ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Ανίχνευση παραβιασμένων λογαριασμών σε κοινωνικά δίκτυα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Σαραφίδη Λάζαρου

Επιβλέπων Καθηγητής: Σταματάτος Ευστάθιος

Μέλη εξεταστικής επιτροπής: Κωστούλας Θεόδωρος, Συμεωνίδης Παναγιώτης

Σάμος, Σεπτέμβριος 2021

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πρόλογος και ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο ολοκλήρωσης των σπουδών μου στο τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων του Πανεπιστημίου Αιγαίου.

Η επιτυχής ολοκλήρωση αυτής δεν θα μπορούσε να έχει γίνει χωρίς την πολύτιμη βοήθεια του επιβλέποντα καθηγητή μου κ. Ευστάθιου Σταματάτου, τον οποίο και θέλω να ευχαριστήσω θερμά, που με εμπιστεύτηκε γι' αυτό το θέμα και μου παρείχε όλες τις γνώσεις και την σωστή καθοδήγηση για την ολοκλήρωση της εργασίας.

Ακόμη, θα ήθελα να ευχαριστήσω την μητέρα μου και τον αδερφό μου Μιχάλη, για την πολύτιμη βοήθειά τους καθ' όλη την διάρκεια των σπουδών μου.

του Σαραφίδη Λάζαρου

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Η σελίδα αυτή είναι σκόπιμα λευκή.

Περιεχόμενα

Λίστα Εικόνων

Ακρωνύμια

XSS	Cross-Site Scripting
AV	Authorship Verification
NLP	Natural Language Processing
AI	Artificial Intelligence
SVM	Support Vector Machine
KDD	Knowledge Discovery from Databases
kNN	k-Nearest Neighbors
ROC	Receiver Operating Characteristic curve
AUC	Area Under the Curve
RT	Retweet

Η σελίδα αυτή είναι σκόπιμα λευκή.

Περίληψη

Η περιοχή της ανάλυσης συγγραφέα (authorship analysis) και, πιο συγκεκριμένα, της επαλήθευσης συγγραφέα (authorship verification) είναι ένα πεδίο που βρίσκει εφαρμογή σε ολοένα και περισσότερους τομείς, καθώς είναι εφικτό να χρησιμοποιηθεί για την ανάλυση οποιουδήποτε είδους κειμένων (π.χ. λογοτεχνικών έργων, άρθρων, αναρτήσεων σε κοινωνικά δίκτυα κ.λπ). Κατά την επαλήθευση συγγραφέα, δεδομένου ενός συνόλου κειμένων (γνωστά κείμενα) από τον ίδιο συγγραφέα, καλούμαστε να αποφασίσουμε αν ένα νέο κείμενο (άγνωστο κείμενο) έχει γραφτεί από τον συγγραφέα αυτόν ή όχι.

Ένα από τα πεδία που βρίσκει εφαρμογή η επαλήθευση συγγραφέα, είναι στην ανίχνευση παραβιασμένων λογαριασμών στα κοινωνικά δίκτυα. Είναι υψίστης σημασίας οι λογαριασμοί αυτοί να εντοπίζονται έγκαιρα και να «επιστρέφονται» στους έγκυρους χρήστες τους, γιατί οι παραβιασμένοι λογαριασμοί αποτελούν μέσο επιθέσεων των κακόβουλων χρηστών.

Στόχος της παρούσας διπλωματικής εργασίας είναι να μελετήσει την δυνατότητα εφαρμογής της επαλήθευσης συγγραφέα σε συνδυασμό με διάφορες μεθόδους μηχανικής μάθησης, στην ανίχνευση παραβιασμένων λογαριασμών σε κοινωνικά δίκτυα. Στο πλαίσιο αυτό, αξιοποιήθηκε ένα σύνολο δεδομένων που περιλάμβανε αναρτήσεις χρηστών από το Twitter (tweets), κι έπειτα, μετά από κατάλληλη προ-επεξεργασία και ανάλυση, έγιναν κάποια πειράματα εφαρμογής των ανωτέρω μεθόδων και παρουσιάστηκαν τα αποτελέσματά τους, τα οποία ήταν αρκετά ικανοποιητικά. Τέλος, σχολιάστηκαν τα αποτελέσματα αυτά, καθώς και μελλοντικές προοπτικές του συγκεκριμένου πεδίου.

Λέξεις Κλειδιά: *Παραβιασμένοι λογαριασμοί, Επαλήθευση συγγραφέα, Κοινωνικά δίκτυα*

Abstract

Authorship analysis and, more specifically, authorship verification is a research area that is being applied in more and more fields, as it can be used for the analysis of any genre of texts (e.g. literary works, articles, posts on social networks, etc.). In the authorship verification process, given a set of texts (known texts) by the same author, we have to decide whether a new text (unknown text) was written by that author or not.

Compromised account detection in social networks, is a field in which authorship verification can be applied. It is of utmost importance for these accounts to be identified quickly and "returned" to their legitimate users, because compromised accounts are a means of attack by malicious users.

The goal of this dissertation is to study the possibility of applying author verification in combination with various machine learning methods, in compromised account detection. In this context, a dataset that included posts from Twitter (tweets) was deployed and then, after proper pre-processing and analysis, some experiments were performed and their results were presented, which were very promising. Finally, these results were commented, as well as future perspectives in this field.

Keywords: *Compromised accounts, Authorship verification, Online social networks*

1

Εισαγωγή

Τα **μέσα κοινωνικής δικτύωσης** έχουν υπερισχύσει στην επιρροή έναντι των παραδοσιακών μέσων επικοινωνίας, όπως τα γράμματα, τα τηλεγραφήματα ή άλλες πρακτικές εκτός διαδικτύου. Το διαδίκτυο αποτελεί πλέον αναγκαίο μέρος της ανθρώπινης ζωής, έχοντας κάνει τους ανθρώπους να βασίζονται περισσότερο σε διαδικτυακές πηγές για την επικοινωνία και την διάδοση πληροφοριών, όπως εφαρμογές ανταλλαγής άμεσων μηνυμάτων, συγγραφή ιστολογίων (blogs), ιστότοπους κοινωνικής δικτύωσης κ.λπ.

Με τον τεράστιο όγκο δεδομένων που υπάρχουν στο διαδίκτυο στις μέρες μας, έχει παρατηρηθεί και το φαινόμενο της **αύξησης κακής χρήσης των πληροφοριών**, όπως η διάδοση ανεπιθύμητων μηνυμάτων, προσβλητικού περιεχομένου και απειλών. Ως εκ τούτου, έχει καταστεί απαραίτητο να γνωρίζουμε την αυθεντικότητα ενός μηνύματος που δημοσιεύεται σε διάφορες διαδικτυακές πλατφόρμες όπως ιστολόγια, ιστότοπους κοινωνικών δικτύων, φόρουμ ειδήσεων ή άλλες τέτοιες πλατφόρμες. Η ανώνυμη εξάπλωση ορισμένων ευαίσθητων και επικίνδυνων δραστηριοτήτων, όπως το «τρολινγκ» (trolling), τα σαρκαστικά σχόλια, οι φήμες, η διασπορά αρνητικών πληροφοριών, οι απειλές, οι άσχημες κριτικές και οι άσεμνες δημοσιεύσεις συνήθως περιλαμβάνονται σε ποινικές έρευνες όπου καθίσταται εξαιρετικά σημαντικό να γνωρίζουμε τον αρχικό/πραγματικό συγγραφέα της ανάρτησης.

Για την αντιμετώπιση του προβλήματος της κακόβουλης δραστηριότητας στα κοινωνικά δίκτυα, οι ερευνητές, ασχολήθηκαν με τον **εντοπισμό ψεύτικων λογαριασμών** (fake accounts - δηλαδή, λογαριασμοί που έχουν δημιουργηθεί αυτόματα, των οποίων ο κύριος στόχος είναι να κάνουν κάποια κακόβουλη ενέργεια). Ωστόσο, το πρόβλημα παραμένει αφού τα συστήματα που εντοπίζουν αποκλειστικά ψεύτικους λογαριασμούς δεν μπορούν να τους ξεχωρίσουν τους ψεύτικους από τους παραβιασμένους λογαριασμούς (compromised accounts).

Ένας **παραβιασμένος λογαριασμός** είναι ένας έγκυρος λογαριασμός που έχει πάρει τον έλεγχο ένας εισβολέας για να δημοσιεύσει ψεύτικο ή/και επιβλαβές περιεχόμενο. Υπάρχουν πολλά

παραδείγματα λογαριασμών διάσημων προσώπων και οργανισμών, οι οποίοι έχουν μεγάλη επιρροή, που μεταδίδουν fake news.

Οι λογαριασμοί μπορούν να παραβιαστούν με πολλούς διαφορετικούς τρόπους, μερικοί από αυτούς περιλαμβάνουν την εκμετάλλευση μιας ευπάθειας των ιστοσελίδων (cross-site scripting – XSS) ή χρησιμοποιώντας «ηλεκτρονικό ψάρεμα» (phishing) για να κλέψουν τα διαπιστευτήρια (credentials) των χρηστών. Μια άλλη επιλογή των επιτιθέμενων για την απόκτηση ελέγχου σε έγκυρους λογαριασμούς, είναι η χρήση bots για την απόκτηση των διαπιστευτηρίων των χρηστών, για πρόσβαση σε ιστότοπους κοινωνικών δικτύων με μολυσμένους εξυπηρετητές (hosts).

Δεδομένου ότι οι ψεύτικοι λογαριασμοί δημιουργούνται κυρίως για να προκαλέσουν ζημιά στα κοινωνικά δίκτυα, η απλούστερη λύση είναι να βρεθούν και να διαγραφούν. Από την άλλη, οι παραβιασμένοι λογαριασμοί αποτελούν ένα πιο περίπλοκο πρόβλημα, αφού πρέπει να ανακτηθούν τα διαπιστευτήρια των χρηστών, για να επιστρέψουν τον έλεγχο των λογαριασμών στους αντίστοιχους κατόχους τους [1]. Στην πραγματικότητα, μια έρευνα έδειξε ότι το 27% των χρηστών των οποίων παραβιάστηκε ο λογαριασμός, έκανε νέο λογαριασμό [2]. Αυτό το ποσό είναι αξιοσημείωτο, καθώς τα κοινωνικά δίκτυα, όπως το Twitter, συνιστά στους χρήστες που έχει παραβιαστεί ο λογαριασμός τους, να προσπαθούν να ανακτήσουν τα διαπιστευτήριά τους.

Εν συνεχεία, η παραπάνω έρευνα έδειξε ότι οι παραβιασμένοι λογαριασμοί είναι ο πιο χρησιμοποιούμενος τρόπος για τη διάδοση δόλιου περιεχομένου. Επιπλέον, μια άλλη έρευνα έδειξε ότι μόνο το 16% των λογαριασμών που στέλνουν ανεπιθύμητα μηνύματα (spamming accounts) ήταν όντως ψεύτικοι λογαριασμοί [3]. Όλοι οι υπόλοιποι ήταν παραβιασμένοι λογαριασμοί. Το ίδιο παρατηρήθηκε στο Facebook όπου το 97% των κακόβουλων λογαριασμών δεν δημιουργήθηκαν αποκλειστικά για spamming, στην πραγματικότητα, ήταν πρώην έγκυροι λογαριασμοί [4]. Όλες αυτές οι έρευνες επιβεβαιώνουν ότι **ένας από τους πιο συχνούς τρόπους διάδοσης κακόβουλου περιεχομένου είναι μέσω παραβιασμένων λογαριασμών**.

Γι’ αυτόν τον λόγο, πολλές έρευνες έχουν στρέψει το ενδιαφέρον τους σε τέτοιες περιπτώσεις για την ανίχνευση τέτοιων λογαριασμών. Το πεδίο της ανάλυσης συγγραφέα, είναι ένα από τα υποσχόμενα πεδία, που μπορεί να βοηθήσει ιδιαίτερα στην ανίχνευση παραβιασμένων λογαριασμών.

Ο όρος που επινοήθηκε ως **ανάλυση συγγραφέα** (**authorship analysis**) περιλαμβάνει τη διερεύνηση διάφορων χαρακτηριστικών ενός κειμένου για τον προσδιορισμό της συγγραφικής ιδιότητάς του. Αν και η ανάλυση συγγραφέα έχει μια ευρεία βάση εφαρμογών, όπως σε λογοτεχνικά έργα, προβλήματα απο-ανωνυμοποίησης (De-anonymization), ταυτοποίηση συγγραφέα πηγαίου κώδικα, και πολλά άλλα, πρόσφατα, έχει λάβει μεγάλη προσοχή στην ανάλυση μέσων κοινωνικών δικτύωσης (π.χ. tweets, μηνύματα, δημοσιεύσεις) όπου μια από τις κύριες προκλήσεις είναι να μπορεί να λειτουργεί με σύντομα κείμενα. Επίσης, σε εγκληματικές δραστηριότητες συνήθως χρησιμοποιούνται σύντομα μηνύματα και ως εκ τούτου η εφαρμογή

της ανάλυσης συγγραφέα στις έρευνες εγκλήματος φαίνεται να είναι ιδιαίτερα σημαντική. Οι έρευνες εγκλημάτων μπορεί να περιλαμβάνουν την ανίχνευση εισβολέων (hackers), παραβιασμένων λογαριασμών, αναρτήσεων ηλεκτρονικού ψαρέματος (phishing), παράνομων δραστηριοτήτων κ.λπ.

Ένας από τους πιο πρόσφατους ερευνητικούς τομείς στην επεξεργασία φυσικής γλώσσας (Natural Language Processing – NLP) είναι η ανάλυση συγγραφέα που προσπαθεί να αξιοποιήσει την υπολογιστική δύναμη των μεγάλων δεδομένων (big data) και της τεχνητής νοημοσύνης (artificial intelligence – AI) σε συνδυασμό με τη γλωσσολογία και τη γνωστική ψυχολογία για να κωδικοποιήσει την αυτόματη ταξινόμηση των κειμένων, τον προσδιορισμό των προφίλ συγγραφέων και την επίλυση των συγχύσεων μεταξύ συγγραφέων [5].

Στην συνέχεια, θα αποσαφηνιστούν οι όροι οι οποίοι κρίνονται απαραίτητοι για την καλύτερη κατανόηση της παρούσας διπλωματικής εργασίας, και οι οποίοι δεν έχουν εξηγηθεί προηγουμένως.

2

Μηχανική Μάθηση

Η **μηχανική μάθηση** (**machine learning**) είναι μια εφαρμογή της τεχνητής νοημοσύνης (artificial intelligence – AI) που παρέχει στα συστήματα τη δυνατότητα αυτόματης μάθησης και βελτίωσης από την εμπειρία χωρίς να προγραμματίζονται ρητά. Η μηχανική μάθηση επικεντρώνεται στην ανάπτυξη προγραμμάτων που μπορούν να έχουν πρόσβαση σε δεδομένα και να τα χρησιμοποιούν για να μάθουν. Η διαδικασία της μάθησης ξεκινά με παρατηρήσεις ή δεδομένα, όπως παραδείγματα, άμεση εμπειρία ή οδηγίες, προκειμένου να αναζητηθούν μοτίβα στα δεδομένα και να ληφθούν καλύτερες αποφάσεις στο μέλλον με βάση τα παραδείγματα που παρέχουμε. Ο πρωταρχικός στόχος είναι να επιτρέπεται στους υπολογιστές να μαθαίνουν αυτόματα χωρίς ανθρώπινη παρέμβαση ή βοήθεια και να προσαρμόζουν ανάλογα τις ενέργειες.

Οι αλγόριθμοι μηχανικής μάθησης κατηγοριοποιούνται συχνά σε επιβλεπόμενη μάθηση (supervised learning), μη-επιβλεπόμενη μάθηση (unsupervised learning), ημι-επιβλεπόμενη μάθηση (semi-supervised learning) και ενισχυτική μάθηση (reinforcement learning).

- Οι αλγόριθμοι που βασίζονται στην **επιβλεπόμενη μάθηση** μπορούν να εφαρμόσουν κάτι που έχουν μάθει στο παρελθόν, σε νέα δεδομένα χρησιμοποιώντας επισημασμένα παραδείγματα για την πρόβλεψη μελλοντικών γεγονότων. Ξεκινώντας από την ανάλυση ενός γνωστού συνόλου δεδομένων εκπαίδευσης (training set), ο αλγόριθμος εκμάθησης παράγει μια συνάρτηση για να κάνει προβλέψεις σχετικά με τις τιμές εξόδου. Το σύστημα είναι σε θέση να προβλέψει για οποιαδήποτε νέα είσοδο μετά από επαρκή εκπαίδευση. Ο αλγόριθμος εκμάθησης μπορεί επίσης να συγκρίνει την έξοδό του με τη σωστή προοριζόμενη έξοδο και να βρει σφάλματα προκειμένου να τροποποιήσει ανάλογα το μοντέλο.

  Η επιβλεπόμενη μάθηση απαιτεί από τον επιστήμονα να εκπαιδεύσει τον αλγόριθμο τόσο με δεδομένα με ετικέτα (labeled data) όσο και με τα επιθυμητά αποτελέσματα. Οι αλγόριθμοι επιβλεπόμενης μάθησης είναι καλοί για τις ακόλουθες εργασίες:

- Δυαδική ταξινόμηση (Binary classification): Διαίρεση δεδομένων σε δύο κατηγορίες.
- Ταξινόμηση πολλαπλών κλάσεων (Multi-class classification): Επιλογή μεταξύ περισσότερων από δύο τύπων απαντήσεων.
- Μοντελοποίηση με χρήση παλινδρόμησης (regression modeling): Πρόβλεψη συνεχών τιμών.
- Συνδυασμό προβλέψεων (Ensembling): Στον συνδυασμό προβλέψεων από διάφορα μοντέλα μηχανικής μάθησης για την παραγωγή μιας πιο ακριβής πρόβλεψης.

- Αντίθετα, οι αλγόριθμοι **μη-επιβλεπόμενης μάθησης** χρησιμοποιούνται όταν τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση είναι χωρίς ετικέτα (unlabeled data) ούτε έχει αποδοθεί κάποια κλάση γι' αυτά. Η μη-επιβλεπόμενη μάθηση μελετά πώς τα συστήματα μπορούν να συμπεράνουν μια συνάρτηση για να περιγράψουν μια κρυφή δομή από μη κατηγοριοποιημένα δεδομένα. Το σύστημα δεν πάει να προβλέψει τη σωστή έξοδο, αλλά διερευνά τα δεδομένα και μπορεί να εξαγάγει συμπεράσματα από σύνολα δεδομένων για να περιγράψει κρυφές δομές από μη κατηγοριοποιημένα δεδομένα.

  Οι αλγόριθμοι μη-επιβλεπόμενης μάθησης δεν απαιτούν την κατοχή δεδομένων με ετικέτα. Φιλτράρουν δεδομένα χωρίς ετικέτα για να αναζητήσουν μοτίβα που μπορούν να χρησιμοποιηθούν για την ομαδοποίηση δεδομένων σε υποσύνολα. Οι περισσότεροι τύποι βαθιάς μάθησης (deep learning), συμπεριλαμβανομένων των νευρωνικών δικτύων (neural networks), είναι αλγόριθμοι μη-επιβλεπόμενης μάθησης. Οι αλγόριθμοι μη-επιβλεπόμενης μάθησης είναι καλοί για τις ακόλουθες εργασίες:

  - Ομαδοποίηση (Clustering): Διαχωρισμός του συνόλου δεδομένων σε ομάδες βάσει ομοιότητας.
  - Ανίχνευση ανωμαλιών (Anomaly detection): Προσδιορισμός ασυνήθιστων τιμών/σημείων σε ένα σύνολο δεδομένων.
  - Εξόρυξη συσχέτισης (Association mining): Προσδιορισμός συνόλων στοιχείων σε ένα σύνολο δεδομένων που συμβαίνουν συχνά μαζί.
  - Μείωση διαστάσεων (Dimensionality reduction): Μείωση του αριθμού των μεταβλητών σε ένα σύνολο δεδομένων.

