' ένα αντικείμενο κλάσης `graph` στο οποίο τρέχει ο αλγόριθμός μας, όσο και με την βιβλιοθήκη `networkx` της Python για την οπτική αποτύπωση του γραφήματος. Πιο συγκεκριμένα εφαρμόζουμε στα γραφήματα των ιστοριών για κάθε γκρουπ τους τρεις αλγορίθμους που αναλύσαμε παραπάνω με στόχο να εντοπίσουμε τους σημαντικούς κόμβους με βάση τις δομικές τους ιδιότητες. Για κάθε γράφημα, βρίσκουμε τους κόμβους που επηρεάζουν το δίκτυο όπως ορίζεται από κάθε μετρική αλγορίθμου και κατασκευάζουμε το αντίστοιχο διάνυσμα. Τα διανύσματα αυτά δεν είναι τίποτα άλλο από λεξικά με κλειδιά τα `id` των κόμβων και τιμές τη σημαντικότητά τους. Στην περίπτωση του `CumulativeRank` όπου οι σημαντικότερες τιμές είναι οι μικρότερες, αντιστρέψαμε τις τιμές εξόδου του αλγορίθμου έτσι ώστε οι μεγαλύτερες τιμές να είναι οι σημαντικότερες και να συμπίπτουν στην λογική των διανυσμάτων των δυο άλλων αλγορίθμων. Έπειτα, κανονικοποιούμε τα διανύσματα ώστε οι τιμές των λεξικών να αθροίζουν στην μονάδα και διατάσσουμε και αποθηκεύουμε μόνο τις τιμές σε φθίνουσα σειρά.

Για την πειραματική αξιολόγηση, θέτουμε δύο κριτήρια διαχωρισμού. Το πρώτο κριτήριο είναι το πλήθος των κόμβων που συγκεντρώνει ένα συγκεκριμένο ποσοστό αθροιστικής σημαντικότητας ως προς το μέγεθος της ιστορίας λαμβάνοντας υπόψη τους κόμβους του γραφήματος με συγκεκριμένη σειρά, από τον πιο σημαντικό προς τον

λιγότερο σημαντικό. Το εκάστοτε ελάχιστο πλήθος κόμβων το υπολογίζουμε, για κάθε ιστορία και για όλες τις τιμές συσσωρευμένης σημαντικότητας, από 0.1 έως και 0.9 με βήμα 0.05 ως εξής:

Διατάσσουμε τις κανονικοποιημένες τιμές του διανύσματος σημαντικοτήτων σε φθίνουσα σειρά και για ένα ποσοστό έστω $X \in [0.1, 0.9]$ αρχίζουμε να προσθέτουμε σημαντικότητες από τη μεγαλύτερη τιμή προς τη πιο μικρή σε μια μεταβλητή μέχρι η τιμή της να είναι μεγαλύτερη ή ίση της τιμής σημαντικότητας X που έχουμε επιλέξει. Ο αριθμός που ψάχνουμε είναι το πλήθος των κόμβων που οι τιμές τους προστέθηκαν στην μεταβλητή και φανερώνει σε πόσους κόμβους διοχετεύεται το X ποσοστό σημαντικότητας της ιστορίας. Αν κρατήσουμε αυτόν τον αριθμό αυτούσιο, τα αποτελέσματα θα επηρεάζονται από τα μεγέθη των ιστοριών (πλήθος κόμβων στο γράφημα διάδοσης) πράγμα που δεν το καθιστά αξιόπιστο κριτήριο. Για να ξεπεράσουμε αυτό το πρόβλημα, διαιρούμε επιπρόσθετα τον αριθμό αυτόν με το μέγεθος του γραφήματος διάδοσης της εν λόγω ιστορίας. Στη συνέχεια, ο κλασματικός αριθμός που προκύπτει στρογγυλοποιείται στα τρία δεκαδικά ψηφία και πολλαπλασιάζεται επί χίλια, ώστε το αποτέλεσμα που προκύπτει να αποθηκευτεί ως ένας ακέραιος αριθμός που εκφράζει το ποσοστό των κόμβων της ιστορίας που συγκεντρώνουν τη ζητούμενη συσσωρευτική σημαντικότητα X και είναι ανεξάρτητο από τα απόλυτα μεγέθη των γραφημάτων διάδοσης των ιστοριών.

Βασικός στόχος είναι να υπολογιστούν κατάλληλες «τυπικές τιμές» σωρευτικής σημαντικότητας οι οποίες είναι σε θέση να διαχωρίσουν όσο το δυνατόν πιο ευκρινώς τις πραγματικές από τις ψευδείς ιστορίες. Προκειμένου να γίνει αυτό, ακολουθούμε μια λογική μηχανικής μάθησης, όπου χρησιμοποιούμε ένα σημαντικό ποσοστό (πχ, τις μισές) ιστοριών για «μάθηση» των καταλλήλων κατωφλίων διαχωρισμού πραγματικών από ψευδείς ειδήσεις. Στη συνέχεια, οι υπόλοιπες ιστορίες «φιλτράρονται» (δηλαδή, κατηγοριοποιούνται) με βάση τα επιλεγμένα κατώφλια κατά τη φάση της μάθησης.

Εκτός από τον παραπάνω τρόπο υπολογισμού κατωφλίων διαχωρισμού για τις σωρευτικές σημαντικότητες στο γράφημα διάδοσης, μια ιστορία μπορεί επίσης να χαρακτηριστεί και με ένα δεύτερο κριτήριο που είναι η K -στη «στιγμή» του διανύσματος των σημαντικοτήτων. Πρόκειται για ένα μέτρο κατανομής μιας μονάδας σημαντικότητας στις συνιστώσες ενός διανύσματος σημαντικοτήτων, και μπορεί να οριστεί για διάφορες τιμές στιγμών (από 1 μέχρι k). Για αυτό το κριτήριο η διάκριση θα επιτευχθεί συνδυάζοντας την 2η, 3η και 5η στιγμή. Παρακάτω ορίζεται και αναλύεται η έννοια των στιγμών:

k-οστή στιγμή: Ως k-οστή στιγμή ορίζεται η k-οστή ρίζα του αθροίσματος των k-στών δυνάμεων συνιστωσών του διανύσματος.

Η 0-ική στιγμή μας δείχνει το πλήθος των διακεκριμένων στοιχείων του διανύσματος και η 1η – στιγμή το μήκος του. Εμείς υπολογίσαμε τη 2η, 3η και 5η στιγμή κάθε σημαντικού διανύσματος. Η 2η - στιγμή ορίζεται ως η Ευκλείδεια νόρμα. Η 2η – στιγμή (2nd – moment / surprise number) μπορεί να μας δείξει πόσο ανισομερής είναι η κατανομή των τιμών του διανύσματος. Όταν η 2η – στιγμή παίρνει υψηλές τιμές, σημαίνει ότι υπάρχει ένας σημαντικός κόμβος στο δίκτυο, ενώ όταν παίρνει χαμηλότερες τιμές, σημαίνει η σημαντικότητα διαμοιράζεται μεταξύ περισσότερων κορυφών. Οι στιγμές τάξης μεγαλύτερης του 2 υπολογίζονται με τον αντίστοιχο τρόπο. Το μόνο που αλλάζει είναι η εκτίμηση που αντλούμε. Η 3η στιγμή ορίζεται ως η κυβική ρίζα των αθροισμάτων των κύβων των συνιστωσών του διανύσματος και η 5η στιγμή ως η πέμπτη ρίζα των αθροισμάτων των συνιστωσών του διανύσματος υψωμένων στην 5η.

4.1.1 Μετρικές ποιότητας για αξιολόγηση αλγορίθμων κατηγοριοποίησης ειδήσεων

Στη συνέχεια περιγράφεται ο τρόπος αξιολόγησης κάθε αλγορίθμου, ανάλογα με την επίδοσή του κατά τη φάση του φιλτραρίσματος του κατηγοριοποιητή που υλοποιήσαμε. Συγκεκριμένα, μετράμε τις τιμές των True Positive, True Negative, False Positive, False Negative τόσο για κάθε γκρουπ ξεχωριστά όσο και συγκεντρωτικά στα δεδομένα μας. Θεωρούμε ως:

- True Positive: αληθείς ειδήσεις που το μοντέλο αναγνωρίζει ορθά (True True)
- True Negative: ψευδείς ειδήσεις που το μοντέλο αναγνωρίζει ορθά (True Fake)
- False Positive: ψευδείς ειδήσεις που λανθασμένα αξιολογούνται ως αληθείς (False True)
- False Negative: αληθείς ειδήσεις που λανθασμένα αξιολογούνται ως ψευδείς (False Fake)

Στη συνέχεια, εισάγουμε τις έννοιες των μετρικών precision, recall, f1 – score, selectivity και negative predicted value, f1 for false και f1 all που χρησιμοποιούνται για τη μέτρηση της ακρίβεια του μοντέλου μας στην ανίχνευση των αληθών ειδήσεων κατά τη φάση του φιλτραρίσματος (αξιολόγηση).

Αρχικά περιγράφουμε τις μετρικές ποιότητας θεωρώντας ως ζητούμενο την ανίχνευση των αληθών ειδήσεων:

Precision: (positive predicted value) μας δίνει το ποσοστό των ορθά ταυτοποιημένων αληθών ειδήσεων, μεταξύ των ειδήσεων που χαρακτηρίζονται ως αληθείς από τον κατηγοριοποιητή, και δίνεται από τον τύπο:

$$precision = \frac{true\ positive}{true\ positive + false\ positive} = \frac{true\ positive}{total\ predicted\ positive}$$

Recall: υπολογίζει το ποσοστό των αληθών ειδήσεων που ταυτοποιούνται σωστά, μεταξύ όλων των αληθών ειδήσεων, και δίνεται από τον τύπο:

$$recall = \frac{true\ positive}{true\ positive + false\ negative} = \frac{true\ positive}{total\ actual\ positive}$$

F1 – score: είναι ένα μέτρο ισορροπίας των precision και recall και υπολογίζεται ως εξής:

$$f1 = 2 * \frac{precision * recall}{precision + recall}$$

Στη συνέχεια, περιγράφουμε τις μετρικές ποιότητας θεωρώντας ως ζητούμενο την ανίχνευση των ψευδών ειδήσεων.

Συγκεκριμένα, εισάγουμε τις έννοιες των μετρικών negative predicted value , selectivity, F1- for false, για να αξιολογήσουμε πόσο ακριβής είναι η αξιολόγηση των ψευδών ειδήσεων από το μοντέλο μας.

Negative predicted value: αντίστοιχο του precision αλλά για τις ψευδείς ειδήσεις, μας δίνει το ποσοστό των σωστά ταυτοποιημένων ως ψευδών ειδήσεων και δίνεται από τον τύπο:

$$negative\ predicted\ value = \frac{true\ negative}{true\ negative + false\ negative} = \frac{true\ negative}{total\ predicted\ negative}$$

Selectivity: (true negative rate) αντίστοιχο του recall αλλά για τις ψευδείς ειδήσεις, υπολογίζει το ποσοστό των ψευδών ειδήσεων που ταυτοποιούνται σωστά και υπολογίζεται από τη σχέση:

$$selectivity = \frac{true\ negative}{true\ negative + false\ positive} = \frac{true\ negative}{total\ actual\ negative}$$

F1 for false: αντίστοιχο του F1 score αλλά για τις ψευδείς ειδήσεις, είναι ένα μέτρο ισορροπίας των Negative predicted value και Selectivity και υπολογίζεται από τη σχέση:

$$F1\ for\ false = 2 * \frac{Negative\ predicted\ value * Selectivity}{Negative\ predicted\ value + Selectivity}$$

Πλέον είναι ξεκάθαρο ότι για την καλύτερη κατηγοριοποίηση και διαχωρισμό των πραγματικών και των ψευδών ειδήσεων πρέπει να συνδυαστούν τα δυο κριτήρια F1-score και F1 score for false σε ένα ενιαίο, για να επιτευχθεί η καλύτερη δυνατή ισορροπία των δύο και κατά επέκταση καλύτερος διαχωρισμός. Έτσι εισάγουμε την μετρική **F1 all** που δίνεται από τον παρακάτω τύπο:

$$F1\ all = F1\ score + F1\ score\ for\ false$$

Οι τιμές της μετρικής F1 all κυμαίνονται στο διάστημα από 0 έως 2.