- Η **ημι-επιβλεπόμενη μάθηση** εμπίπτει κάπου μεταξύ της επιβλεπόμενης και της μη-επιβλεπόμενης μάθησης, δεδομένου ότι χρησιμοποιούν τόσο δεδομένα με ετικέτα όσο και χωρίς για την εκπαίδευση – συνήθως μια μικρή ποσότητα δεδομένων με ετικέτα και μια μεγάλη ποσότητα δεδομένων χωρίς ετικέτα. Τα συστήματα που χρησιμοποιούν

αυτήν τη μέθοδο είναι σε θέση να βελτιώσουν σημαντικά την ακρίβεια της μάθησης. Συνήθως, η ημι-επιβλεπόμενη μάθηση επιλέγεται όταν τα δεδομένα με ετικέτα που έχουμε, απαιτούν εξειδικευμένους και σχετικούς πόρους για να την εκπαιδεύσουν/μάθουν από αυτήν. Διαφορετικά, η απόκτηση δεδομένων χωρίς ετικέτα γενικά δεν απαιτεί πρόσθετους πόρους.

Η απόδοση των αλγορίθμων συνήθως βελτιώνεται όταν εκπαιδεύονται σε σύνολα δεδομένων με ετικέτα. Αλλά τα δεδομένα με ετικέτα μπορεί να είναι χρονοβόρα να δημιουργηθούν ή να βρεθούν. Η ημι-επιβλεπόμενη μάθηση εκμεταλλεύεται τα προτερήματα τόσο της επιβλεπόμενης μάθησης όσο και της μη-επιβλεπόμενης μάθησης και σε συγκεκριμένα είδη προβλημάτων μπορεί να αποδίδει καλύτερα. Ορισμένοι τομείς στους οποίους χρησιμοποιείται ημι-επιβλεπόμενη μάθηση περιλαμβάνουν:

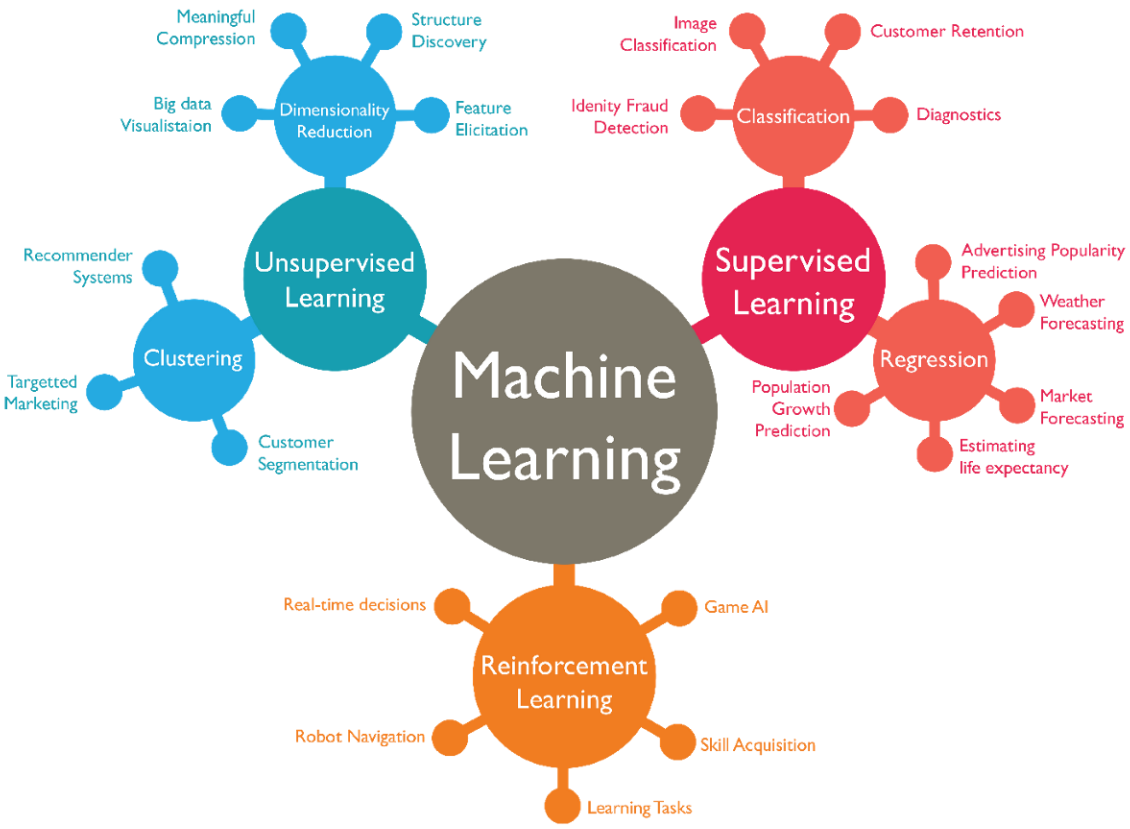
- Μηχανική μετάφραση (Machine translation): Αλγόριθμοι διδασκαλίας για τη μετάφραση γλώσσας.
- Ανίχνευση απάτης (Fraud detection): Προσδιορισμός περιπτώσεων απάτης όταν υπάρχουν μόνο μερικά θετικά παραδείγματα.
- Δεδομένα με ετικέτα (Labeled data): Οι αλγόριθμοι που εκπαιδεύονται σε μικρά σύνολα δεδομένων μπορούν να μάθουν να εφαρμόζουν αυτόματα ετικέτες σε μεγαλύτερα σύνολα δεδομένων.

- Οι αλγόριθμοι **ενισχυτικής μάθησης** είναι μια μέθοδος μάθησης που αλληλεπιδρά με το περιβάλλον της παράγοντας ενέργειες και ανακαλύπτει λάθη ή ανταμοιβές. Η αναζήτηση δοκιμών και σφαλμάτων (trial and error) και η καθυστερημένη ανταμοιβή είναι τα πιο σχετικά χαρακτηριστικά ενισχυτικής μάθησης. Αυτή η μέθοδος επιτρέπει σε μηχανές και πράκτορες λογισμικού να προσδιορίζουν αυτόματα την ιδανική συμπεριφορά σε ένα συγκεκριμένο πλαίσιο, προκειμένου να μεγιστοποιήσουν την απόδοσή της. Απαιτείται απλή ανατροφοδότηση ανταμοιβής για να μάθει ο πράκτορας ποια ενέργεια είναι καλύτερη. Αυτό είναι γνωστό ως σήμα ενίσχυσης (reinforcement signal).

  Η ενισχυτική μάθηση λειτουργεί προγραμματίζοντας έναν αλγόριθμο με διακριτό στόχο και ένα καθορισμένο σύνολο κανόνων για την επίτευξη αυτού του στόχου. Οι επιστήμονες προγραμματίζουν επίσης τον αλγόριθμο για να αναζητήσουν θετικές ανταμοιβές – τις οποίες λαμβάνει όταν εκτελεί μια ενέργεια που είναι ευεργετική προς τον τελικό στόχο – και αποφεύγει τις τιμωρίες – τις οποίες λαμβάνει όταν εκτελεί μια ενέργεια που τον πηγαίνει μακρύτερα από το τελικό στόχο. Η ενισχυτική μάθηση χρησιμοποιείται συχνά σε τομείς όπως:

  - Ρομποτική (Robotics): Τα ρομπότ μπορούν να μάθουν να εκτελούν εργασίες στον φυσικό κόσμο χρησιμοποιώντας αυτήν την τεχνική.

- Βιντεοπαιχνίδια (Gameplay): Η ενισχυτική μάθηση μπορεί να χρησιμοποιηθεί για να μάθει σε bots να παίζουν διάφορα βιντεοπαιχνίδια.
- Διαχείριση πόρων (Resource management): Δεδομένων πεπερασμένων πόρων και καθορισμένου στόχου, η ενισχυτική μάθηση μπορεί να βοηθήσει τις επιχειρήσεις να σχεδιάσουν τον τρόπο κατανομής των πόρων.



Εικόνα 1 Διάγραμμα με τα είδη της μηχανικής μάθησης και τις εφαρμογές κάθε είδους [6]

Η μηχανική μάθηση βοηθάει στην ανάλυση τεράστιων ποσοτήτων δεδομένων. Αν και γενικά παρέχει ταχύτερα και πιο ακριβή αποτελέσματα για τον εντοπισμό κερδοφόρων ευκαιριών ή επικίνδυνων ρίσκων, μπορεί επίσης να απαιτεί επιπλέον χρόνο και πόρους για να εκπαιδευτεί σωστά. Ο συνδυασμός της μηχανικής μάθησης με την τεχνητή νοημοσύνη και τις γνωστικές τεχνολογίες μπορεί να την κάνει ακόμη πιο αποτελεσματική στην επεξεργασία μεγάλου όγκου πληροφοριών.

3

Εξόρυξη Δεδομένων

Εξόρυξη δεδομένων (**Data Mining**) είναι η διαδικασία που περιλαμβάνει τον εντοπισμό καίριων και καινοτόμων μοτίβων, ή προτύπων, τα οποία παρουσιάζουν ενδιαφέρον, καθώς και τη δημιουργία περιγραφικών, κατανοητών και προβλεπτικών μοντέλων από δεδομένα μεγάλης κλίμακας [7].

Οι ρίζες της εξόρυξης δεδομένων βρίσκονται στους πολλούς διαφορετικούς ερευνητικούς τομείς, κι έτσι φαίνεται ο διεπιστημονικός χαρακτήρας αυτού του τομέα. Τρεις από τους ερευνητικούς τομείς που συνδέεται άμεσα: Βάσεις δεδομένων, μηχανική μάθηση και στατιστική.

- Οι βάσεις δεδομένων είναι απαραίτητες για την αποτελεσματική ανάλυση μεγάλων ποσοτήτων δεδομένων. Στο πλαίσιο αυτό, μια βάση δεδομένων αντιπροσωπεύει όχι μόνο το μέσο για συνεπή αποθήκευση και πρόσβαση, αλλά έχει περισσότερο ερευνητικό ενδιαφέρον, αφού η ανάλυση δεδομένων με την χρήση αλγορίθμων εξόρυξης δεδομένων μπορούν να υποστηριχτούν από τις βάσεις δεδομένων και γι’ αυτό είναι σημαντικός παράγοντας στην εξόρυξη δεδομένων.

- Η μηχανική μάθηση είναι ένας τομέας τεχνητής νοημοσύνης που σχετίζεται με την ανάπτυξη τεχνικών που επιτρέπουν στους υπολογιστές να «μάθουν» με την ανάλυση συνόλων δεδομένων, όπως παρουσιάστηκε νωρίτερα.

- Η στατιστική, που βασίζεται στα μαθηματικά και ασχολείται με την επιστήμη και την πρακτική για την ανάλυση εμπειρικών δεδομένων. Βασίζεται στη στατιστική θεωρία που είναι ένας κλάδος των εφαρμοσμένων μαθηματικών.

Παρακάτω θα μελετήσουμε μια ειδικότερη κατηγορία της εξόρυξης δεδομένων, την εξόρυξη κειμένου, που αφορά περισσότερο την παρούσα διπλωματική εργασία.

3.1. Εξόρυξη Κειμένου

Η **εξόρυξη κειμένου** (**text mining**), επίσης γνωστή ως εξόρυξη δεδομένων κειμένου ή ανακάλυψη γνώσεων από βάσεις δεδομένων κειμένου (knowledge discovery from textual databases), αναφέρεται στη διαδικασία εξαγωγής ενδιαφερόντων και σημαντικών προτύπων ή γνώσεων από μη δομημένα έγγραφα κειμένου. Μπορεί να θεωρηθεί ως επέκταση της εξόρυξης δεδομένων ή της ανακάλυψης γνώσεων από βάσεις δεδομένων (knowledge discovery from databases - KDD).

Καθώς η πιο φυσική μορφή αποθήκευσης πληροφοριών είναι σε κείμενο, η εξόρυξη κειμένου πιστεύεται ότι έχει υψηλότερο εμπορικό δυναμικό από αυτό της εξόρυξης δεδομένων. Στην πραγματικότητα, μια πρόσφατη μελέτη έδειξε ότι το 80% των πληροφοριών μιας εταιρείας περιέχεται σε έγγραφα κειμένου. Η εξόρυξη κειμένου, ωστόσο, είναι μια πολύ πιο περίπλοκη εργασία (από την εξόρυξη δεδομένων) καθώς περιλαμβάνει την αντιμετώπιση δεδομένων κειμένου που είναι εγγενώς αδόμητα και ασαφή. Η εξόρυξη κειμένου είναι ένα διεπιστημονικό πεδίο, που περιλαμβάνει τομείς όπως η ανάκτηση πληροφορίας (information retrieval), η ανάλυση κειμένου (text analysis), η εξαγωγή πληροφοριών (information extraction), η ομαδοποίηση (clustering), η κατηγοριοποίηση (categorization), η οπτικοποίηση (visualization), η τεχνολογία βάσεων δεδομένων, η μηχανική μάθηση και η εξόρυξη δεδομένων [8].

Όπως αναφέρθηκε προηγουμένως, το κείμενο είναι ένας από τους πιο συνηθισμένους τύπους δεδομένων στις βάσεις δεδομένων. Ανάλογα με τη βάση δεδομένων, αυτά τα δεδομένα μπορούν να οργανωθούν ως:

- **Δομημένα δεδομένα** (**Structured data**): Αυτά τα δεδομένα τυποποιούνται σε μορφή πίνακα με πολλές σειρές και στήλες, καθιστώντας ευκολότερη την αποθήκευση και την επεξεργασία για ανάλυση και αλγόριθμους μηχανικής μάθησης. Τα δομημένα δεδομένα μπορούν να περιλαμβάνουν εισόδους όπως ονόματα, διευθύνσεις και αριθμούς τηλεφώνου.

- **Μη δομημένα δεδομένα** (**Unstructured data**): Αυτά τα δεδομένα δεν έχουν προκαθορισμένη μορφή. Μπορεί να περιλαμβάνει κείμενο από πηγές, όπως από κοινωνικά δίκτυα ή κριτικές προϊόντων ή μορφές μέσων όπως αρχεία βίντεο και ήχου.

- **Ημι-δομημένα δεδομένα** (**Semi-structured data**): Όπως υποδηλώνει το όνομα, αυτά τα δεδομένα είναι ένα μείγμα μεταξύ δομημένων και μη δομημένων δεδομένων. Αν και έχουν κάποια οργάνωση, δεν είναι αρκετά δομημένα για να ικανοποιούν τις απαιτήσεις μιας σχεσιακής βάσης δεδομένων. Παραδείγματα ημι-δομημένων δεδομένων περιλαμβάνουν αρχεία XML, JSON και HTML.

Δεδομένου ότι το 80% των δεδομένων στον κόσμο βρίσκεται σε μη δομημένη μορφή, η εξόρυξη κειμένων είναι μια εξαιρετικά πολύτιμη πρακτική. Εργαλεία εξόρυξης κειμένων και τεχνικές

επεξεργασίας φυσικής γλώσσας, όπως εξαγωγή πληροφοριών, μας επιτρέπει να μετατρέψουμε τα μη δομημένα αρχεία σε δομημένη μορφή για να μπορέσουμε να τα αναλύσουμε και να δημιουργήσουμε πληροφορία υψηλής ποιότητας. Αυτό, με τη σειρά του, βελτιώνει τη λήψη αποφάσεων των οργανισμών, οδηγώντας σε καλύτερα επιχειρηματικά αποτελέσματα [9].

Η διαδικασία της εξόρυξης κειμένων περιλαμβάνει διάφορες δραστηριότητες που επιτρέπουν να συνάγει κανείς πληροφορίες από μη δομημένα δεδομένα κειμένου. Προτού εφαρμοστούν διάφορες τεχνικές εξόρυξης κειμένου, πρέπει να γίνει μια προ-επεξεργασία κειμένου (pre-processing), η οποία είναι η πρακτική του καθαρισμού και της μετατροπής των δεδομένων κειμένου σε μορφή που μπορούν να χρησιμοποιηθούν πιο αποδοτικά. Αυτή η πρακτική είναι μια βασική πτυχή του NLP και συνήθως περιλαμβάνει τη χρήση τεχνικών όπως η αναγνώριση γλώσσας, η δημιουργία λεξικογραφικής μονάδας (tokenization), η επισήμανση μέρους του λόγου, η κατάτμηση (chunking) και η ανάλυση της σύνταξης για τη σωστή μορφοποίηση δεδομένων για ανάλυση. Όταν ολοκληρωθεί η προ-επεξεργασία κειμένου, μπορεί να εφαρμοστούν αλγόριθμοι εξόρυξης κειμένου για να αντληθούν πληροφορίες από τα δεδομένα. Μερικές από αυτές τις κοινές τεχνικές εξόρυξης κειμένων περιλαμβάνουν η ανάκτηση πληροφορίας, το NLP και η εξαγωγή πληροφοριών [9].

3.2. Ανάκτηση Πληροφορίας

Η έννοια του όρου **ανάκτηση πληροφορίας** είναι ιδιαίτερα ευρεία. Ακόμα και όταν απλώς βγάζουμε την πιστωτική μας κάρτα για να πληκτρολογήσουμε τον αριθμό της έχουμε μια μορφή ανάκτησης πληροφοριών. Η ανάκτηση πληροφορίας επιστρέφει σχετικές πληροφορίες ή έγγραφα βάσει ενός προκαθορισμένου συνόλου ερωτημάτων ή φράσεων. Τα συστήματα ανάκτησης πληροφορίας χρησιμοποιούν αλγόριθμους για την παρακολούθηση των συμπεριφορών των χρηστών και την αναγνώριση σχετικών δεδομένων. Η ανάκτηση πληροφοριών χρησιμοποιείται συνήθως σε συστήματα καταλόγων βιβλιοθηκών και σε δημοφιλείς μηχανές αναζήτησης, όπως το Google. Όμως, ως πεδίο ακαδημαϊκής μελέτης, η ανάκτηση πληροφοριών μπορεί να οριστεί ως:

Ανάκτηση πληροφοριών είναι η εύρεση υλικού (συνήθως εγγράφων) αδόμητης φύσης (συνήθως κειμένου) μέσα σε μεγάλες συλλογές (που βρίσκονται συνήθως αποθηκευμένες σε υπολογιστές), το οποίο ικανοποιεί μια ανάγκη πληροφόρησης [10].

3.3. Επεξεργασία Φυσικής Γλώσσας

Η **επεξεργασία φυσικής γλώσσας** είναι ένας τομέας έρευνας και εφαρμογής που διερευνά τον τρόπο με τον οποίο οι υπολογιστές μπορούν να χρησιμοποιηθούν για την κατανόηση και τον χειρισμό κειμένου ή ομιλίας φυσικής γλώσσας για να κάνουν χρήσιμα πράγματα. Οι ερευνητές του NLP στοχεύουν στη συγκέντρωση γνώσης για το πώς οι άνθρωποι κατανοούν και χρησιμοποιούν τη γλώσσα, έτσι ώστε να μπορούν να αναπτυχθούν κατάλληλα εργαλεία και τεχνικές για να κάνουν τους υπολογιστές να κατανοούν και να χειρίζονται φυσικές γλώσσες για

την εκτέλεση των επιθυμητών εργασιών. Τα θεμέλια του NLP βρίσκονται σε διάφορους κλάδους, δηλαδή, επιστήμες πληροφορικής και πληροφοριών, γλωσσολογία, μαθηματικά, τεχνητή νοημοσύνη, ρομποτική και ψυχολογία. Οι εφαρμογές του NLP περιλαμβάνουν έναν αριθμό τομέων μελέτης, όπως η μετάφραση, επεξεργασία κειμένου φυσικής γλώσσας και περίληψη αυτής, διεπαφές χρήστη, πολύγλωσση ανάκτηση πληροφοριών (Council on Library and Information Resources – CLIR), αναγνώριση ομιλίας και τεχνητή νοημοσύνη [11].

Σε αυτό το σημείο αξίζει να αναφερθεί ότι οι παραπάνω τεχνολογίες για να εφαρμοστούν πιο αποδοτικά απαιτούν την προ-επεξεργασία του κειμένου. Παρακάτω θα αναφερθούν οι πιο σημαντικές εργασίες όσον αφορά την προ-επεξεργασία κειμένου.

3.4. Προ-επεξεργασία κειμένου

Η προ-επεξεργασία κειμένου είναι μια μέθοδος για τον καθαρισμό των δεδομένων κειμένου και την προετοιμασία για την τροφοδοσία των δεδομένων στο μοντέλο. Τα δεδομένα κειμένου περιέχουν θόρυβο σε διάφορες μορφές όπως συναισθήματα, σημεία στίξης ή απλά ανεπιθύμητο κείμενο. Όταν μιλάμε για την ανθρώπινη γλώσσα τότε, υπάρχουν διαφορετικοί τρόποι για να πούμε το ίδιο πράγμα, και αυτό είναι μόνο το κύριο πρόβλημα που πρέπει να αντιμετωπίσουμε, αφού οι μηχανές δεν θα καταλάβουν λέξεις, χρειάζονται αριθμούς. Επομένως πρέπει η μετατροπή του κείμενου σε αριθμούς να γίνει με έναν αποτελεσματικό τρόπο.

Στην συνέχεια παρουσιάζονται, οι βασικότερες ενέργειες που –συνήθως– απαιτούνται στην προ-επεξεργασία κειμένου:

3.4.1. Δημιουργία λεξικογραφικής μονάδας (Tokenization)

Το **tokenization** είναι η διαδικασία διάσπασης μιας ροής κειμένου σε λέξεις, φράσεις, σύμβολα ή άλλα σημαντικά στοιχεία που ονομάζονται tokens. Ο στόχος του tokenization είναι η διερεύνηση των λέξεων σε μια πρόταση. Η λίστα των tokens γίνεται είσοδος για περαιτέρω επεξεργασία, όπως συντακτική ανάλυση (parsing) ή εξόρυξη κειμένου. Το tokenization είναι χρήσιμο τόσο στη γλωσσολογία, όσο και στην επιστήμη των υπολογιστών, όπου αποτελεί μέρος της λεξικής ανάλυσης (lexical analysis). Τα δεδομένα κειμένου είναι μόνο ένα μπλοκ χαρακτήρων στην αρχή. Όλες οι διαδικασίες της ανάκτησης πληροφορίας απαιτούν τις λέξεις του dataset. Συνεπώς, το tokenization είναι απαραίτητη προϋπόθεση σε ένα αρχείο για την συντακτική ανάλυση. Αυτό μπορεί να ακούγεται ασήμαντο καθώς το κείμενο είναι ήδη αποθηκευμένο σε μορφές αναγνώσιμες από μηχανή. Ωστόσο, ορισμένα προβλήματα παραμένουν, όπως η αφαίρεση των σημείων στίξης. Άλλοι χαρακτήρες όπως αγκύλες, παύλες κ.λπ. απαιτούν επίσης επεξεργασία. Επιπλέον, το tokenization μπορεί να φροντίσει για την συνέπεια στα έγγραφα. Η κύρια χρήση του tokenization είναι ο προσδιορισμός των σημαντικών λέξεων-κλειδιών. Η ασυνέπεια μπορεί να αφορά την μορφή αριθμών ή ώρας. Ένα άλλο

πρόβλημα μπορεί να είναι οι συντομογραφίες και τα ακρωνύμια που πρέπει να μετατραπούν στην πλήρη τους μορφή [12].

Στην συνέχεια, παρατίθεται ένα απλό παράδειγμα για το tokenization:

Πρόταση:
«Το Τμήμα ΜΠΕΣ της Πολυτεχνικής Σχολής του Πανεπιστημίου Αιγαίου ιδρύθηκε το 1997»

Tokens:
«Το», « Τμήμα», «ΜΠΕΣ», «της», «Πολυτεχνικής», «Σχολής», «του», «Πανεπιστημίου», «Αιγαίου», «ιδρύθηκε», «το», «1997»

3.4.2. Αφαίρεση συνήθων λέξεων (Stop word removal)

Συνήθως ο όρος stop word αναφέρεται στις λέξεις που χρησιμοποιούνται πιο συχνά σε μια γλώσσα. Είναι οι λέξεις σε οποιαδήποτε γλώσσα που δεν προσθέτουν πολύ νόημα σε μια πρόταση. Μπορούν δηλαδή, να αγνοηθούν με ασφάλεια χωρίς να αλλοιώνεται το νόημα της πρότασης. Για ορισμένες μηχανές αναζήτησης, αυτές είναι μερικές από τις πιο κοινές: “the, is, at, which, and on” – για τα αγγλικά. Σε κάποιες περιπτώσεις, μπορεί να προκαλέσουν προβλήματα κατά την αναζήτηση φράσεων που συμπεριλαμβάνουν τέτοιες λέξεις.

Δεν υπάρχει συγκεκριμένος κανόνας για το πότε πρέπει να αφαιρεθούν τα stop words. Γενικότερα, προτείνεται να αφαιρούνται εάν πρέπει να κάνουμε ταξινόμηση γλώσσας (language classification), το φιλτράρισμα ανεπιθύμητων μηνυμάτων (spam filtering), τη δημιουργία υπότιτλων, τη δημιουργία αυτόματων ετικετών, την ανάλυση συναισθημάτων (sentiment analysis) ή οτιδήποτε που σχετίζεται με την ταξινόμηση κειμένου.

Στην συνέχεια, παρατίθεται ένα απλό παράδειγμα για την αφαίρεση stop word:

Πριν την αφαίρεση των stop words:
«Το», « Τμήμα», «ΜΠΕΣ», «της», «Πολυτεχνικής», «Σχολής», «του», «Πανεπιστημίου», «Αιγαίου», «ιδρύθηκε», «το», «1997»

Μετά την αφαίρεση των stop words:
«Τμήμα», «ΜΠΕΣ», «Πολυτεχνικής», «Σχολής», «Πανεπιστημίου», «Αιγαίου», «ιδρύθηκε», «1997»

3.4.3. Αποκοπή καταλήξεων (Stemming)

Στόχος του stemming είναι η διαδικασία απλοποίησης των λέξεων σε μια μορφή (root form) που είναι ίδια ανεξαρτήτως κατάληξης, και συνήθως επιτυγχάνεται αφαιρώντας την κατάληξη της λέξης.

Το stemming χρησιμοποιεί μια απλή ευρετική διαδικασία που κόβει τις καταλήξεις των λέξεων με την ελπίδα να μετατρέψει σωστά τις λέξεις στη αρχική τους μορφή. Πολλές φορές οι

stemmers μπορεί να μην έχουν ακριβώς το επιθυμητό αποτέλεσμα όπως φαίνεται στο παράδειγμα πιο κάτω. Υπάρχουν διαφορετικοί αλγόριθμοι για το stemming. Ο πιο κοινός αλγόριθμος, είναι ο αλγόριθμος Porters.

π.χ. *studies → studi, studying → study*

3.4.4. Λημματοποίηση (Lemmatization)

Η λημματοποίηση με μια πρώτη ματιά, μοιάζει πολύ με το stemming, προσπαθεί δηλαδή να φέρει την λέξη στην αρχική της μορφή (root form). Η μόνη διαφορά είναι ότι, η λημματοποίηση προσπαθεί να το κάνει με πιο «σωστό» τρόπο. Δεν κόβει απλώς τις καταλήξεις, αλλά αλλάζει τις λέξεις στην πραγματική τους ρίζα. Παράδειγμα:

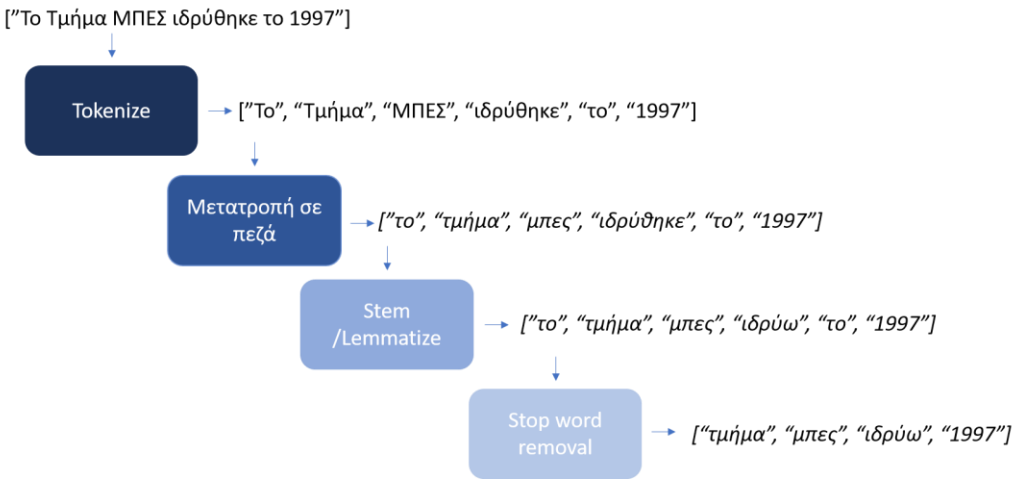
π.χ. *“studies” → “study”, “knives” → “knife”*

3.4.5. Μετατροπή σε πεζά (Lowercase conversion)

Το lowercase conversion αναφέρεται στην διαδικασία μετατροπής όλων των γραμμάτων σε πεζά/μικρά γράμματα.

Έτσι για παράδειγμα τα tokens:
«Το», « Τμήμα», «ΜΠΕΣ», «της», «Πολυτεχνικής», «Σχολής», «του», «Πανεπιστημίου», «Αιγαίου», «ιδρύθηκε», «το», «1997»

θα γίνουν:
«τμήμα», «μπες», «πολυτεχνικής», «σχολής», «πανεπιστημίου», «αιγαίου», «ιδρύθηκε», «1997»



Εικόνα 2 Ολοκληρωμένο παράδειγμα preprocessing

3.5 Κατηγοριοποίηση κειμένου (Text categorization)

Η κατηγοριοποίηση κειμένου, γνωστή και ως ταξινόμηση κειμένου (text classification) είναι η εργασία αυτόματης ταξινόμησης ενός συνόλου εγγράφων σε κατηγορίες από ένα προκαθορισμένο σύνολο. Αυτή η εργασία έχει πολλές εφαρμογές, συμπεριλαμβανομένης της αυτοματοποιημένης δεικτοδότησης επιστημονικών άρθρων, το φιλτράρισμα ειδήσεων, την οργάνωση αρχείων και την ανάκτηση, την εξόρυξη γνώμης, την ταξινόμηση μηνυμάτων ηλεκτρονικού ταχυδρομείου και το φιλτράρισμα αυτών κ.λπ.

Μια μεγάλη ποικιλία τεχνικών έχει σχεδιαστεί για την ταξινόμηση κειμένου. Στην συνέχεια, θα παρουσιαστούν οι ευρείες κατηγορίες τεχνικών και οι χρήσεις τους για εργασίες ταξινόμησης. Σημειώνουμε ότι αυτές οι κατηγορίες τεχνικών υπάρχουν και για άλλους τομείς δεδομένων, όπως ποσοτικά ή κατηγορικά δεδομένα. Δεδομένου ότι τα κείμενα μπορούν να μοντελοποιηθούν ως ποσοτικά δεδομένα με συχνότητες στα χαρακτηριστικά της λέξης, είναι δυνατό να χρησιμοποιηθούν οι περισσότερες μέθοδοι που χρησιμοποιούνται για ποσοτικά δεδομένα, απευθείας σε κείμενο. Ωστόσο, το κείμενο είναι ένα συγκεκριμένο είδος δεδομένων στα οποία τα χαρακτηριστικά της λέξης είναι αραιά και υψηλών διαστάσεων, με χαμηλές συχνότητες στις περισσότερες λέξεις. Ως εκ τούτου, είναι ζωτικής σημασίας να σχεδιαστούν μέθοδοι ταξινόμησης που θα λαμβάνουν αποτελεσματικά υπόψη αυτά τα χαρακτηριστικά του κειμένου [13]. Ορισμένες βασικές μέθοδοι που χρησιμοποιούνται για την ταξινόμηση κειμένου είναι οι εξής:

Αφελής ταξινομητής Bayes (Naïve Bayes classifiers)

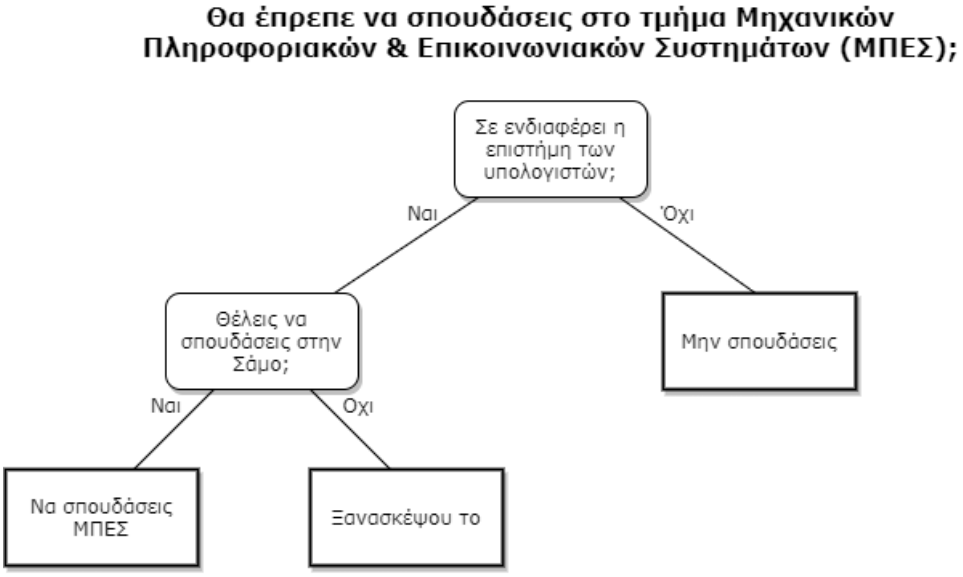
Ο Naive Bayes είναι ένας πιθανολογικός αλγόριθμος μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί σε μια μεγάλη ποικιλία εργασιών ταξινόμησης. Οι τυπικές εφαρμογές περιλαμβάνουν φιλτράρισμα ανεπιθύμητων μηνυμάτων, ταξινόμηση εγγράφων, πρόβλεψη συναισθημάτων κ.λπ. Βασίζεται στα έργα του Thomas Bayes, όπου πήρε και το όνομά του.

Ο χαρακτηρισμός «αφελής» χρησιμοποιείται επειδή υποθέτει ότι τα χαρακτηριστικά που μπαίνουν στο μοντέλο είναι ανεξάρτητα το ένα από το άλλο. Αυτό σημαίνει ότι αλλάζοντας την τιμή ενός χαρακτηριστικού, δεν επηρεάζει ή δεν αλλάζει άμεσα την τιμή οποιουδήποτε άλλου χαρακτηριστικού που χρησιμοποιείται στον αλγόριθμο.

Δέντρα αποφάσεων (Decision trees)

Τα Δέντρα αποφάσεων είναι μια μη-παραμετρική μέθοδος επιβλεπόμενης μάθησης που χρησιμοποιείται για την ταξινόμηση και την παλινδρόμηση. Ο στόχος είναι να δημιουργηθεί ένα μοντέλο που θα προβλέπει την τιμή μιας μεταβλητής-στόχου, μαθαίνοντας απλούς κανόνες απόφασης που συνάγονται από τα χαρακτηριστικά των δεδομένων. Η ιδέα πίσω από τα Δέντρα απόφασης είναι ότι χρησιμοποιείτε τις δυνατότητες του συνόλου δεδομένων για να δημιουργήσετε ερωτήσεις τύπου ναι/όχι και να διαχωρίζετε συνεχώς το σύνολο δεδομένων

έως ότου απομονώσετε όλα τα σημεία δεδομένων που ανήκουν σε κάθε κλάση. Με αυτήν τη διαδικασία τα δεδομένα οργανώνονται σε μια δομή δέντρου. Στην Εικόνα 3 φαίνεται ένα απλό παράδειγμα δέντρου απόφασης.



Εικόνα 3 Απλό παράδειγμα δέντρου απόφασης

Αλγόριθμος τυχαίων δασών (Random forests)

Ο αλγόριθμος τυχαίων δασών είναι ένας αλγόριθμος επιβλεπόμενης μάθησης. Το "δάσος" που χτίζει, είναι ένα σύνολο δέντρων αποφάσεων, που συνήθως εκπαιδεύονται με τη μέθοδο "bagging". Η γενική ιδέα αυτής της μεθόδου είναι ότι ένας συνδυασμός μοντέλων μάθησης αυξάνει το συνολικό αποτέλεσμα. Στην ουσία, ο αλγόριθμος τυχαίων δασών αποτελείται από πολλά δέντρα αποφάσεων και τα συγχωνεύει για να έχει μια πιο ακριβή και σταθερή πρόβλεψη.

Μηχανές Διανυσμάτων Υποστήριξης (Support vector machines)

Ο SVM είναι ένας από τους πιο δημοφιλείς αλγόριθμους επιβλεπόμενης μάθησης, ο οποίος χρησιμοποιείται για την ταξινόμηση καθώς και για προβλήματα παλινδρόμησης. Ωστόσο, κατά κύριο λόγο, χρησιμοποιείται για προβλήματα ταξινόμησης στη μηχανική μάθηση. Ο στόχος του αλγορίθμου SVM είναι να δημιουργήσει την καλύτερη γραμμή ή όριο απόφασης, που μπορεί να διαχωρίσει τον n-διάστατο χώρο σε κλάσεις, ώστε να μπορούμε εύκολα να τοποθετήσουμε ένα νέο σημείο δεδομένων στη σωστή κατηγορία στο μέλλον. Αυτό το καλύτερο όριο απόφασης ονομάζεται “hyperplane”. Ο SVM επιλέγει τα ακραία σημεία/διανύσματα που

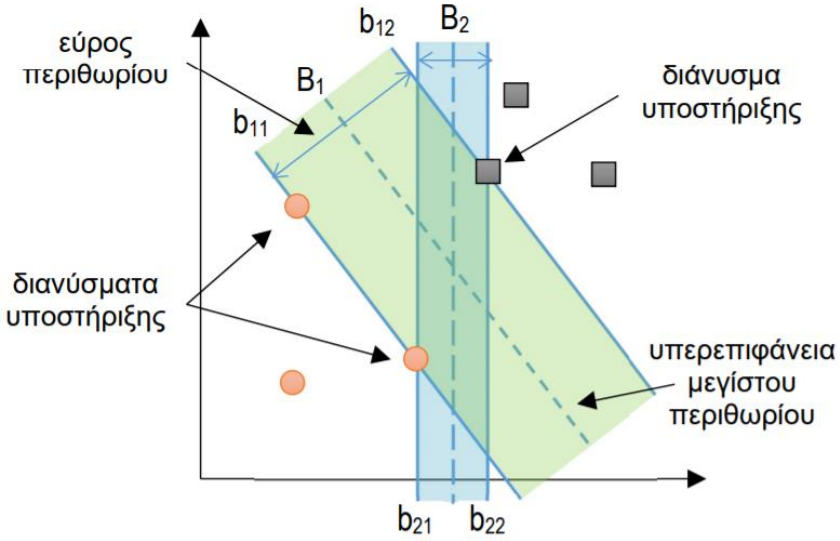
βοηθούν στη δημιουργία του hyperplane. Αυτές οι ακραίες περιπτώσεις ονομάζονται διανύσματα υποστήριξης, γι' αυτό και ο αλγόριθμος ονομάζεται έτσι.