4.1.2 Περιγραφή Κατηγοριοποιητή Ιστοριών

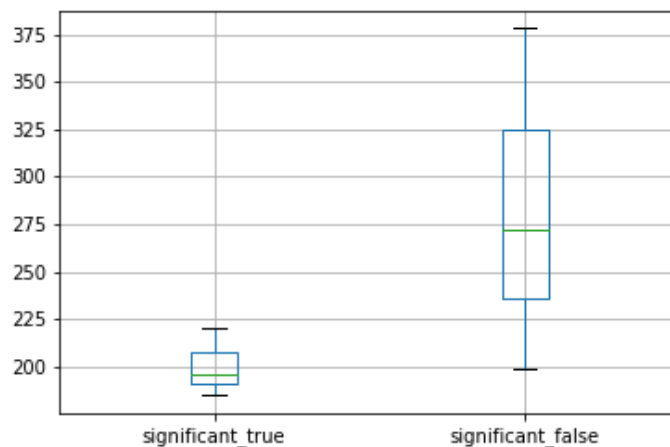
Προκειμένου να καταφέρουμε να εξάγουμε τα δικά μας συμπεράσματα, προχωράμε με την κατασκευή ενός εργαλείου κατηγοριοποίησης που θα αξιοποιεί τα κριτήρια που αναφέραμε αλλά και την ομαδοποίηση των δεδομένων σε γκρουπ με στόχο την επιτυχή κατηγοριοποίηση μιας οποιασδήποτε ιστορίας. Το εργαλείο αυτό περιλαμβάνει δύο φάσεις, τη φάση της εκπαίδευσης και τη φάση της αξιολόγησης. Εκπαιδεύεται και αξιολογείται αξιοποιώντας κάθε έναν αλγόριθμο για υπολογισμό του διανύσματος σημαντικότητας, και κάθε κριτήριο αξιολόγησης (σωρευτική σημαντικότητα, ή στιγμές του διανύσματος σημαντικότητας) ξεχωριστά.

Κατά το βήμα της εκπαίδευσης το εργαλείο αξιοποιεί τις μισές ιστορίες από όλα τα γκρουπ με σκοπό καταλήξει στις τιμές εκείνες που ταιριάζουν καλύτερα στα δεδομένα και η εκπαίδευση γίνεται και με το κριτήριο F1 αλλά και με το F1all ξεχωριστά. Για κάθε ένα από τα κριτήρια και για κάθε μετρική F1 και F1 all , δίνει τη δυνατότητα κατηγοριοποίησης είτε αξιοποιώντας το περιεχόμενο όπου κάθε ιστορία αξιολογείται με βάση τη τυπική τιμή κατωφλιού του κοντινότερου γκρουπ σε επίπεδο κειμένου, είτε όχι (with linguistics ή no linguistics). Συγκεκριμένα:

- Η επιλογή «with linguistics» σημαίνει ότι κάθε ιστορία, πριν κατηγοριοποιηθεί, εξετάζεται ως προς το περιεχόμενό της, προκειμένου να ληφθεί υπόψη μόνο το γκρουπ με τις κοντινότερες σε αυτήν ιστορίες, βάσει περιεχομένου.
- Η επιλογή «no linguistics» σημαίνει ότι όλες οι ιστορίες εκλαμβάνονται ότι ανήκουν στην ίδια ομάδα, οπότε δεν εξετάζεται καθόλου το περιεχόμενό τους.

Για το πρώτο κριτήριο το ποσοστό των κόμβων που συγκεντρώνουν μια συγκεκριμένη τιμή συσσωρευμένης σημαντικότητας, το μοντέλο υπολογίζει το αριθμό που επιτυγχάνει το βέλτιστο διαχωρισμό για κάθε μια από τις τιμές σωρευτικής σημαντικότητας στο διάστημα [0.1, 0.9] με βήμα 0.05 ξεχωριστά. Στη συνέχεια, από αυτά, βρίσκει το βέλτιστο ποσοστό, εκείνο δηλαδή που επιτυγχάνει την καλύτερη κατηγοριοποίηση για τις ιστορίες της εκπαίδευσης, είτε βάσει του F1-score είτε βάσει του F1 all ξεχωριστά. Ο συγκεκριμένος υπολογισμός της βέλτιστης τιμής για τη διάκριση των ιστοριών γίνεται, για κάθε αλγόριθμο σημαντικότητας, ξεχωριστά για κάθε ένα γκρουπ αλλά και συγκεντρωτικά για όλες τις ιστορίες του συνόλου εκπαίδευσης συνολικά.

Η διαδικασία εύρεσης της βέλτιστης τιμής διάκρισης αληθών από ψευδείς ειδήσεις ιστοριών, για ένα ποσοστό συσσωρευτικής σημαντικότητας X, είναι η εξής:



Εικόνα 17 box-plots των ποσοστιαίων τιμών του κριτηρίου *significance* των πραγματικών και των ψευδών ιστοριών για ένα ποσοστό συσσωρευμένης σημαντικότητας *X*.

Πρώτα πρέπει να έχουμε υπολογίσει από τα διανύσματα, τα ελάχιστα ποσοστά των κόμβων (significance percentage) που συγκεντρώνουν το συγκεκριμένο ποσοστό συσσωρευμένης σημαντικότητας *X* για το οποίο ψάχνουμε τη βέλτιστη τιμή διαχωρισμού για τις πραγματικές και για τις ψευδείς ιστορίες. Στη συνέχεια, υπολογίζουμε τον διάμεσο (median) του διανύσματος για τις ψευδείς και του διανύσματος για τις πραγματικές ιστορίες, το οποίο φαίνεται στο παραπάνω γράφημα με πράσινη γραμμή. Επίσης τα κάτω και τα άνω όρια των ορθογωνίων περιγράφονται με τους όρους **χαμηλότερο τεταρτημόριο** (lower quartile), **υψηλότερο τεταρτημόριο** (upper quartile) και αντιπροσωπεύουν τα 25% της κατανομής και το 75% της κατανομής με τον μέσο να είναι στο 50%. Για τον υπολογισμό του κάτω ορίου των ορθογωνίων (lower quartile), για κάθε διάνυσμα εφόσον βρούμε τον διάμεσο (median) του διανύσματος, αρκεί να βρούμε και το διάμεσο για τα στοιχεία που είναι μικρότερα του διαμέσου (median). Αντίστοιχα για τον υπολογισμό του άνω ορίου των ορθογωνίων (upper quartile), για κάθε διάνυσμα εφόσον βρούμε τον διάμεσο (median) του διανύσματος, αρκεί να βρούμε και το διάμεσο για τα στοιχεία που είναι μεγαλύτερα του διαμέσου (median). Επίσης η Python παρέχει τη συνάρτηση `quantile` για `Dataframe` που υπολογίζει το μέσο και τα άνω και κάτω όρια (lower/ upper quartile) των ορθογωνίων, μέσω της βιβλιοθήκης `Pandas`.

Έχοντας βρει τους διαμέσους και τα όρια των ορθογωνίων για κάθε ένα από τα διανύσματα των πραγματικών και των ψευδών ιστοριών, η διαδικασία είναι (στη φάση της εκπαίδευσης) να σκανάρουμε το διάστημα από τον μικρότερο μέχρι τον μεγαλύτερο

διάμεσο ελέγχοντας κάθε μια ακέραια τιμή ξεχωριστά. Για κάθε ακέραια τιμή από το κάτω ακέραιο μέρος του μικρότερου μέσου μέχρι το άνω ακέραιο μέρος του μεγαλύτερου, υπολογίζουμε τις τιμές true positive, true negative, false positive, false negative και στη συνέχεια precision, recall, negative predicted value, selectivity, F1 score, F1 for false, το F1 all. Ανάλογα με το αν η εκπαίδευση γίνεται με βάση το F1 all ή το F1 score, λαμβάνεται υπ' όψιν η μεγιστοποίηση της αντίστοιχης τιμής της μετρικής. Εδώ πρέπει να τονιστεί ότι κάθε ακέραια τιμή διαχωρισμού ορίζει δύο περιοχές, την πάνω από τον ακέραιο και την κάτω. Έτσι θα επιλέξουμε κάθε μία περιοχή ξεχωριστά για περιοχή πραγματικών ιστοριών με την αντίστοιχη άλλη περιοχή για τις ψευδείς ιστορίες και πάλι θα υπολογίσουμε ποια περίπτωση από τις δύο είναι η βέλτιστη και θα κρατήσουμε εκείνη. Στο τέλος κρατάμε, ξεχωριστά για κάθε αλγόριθμο και ομάδα ιστοριών, τη βέλτιστη τιμή διαχωρισμού, αυτή δηλαδή που πέτυχε το μεγαλύτερο F1 score ή F1 all ανάλογα, για το συγκεκριμένο ποσοστό X συσσωρευμένης σημαντικότητας που επιζητάμε.

Στην περίπτωση που τα ορθογώνια δεν επικαλύπτονται, δηλαδή η μικρότερη τιμή υψηλότερου τεταρτημόριου (upper quartile) είναι μικρότερη από την μεγαλύτερη τιμή χαμηλότερου τεταρτημόριου (lower quartile), μπορούμε να μικρύνουμε το διάστημα που ψάχνουμε τη βέλτιστη τιμή διαχωρισμού, στο διάστημα ανάμεσα στη μικρότερη τιμή υψηλότερου τεταρτημόριου (upper quartile) και στη μεγαλύτερη τιμή χαμηλότερου τεταρτημόριου (lower quartile), δηλαδή στο κενό ανάμεσα στα ορθογώνια. Στο παραπάνω γράφημα όπου ισχύει κάτι τέτοιο, αντί να πάρουμε το διάστημα από τον διάμεσο των αληθών ιστοριών (significant_true) μέχρι τον διάμεσο ψευδών ιστοριών (significant_false), παίρνουμε το διάστημα από το άνω όριο του ορθογωνίου του significant_true μέχρι το κάτω όριο του ορθογωνίου του significant_false. Οι τυπικές τιμές έχουν δομή:

(Για κάθε τιμή $X\%$, αριθμός διαχωρισμού που έχει μέγιστο $F1/F1_{all}$, τιμή $F1/F1_{all}$, «<» ή «>» ανάλογα αν η περιοχή πραγματικών έχει κλειδώσει κάτω από την τιμή ή πάνω).

Έχοντας κάνει την παραπάνω διαδικασία για όλες τις υποψήφιες τιμές X στο $[0.1, 0.9]$ με βήμα 0.05, παίρνουμε τον βέλτιστο συνδυασμό τυπικών τιμών, δηλαδή το συνδυασμό που πέτυχε τον βέλτιστο διαχωρισμό βάσει του κριτηρίου αξιολόγησης που θεωρούμε, από όλα τα ποσοστά σύμφωνα με τις τιμές των μετρικών (τυπικές τιμές του βέλτιστου ποσοστού).

Έτσι θα υπολογιστούν βέλτιστες τυπικές τιμές: (ποσοστό X , τιμή διαχωρισμού, πάνω ή κάτω περιοχή των πραγματικών) τόσα όσα είναι τα γκρουπ με τιμές το

βέλτιστο ποσοστό σημαντικότητας με τη βέλτιστη τιμή διάκρισης και μια επιπλέον τέτοια τυπική τιμή για όλες τις ιστορίες από το σύνολο εκπαίδευσης σε συνολικό επίπεδο. Μ' αυτόν τον τρόπο καταφέρνουμε κατά τη φάση του φιλτραρίσματος να γίνει κατηγοριοποίηση είτε με βάση το περιεχόμενο, χρησιμοποιώντας για τιμή διάκρισης την τιμή διάκρισης του γκρουπ εκείνου που σε επίπεδο κειμένου είναι πιο κοντά στο κείμενο της ιστορίας που θέλουμε να κατηγοριοποιήσουμε, ή να επιλέξουμε να μην χρησιμοποιήσουμε το περιεχόμενο της ίδιας της ιστορίας και να βασιστούμε στην τυπική τιμή που υπολογίστηκε για όλες τις ιστορίες από το σύνολο εκπαίδευσης εκλαμβάνοντάς τις ως ένα ενιαίο γκρουπ ιστοριών.

Για το δεύτερο κριτήριο, αυτό των στιγμών, για κάθε γκρουπ παίρνουμε πάλι τις μισές του ιστορίες και υπολογίζουμε τον μέσο όρο των 2ων, 3ων και 5ων στιγμών για τις πραγματικές ιστορίες του γκρουπ και τον μέσο όρο για τις ψευδείς. Έτσι θα υπολογιστούν για κάθε γκρουπ δύο διανύσματα που το καθένα θα περιέχει τρεις τιμές. Για το πρώτο διάνυσμα, αυτό των πραγματικών ιστοριών, η πρώτη τιμή θα είναι ο μέσος όρος των δεύτερων στιγμών, η δεύτερη τιμή θα είναι ο μέσος όρος των τρίτων στιγμών και η τρίτη τιμή θα είναι ο μέσος όρος των πέμπτων στιγμών των πραγματικών ιστοριών και το δεύτερο διάνυσμα υπολογίζεται αντίστοιχα για τις ψευδείς ιστορίες. Επιπλέον για όλες τις ιστορίες του συνόλου εκπαίδευσης συνολικά, υπολογίζουμε τα δύο διανύσματα με μέγεθος τρία από τους μέσους όρους πάλι των 2ων, 3ων και 5ων στιγμών των πραγματικών και των ψευδών ιστοριών σε συνολικό επίπεδο. Και πάλι έχουμε τη δυνατότητα κατηγοριοποίησης με βάση το περιεχόμενο, χρησιμοποιώντας τα διανύσματα του γκρουπ εκείνου που σε επίπεδο κειμένου είναι πιο κοντά στο κείμενο της ιστορίας που θέλουμε να κατηγοριοποιήσουμε, υπολογίζοντας τις αποστάσεις της ιστορίας από τα διανύσματα των αληθών και των ψευδών ιστοριών. Η τελική απόφαση λαμβάνεται με βάση ποια απόσταση είναι μικρότερη δηλαδή χαρακτηρίζεται με την αληθοτιμή του διανύσματος που είναι πιο κοντά στην ιστορία. Επιπλέον μπορεί να γίνει κατηγοριοποίηση χωρίς να αξιοποιήσουμε το περιεχόμενο της προς κατηγοριοποίηση ιστορίας, απλά χρησιμοποιώντας τα δύο διανύσματα των πραγματικών και των ψευδών ιστοριών σε συνολικό επίπεδο επιλέγοντας ανάλογα πάλι με βάση τη διανυσματική απόσταση, την αληθοτιμή του διανύσματος που είναι πιο κοντά.