Συγκεκριμένα, θεωρώντας ένα πρόβλημα δυαδικής ταξινόμησης με θετικά και αρνητικά παραδείγματα, μια επιφάνεια που λειτουργεί ως σύνορο μεταξύ των κλάσεων χαρακτηρίζεται από το γεγονός ότι χωρίζει τα θετικά από τα αρνητικά παραδείγματα. Τέτοια σύνορα υπάρχουν εν γένει πολλά, όπως για παράδειγμα τα B1 και B2 στην Εικόνα 4.

Η μέθοδος SVM επιδιώκει να βρει το σύνορο που απέχει όσο το δυνατόν περισσότερο από τα παραδείγματα των κλάσεων που διαχωρίζει. Η υπερεπιφάνεια αυτή ονομάζεται υπερεπιφάνεια μεγίστου περιθωρίου (maximum margin hypersurface) και σε γραμμικώς διαχωρίσιμα προβλήματα ορίζεται από έναν πεπερασμένο αριθμό παραδειγμάτων του συνόλου εκπαίδευσης που ονομάζονται διανύσματα υποστήριξης (support vectors). Με την παραπάνω προσέγγιση, στην Εικόνα 4, το σύνορο B1 είναι το σύνορο με το μέγιστο εύρος περιθωρίου.

Επιπλέον, μέσω των συναρτήσεων πυρήνα (kernel functions), οι SVM μπορούν να μετασχηματίσουν τον αρχικό χώρο υποθέσεων έτσι ώστε μη-γραμμικώς διαχωρίσιμα προβλήματα να μετατραπούν σε γραμμικώς διαχωρίσιμα και τελικά να λυθούν με την ίδια μεθοδολογία.

Επομένως, το κύριο πλεονέκτημα των SVM είναι ότι όχι μόνο δημιουργούν ένα διαχωριστικό σύνορο μεταξύ των επιμέρους κλάσεων (όπως άλλωστε κάνουν οι περισσότερες μέθοδοι ταξινόμησης), αλλά φροντίζουν αυτό το σύνορο να απέχει όσο το δυνατό περισσότερο από τα παραδείγματα των κλάσεων που διαχωρίζουν.



Εικόνα 4 Δύο σύνορα απόφασης (διακεκομμένες γραμμές) σε πρόβλημα δυαδικής ταξινόμησης. Το σύνορο B1 χωρίζει τέλεια τις δύο κλάσεις και εξασφαλίζει επιπλέον και μέγιστο εύρος περιθωρίου [14]

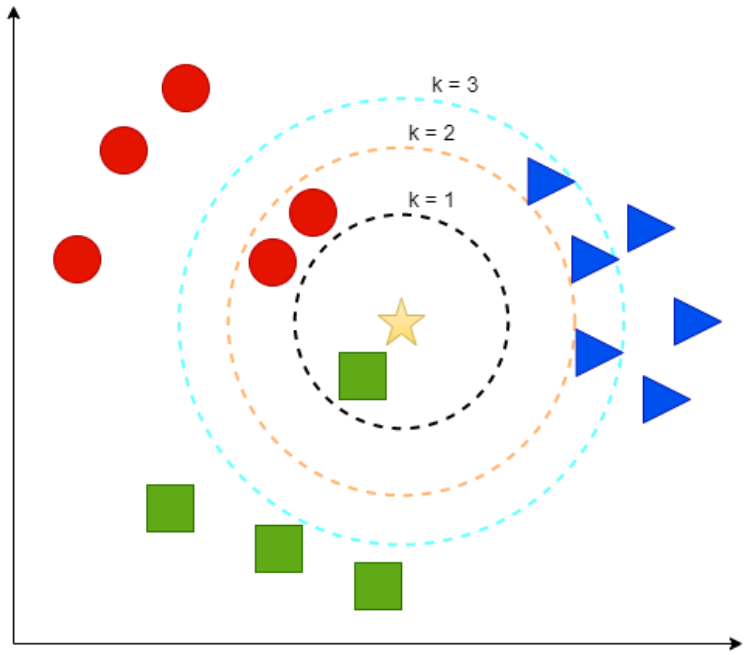
Τα σύνορα απόφασης με μεγάλα περιθώρια έχουν κατά βάση μεγαλύτερη ανοχή σε φαινόμενα υπερπροσαρμογής. Αν το περιθώριο είναι μικρό, τότε μια μικρή διαταραχή (π.χ. εξαιτίας θορύβου στα δεδομένα εκπαίδευσης) μπορεί να προκαλέσει σημαντική πτώση στην απόδοση του ταξινομητή. Στα μεγάλα περιθώρια, ακόμη κι αν τα διανύσματα υποστήριξης είναι παραδείγματα με λίγο θόρυβο, τα σφάλματα γενίκευσης είναι κατά βάση μικρότερα [14].

k-κοντινότεροι γείτονες (k-nearest neighbors)

Ο KNN είναι ένας μη παραμετρικός και τεμπέλης αλγόριθμος μάθησης. Μη παραμετρικός σημαίνει ότι δεν υπάρχει υπόθεση για υποκείμενη κατανομή δεδομένων. Με άλλα λόγια, η δομή του μοντέλου καθορίζεται από το σύνολο δεδομένων. Αυτό είναι πολύ χρήσιμο στην πράξη, όπου τα περισσότερα σύνολα δεδομένων του πραγματικού κόσμου δεν ακολουθούν μαθηματικές θεωρητικές υποθέσεις. Ο όρος «τεμπέλης αλγόριθμος» σημαίνει ότι δεν χρειάζεται δεδομένα εκπαίδευσης για τη δημιουργία του μοντέλου. Όλα τα δεδομένα εκπαίδευσης χρησιμοποιούνται στη φάση των δοκιμών. Αυτό καθιστά την εκπαίδευση γρηγορότερη και τη φάση δοκιμών πιο αργή και δαπανηρή (χρόνο και μνήμη).

Στον KNN, το K είναι ο αριθμός των πλησιέστερων γειτόνων που θα λάβει υπόψη ο αλγόριθμος. Ο αριθμός των γειτόνων είναι ο βασικός καθοριστικός παράγοντας. Το K συνήθως είναι περιττός αριθμός εάν ο αριθμός των κλάσεων είναι 2. Όταν K = 1, τότε ο αλγόριθμος είναι γνωστός ως ο αλγόριθμος του πλησιέστερου γείτονα. Αυτή είναι η απλούστερη περίπτωση.

Στην Εικόνα 5 βλέπουμε ένα παράδειγμα όπου έχουμε ένα δείγμα προς μελέτη (το κίτρινο αστέρι) όπου ανάλογα της τιμής του k, μπορεί να μπει σε διαφορετική ομάδα.



Εικόνα 5 Παράδειγμα knn για διαφορετικές τιμές του k.

Βλέπουμε ότι στην περίπτωση για k=1 θα μπει στην κλάση με τα πράσινα τετράγωνα, για k=2 στην κλάση με τους κόκκινους κύκλος ενώ τέλος, για k=3 θα μπει στην κλάση με τα μπλε τρίγωνα.

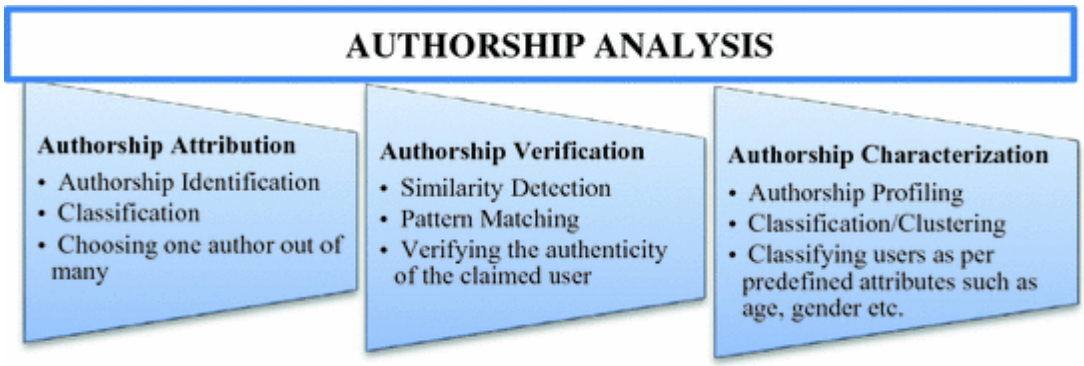
Αλγόριθμοι βαθιάς μάθησης (Deep learning classifiers)

Τα μοντέλα βασισμένα στη βαθιά μάθηση έχουν ξεπεράσει τις κλασικές προσεγγίσεις που βασίζονται στην μηχανική μάθηση σε διάφορες εργασίες ταξινόμησης κειμένου, όπως ανάλυση συναισθημάτων, κατηγοριοποίηση ειδήσεων, απάντηση ερωτήσεων και συμπέρασμα φυσικής γλώσσας. Υπάρχουν πολλά μοντέλα και συνδυασμός αυτών για την εργασία της ταξινόμησης του κειμένου με την χρήση βαθιάς μάθησης, που έχουν αναλυθεί σε περιεκτικές έρευνες [15].

4

Ανάλυση Συγγραφέα

Η **ανάλυση συγγραφέα** είναι η διαδικασία της μελέτης των χαρακτηριστικών ενός κειμένου προκειμένου να εξαχθούν συμπεράσματα σχετικά με τον συγγραφέα. Η ανάλυση συγγραφέα έχει της ρίζες της σε ένα ερευνητικό πεδίο της γλωσσολογίας το οποίο καλείται στυλομετρία, η οποία αναφέρεται στην στατιστική ανάλυση του γλωσσικού ύφους. Η ανάλυση συγγραφέα υπολογίζει ορισμένα χαρακτηριστικά κειμένου και αποφεύγει την διάκριση μεταξύ κειμένων γραμμένα από τον ίδιο συγγραφέα. Οι πρώτες μελέτες χρονολογούνται στον 19^{ο} αιώνα με την μελέτη των έργων του Σαίξπηρ το 1887 ακολουθούμενες από στατιστικές μελέτες στο πρώτο μισό του 20^{ου} αιώνα. Η μελέτη των Ομοσπονδιακών Κειμένων των ΗΠΑ το 1964 θεωρήθηκε η πιο σημαντική στο πεδίο της απόδοσης συγγραφέα (authorship attribution). Η ανάλυση συγγραφέα μπορεί χωριστεί στις εξής κατηγορίες [16]:

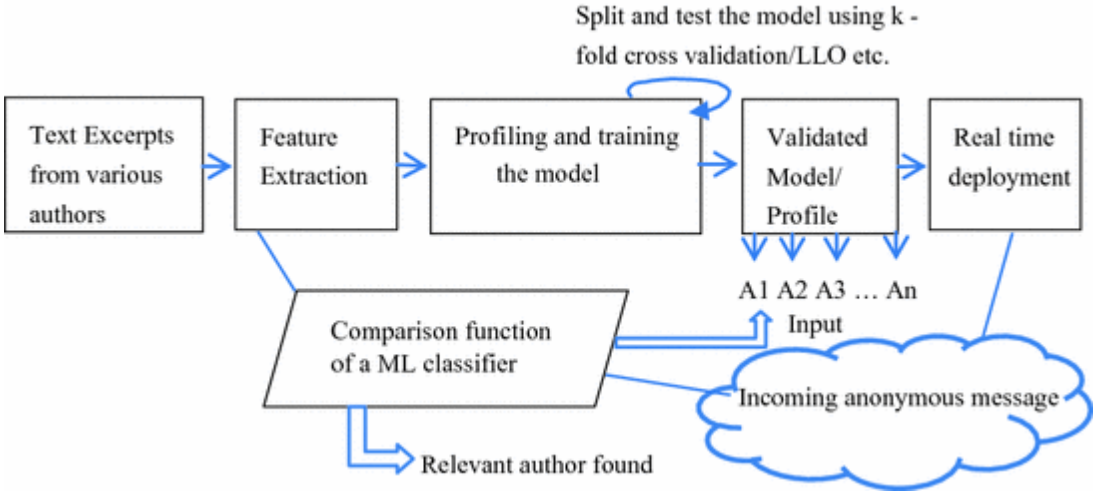


Εικόνα 6 Κατηγορίες της ανάλυσης συγγραφέα [5]

4.1. Απόδοση Συγγραφέα ή Ταυτοποίηση (Authorship attribution or identification)

Η **απόδοση συγγραφέα** ή ταυτοποίηση είναι ένα πρόβλημα ταξινόμησης πολλά-προς-ένα, όπου μεταξύ ενός δεδομένου συνόλου χρηστών και ενός άγνωστου κειμένου, ο στόχος είναι να προσδιοριστεί ο σωστός χρήστης για το δεδομένο κείμενο. Δεδομένου "n" χρηστών, το άγνωστο κείμενο αποδίδεται στον πλησιέστερο χρήστη. Σε αντίθεση με την κλασική παράδοση της μελέτης του προβλήματος σαν closed-set, όπου ακόμη και στο χειρότερο σενάριο σίγουρα το κείμενο αντιστοιχίζεται σε κάποιον χρήστη, οι πρόσφατες έρευνες έχουν μελετήσει το πρόβλημα ως ένα ανοιχτό πρόβλημα όπου το κείμενο χωρίς ετικέτα μπορεί να μην αποδοθεί απαραίτητα στους δεδομένους χρήστες. Εν ολίγοις, το πρώτο σενάριο υποθέτει ότι ο συγγραφέας είναι παρών στο σύνολο δεδομένων εκπαίδευσης και ως εκ τούτου αποδίδει το κείμενο σε κάποιον χρήστη, ενώ στην τελευταία περίπτωση εάν δεν βρεθεί ταίριασμα από το σύνολο δεδομένων των χρηστών, ακολουθείται η απόρριψη, που δηλώνει ότι ο συγγραφέας είναι κάποιος εκτός του συνόλου δεδομένων.

Οι ερευνητές έχουν χρησιμοποιήσει διάφορα αποσπάσματα κειμένου που κυμαίνονται από μεγάλα έγγραφα (π.χ. λογοτεχνικά έργα) έως και μικρότερα κείμενα (π.χ. tweets) που περιορίζονται σε μόλις 140 χαρακτήρες για την αναγνώριση του συγγραφέα. Μια γενική μεθοδολογία που υιοθετήθηκε για την απόδοση συγγραφέα φαίνεται στην Εικόνα 7Εικόνα 7.



Εικόνα 7 Ένα απλοποιημένο πλαίσιο για την απόδοση συγγραφέα [5]

Η Εικόνα 7Εικόνα 7 αντικατοπτρίζει τα γενικευμένα βήματα που ακολουθούνται για το έργο της απόδοσης συγγραφέα. Για ένα άγνωστο δείγμα, πρέπει να προσδιοριστεί ποιος ανάμεσα σε ένα δεδομένο σύνολο χρηστών (A1, A2,…, An) είναι ο πραγματικός συγγραφέας του άγνωστου κειμένου. Αυτό επιτυγχάνεται ανεξάρτητα, μοντελοποιώντας τα προφίλ των χρηστών που

αναφέρονται και συγκρίνοντας τα άγνωστα δεδομένα με κάθε προφίλ για να προσδιορίσουμε τον πιθανό συγγραφέα. Το πρόβλημα εάν μελετηθεί σαν closed-set, επιστρέφει τον πιο πιθανό χρήστη. Αλλά για το σενάριο open-set, αν δεν βρεθεί πιθανή αντιστοιχία πέρα από ένα όριο (threshold), σημαίνει ότι ο συγγραφέας είναι εκτός του συνόλου [5].

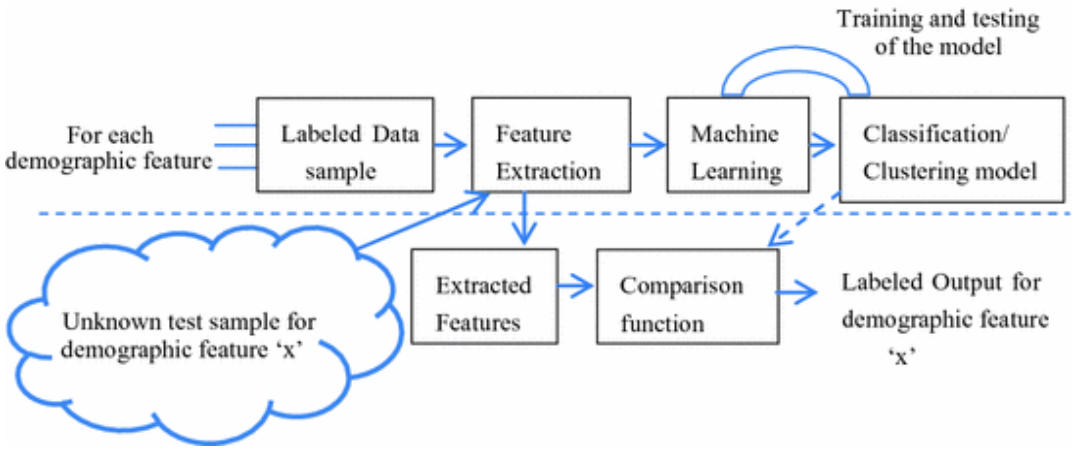
Στην Εικόνα 8 βλέπουμε ένα απλό παράδειγμα για την κατανόηση της απόδοσης συγγραφέα.

Εικόνα 8 Απόδοση συγγραφέα [17]

4.2. Δημιουργία προφίλ ή χαρακτήρα (Authorship profiling or characterization)

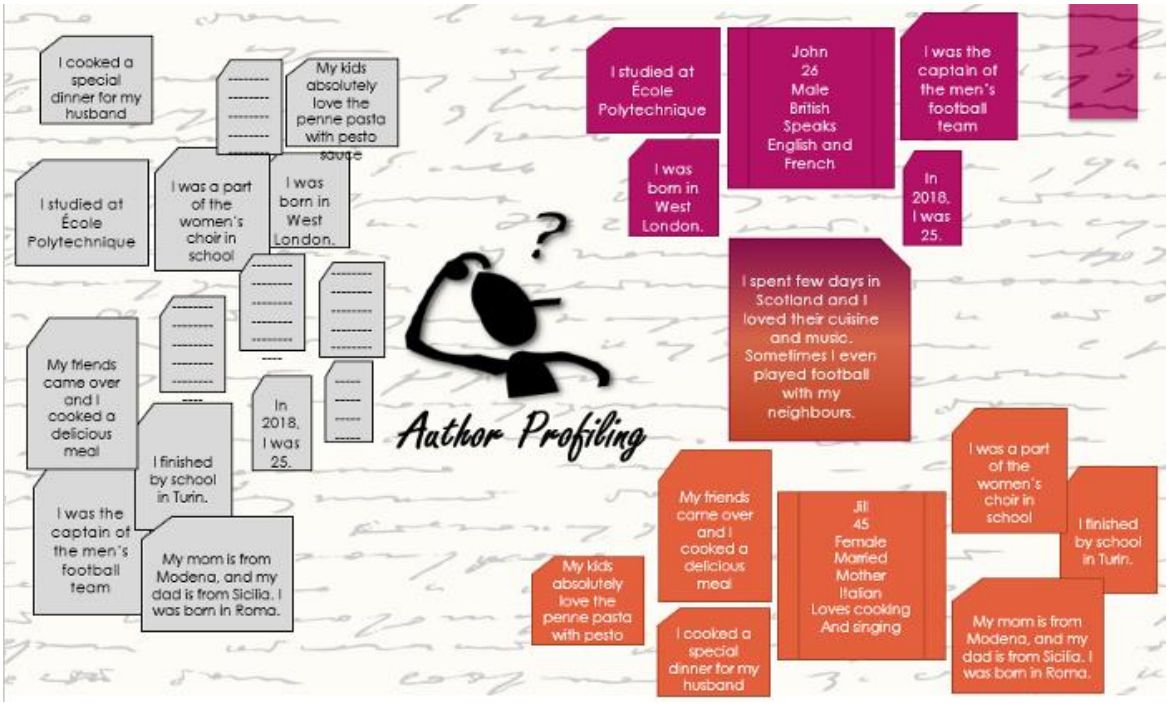
Η **δημιουργία προφίλ** ή χαρακτήρα εστιάζει στον προσδιορισμό των χαρακτηριστικών όπως η ηλικία, το φύλο, οι γλωσσικές πληροφορίες κ.λπ. των συγγραφέων του κειμένου. Δεδομένου ενός άγνωστου κειμένου και ενός συνόλου προκαθορισμένων κατηγοριοποιήσεων, ο στόχος είναι να προσδιοριστούν τα απαιτούμενα δημογραφικά χαρακτηριστικά ενός χρήστη.