Στόχος μας είναι να δούμε για κάθε περίπτωση εκτέλεσης των αλγόριθμων και για κάθε ένα από τα τρία κριτήρια χωριστά, αυτό των στιγμών, αυτό των ποσοστιαίων τιμών σημαντικότητας με εκπαίδευση F1 και για αυτό των ποσοστιαίων τιμών σημαντικότητας με εκπαίδευση F1 all, τόσο με χρήση του περιεχομένου της ιστορίας

προς κατηγοριοποίηση όσο και χωρίς, πως επιτυγχάνεται ο καλύτερος διαχωρισμός των πραγματικών από τις ψευδείς ειδήσεις.

Να σημειωθεί ότι, υπάρχουν 6 περιπτώσεις εκτέλεσης αλγορίθμων, του CumulativeRank, του Collective Influence για radius=1,2,3,4 και του SimRank. Εξετάζουμε συνολικά 6 εκτελέσεις αλγορίθμων * 3 κριτήρια * 2 περιπτώσεις της χρήσης λεκτικής πληροφορίας και της μη χρήσης λεκτικής, άρα συνολικά 36 περιπτώσεις.

4.1.3 Μεθοδολογία αναζήτησης γκρουπ των πιο κοντινών ιστοριών

Κατά την φάση του φιλτραρίσματος (ή φάση αξιολόγησης), η διαδικασία εύρεσης του γκρουπ που σε επίπεδο κειμένου είναι πιο κοντά στο κείμενο μιας συγκεκριμένης ιστορίας γίνεται ως εξής:

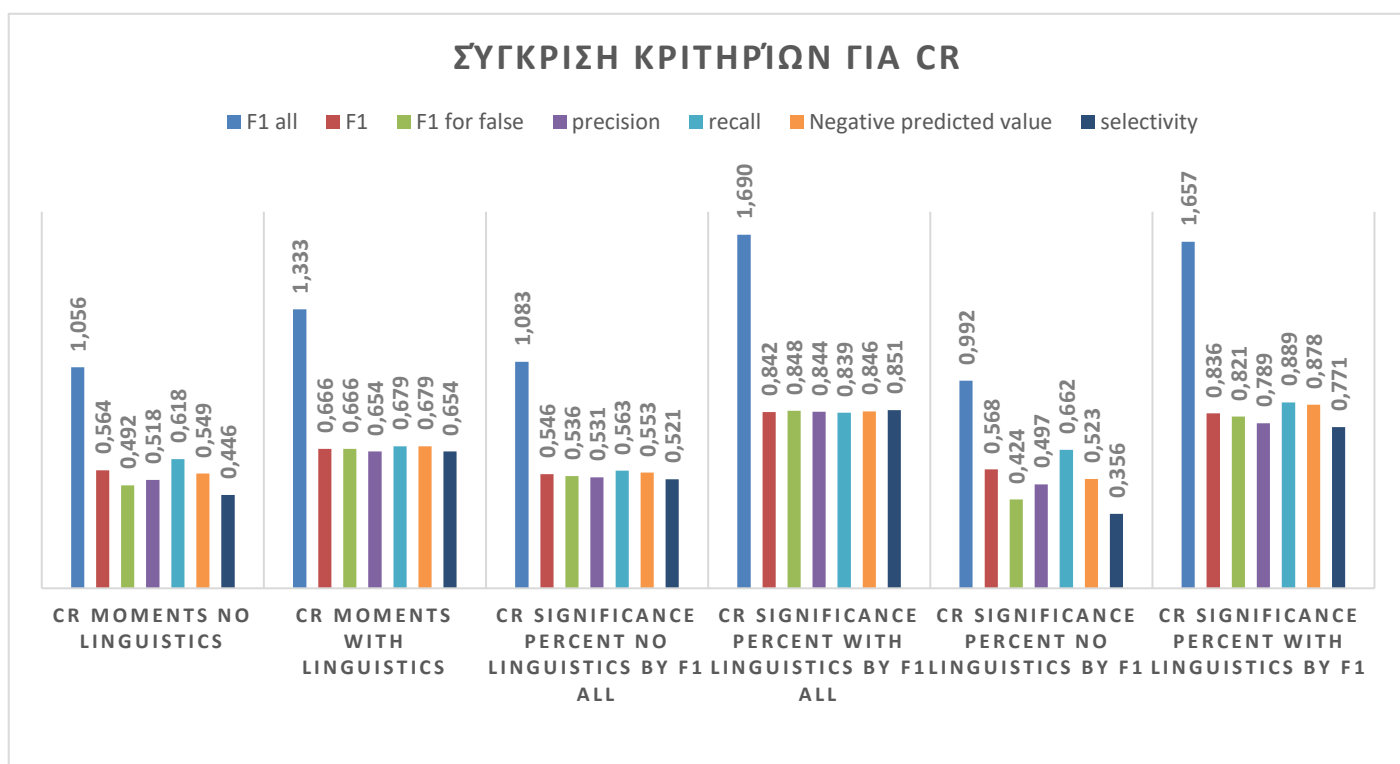
Ξεκινάμε με απαλοιφή ειδικών χαρακτήρων, μετατροπή σε μικρά, απαλοιφή τετριμμένων λέξεων τερματισμού (stop words), στη συνέχεια πραγματοποιούμε λημματοποίηση των υπόλοιπων λέξεων (lemmatization), υπολογισμό της tf-idf των ιστοριών, και τέλος υπολογισμό της ομοιότητας συνημίτονου (cosine similarity) μεταξύ των κειμένων της κάθε ιστορίας με το κείμενο του κάθε γκρουπ που δεν είναι άλλο από ένα αλφαριθμητικό όλων των κειμένων των ιστοριών που περιέχει. Επιπλέον, μπορεί να γίνει κατηγοριοποίηση της υπό εξέταση ιστορίας και χωρίς να αξιοποιήσουμε την λεκτική πληροφορία απλά χρησιμοποιώντας το ζεύγος κατηγοριοποίησης σε συνολικό επίπεδο.

4.2 Αναλυτική παρουσίαση αποτελεσμάτων

Για να μπορέσουμε να εξάγουμε ποιοτικά συμπεράσματα, έπρεπε να εξετάσουμε ενδελεχώς κάθε μια από τις έξι περιπτώσεις εκτέλεσης των αλγορίθμων που αξιολογούμε, τον CumulativeRank, τον Collective Influence για radius=1, τον Collective Influence για radius=2, τον Collective Influence για radius=3, τον Collective Influence για radius=4, και τέλος τον SimRank. Για κάθε μια από αυτές τις έξι περιπτώσεις έπρεπε να εξετάσουμε, τόσο με αξιοποίηση του περιεχομένου της εκάστοτε ιστορίας όσο και χωρίς, το κριτήριο των στιγμών, το κριτήριο του Significance percentage με κριτήριο εκπαίδευσης τη μετρική F1 all, και το κριτήριο του Significance percentage αλλά αυτή την φορά με κριτήριο εκπαίδευσης τη μετρική F1. Έτσι παραθέτουμε παρακάτω τα αποτελέσματα του κατηγοριοποιητή, σε επίπεδο αλγορίθμου και για κάθε κριτήριο. Πρέπει να τονιστεί ότι τα αποτελέσματα βγήκαν για τυχαία επιλεγμένα σύνολα εκπαίδευσης και αξιολόγησης.

4.2.1 Σύγκριση ανά αλγόριθμο

Αρχικά προσδιορίζουμε τη χρωματική κωδικοποίηση των αποτελεσμάτων που παρατίθενται στα γραφήματα που ακολουθούν. Η σειρά 1, με μπλε χρώμα, αντιπροσωπεύει τη μετρική F1 all, η σειρά 2, με κόκκινο χρώμα, αντιπροσωπεύει τη μετρική F1, η σειρά 3, με πράσινο χρώμα, τη μετρική F1 for false, η σειρά 4, με μωβ χρώμα, αντιπροσωπεύει τη μετρική precision, η σειρά 5, με γαλάζιο χρώμα, αντιπροσωπεύει τη μετρική recall, η σειρά 6, με πορτοκαλί χρώμα, τη μετρική negative predicted value, και τέλος η σειρά 7 με σκούρο μπλε χρώμα, τη μετρική selectivity.



Γράφημα σύγκρισης κριτηρίων για τον CumulativeRank.

Στο παραπάνω γράφημα βλέπουμε πώς διαμορφώνονται οι τιμές των μετρικών F1 all, F1, F1 for false, precision, recall, negative predicted value και selectivity με αυτή την σειρά για τον CumulativeRank και για όλα τα κριτήρια, είτε χωρίς χρήση λεκτικής πληροφορίας ή με χρήση λεκτικής πληροφορίας (no linguistics/with linguistics).

Αρχικά μπορούμε εύκολα να διακρίνουμε ότι για όλα τα κριτήρια στην εκτέλεση για τον CumulativeRank ισχύει ότι η χρήση λεκτικής πληροφορίας

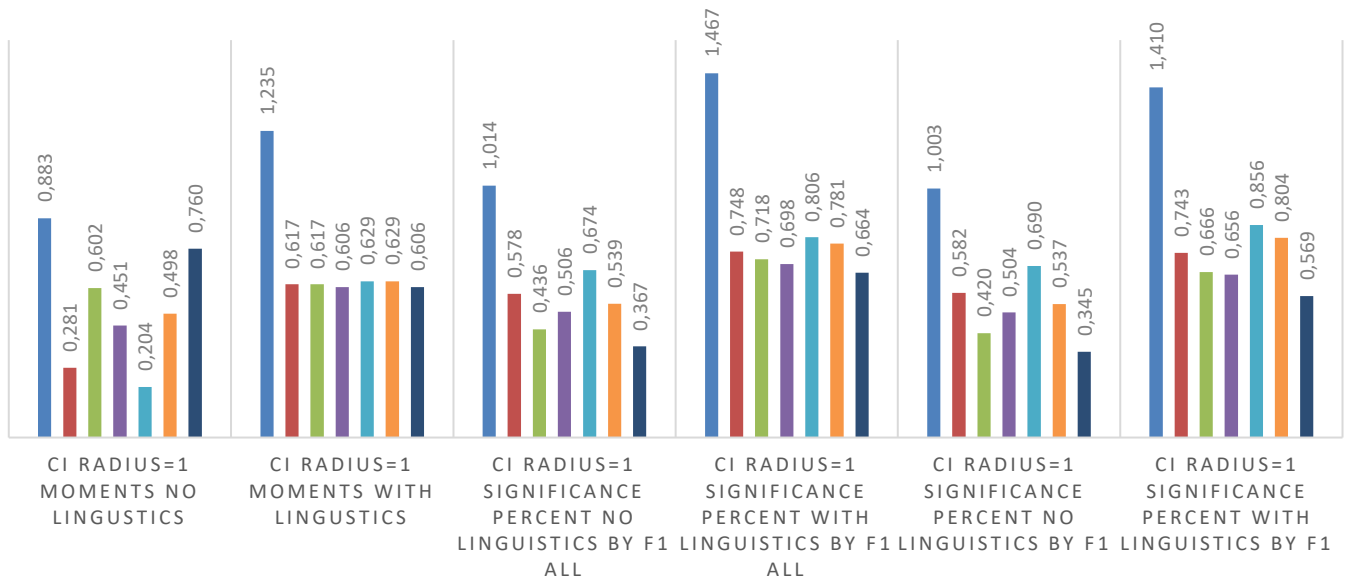
δίνει μεγαλύτερες τιμές σε όλες τις μετρικές και αρά σημαντικά καλύτερα αποτελέσματα. Αυτό συμβαίνει γιατί τα δομικά χαρακτηριστικά της διασποράς των ειδήσεων επηρεάζονται σε σημαντικό βαθμό από το είδος της θεματολογίας της είδησης. Για να μπορέσουμε να κατανοήσουμε καλύτερα γιατί συμβαίνει αυτό, μπορούμε να σκεφτούμε ότι παρ' όλο που οι ψευδείς ειδήσεις έχουν την τάση να είναι μεγαλύτερες σε αντιδράσεις από τις πραγματικές, μια πραγματική είδηση για κάποιο πολύ δημοφιλές θέμα, όπως για τον Αντετοκούνμπο μπορεί να έχει περισσότερες αντιδράσεις από μια ψευδή είδηση με επιστημονική θεματολογία. Έτσι, εφαρμόζοντας διαφορετική κατηγοριοποίηση για κάθε θεματολογία-γκρουπ και χρησιμοποιώντας κάθε φορά την κατηγοριοποίηση του πιο κοντινού γγκρουπ σε θεματολογία για την είδηση-ιστορία που αξιολογούμε, προκύπτουν αρκετά καλύτερα αποτελέσματα.

Στη συνέχεια, μια πολύ σημαντική παρατήρηση είναι η σύγκριση των μετρικών για το κριτήριο του significance percent με εκπαίδευση με F1 all με το ίδιο κριτήριο αλλά με εκπαίδευση με F1 τόσο με λεκτική πληροφορία όσο και χωρίς. Στην περίπτωση της λεκτικής πληροφορίας βλέπουμε ότι η εκπαίδευση με F1 all δίνει υψηλότερες τιμές σε F1 all, F1, F1 for false, precision και selectivity καθώς χάνει λίγο σε recall και negative predicted value. Αυτό μας δείχνει ότι η μεταβολή στο precision είναι μεγαλύτερη από την μεταβολή στο recall αφού οι τιμές του F1 με εκπαίδευση F1 all παραμένουν μεγαλύτερες του F1 με εκπαίδευση με F1. Στην περίπτωση της μη χρήσης λεκτικής πληροφορίας η εκπαίδευση με F1 all κερδίζει σε F1 all, F1 for false, precision, negative predicted value και selectivity ενώ χάνει λίγο σε recall και σε F1 σε σχέση με την εκπαίδευση με F1. Αυτό συμβαίνει γιατί η εκπαίδευση με το κριτήριο F1 all αποσκοπεί στην βέλτιστη ισορροπία αναμεσα στα F1 και F1 for false, έτσι για να την πετύχει, μπορεί να θυσιάσει ένα μικρό ποσοστό από αληθινές ιστορίες για να εξασφαλίσει ότι το μεγαλύτερο ποσοστό των ψευδών ειδήσεων (selectivity) θα χαρακτηριστεί ψευδής.

Τέλος παρατηρούμε ότι, καλύτερα αποτελέσματα με χρήση λεκτικής πληροφορίας δίνει το κριτήριο του significance percent και πιο συγκεκριμένα με εκπαίδευση με F1 all με βάση τις τιμές F1, F1 for false και F1 all. Στην περίπτωση μη χρήσης λεκτικής πληροφορίας το significance percent με εκπαίδευση με F1 all δίνει καλύτερο F1 for false και F1 all σκορ ενώ το significance percent με εκπαίδευση με F1 δίνει καλύτερο F1 σκορ.

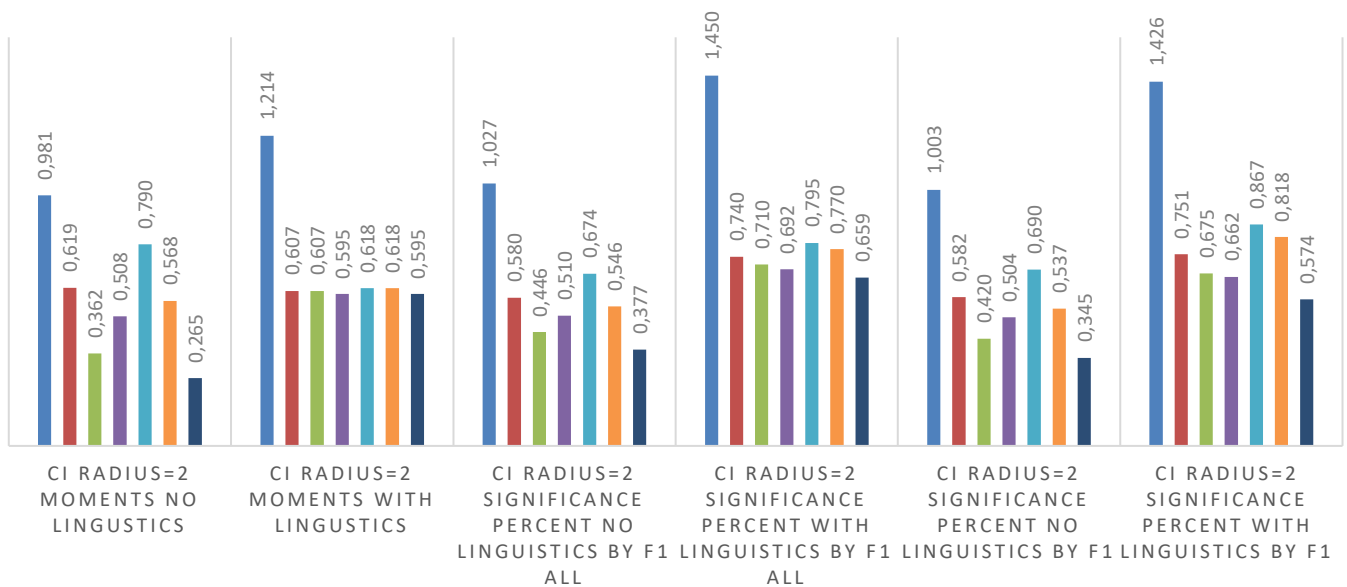
ΣΥΓΚΡΙΣΗ ΚΡΙΤΗΡΙΩΝ ΓΙΑ CI RADIUS=1

■ F1 all ■ F1 ■ F1 for false ■ precision ■ recall ■ Negative predicted value ■ selectivity

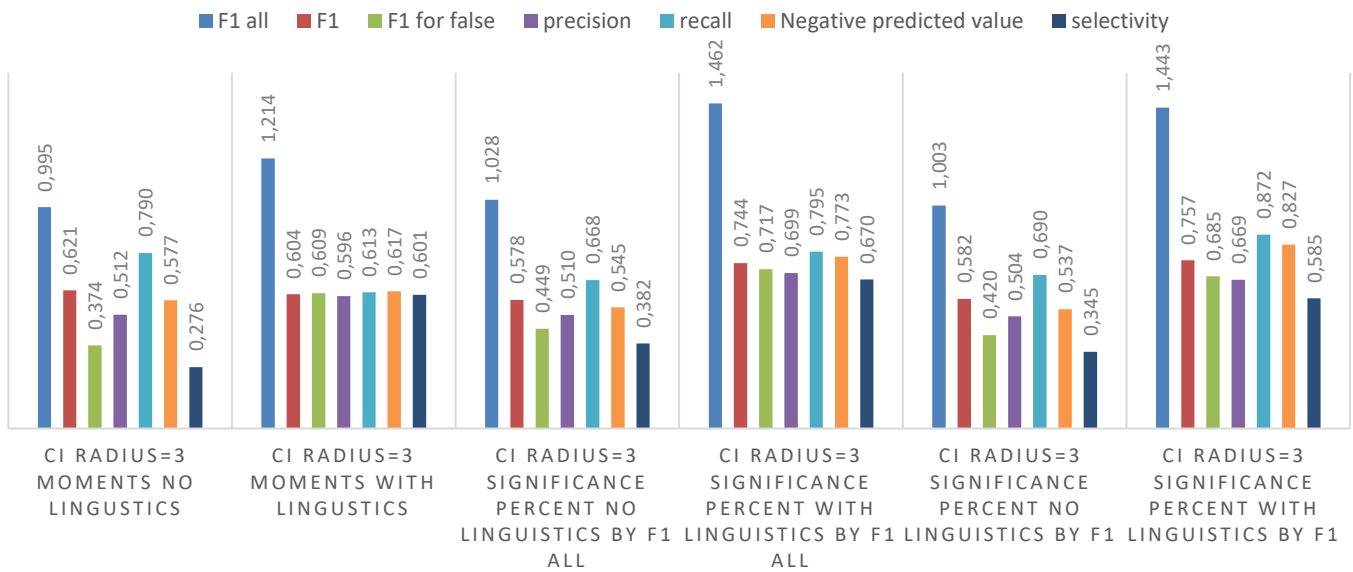


ΣΥΓΚΡΙΣΗ ΚΡΙΤΗΡΙΩΝ ΓΙΑ CI RADIUS=2

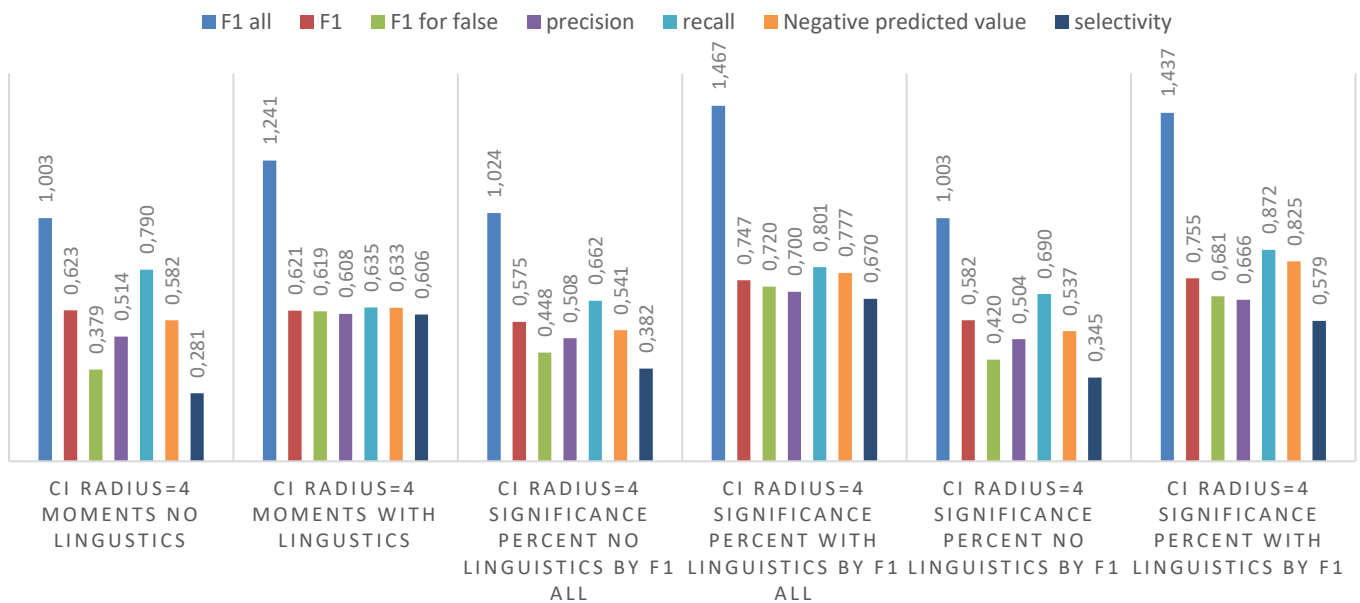
■ F1 all ■ F1 ■ F1 for false ■ precision ■ recall ■ Negative predicted value ■ selectivity



ΣΥΓΚΡΙΣΗ ΚΡΙΤΗΡΙΩΝ ΓΙΑ CI RADIUS=3



ΣΥΓΚΡΙΣΗ ΚΡΙΤΗΡΙΩΝ ΓΙΑ CI RADIUS=4



Γραφήματα του αλγορίθμου Collective Influence για radius=1,2,3,4.

Στα παραπάνω γραφήματα βλέπουμε πώς διαμορφώνονται οι τιμές των μετρικών F1 all , F1, F1 for false, precision, recall, negative predicted value και selectivity με αυτή τη σειρά, για τον Collective Influence με radius=1,2,3,4 και για όλα τα κριτήρια χωρίς χρήση λεκτικής πληροφορίας αλλά και με χρήση λεκτικής πληροφορίας.

Αρχικά μπορούμε εύκολα να διακρίνουμε πως, για όλα τα κριτήρια πως ισχύει ότι η χρήση λεκτικής πληροφορίας δίνει μεγαλύτερες τιμές σε όλες τις μετρικές και άρα σημαντικά καλύτερα αποτελέσματα. Συγκεκριμένα, **για την εκτέλεση με radius=1**, ισχύει με μόνη εξαίρεση στο κριτήριο των στιγμών όπου υπάρχει μείωση του selectivity, ενώ **για τις εκτελέσεις με radius=2,3,4** , ισχύει με μόνη εξαίρεση στο κριτήριο των στιγμών όπου υπάρχει μείωση του recall και κατά επέκταση και του F1. Αυτό συμβαίνει γιατί τα δομικά χαρακτηριστικά της διασποράς των ειδήσεων επηρεάζονται σε μεγάλο βαθμό από τον είδος της θεματολογίας της είδησης. Έτσι, εφαρμόζοντας διαφορετική κατηγοριοποίηση για κάθε θεματολογία-γκρουπ και χρησιμοποιώντας κάθε φορά την κατηγοριοποίηση του πιο κοντινού γκρουπ σε θεματολογία, δίνει αρκετά καλύτερα αποτελέσματα.