Όπως και με τις άλλες τεχνικές ανάλυσης συγγραφέα, τα χαρακτηριστικά (features) που εξάγονται από τα δείγματα δεδομένων με ετικέτα για κάθε δημογραφικό χαρακτηριστικό αναλύονται και σχεδιάζονται ανεξάρτητα. Χρησιμοποιούνται μηχανή μάθηση ή άλλες τεχνικές αυτό-μάθησης για την ανάλυση της αποτελεσματικότητας διαφορετικών ταξινομητών ή μοντέλων ταξινόμησης. Οι εκτιμήσεις ακρίβειας διαφορετικών ταξινομητών αντικατοπτρίζουν την αποτελεσματικότητα κάθε μεθόδου. Ομοίως, η στάθμιση των χαρακτηριστικών σύμφωνα με τη σημασία τους δίνει μια εκτίμηση της σημασίας κάθε χαρακτηριστικού για τη διάκριση διαφορετικών συγγραφέων Εικόνα 9 [5].



Εικόνα 9 Ένα απλοποιημένο πλαίσιο για την δημιουργία προφίλ [5]

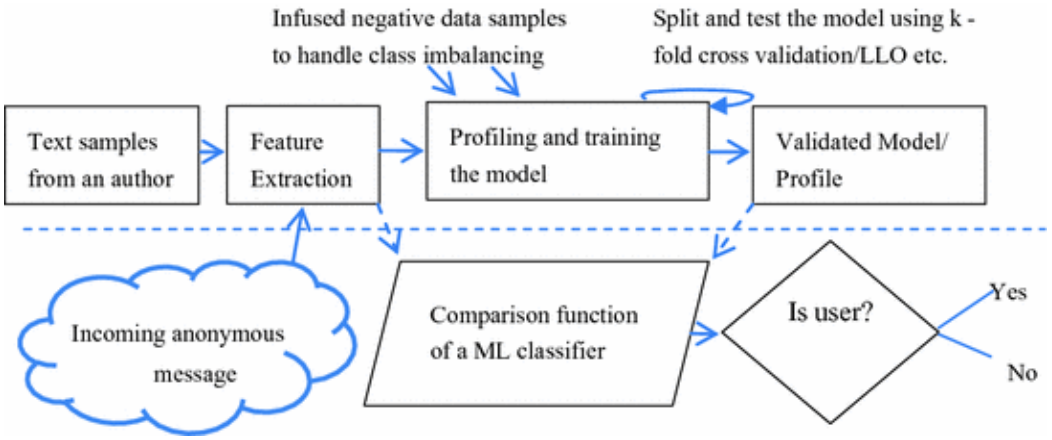
Στην Εικόνα 10 βλέπουμε ένα απλό παράδειγμα για την κατανόηση της δημιουργίας προφίλ συγγραφέα.



Εικόνα 10 Δημιουργία προφίλ συγγραφέα [17]

4.3. Επαλήθευση Συγγραφέα (Authorship Verification)

Η **επαλήθευση συγγραφέα**, η οποία επίσης αναφέρεται ως ανίχνευση ομοιότητας (similarity detection), είναι ένα πρόβλημα δυαδικής ταξινόμησης ένα-προς-ένα και θα μπορούσε να θεωρηθεί ως υποσύνολο του open-set προβλήματος απόδοσης συγγραφέα, όπου για ένα δεδομένο άγνωστο κείμενο πρέπει να προσδιοριστεί αν είναι γραμμένο από έναν χρήστη ή όχι. Η βασική μεθοδολογία που ακολουθήθηκε για την εργασία επαλήθευσης συγγραφέα φαίνεται στην Εικόνα 11.



Εικόνα 11 Ένα απλοποιημένο πλαίσιο για την επαλήθευση συγγραφέα [5]

Δημιουργείται ένα προφίλ συμπεριφοράς επιλέγοντας και εξάγοντας τα σχετικά χαρακτηριστικά από τα δείγματα δεδομένων ενός χρήστη. Τα ίδια σύνολα χαρακτηριστικών που λαμβάνονται από το άγνωστο δείγμα συγκρίνονται με τα μοντέλα που έχουν μάθει. Μετά τη σύγκριση με τα ήδη γνωστά μοτίβα τόσο των θετικών όσο και των αρνητικών κλάσεων, μια κλάση με την οποία οι τιμές χαρακτηριστικών του άγνωστου δείγματος ταιριάζουν περισσότερο θεωρείται ως ετικέτα [5].

Στην Εικόνα 12 βλέπουμε ένα απλό παράδειγμα για την κατανόηση της επαλήθευσης συγγραφέα.

Εικόνα 12 Επαλήθευση Συγγραφέα [17]

Σε σχέση με τις άλλες προσεγγίσεις ανάλυσης συγγραφέα**, αυτή της επαλήθευσης συγγραφέα έχει ορισμένα σημαντικά πλεονεκτήματα**. Πρώτον, είναι πιο γενική, καθώς οποιοδήποτε πρόβλημα απόδοσης μπορεί να αποσυντεθεί σε μια σειρά περιπτώσεων επαλήθευσης. Στη συνέχεια, ορισμένοι παράγοντες που επηρεάζουν την απόδοση των closed και open-set απόδοσης συγγραφέα, όπως το μέγεθος του σετ υποψήφιων χρηστών και η διανομή των κειμένων εκπαίδευσης στους υποψήφιους συγγραφείς έχουν περιορισμένο αντίκτυπο στην επαλήθευση συγγραφέα. Επομένως, είναι πιο εφικτό να εκτιμηθεί το ποσοστό σφάλματος της απόδοσης συγγραφέα, που απαιτείται στο πλαίσιο εγκληματολογικών εφαρμογών, όταν εστιάζεται στην λογική της επαλήθευσης. Τα τελευταία χρόνια, υπάρχει αυξανόμενο ενδιαφέρον για την επαλήθευση συγγραφέα, κυρίως λόγω των διαγωνισμών PAN [18]. Έχουν αναπτυχθεί και δοκιμαστεί πολλές μέθοδοι σε σύνολα κειμένων που καλύπτουν διάφορες γλώσσες και είδη. [19]

Κάθε μια από αυτές τις κατηγορίες της ανάλυσης συγγραφέα είναι επεκτάσιμες ανάλογα με το είδος των προβλημάτων για τα οποία χρησιμοποιούνται στον πραγματικό κόσμο. Μερικές φορές, αυτές αλληλεπικαλύπτονται. Οι προσεγγίσεις αυτές δεν περιορίζονται μόνο στα αγγλικά στην αυτόματη ανάλυση συγγραφέα. Ερευνώνται αυτοματοποιημένες εφαρμογές και αναπτύσσονται και για άλλες γλώσσες όπως ελληνικά, γαλλικά, ολλανδικά, ισπανικά, αραβικά, πορτογαλικά και ιταλικά.

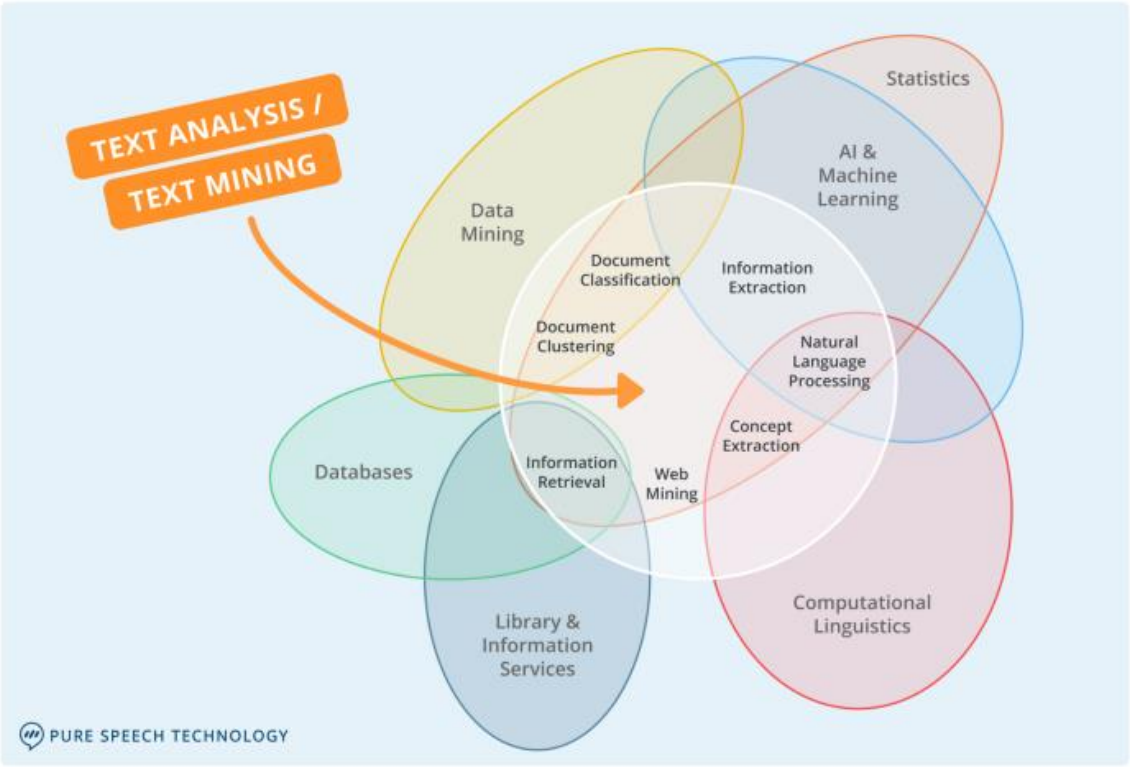
4.4. Απόδοση Συγγραφέα ή Επαλήθευση Συγγραφέα;

Η επαλήθευση συγγραφέα είναι ένα θεμελιώδες πρόβλημα στην απόδοση συγγραφέα δεδομένου ότι οποιοδήποτε πρόβλημα, είτε closed-set ή open-set, μπορεί να αποσυντεθεί σε ένα σύνολο προβλημάτων επαλήθευσης. Ωστόσο, είναι αρκετά δύσκολο σε σύγκριση τόσο με την closed-set προσέγγιση όσο και με την open-set, καθώς το μοντέλο επαλήθευσης πρέπει να εκτιμήσει εάν το άγνωστο κείμενο είναι αρκετά παρόμοιο σε σχέση με τα δοθέντα κείμενα από έναν συγκεκριμένο συγγραφέα, ενώ ένα μοντέλο απόδοσης πρέπει να εκτιμήσει ποιος είναι ο πιο πιθανός υποψήφιος συγγραφέας του κειμένου.

Όπως έχει ήδη αναφερθεί, η απόδοση συγγραφέα σχετίζεται με σημαντικές εγκληματολογικές εφαρμογές. Ωστόσο, είναι αμφισβητήσιμο εάν μπορεί να χρησιμοποιηθεί ως αποδεικτικό στοιχείο στο δικαστήριο. Σίγουρα, αυτή η τεχνολογία μπορεί να χρησιμοποιηθεί από τους ερευνητές για να καθοδηγήσουν την εστίασή τους σε συγκεκριμένους υπόπτους και στη συνέχεια να συλλέξουν άλλα αποδεκτά αποδεικτικά στοιχεία (π.χ. δείγματα DNA) για παρουσίαση στο δικαστήριο. Υπάρχουν διάφοροι παράγοντες που επηρεάζουν την απόδοση ενός μοντέλου απόδοσης συγγραφέα, όπως ο αριθμός των υποψηφίων συγγραφέων, η κατανομή των κειμένων που χρησιμοποιήθηκαν για εκπαίδευση στους υποψήφιους συγγραφείς, το μήκος των δειγμάτων κειμένου και αν τα υπό έρευνα κείμενα ταιριάζουν στο είδος και το θέμα, για να μην αναφέρουμε παράγοντες όπως η ωρίμανση του στυλ γραφής (όταν

το προσωπικό στυλ κάποιου αλλάζει με την πάροδο του χρόνου). Η εκτίμηση του ποσοστού σφάλματος της τεχνολογίας απόδοσης συγγραφέα είναι σίγουρα πιο εφικτή εάν υιοθετήσουμε την λογική της επαλήθευσης. Ένα εγγενές πρόβλημα στην απόδοση closed-set και open-set είναι ότι η απόδοση επιδεινώνεται αυξάνοντας το μέγεθος του συνόλου υποψηφίων. Από την άλλη, στην επαλήθευση συγγραφέα το σύνολο υποψηφίων είναι πάντα μεμονωμένος (singleton) και ως εκ τούτου το ποσοστό σφάλματος είναι ευκολότερο να εκτιμηθεί.

Ένα άλλο κρίσιμο ζήτημα στην απόδοση closed-set και multi-class open-set είναι ότι η απόδοση του μοντέλου εξαρτάται από τη διανομή των κειμένων εκπαίδευσης στους υποψήφιους συγγραφείς. Το λεγόμενο πρόβλημα ανισοκατανομής μεταξύ των κατηγοριών (class imbalance) αναγκάζει τα μοντέλα απόδοσης συγγραφέα να προτιμούν στην πρόβλεψή τους τούς συντάκτες που είναι περισσότεροι. Ωστόσο, σε μια εγκληματολογική υπόθεση, το γεγονός ότι είναι διαθέσιμα πολλά δείγματα κειμένου για έναν συγκεκριμένο υποψήφιο συγγραφέα δεν θα πρέπει να κάνει αυτόν τον ύποπτο ως τον πιο πιθανό συγγραφέα των κειμένων υπό έρευνα. Η επαλήθευση συγγραφέα είναι πιο ισχυρή ως προς ανισοκατανεμημένα μεταξύ των κατηγοριών, καθώς κάθε υποψήφιος συγγραφέας εξετάζεται ξεχωριστά.



Εικόνα 13 Διάγραμμα Venn που δείχνει τη σχέση της Εξόρυξης Κειμένου με έξι συναφή πεδία: στατιστική, AI και μηχανική μάθηση, υπολογιστική γλωσσολογία, υπηρεσίες βιβλιοθήκης και πληροφοριών, βάσεις δεδομένων και εξόρυξη δεδομένων [20]

4.5. Πεδία Εφαρμογής

Η ανάλυση συγγραφέα παίζει καθοριστικό ρόλο στην εγκληματολογική ανάλυση και την έρευνα εγκλημάτων. Εκτός αυτού, τα μέσα κοινωνικής δικτύωσης και οι ανοιχτοί πόροι διαδικτύου έχουν φέρει ένα ευρύ φάσμα εγκλημάτων στον κυβερνοχώρο – ψεύτικα προφίλ, ψεύτικες κριτικές από bots, λογοκλοπή, ιστοσελίδες του dark web που διευκολύνουν την τρομοκρατία, την παρενόχληση και τον εκφοβισμό μέσω μηνυμάτων στα κοινωνικά δίκτυα.

Επίσης, η κατανόηση των προφίλ καταναλωτών και η ανάλυση των σχολίων είναι υψίστης σημασίας για την ανάλυση αγοράς (market analysis) και σκοπεύει να εξετάσει τα δημογραφικά στοιχεία συγγραφέων ανώνυμων σχολίων. Η ανάλυση συγγραφέα βοηθά στη δημιουργία προφίλ καταναλωτών, στον εντοπισμό ψευδών κριτικών και στην τμηματοποίηση πελατών.

Άλλοι τομείς εφαρμογής περιλαμβάνουν την επίλυση διαφορών στην συγγραφή μυθιστορημάτων, την ανίχνευση λογοκλοπής, τη χρονολόγηση εγγράφων, την εξέταση κοινωνικοοικονομικών παραγόντων και την εξέταση ψυχικής υγείας [17].

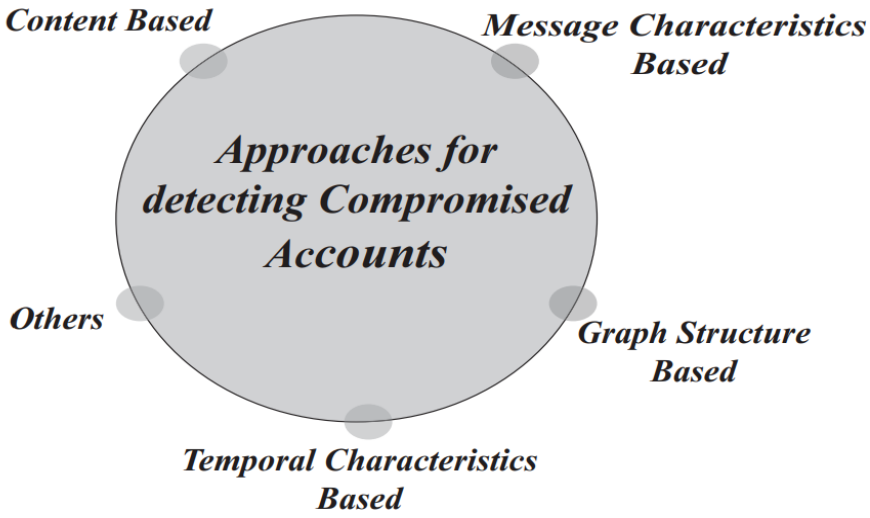
5

Ανίχνευση παραβιασμένων λογαριασμών (Compromised Account Detection)

Τα κοινωνικά δίκτυα αποτελούν αναπόσπαστο μέρος της σύγχρονης κοινωνίας, και η χρήση τους τα τελευταία χρόνια έχει αυξηθεί εκθετικά. Γι' αυτό και οι απαιτήσεις ασφάλειας αυτών, από κακόβουλους χρήστες, είναι πολύ υψηλές. Ένας από τους στόχους των επιτιθέμενων είναι η παραβίαση έγκυρων λογαριασμών. Άρα, μπορούμε να πούμε ότι το "compromised account detection" στα κοινωνικά δίκτυα, είναι η διαδικασία επιβεβαίωσης εάν ένας λογαριασμός που χρησιμοποιείται από ένα άτομο (ή από έναν οργανισμό) έχει παραβιαστεί ή όχι.