Στη συνέχεια, μια πολύ σημαντική παρατήρηση είναι η σύγκριση των μετρικών για το κριτήριο του significance percent με εκπαίδευση με F1 all με το ίδιο κριτήριο αλλά με εκπαίδευση με F1 τόσο με λεκτική πληροφορία όσο και χωρίς. Στην περίπτωση της λεκτικής πληροφορίας για την **εκτέλεση με radius=1** βλέπουμε ότι η εκπαίδευση με F1 all δίνει υψηλότερες τιμές σε F1 all ,F1 ,F1 for false, precision και selectivity καθώς χάνει λίγο σε negative predicted value, και recall. Αυτό μας δείχνει ότι η μεταβολή στο precision είναι μεγαλύτερη από τη μεταβολή στο recall, αφού οι τιμές του F1 παραμένουν μεγαλύτερες του F1 με εκπαίδευση με F1. **Για τις υπόλοιπες εκτελέσεις για radius=2,3,4**, παρατηρούμε ότι η εκπαίδευση με F1 all δίνει υψηλότερες τιμές σε F1 all, F1 for false, precision και selectivity καθώς χάνει λίγο σε negative predicted value, recall αλλά και F1. Αυτό μας δείχνει ότι η μεταβολή στο recall είναι μεγαλύτερη από τη μεταβολή στο precision, αφού οι τιμές του F1 με εκπαίδευση F1 all είναι μικρότερες από του F1 με εκπαίδευση με F1.

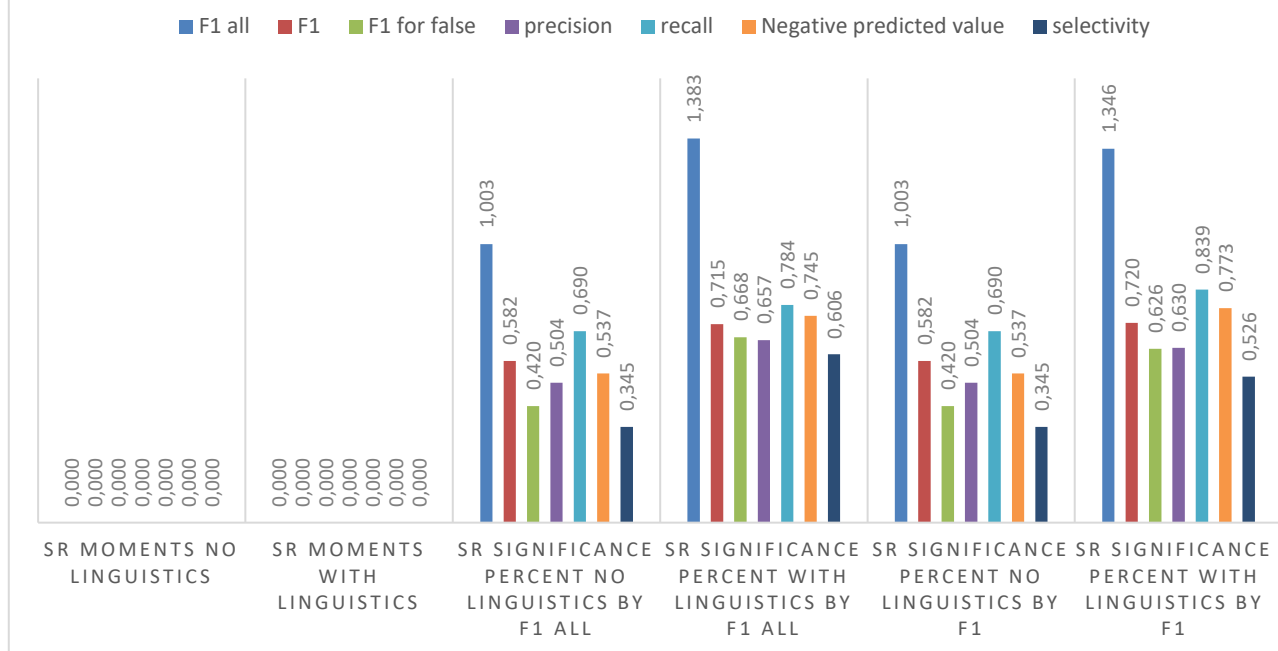
Στην περίπτωση της μη χρήσης λεκτικής πληροφορίας για όλες τις εκτελέσεις του αλγορίθμου Collective Influence, η εκπαίδευση με F1 all κερδίζει σε F1 all ,F1 for false, precision, negative predicted value και selectivity ενώ χάνει λίγο σε F1 και recall σε σχέση με την εκπαίδευση με F1. Αυτό συμβαίνει

γιατί η εκπαίδευση με το κριτήριο F1 all αποσκοπεί στην βέλτιστη ισορροπία ανάμεσα στα F1 και F1 for false, έτσι για να την πετύχει μπορεί να θυσιάσει ένα μικρό ποσοστό από αληθινές ιστορίες για να εξασφαλίσει ότι το μεγαλύτερο ποσοστό των ψευδών ειδήσεων (selectivity) θα χαρακτηριστούν ψευδείς.

Για την εκτέλεση του αλγορίθμου **Collective Influence με radius=1**, παρατηρούμε ότι, καλύτερα αποτελέσματα με χρήση λεκτικής πληροφορίας δίνει το κριτήριο του significance percent, συγκεκριμένα με εκπαίδευση με F1 all με βάση τις τιμές F1 , F1 for false και F1 all. Χωρίς χρήση λεκτικής πληροφορίας το καλύτερο F1 all σκορ το έχει πάλι ο significance percent με εκπαίδευση με F1 all , το καλύτερο F1 το έχει ο πάλι ο significance percent με εκπαίδευση με F1 και το καλύτερο F1 for false σκορ το έχει το κριτήριο των στιγμών και το αμέσως επόμενο το significance percent με εκπαίδευση με F1 all.

Τέλος, για τις εκτελέσεις του αλγορίθμου **Collective Influence με radius=2,3,4**, παρατηρούμε ότι καλύτερα αποτελέσματα με χρήση λεκτικής πληροφορίας δίνει το κριτήριο του significance percent, πιο συγκεκριμένα με εκπαίδευση με F1 all με βάση τις τιμές F1 for false και F1 all ενώ καλύτερο F1 σκορ πετυχαίνει η εκπαίδευση με F1 με χρήση λεκτικής πληροφορίας. Χωρίς λεκτική πληροφορία, το καλύτερο F1 all και F1 for false το επιτυγχάνει το κριτήριο significance percent με εκπαίδευση με F1 all και το καλύτερο F1 σκορ το κριτήριο των στιγμών.

ΣΥΓΚΡΙΣΗ ΚΡΙΤΗΡΙΩΝ ΓΙΑ SR



Γράφημα σύγκρισης κριτηρίων για τον αλγόριθμο SimRank.

Στο παραπάνω γράφημα βλέπουμε πώς διαμορφώνονται οι τιμές των μετρικών F1 all , F1, F1 for false, precision, recall, negative predicted value και selectivity με αυτή την σειρά για τον SimRank και για όλα τα κριτήρια χωρίς χρήση λεκτικής πληροφορίας αλλά και με χρήση λεκτικής πληροφορίας.

Αρχικά μπορούμε εύκολα να διακρίνουμε ότι για το κριτήριο των στιγμών όλες οι τιμές των μετρικών βγαίνουν 0. Αυτό συμβαίνει γιατί σχεδόν όλες οι τιμές των στιγμών με βάση τον αλγόριθμο SimRank, τόσο σε επίπεδο γκρουπ όσο και συνολικά για όλες τις ιστορίες συμπίπτουν στην τιμή 1 και έτσι δεν μπορεί να γίνει διαχωρισμός με βάση τις τιμές των στιγμών του SimRank.

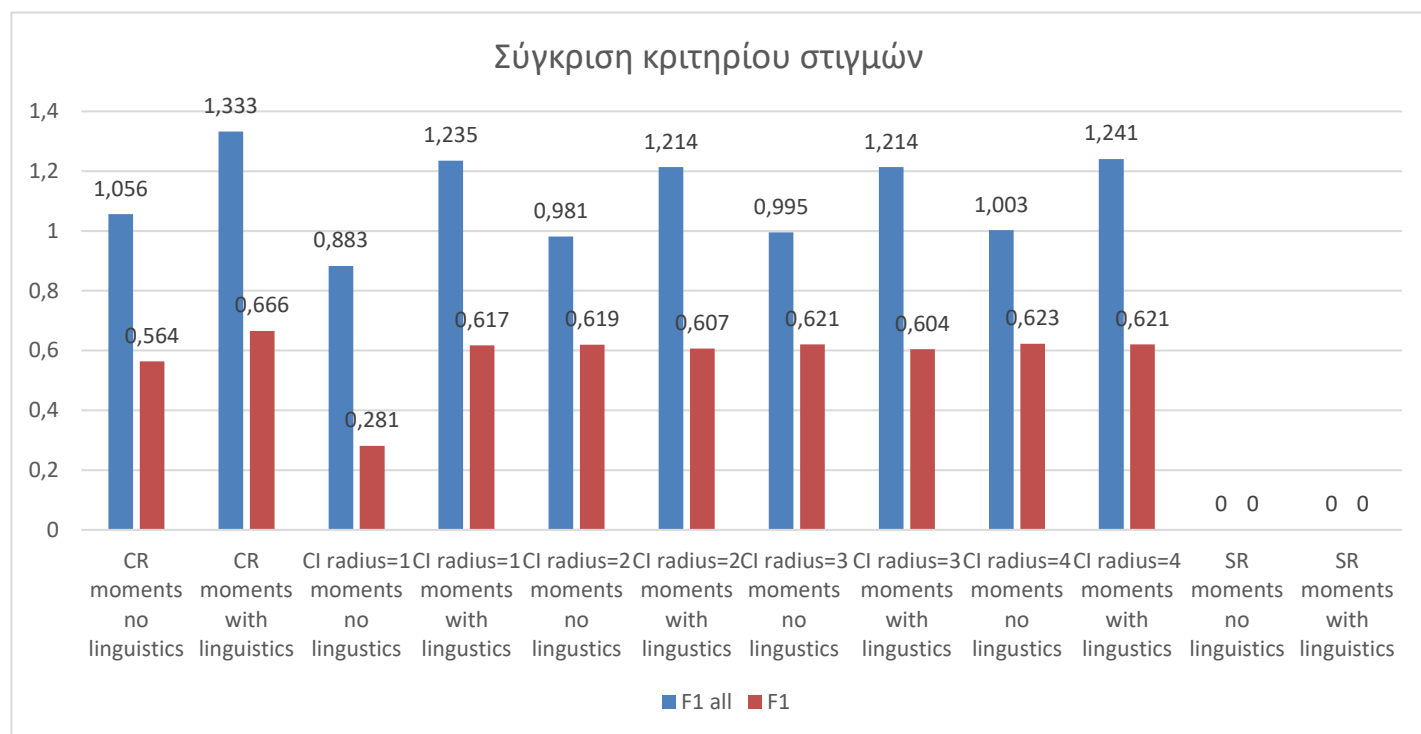
Έπειτα, φαίνεται ότι για όλα τα κριτήρια ισχύει ότι η χρήση λεκτικής πληροφορίας δίνει μεγαλύτερες τιμές σε όλες τις μετρικές και αρά σημαντικά καλύτερα αποτελέσματα. Αυτό συμβαίνει γιατί τα δομικά χαρακτηριστικά της διασποράς των ειδήσεων επηρεάζονται σε μεγάλο βαθμό από το είδος της θεματολογίας της είδησης. Έτσι εφαρμόζοντας διαφορετική κατηγοριοποίηση

για κάθε θεματολογία-γκρουπ και χρησιμοποιώντας κάθε φορά την κατηγοριοποίηση του πιο κοντινού γκρουπ σε θεματολογία προκύπτουν αρκετά καλύτερα αποτελέσματα.

Στη συνέχεια, μια πολύ σημαντική παρατήρηση είναι η σύγκριση των μετρικών για το κριτήριο του significance percent με εκπαίδευση με F1 all με το ίδιο κριτήριο αλλά με εκπαίδευση με F1 τόσο με λεκτική πληροφορία όσο και χωρίς. Στην περίπτωση της μη χρήσης λεκτικής πληροφορίας η εκπαίδευση με F1 all δεν παρουσιάζει καμία διαφορά στις τιμές των μετρικών σε σχέση με την εκπαίδευση με F1. Στη περίπτωση της λεκτικής πληροφορίας βλέπουμε ότι η εκπαίδευση με F1 all δίνει υψηλότερες τιμές σε F1 all, F1 for false, precision και selectivity καθώς χάνει λίγο σε recall, negative predicted value και F1 αφού έχει άμεση εξάρτηση από το recall. Αυτό συμβαίνει γιατί η εκπαίδευση με το κριτήριο F1 all αποσκοπεί στην βέλτιστη ισορροπία αναμεσά στα F1 και F1 for false, έτσι για να την πετύχει μπορεί να θυσιάσει ένα μικρό ποσοστό από αληθινές ιστορίες για να εξασφαλίσει ότι το μεγαλύτερο ποσοστό των ψευδών ειδήσεων (selectivity) θα χαρακτηριστεί ψευδής.

Τέλος παρατηρούμε ότι, καλύτερα αποτελέσματα με χρήση λεκτικής πληροφορίας δίνει το κριτήριο του significance percent και πιο συγκεκριμένα με εκπαίδευση με F1 all με βάση τις τιμές F1 for false και F1 all ενώ καλύτερο F1 σκορ πετυχαίνει το κριτήριο significance percent με εκπαίδευση με F1. Χωρίς χρήση λεκτικής πληροφορίας, το κριτήριο significance percent δίνει τα ίδια σκορ ανεξάρτητα με τη μετρική εκπαίδευσης.

4.2.2 Σύγκριση ανά κριτήριο



Στο παραπάνω διάγραμμα βλέπουμε τα σκορ F1 all με μπλε και τα σκορ F1 με κόκκινο των στιγμών τόσο με χρήση λεκτικής πληροφορίας όσο χωρίς χρήση λεκτικής για το κριτήριο των στιγμών και για όλες τις εκτελέσεις των αλγορίθμων μας.

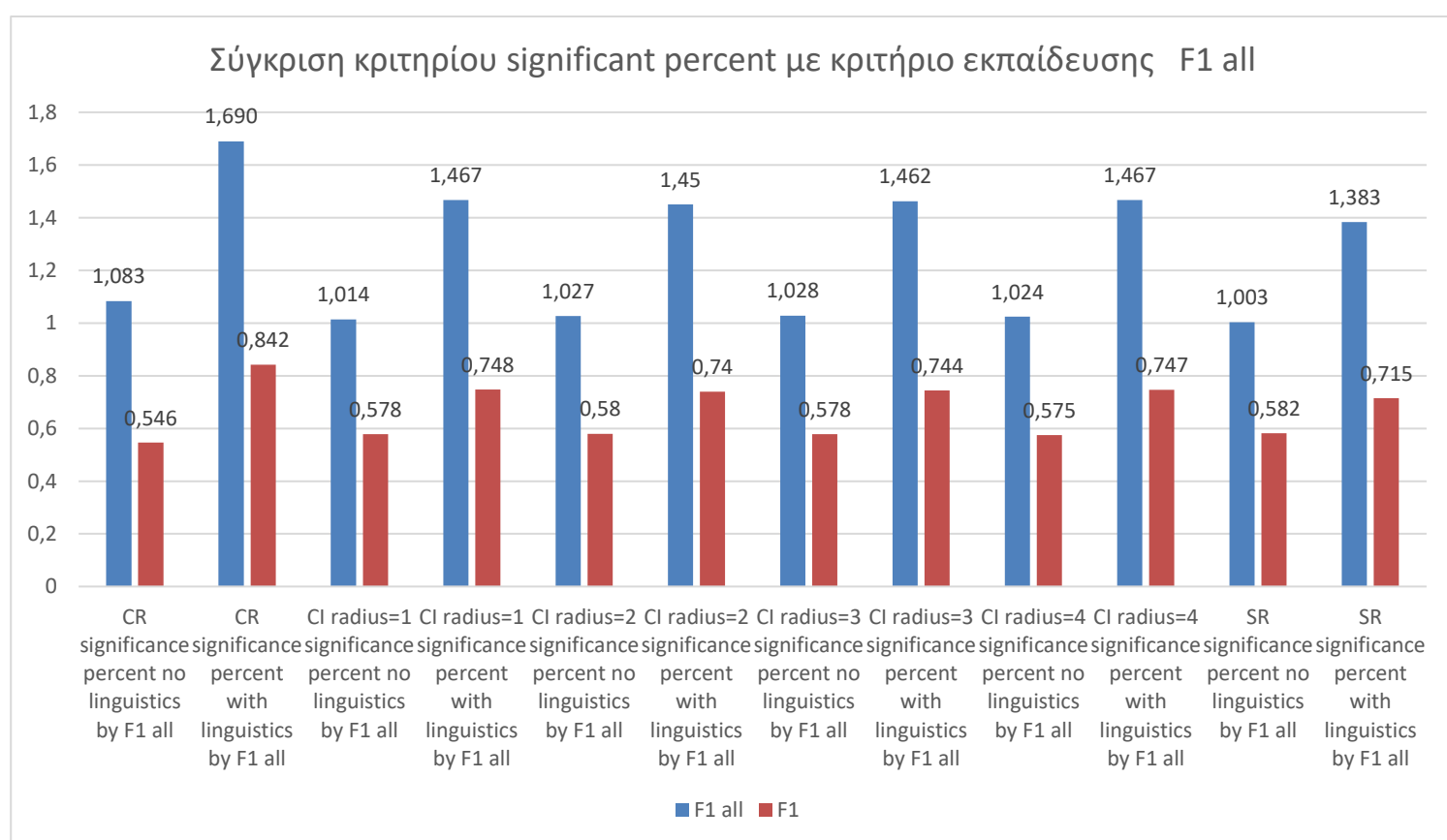
Αυτό που παρατηρείται πρώτα είναι ότι το κριτήριο των στιγμών για τον αλγόριθμο SimRank έχει τιμές 0. Αυτό συμβαίνει γιατί σχεδόν όλες οι τιμές των στιγμών με βάση τον αλγόριθμο SimRank, τόσο σε επίπεδο γκρουπ όσο και συνολικά για όλες τις ιστορίες συμπίπτουν στην τιμή 1 και έτσι δεν μπορεί να γίνει διαχωρισμός με βάση τις τιμές των στιγμών του SimRank.

Στη συνέχεια, είναι αρκετά ευδιάκριτο το γεγονός ότι η χρήση της λεκτικής πληροφορίας παρόλο που σε όλες τις περιπτώσεις προσδίδει καλύτερο F1 all, αυτό δεν συμβαίνει και με το F1 σκορ. Στις περιπτώσεις του CumulativeRank και Collective Influence με radius 1 βλέπουμε αύξηση του F1 σκορ ενώ στις περιπτώσεις του Collective Influence με radius 2,3,4 υπάρχει μια

μείωση του F1 σκορ και καθώς αυξάνεται το F1 all αυτό σημαίνει ότι με χρήση λεκτικής πληροφορίας αυξάνεται και το F1 for false.

Στην περίπτωση χρήσης λεκτικής πληροφορίας, ο αλγόριθμος που επιτυγχάνει το υψηλότερο το F1 all σκορ αλλά και το υψηλότερο F1 σκορ είναι ο CumulativeRank και στην συνέχεια ο Collective Influence με radius=4.

Στην περίπτωση μη χρήσης λεκτικής πληροφορίας το καλύτερο F1 all σκορ το επιτυγχάνει πάλι ο CumulativeRank και ακολουθεί πάλι ο Collective Influence με radius=4 ενώ το υψηλότερο F1 σκορ ο Collective Influence με radius=4 και ακολουθεί ο Collective Influence με radius=3.



Στο παραπάνω διάγραμμα βλέπουμε τα σκορ F1 all με μπλε και τα σκορ F1 με κόκκινο των στιγμών τόσο με χρήση λεκτικής πληροφορίας όσο χωρίς

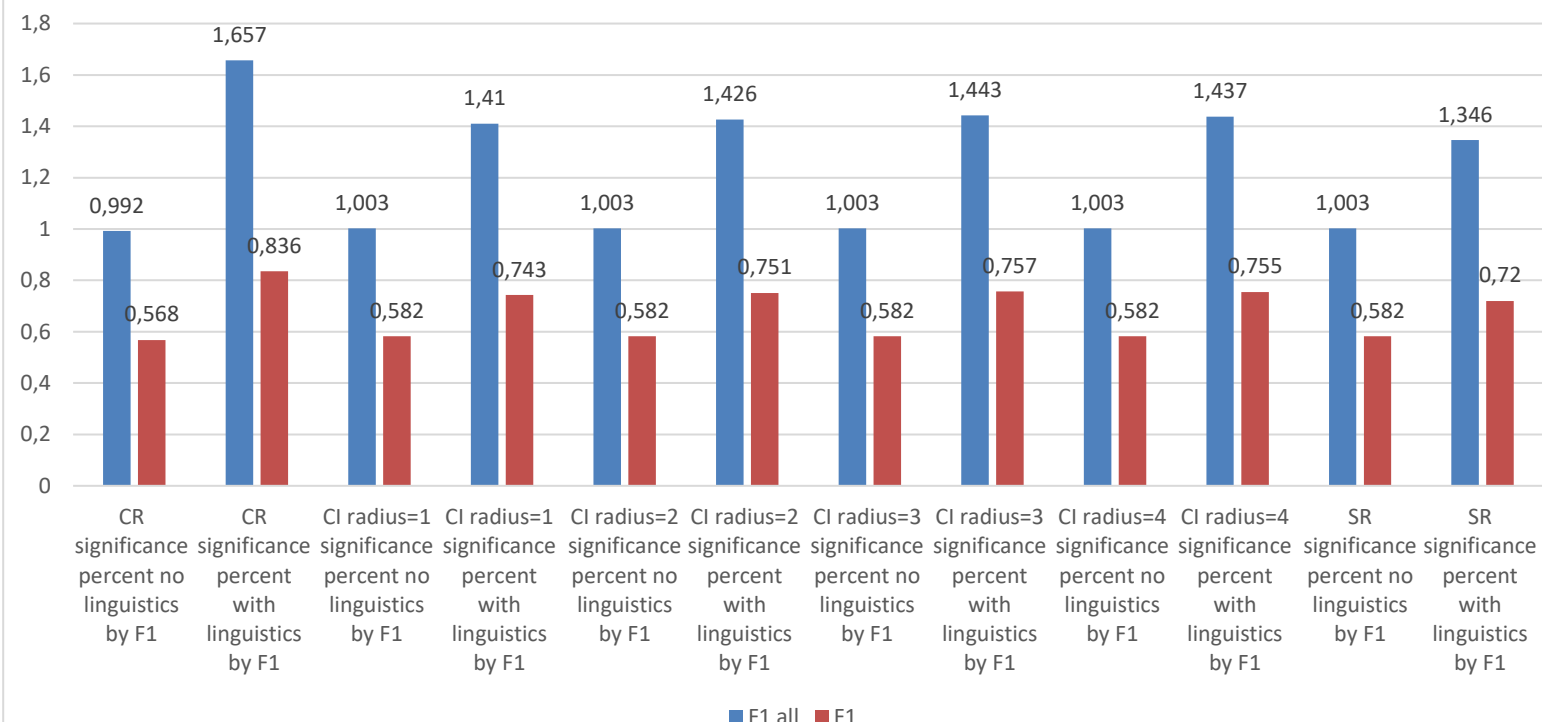
χρήση λεκτικής για το κριτήριο significance percent με εκπαίδευση με την μετρική F1 all και για όλες τις εκτελέσεις των αλγορίθμων μας.

Αρχικά αυτό που παρατηρούμε με μια πρώτη ματιά, είναι το γεγονός ότι η χρήση της λεκτικής πληροφορίας σε όλες τις περιπτώσεις προσδίδει καλύτερο F1 all αλλά και F1 σκορ. Με μια πιο προσεκτική ματιά αυτό που διακρίνεται είναι ότι οι τιμές F1 και F1 for false αυξάνονται ομοιόμορφα. Συγκεκριμένα η κάθε αύξηση είναι περίπου το μισό της αύξησης του F1 all σκορ αν λάβουμε υπόψιν τις αυξήσεις των τιμών F1 και F1 all , με την χρήση λεκτικής πληροφορίας αλλά και χωρίς.

Στην περίπτωση χρήσης λεκτικής πληροφορίας, ο αλγόριθμος που επιτυγχάνει το υψηλότερο το F1 all σκορ αλλά και το υψηλότερο F1 σκορ είναι ο CumulativeRank και στην συνέχεια ο Collective Influence με radius=1 που προηγείται με μικρή διαφορά λόγω του F1 σκορ από τον Collective Influence με radius=4.

Στην περίπτωση μη χρήσης λεκτικής πληροφορίας το καλύτερο F1 all σκορ το επιτυγχάνει πάλι ο CumulativeRank και ακολουθεί πάλι ο Collective Influence αλλά με radius=3, ενώ το υψηλότερο F1 σκορ επιτυγχάνει αυτή τη φορά ο SimRank που προηγείται με μικρή διαφορά από τον Collective Influence με radius=2. Παρ' όλα αυτά, είναι προτιμότερος ο Collective Influence με radius=2 γιατί η συνολική εικόνα του SimRank έχει δείξει ότι δεν είναι τόσο αξιόπιστος όσο οι άλλοι δύο αλγόριθμοι.