Όπως ήδη συζητήθηκε, οι κακόβουλοι χρήστες, βρίσκουν και εφαρμόζουν νέους τρόπους για να αποφύγουν την ανίχνευσή τους και να εκπληρώσουν τις επιθέσεις τους. Ένας τέτοιος τρόπος είναι η παραβίαση άλλων έγκυρων λογαριασμών. Παρακάτω θα δούμε μερικές προσεγγίσεις των ερευνητών για την ανίχνευση παραβιασμένων λογαριασμών. Η ανίχνευση παραβιασμένων λογαριασμών εμπίπτει σε δύο κατηγορίες, αυτή κατά την εισαγωγή του χρήστη στον λογαριασμό (log-in) ή κατά την διάρκεια της χρήσης (at-use detection). Αυτή κατά το log-in του χρήστη στον λογαριασμό, προσπαθεί να αναγνωρίσει τον παραβιασμένο λογαριασμό πριν οποιαδήποτε ενέργεια του λογαριασμού η οποία θα είναι ορατή στο κοινωνικό δίκτυο, ενώ η ανίχνευση κατά την διάρκεια της χρήσης βασίζεται στον εντοπισμό ασυνήθιστης συμπεριφοράς, όπως η δημοσίευση URLs που είναι spam ή με την παρουσία των ασυνήθιστων προτύπων συμπεριφοράς (π.χ. ασυνήθιστα πολλά αιτήματα φιλίας, μεγάλος αριθμός δημοσιεύσεων κ.λπ.). Η πρώτη, κατά το log-in, δεν έχει ερευνηθεί αρκετά (ως επί το πλείστον πρόβλημα της πρόσβασης των δεδομένων) και χειρίζεται αρκετά καλά από τους παρόχους τέτοιων υπηρεσιών. Η ανίχνευση κατά τη χρήση είναι η κύρια εστίαση των ερευνητών, η οποία εκτελείται με την καταγραφή διάφορων χαρακτηριστικών και με την ανίχνευση ασυνήθιστων αποκλίσεων σε αυτά

τα χαρακτηριστικά. Οι τεχνικές που αφορούν την ανίχνευση παραβιασμένων λογαριασμών χωρίζονται σε πέντε κατηγορίες, όπως φαίνεται στην Εικόνα 14 [21]:



Εικόνα 14 Προσεγγίσεις για τον εντοπισμό παραβιασμένων λογαριασμών [21]

Οι κακόβουλοι χρήστες μπορούν να κλέψουν τα διαπιστευτήρια ενός έγκυρου χρήστη και να εκτελέσουν ασυνήθιστες δραστηριότητες από τους λογαριασμούς τους. Σε κάθε περίπτωση, είναι πολύ δύσκολο να αποκτηθούν και να χειριστούν τα πρότυπα συμπεριφοράς (behavioral patterns) ενός χρήστη. Με τον καιρό, οι χρήστες των κοινωνικών δικτύων δημιουργούν κάποιες χαρακτηριστικές συμπεριφορές που σχετίζονται με τη χρήση που κάνουν και τηρούν ασύνειδα αυτήν τη συμπεριφορά. Για παράδειγμα, ο χρόνος χρήσης του κοινωνικού δικτύου, η γλώσσα στην οποία πραγματοποιείται η επικοινωνία, η διάρκεια της δραστηριότητας, οι συνήθειες περιήγησης, οι ακολουθίες περιήγησης κ.λπ. Κάθε φορά που ο λογαριασμός παραβιάζεται, οι επιτιθέμενοι δεν μπορούν να «συμμορφώνονται» με αυτήν τη συμπεριφορά, ανοίγοντας ένα δρόμο για την ανίχνευσή τους. Εάν, σε περίπτωση που, δημιουργηθεί ένα «προφίλ» για τα πρότυπα συμπεριφοράς κάθε χρήστη εκ των προτέρων, οποιαδήποτε απόκλιση από τα καθορισμένα χαρακτηριστικά αυτά, θα μπορούσε να βοηθήσει στην πρόβλεψη των παραβιασμένων λογαριασμών.

Οι **content based techniques** πιστοποιούν και αναγνωρίζουν τους χρήστες με βάση το στυλ γραφής τους. Τέτοιες τεχνικές είναι κατάλληλες για τους χρήστες που έχουν συνέπεια στην επιλογή των λέξεων και του στυλ γραφής τους, αντιθέτως οι χρήστες που δεν δημοσιεύουν συχνά ή αυτοί με διαφορετικό τρόπο γραφής έχουν συχνά ένα ελλιπές και αναξιόπιστο προφίλ και ως εκ τούτου ενδέχεται να μην εντοπιστούν από το περιεχόμενό τους. Δεύτερον, η επιλογή

των κριτηρίων ορίου απόκλισης είναι αρκετά κουραστική, καθώς κάποια απόκλιση θα πρέπει να εξεταστεί εντός των αποδεκτών ορίων.

Σε αντίθεση με τις τεχνικές που βασίζονται σε περιεχόμενο, οι **message characteristics based techniques** δεν μελετούν τις δημοσιεύσεις και τις πληροφορίες σχετικά με τα συμφραζόμενα που σχετίζονται με τις ίδιες, αλλά αντιθέτως ασχολούνται με τα μεταδεδομένα (meta data) που σχετίζονται με ένα tweet/δημοσίευση. Οι πληροφορίες των μεταδεδομένων που είναι ανεξάρτητες από κείμενο θεωρείται ότι είναι πιο σταθερές και λιγότερο περιορισμένες. Για παράδειγμα, η χρήση μιας συγκεκριμένης γλώσσας, η πηγή εφαρμογής, οι αναφορές κ.λπ. σπάνια υφίσταται οποιαδήποτε αλλαγή και διαπιστώνεται ότι είναι πιο αξιόπιστα. Όπως και στις content based techniques, οι τεχνικές που βασίζονται σε χαρακτηριστικά μηνυμάτων περιορίζονται μόνο σε εκείνους τους χρήστες που χρησιμοποιούν συχνά το κοινωνικό δίκτυο για να κάνουν tweet ή να ενημερώσουν την κατάστασή τους. Ωστόσο, η παραβίαση του λογαριασμού δεν είναι πάντα ορατή μέσω μόνο των χαρακτηριστικών των μηνυμάτων. Οι προθέσεις μπορεί να περιλαμβάνουν την αλλαγή του δικτύου φίλων/ακολούθων ή την παρακολούθηση (stalking) ενός χρήστη, με απλή ανάλυση της συμπεριφοράς περιήγησης ή την αλλαγή των χαρακτηριστικών του προφίλ και πολλά άλλα.

Όπως ήδη αναφέρθηκε, σε περίπτωση παραβίασης ενός λογαριασμού, οι αλλαγές μοτίβου μπορεί επίσης να αφορά «διαταραχές» στη δομή του δικτύου ενός χρήστη. Συνεπώς, η ανάλυση των δομών γραφήματος (**graph structures**) είναι απαραίτητη σε τέτοια σενάρια. Αλλά τεράστιος αριθμός χρηστών και πολλές πτυχές φιλικών σχέσεων μεταξύ τέτοιων κόμβων καθιστούν τη συνεχή παρακολούθηση και ανάλυση αποκλίσεων τέτοιων γραφικών δομών αρκετά δύσκολη και δαπανηρή. Άρα, στη βιβλιογραφία περιλαμβάνονται πολύ λίγα έργα που σχετίζονται με αυτές τις τεχνικές για τον εντοπισμό παραβιασμένων λογαριασμών.

Οι τεχνικές που βασίζονται σε χρονικά χαρακτηριστικά (**temporal characteristics based techniques**) πήραν την ιδέα από την κιρκαδική τοπολογία (circadian topology) που διαφοροποιεί τα άτομα με βάση διαφορετικά χρονικά πρότυπα. Λαμβάνοντας υπόψη αυτό, οι χρήστες βρέθηκαν να υπακούν σε μια τέτοια συμπεριφορά και στα κοινωνικά δίκτυα. Τα διαφορετικά επίπεδα λεπτομερειών επιτρέπουν στους χρήστες να αναλύουν τη συνέπεια σε διαφορετικές χρονικές διαστάσεις. Οι temporal based techniques, αν και έχει παρατηρηθεί ότι είναι αποτελεσματικές, κατά κάποιον τρόπο υστερούν στο να διακρίνουν τους χρήστες μοναδικά. Επιπλέον, οι χάκερς μπορεί να έχουν αλληλεπικαλυπτόμενα χρονικά μοτίβα με τους έγκυρους χρήστες και ως εκ τούτου μπορεί να «διαφύγουν» από τον εντοπισμό.

Όπως περιγράφεται στην Εικόνα 15, κάθε τεχνική ακολουθεί την βασική αρχή της δημιουργίας προφίλ βάσης για τους χρήστες και αναζήτηση τυχόν απόκλισης από αυτό το προφίλ, για να ειδοποιεί τους χρήστες για πιθανή παραβίαση. Λόγω της εμπλεκόμενης υποκειμενικότητας στα κοινωνικά δίκτυα, ένας χρήστης δεν μπορεί ποτέ να περιοριστεί να συμπεριφέρεται με

συγκεκριμένο τρόπο. Μια από τις πιο σημαντικές ενέργειες είναι ο εντοπισμός των χαρακτηριστικών που είναι κατάλληλα για τη δημιουργία προφίλ και στην εκτίμηση της απόκλισης ανωμαλιών. Λόγω της εύκολης διαθεσιμότητας και χρήσης, τα textual, temporal και message specific features έχουν αποκτήσει μεγάλη δημοτικότητα, αλλά περιορίζονται στο να διακρίνουν μόνο τους χρήστες που δημοσιεύουν συχνά. Για τους χρήστες που δεν δημοσιεύουν συχνά υλικό, το profiling και η παρακολούθηση των χαρακτηριστικών συμπεριφοράς αποδεικνύεται πιο αποτελεσματική, αλλά αυτά τα χαρακτηριστικά είναι συνήθως δύσκολο να αποκτηθούν. Οι πάροχοι υπηρεσιών που έχουν πρόσβαση σε όλες τις πληροφορίες ενός χρήστη ενδέχεται να μην αντιμετωπίζουν μεγάλη δυσκολία στην εξαγωγή και ανάλυση τέτοιων μοτίβων συμπεριφοράς. Για να συμβεί αυτό, οι χρήστες μπορούν να επιτρέψουν στους παρόχους υπηρεσιών να εφαρμόσουν τέτοιες τεχνικές, αλλά με το κόστος της απώλειας της ιδιωτικότητάς τους. Σε γενικές γραμμές, είναι εξαιρετικά σημαντικό να υπάρχουν καλοί μηχανισμοί υπεράσπισης έναντι παραβιασμένων λογαριασμών για να διασφαλίσουμε την ελάχιστη ζημιά, ακόμη και αν η πλήρης πρόληψη φαίνεται ανέφικτη.

Τεχνική	**Πλεονεκτήματα**	**Μειονεκτήματα**
Content-based	• Εύκολη διαθεσιμότητα και απόκτηση δεδομένων (content based data) – απαιτεί μόνο μικρή ποσότητα προ-επεξεργασίας • Αποτελεσματική για χρήστες που δημοσιεύουν συχνά	• Αποτυχία σε χρήστες που δεν αναρτούν τόσο συχνά • Η απόφαση για μια γραμμή διαχωρισμού (threshold) μεταξύ της «κανονικής» και της «ανώμαλης» συμπεριφοράς είναι λίγο δύσκολη
Message characteristics based	• Οι πληροφορίες μεταδεδομένων είναι πιο σταθερές και αξιόπιστες για τη δημιουργία προφίλ στη συμπεριφορά του χρήστη • Βοηθάει στο να δημιουργηθούν σταθερά μοτίβα για κάθε χρήστη	• Οι χρήστες κι εδώ πρέπει να δημοσιεύουν συχνά κανονικά μηνύματα
Graph Structure based	• Τα προφίλ τις περισσότερες φορές υφίστανται αλλαγή υπό την επιρροή του spammer • Ενσωματώνει αξιόλογες παραλείψεις από τις content and message characteristic based techniques	• Δύσκολο να αποκτηθεί και να καταγραφεί. Κανένα API δεν παρέχει πληροφορίες σχετικά με τις αλλαγές μοτίβου γραφήματος • Χρειάζεται μη αυτόματη (manual) ανάλυση από την ημέρα δημιουργίας του λογαριασμού
Temporal characteristics based	• Οι θεωρίες ενισχύουν την υπόθεση για την τήρηση των χρηστών σε συγκεκριμένα χρονικά μοτίβα που αντικατοπτρίζουν τη συνέπεια στη συμπεριφορά σύνδεσης και δημοσίευσης.	• Μπορεί να αντιμετωπίσει τυχαία αλληλεπικαλυπτόμενα μοτίβα μεταξύ χρηστών και χάκερς • Συχνά αποτυγχάνουν να ξεχωρίσουν τους έγκυρους χρήστες μοναδικά

Εικόνα 15 Σύγκριση διαφορετικών τεχνικών για τον εντοπισμό παραβιασμένων λογαριασμών

5.1. Γιατί μας ενδιαφέρει;

Οι απειλές που προκαλούνται από αυτούς τους παραβιασμένους λογαριασμούς περιλαμβάνουν διάδοση κακόβουλου λογισμικού, αποστολή σπαμ μηνυμάτων, adwares και spywares. Η ταχύτερη ανίχνευση λοιπόν, αυτών των λογαριασμών, είναι σημαντική προκειμένου να μειωθεί ο αντίκτυπος της ζημιάς.

Ακόμη, η έγκαιρη ανίχνευση αυτών των παραβιασμένων λογαριασμών είναι πολύ σημαντική για την προστασία των χρηστών και την ομαλή λειτουργία των κοινωνικών δικτύων. Οι κίνδυνοι που κρύβονται πίσω από αυτό το πρόβλημα είναι πολλοί και πολύ συχνά μπορεί να αποβούν καταστροφικοί για το άτομο (ή τον οργανισμό).

5.2. Προηγούμενες έρευνες

Η βιβλιογραφία που σχετίζεται με το παρόν θέμα, μπορεί να χωριστεί σε δύο ομάδες: α) Ανίχνευση ανεπιθύμητων μηνυμάτων και κακόβουλου περιεχομένου στα κοινωνικά δίκτυα και β) απόδοση συγγραφέα.

Μπορεί να παρατηρηθεί μια συλλογή πολύ διαφορετικών έργων που αφορούν την ανίχνευση ανεπιθύμητων μηνυμάτων και κακόβουλου περιεχομένου στα κοινωνικά δίκτυα. Καθώς τα κοινωνικά δίκτυα έγιναν πιο δημοφιλείς, κάποιοι προσπάθησαν να συγκρίνουν τους spammers σε e-mail και σε tweets. Θεωρώντας ανεπιθύμητο κάθε μήνυμα που περιέχει σύνδεσμο (link) προς δόλιο ιστότοπο με κακόβουλο περιεχόμενο, οι συγγραφείς ανέλυσαν link domains ως προσπάθεια εύρεσης μοτίβων ανεπιθύμητων μηνυμάτων. Ως αποτέλεσμα, ανακάλυψαν ότι τα links που χρησιμοποιούνται σε tweets και e-mail είναι συνήθως διαφορετικά [22]. Μια διαφορετική μελέτη, εξέτασε μόνο black-listed links για τον εντοπισμό ανεπιθύμητων μηνυμάτων [23]. Το συμπέρασμα των συγγραφέων ήταν δεν είναι αρκετά μόνο τα black-listed links για να λύσουν το πρόβλημα. Ως ελαφρώς διαφορετική προσέγγιση, μια άλλη έρευνα χρησιμοποίησε γραφήματα σχέσεων (relationship graphs) για να ελέγξει εάν ένας λογαριασμός μπορεί να θεωρηθεί κακόβουλος [24].

Μια άλλη έρευνα πρότεινε την μελέτη για τη διάκριση των παραβιασμένων προφίλ από τα έγκυρα προφίλ, λαμβάνοντας υπόψη τα προφίλ των χρηστών. Σε αυτή την προσέγγιση λαμβάνετε υπόψη ο αριθμός των σχέσεων, η «ηλικία» του προφίλ και την πληρότητα του προφίλ ως χαρακτηριστικά του λογαριασμού [25]. Σε μια άλλη έρευνα, προτάθηκε ένα πλήρες σύστημα. Χρησιμοποιήθηκε ένα προ-επισημασμένο σύνολο tweets που περιείχαν κακόβουλους συνδέσμους. Στη συνέχεια, χρησιμοποιήθηκε μηχανική μάθηση για την εξαγωγή μοτίβων από δημοσιεύσεις ανεξάρτητα από την παρουσία κακόβουλων συνδέσμων. Με αυτόν τον τρόπο, οι μελλοντικές αναρτήσεις θα ταξινομούνται χωρίς να χρειάζονται κακόβουλους συνδέσμους από γνωστές black-lists στο περιεχόμενό τους [26].

Στην συνέχεια, μια επόμενη έρευνα πρότεινε μια μέθοδο, βασισμένη σε μηχανική μάθηση και πολλά χαρακτηριστικά κειμένου (textual features). Ο αριθμός των λέξεων ανά δημοσίευση, ο αριθμός των συνδέσμων ανά λέξη, ο αριθμός των χαρακτήρων, οι σύνδεσμοι και τα hashtags αποτελούν το σύνολο των features. Χρησιμοποιήθηκαν μη-κειμενικά χαρακτηριστικά, συμπεριλαμβανομένης της «ηλικίας» του λογαριασμού και του αριθμού των σχέσεων. Ο ταξινομητής που χρησιμοποιήθηκε από τους συγγραφείς ήταν ο SVM [27].

Εφαρμόζοντας μια καθαρά προσέγγιση εξόρυξης κειμένου, η επόμενη έρευνα, πρότεινε μια μέθοδο σε επίπεδο χαρακτήρων. Οι συγγραφείς χρησιμοποίησαν ακολουθίες χαρακτήρων, οι οποίες ονομάστηκαν έγγραφα (documents), ως χαρακτηριστικά για την αναπαράσταση του περιεχομένου που παράγεται από κάθε λογαριασμό. Στη συνέχεια, αυτές οι ακολουθίες χαρακτήρων επεξεργάστηκαν με έναν αλγόριθμο μηχανικής μάθησης. Παρόλο που αυτές οι προσεγγίσεις μπορούν να αντιμετωπίσουν το πρόβλημα της ανίχνευσης κακόβουλων λογαριασμών, δεν διακρίνουν τη διαφορά μεταξύ κακόβουλων λογαριασμών και παραβιασμένων λογαριασμών [28].

Σε μια επόμενη έρευνα, προτάθηκε ένα μοντέλο με βάση τα γραφήματα για την αξιολόγηση της αξιοπιστίας των χρηστών. Οι συγγραφείς δήλωσαν ότι η αξιολόγηση εμπιστοσύνης μεταξύ των χρηστών είναι απαραίτητη επειδή έχουν τη συνήθεια να μοιράζονται συνδέσμους και αρχεία. Με αυτόν τον τρόπο, εφαρμόστηκε μια μικρή θεωρία λέξεων για την προώθηση της κοινής χρήσης μεταξύ αξιόπιστων χρηστών [29].

Ένα άλλο παράδειγμα κακόβουλης δραστηριότητας εκτός από το spam, είναι η ύπαρξη ψεύτικων λογαριασμών. Πολλές προσεγγίσεις έχουν επικεντρωθεί επιτυχώς στην ανάλυση των προφίλ στα κοινωνικά δίκτυα, για τον εντοπισμό ψεύτικων λογαριασμών, που έχουν εφαρμοστεί τόσο στο Twitter όσο και στο Facebook.

Ωστόσο, με την εξέλιξη αυτού του τομέα και την επιτυχή προσπάθεια αυτών των ερευνών να εντοπίζουν τους ψεύτικους λογαριασμούς, οι κακόβουλοι χρήστες κατέφυγαν στο να παραβιάζουν έγκυρους λογαριασμούς. Οι επιτιθέμενοι παρατήρησαν ότι η παραβίαση ενός έγκυρου λογαριασμού για τη διάδοση κακόβουλου περιεχομένου θα μπορούσε να είναι πιο αποτελεσματική για δύο λόγους: α) ένας έγκυρος λογαριασμός έχει καλό ιστορικό και οι φίλοι τους τείνουν να εμπιστεύονται τις δημοσιεύσεις του και β) ένας παραβιασμένος λογαριασμός γίνεται κακόβουλος μόνο όταν παραβιαστεί, οπότε παλαιότερες μέθοδοι θα είχαν προβλήματα για τον εντοπισμό τους [1].

Κάποιες από τις πρόσφατες έρευνες που αφορούσαν παραβιασμένους λογαριασμούς στα κοινωνικά δίκτυα αναλύονται στις [1] και [2]. Η πρώτη εργασία πρότεινε μια μέθοδο ανίχνευσης λογαριασμών που παραβιάστηκαν, από πληροφορίες λογαριασμού, συμπεριλαμβανομένων των περισσότερων ωρών πρόσβασης της ημέρας, των θεμάτων και των φίλων που καλούνται συχνά. Τα αποτελέσματα που επιτεύχθηκαν ήταν ικανοποιητικά και το προτεινόμενο σύστημα φάνηκε να επεκτείνεται εύκολα. Ο μόνος περιορισμός τους είναι εάν ο

εισβολέας μάθει να συμπεριφέρεται όπως ο έγκυρος χρήστης που δημοσιεύει τη συνηθισμένη ώρα της ημέρας και μιλάει με άτομα που συνήθως αναφέρονται. Η δεύτερη εργασία μελέτησε πώς συμπεριφέρεται ένας χρήστης αφού έχει παραβιαστεί ο λογαριασμός του. Παραδόξως το ένα τέταρτο του συνόλου των χρηστών των οποίων οι λογαριασμοί παραβιάστηκαν έτειναν να μετακινηθούν σε νέο λογαριασμό.