Σύγκριση κριτηρίου Significance percent με κριτήριο εκπαίδευσης το F1



Στο παραπάνω διάγραμμα βλέπουμε τα σκορ F1 all με μπλε και τα σκορ F1 με κόκκινο των στιγμών τόσο με χρήση λεκτικής πληροφορίας όσο χωρίς χρήση λεκτικής για το κριτήριο significance percent αλλά με εκπαίδευση αυτή την φορά με την μετρική F1 και για όλες τις εκτελέσεις των αλγορίθμων μας.

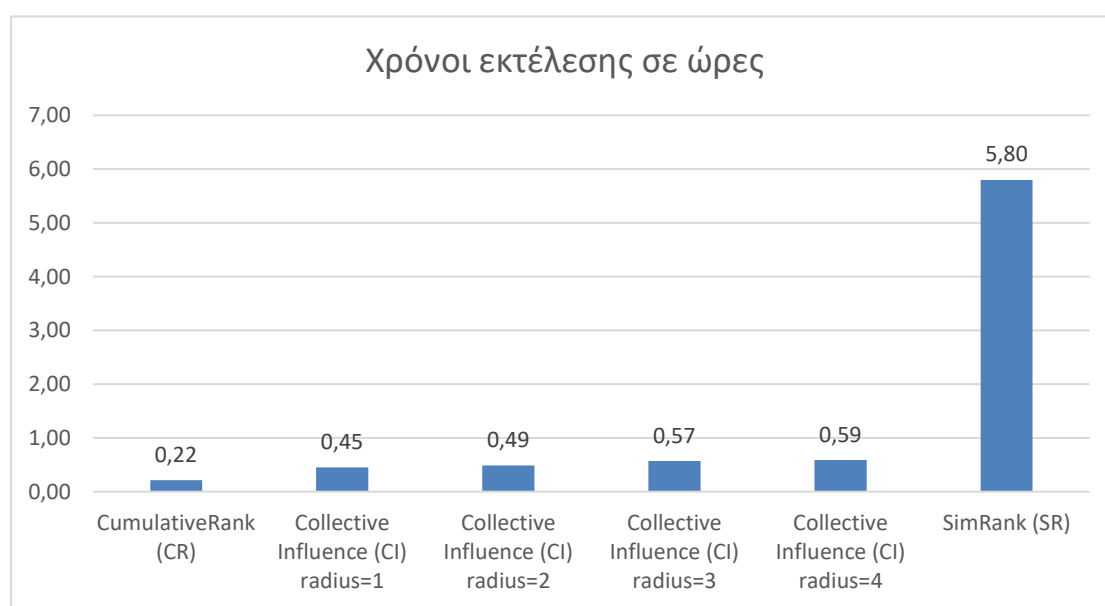
Αρχικά, αυτό που παρατηρούμε με μια πρώτη ματιά, είναι το γεγονός ότι η χρήση της λεκτικής πληροφορίας σε όλες τις περιπτώσεις προσδίδει καλύτερο F1 all αλλά και F1 σκορ. Με μια πιο προσεκτική ματιά αυτό που διακρίνεται είναι ότι οι τιμές F1 και F1 for false αυξάνονται σχεδόν ομοιόμορφα. Συγκεκριμένα παρατηρείται ελαφρώς μεγαλύτερη αύξηση του F1 for false σκορ σε σχέση με το F1, αν λάβουμε υπόψιν τις αυξήσεις των τιμών F1 και F1 all, με τη χρήση λεκτικής πληροφορίας αλλά και χωρίς.

Στην περίπτωση χρήσης λεκτικής πληροφορίας, ο αλγόριθμος που επιτυγχάνει το υψηλότερο το F1 all σκορ αλλά και το υψηλότερο F1 σκορ είναι ο CumulativeRank και στην συνέχεια ο Collective Influence με radius=3.

Στην περίπτωση μη χρήσης λεκτικής πληροφορίας το καλύτερο F1 all αλλά και F1 σκορ το επιτυγχάνουν όλες οι εκτελέσεις αλγορίθμων που επιτυγχάνουν το ίδιο F1 all αλλά και F1 σκορ εκτός από τον CumulativeRank που έρχεται τελευταίος με μικρή διαφορά και στα δύο σκορ.

4.2.3 Σύγκριση χρόνων εκτέλεσης

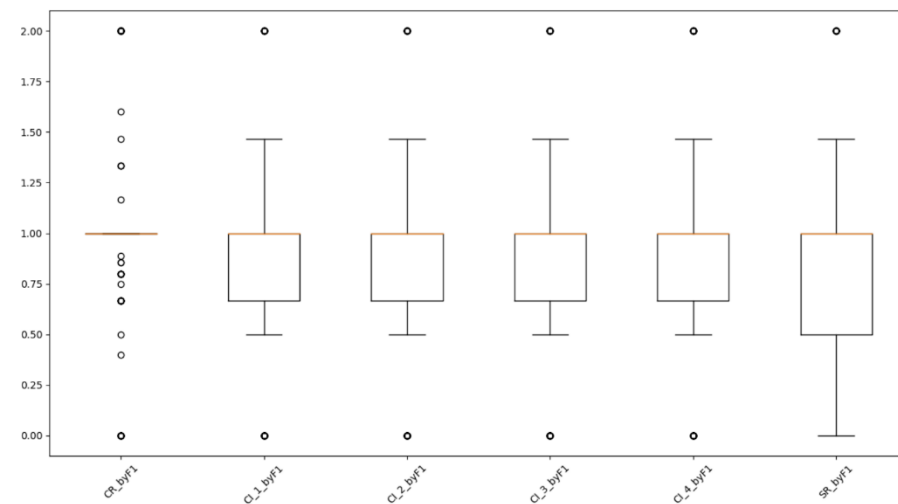
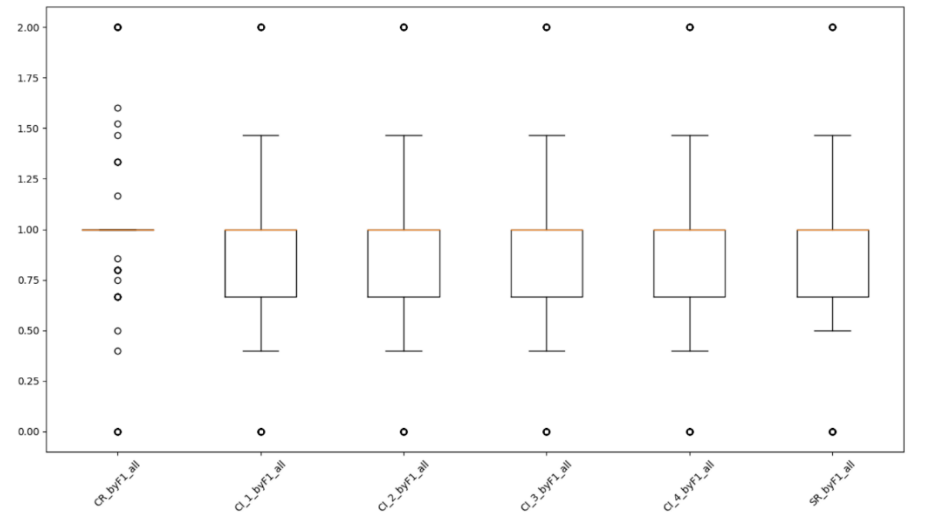
Εκτός από την ποιότητα των προτεινόμενων αλγορίθμων για την κατηγοριοποίηση των αναδυόμενων ιστοριών, σημαντικό ρόλο παίζει και ο χρόνος εκτέλεσής τους, ειδικά όταν έχουμε να διαχειριστούμε πολύ μεγάλα σύνολα δεδομένων. Στο παρακάτω γράφημα μπορούμε να δούμε αναλυτικά τους χρόνους εκτέλεσης όλων των εκδοχών εκτέλεσης των αλγορίθμων σε ώρες.



Όπως μπορούμε εύκολα να διακρίνουμε, τον μικρότερο χρόνο εκτέλεσης επιτυγχάνει ο αλγόριθμος CumulativeRank με 0.22 ώρες που αντιστοιχεί σε 12,94 λεπτά, και στη συνέχεια ακολουθούν οι τέσσερις εκδόσεις του Collective Influence, που όσο αυξάνεται η απόσταση της γειτονιάς τιμή radius, τόσο αυξάνεται με μικρό ρυθμό και ο χρόνος εκτέλεσης με την έκδοση με radius=1 να χρειάζεται 0.45 ώρες και την εκδοχή radius=2 να χρειάζεται 0.49 ώρες, με την έκδοση με radius=3 να χρειάζεται 0.57 ώρες και την έκδοση με radius=4 να χρειάζεται 0.59 ώρες.

Τέλος, επιβεβαιώνεται ότι ο αλγόριθμος SimRank δεν είναι καλή λύση ούτε από πλευρά χρόνου εκτέλεσης καθώς χρειάζεται 5.8 ώρες για να εκτελέσει τα 340 γκρουπ από το σύνολο δεδομένων μας.

4.2.4 Συνεισφορά των γκρουπ κατά την χρήση λεκτικής πληροφορίας



Box plots με τα σκορ F1all όλων των γκρουπ που συμμετέχουν στην εκτέλεση με κριτήριο significance percentage με χρήση λεκτικής πληροφορίας με εκπαίδευση με F1all στο πάνω γράφημα και με εκπαίδευση με F1 στο κάτω γράφημα.

Στο πρώτο γράφημα βλέπουμε τα box plots με τα σκορ F1all όλων των γκρουπ που συμμετέχουν στην εκτέλεση με κριτήριο significance percent με χρήση λεκτικής πληροφορίας με εκπαίδευση με F1all ενώ στο δεύτερο γράφημα βλέπουμε τα box plots με τα σκορ F1all όλων των γκρουπ που συμμετέχουν πάλι στην εκτέλεση με κριτήριο significance percent με χρήση λεκτικής πληροφορίας αλλά με εκπαίδευση με F1 σκορ αντί με F1all. Για τα συγκεκριμένα τυχαία επιλεγμένα σύνολα εκπαίδευσης και αξιολόγησης, τα γκρουπ που συμμετέχουν στην περίπτωση χρήσης λεκτικής πληροφορίας είναι 211 από τα οποία τα 93 συμμετέχουν τουλάχιστον δύο φορές ανεξάρτητα της περίπτωσης του αλγορίθμου. Η επικάλυψη των box plots και στις δύο περιπτώσεις μας δείχνει ότι σε κάθε μια από τις έξι εκτελέσεις τα ίδια γκρουπ ίσως πετυχαίνουν διαφορετικά σκορ και έτσι συμβάλουν ανάλογα το ποσοστό επιτυχίας που πετυχαίνουν. Το τελικό αποτέλεσμα των F1 all σκορ βγαίνει σαν σταθμισμένος μέσος όρος των F1 all σκορ που πετυχαίνουν και τον αριθμό των ιστοριών που καλείται να κατηγοριοποιήσει, λόγω ότι είναι πιο κοντά λεκτικά.

Έτσι, για κάθε ένα από τα γκρουπ που χρησιμοποιούνται, χρησιμοποιώντας κάθε φορά την κατηγοριοποίηση του αλγορίθμου που το γκρουπ στην εκτέλεσή του πετυχαίνει το μέγιστο F1 all σκορ από όλους τους άλλους αλγορίθμους, είναι δυνατόν να επιτευχθεί ακόμα καλύτερο αποτέλεσμα ως προς την ακρίβεια κατηγοριοποίησης των ειδήσεων σε πραγματικές και ψευδείς.

4.3 Σύνοψη και συμπεράσματα

Στην παρούσα διπλωματική προτείνουμε μία νέα network – based μέθοδο υπολογισμού της σημαντικότητας κόμβων στο δίκτυο διάδοσης μιας είδησης στα μέσα κοινωνικής δικτύωσης. Αξιολογήσαμε την προτεινόμενη μέθοδο σε σύγκριση με άλλες μεθόδους υπολογισμού σημαντικότητας, μέσα από τον αλγόριθμο κατηγοριοποίησης που αναπτύξαμε για τις αναδυόμενες ειδήσεις. Μοντελοποιήσαμε το δίκτυο αντιδράσεων των χρηστών και ερευνήσαμε τις τοπολογικές διαφορές με τρεις διαφορετικούς αλγορίθμους που δίνουν έξι περιπτώσεις εκτέλεσης, με τρία κριτήρια και με χρήση ή μη λεκτικής πληροφορίας. Ο CumulativeRank, ο αλγόριθμος που υλοποιήσαμε, αφορά την εύρεση των σημαντικών κόμβων σε ένα δίκτυο. Ο αλγόριθμος αυτός συγκρίθηκε με τον αλγόριθμο Collective Influence, ενώ ο SimRank αλγόριθμος επίσης παράγει μία μετρική ομοιότητας των κόμβων, την οποία χρησιμοποιήσαμε για

να βρούμε τους κόμβους οι οποίοι μοιάζουν με αυτούς που εκ των προτέρων γνωρίζουμε ότι συμβάλλουν πιο έντονα στη διάδοση της πληροφορίας στο δίκτυο (κόμβος πηγή).