Μια τελευταία έρευνα, εστίασε στο στυλ γραφής του χρήστη. Στην προσέγγιση αυτή, ο εντοπισμός παραβιασμένων λογαριασμών βασίζεται σε περιεχόμενο post-textual, σε αντίθεση με την [3], στην οποία χρησιμοποιήθηκα black-lists συνδέσμων για τον εντοπισμό ανεπιθύμητων μηνυμάτων, απαιτώντας ερωτήματα για πόρους εκτός των βάσεων δεδομένων των κοινωνικών δικτύων [30].

Στην συνέχεια, θα παρουσιαστούν διάφορες προσεγγίσεις με τα θετικά και τα αρνητικά τους στοιχεία, σχετικά με τις “content based techniques” γιατί είναι αυτές που αφορούν την παρούσα διπλωματική.

5.2.1. Brocardo et al. (Προσέγγιση No. 1) [31]

Εφαρμόστηκε μάθηση με επίβλεψη για να πραγματοποιηθεί επαλήθευση συγγραφέα σε μηνύματα ηλεκτρονικού ταχυδρομείου μικρού μεγέθους. Αντί να υιοθετηθεί η πιο συχνή πρακτική σύγκρισης της κατανομής συχνότητας των n-gram σε ένα κείμενο, οι ερευνητές εδώ, επικεντρώθηκαν στη δυαδική προσέγγιση, δηλαδή την παρουσία ή την απουσία ενός συγκεκριμένου n-gram. Επεκτείνοντας το έργο, οι ερευνητές χρησιμοποίησαν την αμοιβαία πληροφορία (mutual information) ως μέτρο επιλογής χαρακτηριστικών και ακολούθησαν μια υβριδική προσέγγιση που συνδυάζει τις μηχανές διανυσμάτων υποστήριξης (support vector machines – SVM) και της λογιστικής παλινδρόμησης (logistic regression).

Το προτεινόμενο μοντέλο αποτελείται από δύο βήματα, την εγγραφή (enrollment) και την επαλήθευση (verification). Η διαδικασία της εγγραφής περιλαμβάνει τη δημιουργία προφίλ συμπεριφοράς για έναν χρήστη, ενώ η φάση επαλήθευσης αντιμετωπίστηκε ως πρόβλημα κατηγοριοποίησης δύο κλάσεων. Λήφθηκαν τόσο θετικά όσο και αρνητικά δείγματα από χρήστες και εφαρμόστηκαν προσεγγίσεις ταξινόμησης σε αυτούς για επαλήθευση. Η υπερ-δειγματοληψία των δεδομένων (oversampling) διαχειρίστηκε με την ανάθεση αναλογικών βαρών στην oversampled (αρνητική) κλάση. Το βάρος που αντιστοιχεί στην αναλογία θετικών προς αρνητικών δειγμάτων, αποδόθηκε στα αρνητικά δείγματα.

Θετικά: Μελέτησε διαφορετικό συνδυασμό χαρακτηριστικών, ταξινομητών (classifiers) και διάφορα μεγέθη block για να συμπεράνει τα ποσοστά σφάλματος που αντιμετωπίζει κάθε ταξινομητής.

Αρνητικά: Η τεχνική εξακολουθεί να είναι επιρρεπής στο spoofing από επιτιθέμενους και ως εκ τούτου χρειάζεται μια πιο διεξοδική έρευνα. Ο όρος spoofing είναι ο τρόπος που χρησιμοποιεί κάποιος για να αποκτήσει πρόσβαση σε ένα δίκτυο ή μια υπηρεσία χωρίς να είναι ο νόμιμος χρήστης.

5.2.2. Igawa et al. (Προσέγγιση No. 2) [32]

Εφαρμόστηκε μια προσέγγιση καθαρά βασισμένη στην εξόρυξη κειμένου για την αναγνώριση παραβιασμένων λογαριασμών σε κοινωνικά δίκτυα. Συμπεράθηκε ότι τα hashtags και οι αναφορές φέρουν σημαντικές και σχετικές πληροφορίες και δεν πρέπει να αφαιρεθούν πριν από την επεξεργασία, αφού αν αφαιρεθούν οδηγεί στην μείωση της ακρίβειας σε μεγάλο βαθμό. Πραγματοποιήθηκαν πειράματα με διαφορετικές τιμές μεγέθους του corpus, διαφορετικές τιμές n στα n-grams, ποσότητα προ-επεξεργασίας κ.λπ. Η καλύτερη ακρίβεια επιτεύχθηκε για n = 6, μέγεθος του corpus 100 και έγινε μόνο stop-word removal ως βήμα προ-επεξεργασίας.

Σε συνέχεια της παραπάνω μελέτης, χρησιμοποιήθηκαν n-gram επαλήθευσης συγγραφέα, με βάση το «στυλ» γραφής των χρηστών, για την ανίχνευση παραβιασμένων λογαριασμών. Δημιουργήθηκε ένα προφίλ του «στυλ» γραφής του χρήστη, αναλύοντας προηγούμενα tweets και δημοσιεύσεις του στα κοινωνικά δίκτυα. Για κάθε νέο tweet/δημοσίευση, αν το στυλ γραφής δεν έμοιαζε με την αντίστοιχη τιμή κατωφλίου (threshold), σημειωνόταν ως ύποπτη δραστηριότητα και ενημερωνόντουσαν οι αντίστοιχοι χρήστες, καθώς και γίνονταν μπλοκ τα ύποπτα tweet/δημοσιεύσεις. Μεταβάλλοντας τις παραμέτρους, διαμορφώθηκαν πέντε τιμές κατωφλίου. Δοκιμάζοντας κάθε συνδυασμό, η μέγιστη ακρίβεια ήταν 86,72% και 95,13% για δύο σύνολα δεδομένων του twitter.

Θετικά: Εισήγαγε την έννοια της εξόρυξης κειμένου για τον εντοπισμό παραβιασμένων λογαριασμών. Τα tweets είναι εύκολα προσβάσιμα και η διαδικασία επεξεργασίας τους είναι αρκετά εύκολη και φαίνεται να είναι ένας αποτελεσματικός τρόπος αντιμετώπισης του προβλήματος.

Αρνητικά: Βασίστηκε μόνο στην ανάλυση n-gram, πράγμα που σημαίνει ότι δεν θα είναι αποδοτικό με χρήστες που δεν είναι συνεπείς με τη συμπεριφορά στις δημοσιεύσεις τους. Επομένως, η προσέγγιση χρειάζεται μια περαιτέρω ανάλυση άλλων χαρακτηριστικών.

5.2.3. Jenny et al. (Προσέγγιση No. 3) [33]

Εδώ χρησιμοποιήθηκαν στυλομετρικά χαρακτηριστικά για την επαλήθευση συγγραφέα και συνεπώς, για τον εντοπισμό παραβιασμένων λογαριασμών σε κοινωνικά δίκτυα. Σε αντίθεση με υπάρχουσες μελέτες σχετικά με την επαλήθευση συγγραφέα, που έχουν επικεντρωθεί μόνο σε μηνύματα ηλεκτρονικού ταχυδρομείου μεγάλου μεγέθους, αυτή η έρευνα επικεντρώθηκε σε μηνύματα μικρότερου μεγέθους σε κοινωνικά δίκτυα. Εξετάστηκε η αποτελεσματικότητα

αλγορίθμων μηχανικής μάθησης για το σωστό διαχωρισμό των δημοσιεύσεων ενός αυθεντικού χρήστη από τους κακόβουλους χρήστες, αναλύοντας τα μοτίβα γραφής και το ιστορικό των χρηστών. Λειτούργησε ως συμπληρωματική μέθοδος για τις υπάρχουσες μεθόδους επαλήθευσης συγγραφέα, τα οποία έψαχναν για επαλήθευση μόνο βάσει του ονόματος χρήστη και του κωδικού πρόσβασης κατά τη στιγμή της σύνδεσης. Οι ερευνητές ανέλυσαν 233 χαρακτηριστικά, βασισμένα σε στυλομετρικές μεθόδους και στα κοινωνικά δίκτυα σε συνδυασμό και βρήκαν σωστά τον συγγραφέα ενός μηνύματος με ακρίβεια 79,6%. Η επίπτωση διαφορετικών παραμέτρων, όπως ο τύπος και ο αριθμός των χαρακτηριστικών, ο αριθμός χρηστών, η χρήση κανονικοποίησης, η εφαρμογή διαφόρων αλγορίθμων ταξινόμησης και ο συνδυασμός ταξινομητών αναλύθηκαν και βγήκαν τα σχετικά συμπεράσματα.

Συνήχθη το συμπέρασμα ότι όλα τα χαρακτηριστικά μαζί έδωσαν καλύτερη απόδοση από ότι τα μεμονωμένα χαρακτηριστικά. Επίσης, οι χρήστες που είχαν κάποιο ιδιαίτερο στυλ στις δημοσιεύσεις, ήταν εύκολα προβλέψιμοι με αυτήν την μέθοδο. Η κανονικοποίηση των χαρακτηριστικών δεν βελτίωσε πολύ την απόδοση. Μεταξύ των διαφόρων ταξινομητών, τα δέντρα αποφάσεων (decision trees) και ο SVM απέδωσαν καλύτερα. Αναπτύχθηκε ένας αλγόριθμος ψηφοφορίας (voting algorithm) για την πλειοψηφία, που συνδυάζει την έξοδο του SVM, των δέντρων αποφάσεων και των νευρωνικών δικτύων για τη λήψη απόφασης.

Θετικά: Η απόδοση διαφορετικών ταξινομητών και αλγορίθμων ψηφοφορίας διερευνήθηκαν χρησιμοποιώντας στυλομετρικά χαρακτηριστικά σε κοινωνικά δίκτυα επιτυγχάνοντας 79,6% ως το μέγιστο ποσοστό ακρίβειας.

Αρνητικά: Το μοντέλο αυτό δεν φάνηκε να διακρίνει τους χρήστες με παρόμοιες συμπεριφορές και αλληλεπιδράσεις (πιθανώς φίλοι), επομένως, χρειάζονται υποστήριξη από άλλα είδη χαρακτηριστικών.

5.2.4. Barbon et al. (Προσέγγιση No. 4) [30]

Εδώ μελετήθηκε πόσο αποτελεσματικές είναι οι τεχνικές της επαλήθευσης συγγραφέα στον εντοπισμό παραβιασμένων λογαριασμών στα κοινωνικά δίκτυα. Προτάθηκε μια προσέγγιση με την χρήση n-grams βασισμένη στην εξόρυξη κειμένου. Η χρήση των n-gram προτιμήθηκε για την ευκολία αυτής τη μεθόδου. Ακολουθήθηκε μια διαδικασία 3 βημάτων: η δημιουργία βάσης (baseline), το ταίριασμα (matching) και η ενημέρωση (updating). Προκειμένου να αποφευχθεί το πρόβλημα της εποχικότητας, πραγματοποιήθηκε διαμοιρασμός και το έγγραφο, που περιέχει όλα τα tweets, χωρίστηκε σε baseline set, threshold sets και test sets. Για τη συμπερίληψη δυναμικής ενημέρωσης του προφίλ, χρησιμοποιήθηκε ένας ταξινομητής εκμάθησης με βάση την παρουσία (instance-based learning), ο αλγόριθμος των k-κοντινότερων γειτόνων (k-nearest neighbors – kNN). Σημαντική συμβολή ήταν η συνεχής αναβάθμιση του baseline (adaptive baseline) καθώς εκτός από τη βελτίωση της ακρίβειας (κατά 60%), συνέβαλε στην κάλυψη της τάσης αλλαγών στο «στυλ» γραφής και άλλων σχετικών χαρακτηριστικών. Η προτεινόμενη προσέγγιση αποδείχτηκε εξαιρετικά σημαντική επιτυγχάνοντας ακρίβεια 93%, με

τον μοναδικό περιορισμό ότι εξαρτάται σε μεγάλο βαθμό από την παρουσία «ομοιομορφίας» στις δημοσιεύσεις των χρηστών.

Θετικά: Αποτελεσματική ανάλυση της εφαρμογής των n-grams για την επαλήθευση συγγραφέων, και πιο συγκεκριμένα για την ανίχνευση παραβιασμένων λογαριασμών.

Αρνητικά: Αποτυχία αναγνώρισης των χρηστών που δεν ακολουθούν το μοτίβο n-gram, δηλαδή η ασυνέπεια στο «στυλ» των δημοσιεύσεων ενός χρήστη. Επομένως, χρειάζεται κι άλλα χαρακτηριστικά συμπληρωματικά για τη μείωση των εσφαλμένα θετικών/αρνητικών (FP/FN).

6

Υλικό και μέθοδοι

6.1. Διαγωνισμός PAN

Το PAN είναι μια σειρά επιστημονικών εκδηλώσεων και διαμοιραζόμενων εργασιών σχετικά με την ψηφιακή εγκληματολογία και τη στυλομετρία (stylometry).

Ένα από τα πεδία που απασχόλησε τα τελευταία χρόνια τον συγκεκριμένο διαγωνισμό ήταν και αυτό της επαλήθευσης συγγραφέα [18]. Πιο συγκεκριμένα, στην παρούσα διπλωματική εργασία θα αξιοποιηθούν αλγόριθμοι που προτάθηκαν στον διαγωνισμό "Authorship Verification 2020" [34]. Τα χαρακτηριστικά του διαγωνισμού είναι τα εξής:

6.1.1. Σύνοψη

Εργασία (Task): Λαμβάνοντας υπόψη δύο κείμενα, να προσδιοριστεί εάν είναι γραμμένα από τον ίδιο συγγραφέα.

Είσοδος (Input): Ιστορίες που συλλέχθηκαν από το fanfiction.net. 53k ζεύγη κειμένων συνολικά. Μικρό και μεγάλο training set.

Αξιολόγηση (Evaluation): AUC, F1, c@1, F_0.5u

Υποβολή (Submission): Ανάπτυξη στο TIRA

Βασική μέθοδος (Baseline): TFIDF-weighted char 3-grams cosine similarity. Μέθοδος συμπίεσης για τον υπολογισμό της εγκάρσιας εντροπίας

6.1.2. Εργασία

Όπως αναφέρθηκε παραπάνω, στόχος στον συγκεκριμένο διαγωνισμό ήταν η επαλήθευση συγγραφέα, δηλαδή εάν δύο κείμενα έχουν γραφτεί από τον ίδιο συγγραφέα με βάση τη σύγκριση των στυλ γραφής των κειμένων.

Τα επερχόμενα τρία χρόνια στο PAN 2020 έως το PAN 2022, αναπτύχθηκε μια νέα πειραματική ρύθμιση που αντιμετωπίζει τρία βασικά ερωτήματα στην επαλήθευση συγγραφέα που δεν έχουν μελετηθεί πολύ μέχρι σήμερα:

- 1η χρονιά (PAN 2020): Επαλήθευση κλειστού συνόλου.

  Δεδομένου ενός μεγάλου training dataset που αποτελείται από γνωστούς συγγραφείς που έχουν γράψει για ένα συγκεκριμένο σύνολο θεμάτων, το test dataset περιέχει περιπτώσεις επαλήθευσης από ένα υποσύνολο των συγγραφέων και θέματα που βρίσκονται στα δεδομένα εκπαίδευσης.

- 2η χρονιά (PAN 2021): Επαλήθευση ανοιχτού συνόλου.

  Λαμβάνοντας υπόψη το σύνολο δεδομένων εκπαίδευσης του 1ου έτους, το test dataset περιέχει περιπτώσεις επαλήθευσης από προηγούμενους άγνωστους συγγραφείς και θέματα.

- 3η χρονιά (PAN 2022): Εργασία-έκπληξη.

  Το έργο του τελευταίου έτους αυτού του κύκλου αξιολόγησης θα σχεδιαστεί με επίκεντρο τον ρεαλισμό και την πρακτική εφαρμογή.

Αυτός ο κύκλος αξιολόγησης σχετικά με την επαλήθευση συγγραφέα παρέχει μια νέα πρόκληση αυξανόμενης δυσκολίας στο πλαίσιο μιας μεγάλης κλίμακας αξιολόγησης.

6.1.3. Δεδομένα

Τα σύνολα δεδομένων για την εκπαίδευση αποτελούνται από ζεύγη (αποσπάσματα) δύο διαφορετικά fanfics (ιστορίες των φαν), τα οποία αποκτήθηκαν από το fanfiction.net. Σε κάθε ζεύγος εκχωρήθηκε ένα μοναδικό αναγνωριστικό και διακρίνουμε μεταξύ ζευγών ίδιου συγγραφέα και διαφορετικών συγγραφέων. Επιπλέον, υπάρχουν μεταδεδομένα στο fandom (δηλ. θεματική κατηγορία) για κάθε κείμενο του ζεύγους. Η τυχαία κατανομή σε αυτά τα σύνολα δεδομένων προσεγγίζει κατά πολύ την κατανομή (long-tail) των fandoms στο αρχικό σύνολο δεδομένων.

Το training set διατίθεται σε δύο παραλλαγές: ένα μικρότερο σύνολο δεδομένων, ιδιαίτερα κατάλληλο για “κλασικές” μεθόδους μηχανικής μάθησης και ένα μεγάλο, σύνολο δεδομένων, κατάλληλο για την εφαρμογή αλγορίθμων βαθιάς μάθησης. Τα μοντέλα που χρησιμοποιούν το μικρό σετ θα αξιολογούνται ξεχωριστά από τα μοντέλα που χρησιμοποιούν το μεγάλο σετ.

Τόσο το μικρό όσο και το μεγάλο σύνολο δεδομένων αποτελούνται από δύο αρχεία JSON οριοθετημένα με νέα γραμμή το καθένα (* .jsonl). Το πρώτο αρχείο περιέχει ζεύγη κειμένων (κάθε ζεύγος έχει ένα μοναδικό αναγνωριστικό) και τις ετικέτες τους:

```
1. {"id": "6cced668-6e51-5212-873c-717f2bc91ce6", "fandoms": ["Fandom 1", "Fandom
   2"], "pair": ["Text 1...", "Text 2..."]}
2. {"id": "ae9297e9-2ae5-5e3f-a2ab-ef7c322f2647", "fandoms": ["Fandom 3", "Fandom
   4"], "pair": ["Text 3...", "Text 4..."]}
3. ...
```

Το δεύτερο αρχείο, που τελειώνει σε * _truth.jsonl, περιέχει τη βασική αλήθεια (ground truth) για όλα τα ζεύγη. Αυτή αποτελείται από μια boolean “σημαία” που δείχνει εάν τα κείμενα σε ένα ζεύγος προέρχονται από τον ίδιο συντάκτη και τα αριθμητικά αναγνωριστικά συγγραφέων:

```
1. {"id": "6cced668-6e51-5212-873c-
   717f2bc91ce6", "same": true, "authors": ["1446633", "1446633"]}
2. {"id": "ae9297e9-2ae5-5e3f-a2ab-
   ef7c322f2647", "same": false, "authors": ["1535385", "1998978"]}
3. ...
```

Τα δεδομένα και η βασική αλήθεια είναι στην ίδια σειρά και μπορούν να χρησιμοποιηθούν παράλληλα χωρίς την ανάγκη αλλαγής βάσει του ID ζεύγους. Οι ετικέτες των fandom δίνονται και για το training και για το test set. Το αρχείο βασικής αλήθειας θα είναι διαθέσιμο μόνο για τα δεδομένα εκπαίδευσης.