Η αξιολόγηση των διαφορετικών αλγορίθμων σημαντικότητας κόμβων έγινε με το κριτήριο των στιγμών, καθώς και με το κριτήριο της σωρευτικής σημαντικότητας (significance percentage), με εκπαίδευση με F1all και με το ίδιο κριτήριο significance percent αλλά με εκπαίδευση με F1 τόσο με χρήση λεκτικής πληροφορίας όσο και χωρίς για τον CumulativeRank, για τον Collective Influence με radius=1,2,3,4 και για τον SimRank.

Για το κριτήριο των στιγμών αποδείχθηκε ότι, χάνει σε αποτελεσματικότητα σε σύγκριση με τον κριτήριο significance percent με εκπαίδευση με F1all αλλά και με εκπαίδευση με F1. Επίσης για τον αλγόριθμο SimRank, παρατηρήθηκε ότι αδυνατεί να κατηγοριοποιήσει τις ιστορίες αφού σχεδόν όλες οι τιμές των ιστοριών τόσο ανά γκρουπ όσο και συγκεντρωτικά συμπίπτουν στην μονάδα. Η χρήση λεκτικής πληροφορίας δίνει καλύτερο F1all και F1 για τον CumulativeRank ενώ σε όλες τις περιπτώσεις του Collective Influence δίνει καλύτερο F1all και μικρότερο recall και F1. Την καλύτερη απόδοση και για F1all=1,333 και για F1=0,666 για το κριτήριο των στιγμών με χρήση λεκτικής πληροφορίας την έχει ο CumulativeRank ενώ χωρίς χρήση λεκτικής πληροφορίας για μέγιστο F1all=1,056 ο CumulativeRank και για μέγιστο F1=0,623 ο Collective Influence με radius=4.

Για τα κριτήρια significance percent υπάρχουν δύο περιπτώσεις, με εκπαίδευση με F1all αλλά και με εκπαίδευση με F1. Με χρήση λεκτικής πληροφορίας έχουμε πολύ καλύτερα αποτελέσματα με το κριτήριο με εκπαίδευση με F1all υπερτερεί με F1all=1,690 (στα 2) και F1=0.842 και F1 for false = 0,848 από τον CumulativeRank. Χωρίς λεκτική πληροφορία πάλι υπερτερεί το significance percent με εκπαίδευση με F1all με σκορ F1all=1.083 από τον CumulativeRank ενώ σε μέγιστο F1 και τα δυο κριτήρια ισοβαθμούν, στο κριτήριο εκπαίδευσης F1all με τον SimRank με F1=0,582 με τον Collective Influence με radius=2 να είναι πολύ κοντά με F1=0,582. Στο κριτήριο significance percent με εκπαίδευση με F1 χωρίς λεκτικής πληροφορία όλοι αλγόριθμοι ισοβαθμούν με F1all=1,003 και F1=0,582. Η εκπαίδευση με F1all δίνει υψηλότερες τιμές σε F1all, F1 for false, precision, selectivity και σε κάποιες περιπτώσεις και σε negative predicted value (όχι πάντα) ενώ χάνει λίγο σε recall και προφανώς και σε κάποιες περιπτώσεις σε F1 αφού αλληλεξαρτώνται.

Τελικά, τα αποτελέσματα έδειξαν ότι η χρήση λεκτικής πληροφορίας αυξάνει όλα τα σκορ των μετρικών για όλους τους αλγορίθμους και επιτυγχάνει καλύτερη κατηγοριοποίηση για όλα τα κριτήρια τόσο σε F1all όσο και σε F1 σκορ.

Επιπλέον, η χρήση του F1all σαν μετρική εκπαίδευσης αντί του F1 στο κριτήριο significance percent πετυχαίνει καλύτερα αποτελέσματα σε F1all, F1 for false, precision, selectivity και σε κάποιες περιπτώσεις και σε negative predicted value (όχι πάντα) ενώ χάνει λίγο σε recall και προφανώς και σε κάποιες περιπτώσεις και σε F1. Αυτό συμβαίνει γιατί το F1all επιτυγχάνει την καλύτερη δυνατή ισορροπία ανάμεσα στα F1 και F1 for false, έτσι για να πετύχει καλύτερο διαχωρισμό θυσιάζει ένα μικρό ποσοστό αληθινών ιστοριών για να εξασφαλίσει ότι το μεγαλύτερο ποσοστό των ψευδών ειδήσεων (selectivity) θα χαρακτηριστεί ψευδής, κάτι το οποίο μας ωφελεί καθώς μας πειράζει περισσότερο να θεωρήσουμε ότι μια ψευδής είδηση είναι αληθής από το να θεωρήσουμε ότι μια αληθινή είδηση είναι πιθανόν ψευδής.

Από πλευρά αλγορίθμου σημαντικότητας που υλοποιήσαμε στο πλαίσιο της παρούσας διπλωματικής εργασίας, ο CumulativeRank πετυχαίνει τα καλύτερα αποτελέσματα και σε μικρότερο χρόνο εκτέλεσης, ενώ ακολουθούν οι εκτελέσεις του Collective Influence που σε γενικές γραμμές όσο αυξάνεται το radius αυξάνει την απόδοσή του αλλά και τον χρόνο εκτέλεσης του. Τελευταίος ο αλγόριθμος SimRank, του οποίου το μεγαλύτερο ποσοστό των διανυσμάτων εξόδου βγαίνει στην μονάδα. Για τον SimRank οι στιγμές αδυνατούν να πραγματοποιήσουν κατηγοριοποίηση. Το κριτήριο significance percent υπολογίζεται από τον αριθμό των κόμβων που κρατάνε ένα ποσοστό σημαντικότητας διαιρεμένο με το μέγεθος της ιστορίας για να μην εξαρτάται η κατηγοριοποίηση από τα μεγέθη των ιστοριών. Ο λόγος που το κριτήριο significance percent πετυχαίνει κατηγοριοποίηση στα διανύσματα μονάδας του SimRank είναι επειδή εκμεταλλεύεται τα μεγέθη των ιστοριών. Έτσι παρόλο που βγάζει μια καλή λύση με το κριτήριο significance percent, δεν αποτελεί αξιόπιστο αλγόριθμο για κατηγοριοποίηση καθώς εκμεταλλεύεται κατά κύριο λόγο μόνο τα μεγέθη.

4.4 Μελλοντικές επεκτάσεις

Εκτός των παραπάνω, στα σχέδια μας ήταν να χρησιμοποιήσουμε και ένα ακόμα κριτήριο, την κεντρικότητα *betweenness centrality* το οποίο δεν προλάβουμε να υλοποιήσουμε. Επίσης η επικάλυψη των *boxplots* είτε για F1 είτε για F1 all, για όλους τους αλγόριθμους για το κριτήριο *Significant percent* με χρήση λεκτικής πληροφορίας (*with linguistics*), ανοίγει το δρόμο για την διερεύνηση κατασκευής μιας υβριδικής μεθόδου που για κάθε ιστορία αφού δούμε σε πιο γκρουπ είναι πιο κοντά λεκτικά, θα χρησιμοποιείται η κατηγοριοποίηση του συγκεκριμένου γκρουπ αλλά για την εκδοχή του αλγορίθμου που το συγκεκριμένο γκρουπ έχει καλύτερο F1 all σκορ, επιτυγχάνοντας ακόμα καλύτερα αποτελέσματα.

Συνεπώς, σε επόμενη φάση θα θέλαμε να διερευνήσουμε την πηγή του προβλήματος και με τη χρήση της κεντρικότητας *betweenness centrality* και να προχωρήσουμε στην κατασκευή της υβριδικής μεθόδου με κριτήριο *Significant percent* με χρήση λεκτικής πληροφορίας (*with linguistics*).

Επιπλέον, θα θέλαμε να διερευνήσουμε και την εκπαίδευση του κριτηρίου *significance percentage* με βάση το F1 for fake σε σχέση με την εκπαίδευση του κριτηρίου με βάση το F1 all.

Τέλος, θα θέλαμε να πραγματοποιήσουμε την παραπάνω διαδικασία και για τα γραφήματα που προκύπτουν από τα αρχικά γραφήματα αξιοποιώντας μόνο τους κόμβους που αντιπροσωπεύουν αντιδράσεις: μόνο των πρώτων 100 ή 200 αντιδράσεων, μόνο της πρώτης ώρας, μόνο των πρώτων 6 ωρών, 12 ωρών και τέλος της πρώτης μέρας, έτσι ώστε να αξιολογήσουμε την αποτελεσματικότητα της μεθόδου μας σε πραγματικό χρόνο και στις πρώιμες φάσεις των ειδήσεων (*fake news early detection*).

Βιβλιογραφία

- Alexios Polyzos, I. P. (2019). *Measuring the Propagation Speed of Information Spread in Social Networks for Real-Time Fake News Detection. Diploma Thesis, Department of Computer Science & Engineering, University of Ioannina, September 2019.*
- Arkaitz Zubiaga. (2016). PHEME dataset of rumours and non-rumours 2016 URL https://figshare.com/articles/dataset/PHEME_dataset_of_rumours_and_non-rumours/4010619. *Learning Reporting Dynamics during Breaking News for Rumour Detection in Social Media' for rumours detection.*
- Arkaitz Zubiaga, M. L. (2016). Learning Reporting Dynamics during Breaking News for Rumor Detection in Social Media 2016 URL <https://arxiv.org/pdf/1610.07363.pdf>.
- Arkaitz Zubiaga, M. L. (2016, April). PHEME rumour scheme dataset: journalism use case URL https://figshare.com/articles/dataset/PHEME_rumour_scheme_dataset_journalism_use_case/2068650.
- Bourogianni, k. (2020). *Algorithmic techniques for detecting structural properties of the propagation of news data on social media Diploma Thesis, Department of Computer Science & Engineering, University of Ioannina, august 2020.*
- Carlos Castillo, M. M. (2011, March). Information credibility on twitter. In Proceedings of the 20th international conference on World wide web. ACM. March 2011. pp. 675–684 URL <https://dl.acm.org/doi/10.1145/1963405.1963500>.
- Claire Wardle, H. D. (2017, September). Information disorder: Toward an interdisciplinary framework for research and policy making URL https://scholar.google.com/citations?view_op=view_citation&hl=en&user=vpPqsRIAAAAJ&citation_for_view=vpPqsRIAAAAJ:qjMakFHDy7sC.

- Craig Silverman, Z. (2015). Lies, damn lies and viral content. Tow Center for Digital Journalism, Columbia Journalism School. URL <https://doi.org/10.7916/D8Q81RHH>.
- Dayong Zhang, Y. W. (2019). CumulativeRank Identifying and quantifying potential super-spreaders in social networks URL <https://www.nature.com/articles/s41598-019-51153-5.pdf>. *nature.com*.
- Elena Kochkina, M. L. (2018, june). PHEME dataset for Rumour Detection and Veracity Classification URL https://figshare.com/articles/dataset/PHEME_dataset_for_Rumour_Detection_and_Veracity_Classification/6392078.
- Flaviano Morone, B. M. (2016). Collective Influence Algorithm to find influencers via optimal percolation in massively large social media URL <https://www.nature.com/articles/srep30062>. *nature.com*.
- Kai Shu, A. S. (2017). Fake News Detection on Social Media: A Data Mining Perspective, 2017. URL <https://arxiv.org/pdf/1708.01967.pdf>.
- Karishma Sharma, F. Q. (2018, August). Combating Fake News: A Survey on Identification and Mitigation Techniques. ACM Trans. Intell. Syst. Technol. 37, 4, Article 111, URL <https://arxiv.org/pdf/1901.06437.pdf>.
- Ma et al, I. E. (2017). Twitter Dataset 2017 includes Twitter15 and Twitter16 datasets, URL https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?file_subpath=%2Frumor_detection_acl2017.
- Rao, P. K. (2017). Fake News Challenge: FCN- URL <http://www.fakenewschallenge.org/>.
- S. Vosoughi, D. R. (2018, March). The spread of true and false news online, Science, vol. 359, no. 6380, pp. 1146–1151, March 2018. URL: <http://dx.doi.org/10.1126/science.aap9559>.
- Shu et al. (2018). Shu et al. FakeNewsNet 2018 Github URL <https://github.com/KaiDMML/FakeNewsNet>.
- Vahed Qazvinian, E. R. (2011, july). Rumor has it: Identifying misinformation in microblogs. Conference on Empirical Methods Natural Language Processing, pp 1589–1599, Edinburgh, Scotland, Association for Computational Linguistics <https://aclanthology.org/D11-1147.pdf>.
- Zhe Zhao, P. R. (2015, March). Enquiring minds: Early detection of rumors in social media from enquiry posts. In Proceedings of the 24th International Conference

on World Wide 1395–1405. URL
<https://dl.acm.org/doi/10.1145/2736277.2741637>.