6.1.4. Βασικές Μέθοδοι (Baselines)

Στο PAN 2020 παρέχονται οι ακόλουθες βασικές μέθοδοι:

- Μια απλή μέθοδος που υπολογίζει τις ομοιότητες συνημιτόνων (cosine similarity) μεταξύ των TFIDF-normalized αναπαραστάσεων bag-of-character-tetragrams, των κειμένων σε ένα ζεύγος. Τα αποτελέσματα που προκύπτουν έπειτα αλλάζουν, χρησιμοποιώντας μια απλή αναζήτηση πλέγματος (grid search), για να επιτευχθεί η βέλτιστη απόδοση στα δεδομένα.
- Μια απλή μέθοδος που βασίζεται στη συμπίεση κειμένου, όπου δεδομένου ενός ζεύγους κειμένων υπολογίζει την εγκάρσια εντροπία του κειμένου2 χρησιμοποιώντας το μοντέλο πρόβλεψης με μερική αντιστοίχιση (prediction by partial matching) του κειμένου1 και το αντίστροφο. Στη συνέχεια, η μέση και απόλυτη διαφορά των δύο

εγκάρσιων εντροπιών χρησιμοποιείται από ένα μοντέλο λογιστικής παλινδρόμησης (logistic regression) για την εκτίμηση της βαθμολογίας επαλήθευσης στο διάστημα [0,1].

6.1.5. Αξιολόγηση

Τα συστήματα θα συγκριθούν και θα ταξινομηθούν με βάση τέσσερις, συμπληρωματικές μετρικές:

- AUC: η συμβατική βαθμολογία περιοχής κάτω από την καμπύλη, όπως εφαρμόζεται στο scikit-learn.
- F1-score: το γνωστό μέτρο απόδοσης (χωρίς να λαμβάνονται υπόψη οι μη-απαντήσεις), όπως εφαρμόζεται στο scikit-learning.
- c@1: μια παραλλαγή του συμβατικού F1-score, η οποία επιβραβεύει συστήματα που αφήνουν αναπάντητα δύσκολα προβλήματα (δηλαδή σκορ ακριβώς 0.5).
- F_0.5u: μια νέα πρόσφατη προτεινόμενη μετρική που δίνει μεγαλύτερη έμφαση στη σωστή επιλογή υποθέσεων ίδιου συγγραφέα.

Ένα σενάριο αξιολόγησης αναφοράς είναι διαθέσιμο. Τα συστήματα θα εφαρμοστούν σε ένα ζεύγος-αρχείων και αναμένεται να παράγουν ένα μόνο αρχείο jsonl ως αποτέλεσμα. Σε αυτό το αρχείο πρόβλεψης, κάθε ξεχωριστή γραμμή πρέπει να περιέχει μια έγκυρη συμβολοσειρά json που παρέχει το αναγνωριστικό ενός ζεύγους και ένα πεδίο "τιμής", με μια βαθμολογία που κυμαίνεται στο [0, 1], υποδεικνύοντας την πιθανότητα ότι αυτό το ζεύγος είναι ζεύγος κειμένων ίδιου συγγραφέα. Τα συστήματα επιτρέπεται να αφήνουν ορισμένα προβλήματα αναπάντητα: σε τέτοιες περιπτώσεις, η απάντηση μπορεί να μείνει έξω από το αρχείο πρόβλεψης Ή η τιμή του πρέπει να οριστεί ακριβώς στο 0.5. Όλες οι απαντήσεις που έχουν τιμή 0.5 θα θεωρούνται μη-αποφάσεις.

Στην συνέχεια παρουσιάστηκαν κάποια στοιχεία σχετικά με την υποβολή των αρχείων για τους διαγωνιζόμενους.

6.2. Δημιουργία των dataset

Για τις ανάγκες της παρούσας διπλωματικής εργασίας, αξιοποιήθηκε ένα dataset, το οποίο έχει χρησιμοποιηθεί σε προηγούμενη έρευνα [35]. Το dataset αυτό αποτελείται από συλλογή δημοσιεύσεων χρηστών (tweets) στην πλατφόρμα κοινωνικής δικτύωσης Twitter. Οι δημοσιεύσεις έχουν πραγματοποιηθεί το χρονικό διάστημα από το Σεπτέμβριο 2007 έως και το Μάρτιο 2014. Στην προαναφερθείσα έρευνα το dataset χρησιμοποιήθηκε στο πρόβλημα της απόδοσης συγγραφέα (authorship attribution).

Στην συνέχεια, θα περιγραφεί αναλυτικά η αρχική μορφή και το περιεχόμενο του dataset και τα βήματα προ-επεξεργασίας που ακολουθήθηκαν ώστε να προσαρμοστεί το dataset αυτό στις ανάγκες της παρούσας διπλωματικής.

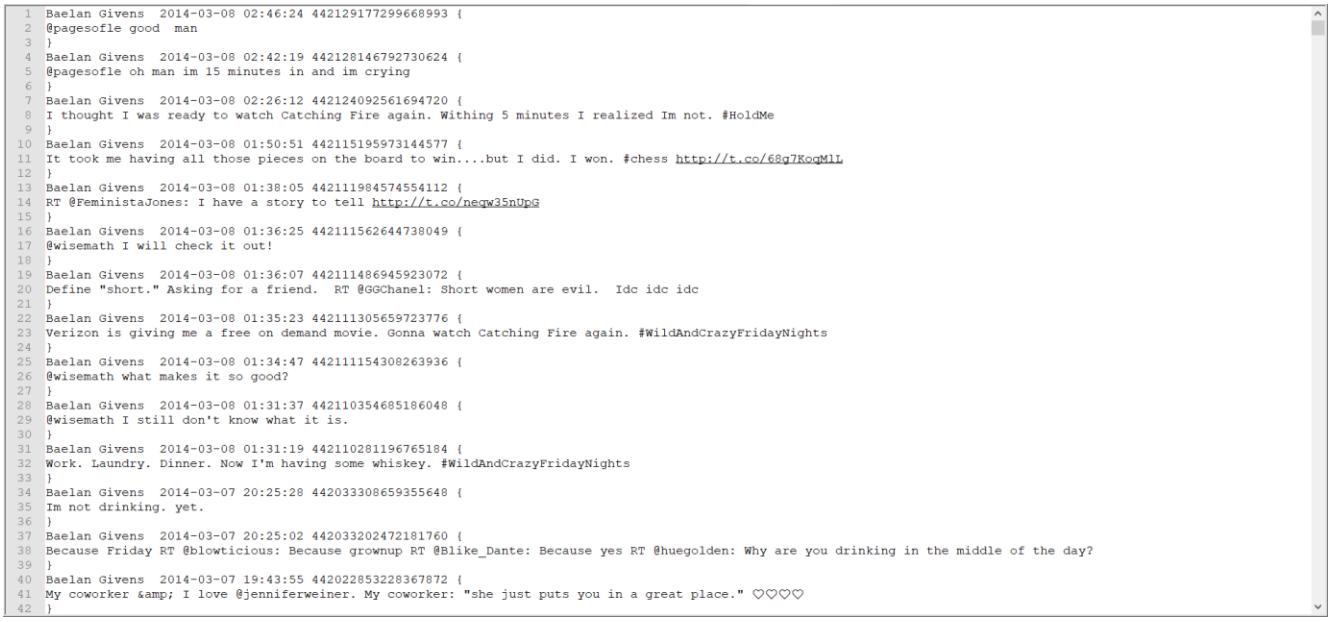
Αρχικά, μελετήθηκε και αναλύθηκε η μορφή και το περιεχόμενο του dataset. Πιο συγκεκριμένα, αυτό αποτελείται από 6.312 αρχεία, όπου κάθε ένα από αυτά αντιστοιχεί σε έναν χρήστη και περιέχει το σύνολο των δημοσιεύσεών του (tweets). Τα ονόματα των αρχείων αποτελούν ουσιαστικά το id του κάθε χρήστη. Μέσα σε κάθε αρχείο υπάρχουν τα tweets των χρηστών, τα οποία συμπεριλαμβάνουν στοιχεία όπως το όνομα του χρήστη στην πλατφόρμα (username), η ημερομηνία και η ώρα δημοσίευσης των tweets, το αναγνωριστικό id του tweet και, τέλος, το περιεχόμενο του tweet. Οι εγγραφές στο εκάστοτε αρχείο έχουν την μορφή που παρουσιάζεται παρακάτω:

```
[όνομα_χρήστη] [ημερομηνία] [ώρα] [id] {
      [tweet]
}
[όνομα_χρήστη] [ημερομηνία2] [ώρα2] [id2] {
      [tweet2]
}
```

Επίσης, στο τέλος κάθε αρχείου υπάρχει μια γραμμή που υποδεικνύει τον συνολικό αριθμό των tweets του συγκεκριμένου χρήστη. Η γραμμή αυτή είναι της μορφής:

```
[number_of_tweets] tweets extracted from [user_id]
```

Στην Εικόνα 16 φαίνεται ενδεικτικά ένα αρχείο ενός χρήστη. Σε πρώτο στάδιο, διαγράφηκαν τα αρχεία των χρηστών τα οποία δεν περιείχαν κανένα tweet (έμειναν 6293 χρήστες). Έπειτα, δημιουργήθηκε ένας πίνακας ο οποίος περιέχει τον αριθμό των tweets για κάθε χρήστη ανά μήνα (Εικόνα 17).



```
Baelan Givens  2014-03-08 02:46:24 442129177299668993 {
@pagesofle good  man
}
Baelan Givens  2014-03-08 02:42:19 442128146792730624 {
@pagesofle oh man im 15 minutes in and im crying
}
Baelan Givens  2014-03-08 02:26:12 442124092561694720 {
I thought I was ready to watch Catching Fire again. Withing 5 minutes I realized Im not. #HoldMe
}
Baelan Givens  2014-03-08 01:50:51 442115195973144577 {
It took me having all those pieces on the board to win....but I did. I won. #chess http://t.co/68g7KoqM1L
}
Baelan Givens  2014-03-08 01:38:05 442111984574554112 {
RT @FeministaJones: I have a story to tell http://t.co/neqw35nUpG
}
Baelan Givens  2014-03-08 01:36:25 442111562644738049 {
@wisemath I will check it out!
}
Baelan Givens  2014-03-08 01:36:07 442111486945923072 {
Define "short." Asking for a friend.  RT @GGChanel: Short women are evil.  Idc idc idc
}
Baelan Givens  2014-03-08 01:35:23 442111305659723776 {
Verizon is giving me a free on demand movie. Gonna watch Catching Fire again. #WildAndCrazyFridayNights
}
Baelan Givens  2014-03-08 01:34:47 442111154308263936 {
@wisemath what makes it so good?
}
Baelan Givens  2014-03-08 01:31:37 442110354685186048 {
@wisemath I still don't know what it is.
}
Baelan Givens  2014-03-08 01:31:19 442110281196765184 {
Work. Laundry. Dinner. Now I'm having some whiskey. #WildAndCrazyFridayNights
}
Baelan Givens  2014-03-07 20:25:28 442033308659355648 {
Im not drinking. yet.
}
Baelan Givens  2014-03-07 20:25:02 442033202472181760 {
Because Friday RT @blowticious: Because grownup RT @Blike_Dante: Because yes RT @huegolden: Why are you drinking in the middle of the day?
}
Baelan Givens  2014-03-07 19:43:55 442022853228367872 {
My coworker &amp;amp; I love @jenniferweiner. My coworker: "she just puts you in a great place." ♡♡♡♡
}
```

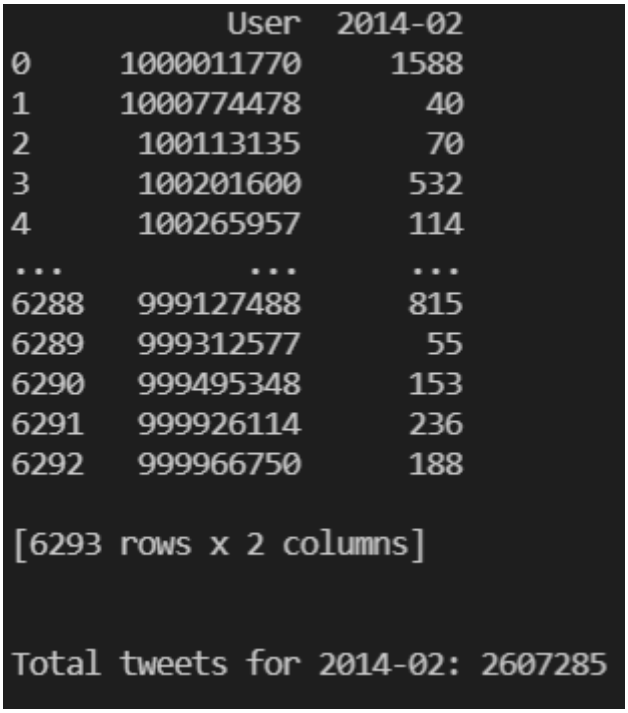
Εικόνα 16 Ένα αρχείο του dataset

User	2007-01	2007-02	2007-03	2007-04	2007-05	2007-06	...	2014-06	2014-07	2014-08	2014-09	2014-10	2014-11	2014-1
1000011770	0	0	0	0	0	0	...	0	0	0	0	0	0	
1000774478	0	0	0	0	0	0	...	0	0	0	0	0	0	
100113135	0	0	0	0	0	0	...	0	0	0	0	0	0	
100201600	0	0	0	0	0	0	...	0	0	0	0	0	0	
100265957	0	0	0	0	0	0	...	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	..
999127488	0	0	0	0	0	0	...	0	0	0	0	0	0	
999312577	0	0	0	0	0	0	...	0	0	0	0	0	0	
999495348	0	0	0	0	0	0	...	0	0	0	0	0	0	
999926114	0	0	0	0	0	0	...	0	0	0	0	0	0	
999966750	0	0	0	0	0	0	...	0	0	0	0	0	0	

Εικόνα 17 Ο πίνακας των tweet για όλους τους μήνες σαν dataframe

Το χρονικό διάστημα όλων των tweets είναι από 2007-09 έως 2014-03. Με βάση αυτόν τον παραπάνω πίνακα, προκύπτει ότι οι μήνες με τα περισσότερα tweets συνολικά, ήταν οι τελευταίοι μήνες του 2013, καθώς και οι μήνες 2014-01, 2014-02 και 2014-03. Για τις ανάγκες της συγκεκριμένης εργασίας, είναι επιθυμητό να εντοπιστούν οι κατά το δυνατόν περισσότερες δημοσιεύσεις σε όσο το δυνατόν μικρότερο χρονικό εύρος.

Για λόγους ευκολίας, επιλέχθηκε το χρονικό αυτό εύρος να είναι ένας μήνας. Πιο συγκεκριμένα, ο μεγαλύτερος αριθμός tweets παρατηρήθηκε στον μήνα 2014-02 (Εικόνα 18). Έτσι, τα tweets του συγκεκριμένου μήνα επιλέχθηκαν για την δημιουργία του νέου dataset στο πλαίσιο της συγκεκριμένης εργασίας.



```
            User  2014-02
0     1000011770      1588
1     1000774478        40
2      100113135        70
3      100201600       532
4      100265957       114
...          ...       ...
6288   999127488       815
6289   999312577        55
6290   999495348       153
6291   999926114       236
6292   999966750       188

[6293 rows x 2 columns]


Total tweets for 2014-02: 2607285
```

Εικόνα 18 Η στήλη των tweets για τον μήνα 2014-02 και το άθροισμα αυτών

Έπειτα, δημιουργήθηκε ένας άλλος πίνακας ο οποίος περιέχει τα ids όλων των χρηστών και τον αντίστοιχο αριθμό tweets που είχαν τον μήνα 2014-02. Στην συνέχεια, έγινε ταξινόμηση με βάση τη φθίνουσα σειρά του πλήθους των tweets των χρηστών για αυτόν τον μήνα. Για λόγους υπολογιστικής πολυπλοκότητας επιλέχθηκαν οι πρώτοι 50 χρήστες με τα περισσότερα tweets και έγινε ο πίνακας στην Εικόνα 19.

	A	B
1	User	No. of Tweets
2	1124513143	2447
3	1130074201	2461
4	1238373806	2450
5	127297106	2521
6	135511135	2349
7	138398982	2472
8	1386804798	2385
9	1419508352	2264
10	14678067	2241
11	1474368482	2705
12	1523922697	2735
13	155991103	2630
14	155991469	2648
15	1564224128	2510
16	18299613	2376
17	1895409524	2284
18	1920932666	2390
19	195966016	2383
20	20993979	2576
21	2209947793	2395
22	2232433148	2463
23	2266980822	2522
24	2270009131	2293
25	23880967	2550
26	25512667	2477
27	266611274	2487
28	292009132	2286
29	299286187	2251
30	301897035	2359
31	311304699	2286
32	37736362	2227

Εικόνα 19 Οι 50 χρήστες με το μεγαλύτερο πλήθος tweet τον μήνα 2014-02

Σε επόμενο βήμα, δημιουργήθηκε ένας πίνακας με το περιεχόμενο των ανωτέρω tweets των 50 χρηστών. Ο πίνακας αποθηκεύτηκε σε μορφή αρχείου .csv και περιέχει το αναγνωριστικό του χρήστη καθώς και την δημοσίευσή του (tweet), όπως φαίνεται στην Εικόνα 20.

	A	B	C
1		User	Tweet
2	0	1124513143	@SecondYearAce > 'excuse you what is so funny. emoticons are cool'
3	1	1124513143	[[Evening.]]
4	2	1124513143	@SecondYearAce > 'force of habit' passed the helmet along with a smiley face sticking its tongue out. That is so childish, Strider
5	3	1124513143	@SecondYearAce + off the shades'
6	4	1124513143	@SecondYearAce > You nodded, raising a thumb to assure her you're okay.
7	5	1124513143	@iSharkRin [[Ah, ha'i. Arigatou!]]
8	6	1124513143	@iSharkRin [[Permisi, bisa minta asrama WE?]]
9	7	1124513143	@xGuardianOfFun_ [[Sip2 makasih mas. /ikut mengamini/ /heh
10	8	1124513143	@xGuardianOfFun_ [[Mas, kalo karakternya dari webcomic boleh join gak? o/]]
11	9	1124513143	@SecondYearAce + that was left untouched, the beep sounding slightly sad
12	10	1124513143	@SecondYearAce > Your helmet beeped like you were happy with her reaction, sitting down again and staring at the bottle of apple juice +
13	11	1124513143	@AkariShiro_OZ + more trees. You decided you should climb up one of them and hide for a moment. But that's a cowardly action
14	12	1124513143	@AkariShiro_OZ + collar of your shirt, your shoulders sagged as you noticed you could see no signs of a lake or a river. Just trees and +
15	13	1124513143	@AkariShiro_OZ > You're not going to answer, you're not that stupid to reveal your location. Shades now folded and left hanging at the +
16	14	1124513143	@ReisenNoUsagi kinda. could be called something like that
17	15	1124513143	@SecondYearAce + bright red passed your helmet
18	16	1124513143	+
19	17	1124513143	[[....What is happening.]]

users_50_tweets

Εικόνα 20 Η αρχή του αρχείου με όλα τα tweets των 50 χρηστών που επιλέχθηκαν για τον μήνα 2014-02

Ο πίνακας αποτελείται από 122.164 tweets και χρησιμοποιήθηκε στη δημιουργία του τελικού dataset, το οποίο θα πρέπει να έχει τη μορφή του dataset όπως χρησιμοποιήθηκε στον διαγωνισμό PAN2020, που περιγράφηκε παραπάνω, δηλαδή της μορφής:

1. `{"id": "1", "tweets": ["Tweet 1", "Tweet 2"], "pair": ["Tweet 1...", "Tweet 2..."]}`
2. `{"id": "2", "tweets": ["Tweet 3", " Tweet 4"], "pair": ["Tweet 3...", " Tweet 4..."]}`
3. `...`

```
1. {"id": "1", "same": true, "authors": ["1124513143", "1124513143"]}
2. {"id": "2", "same": false, "authors": ["2270009131", "266611274"]}
3. ...
```

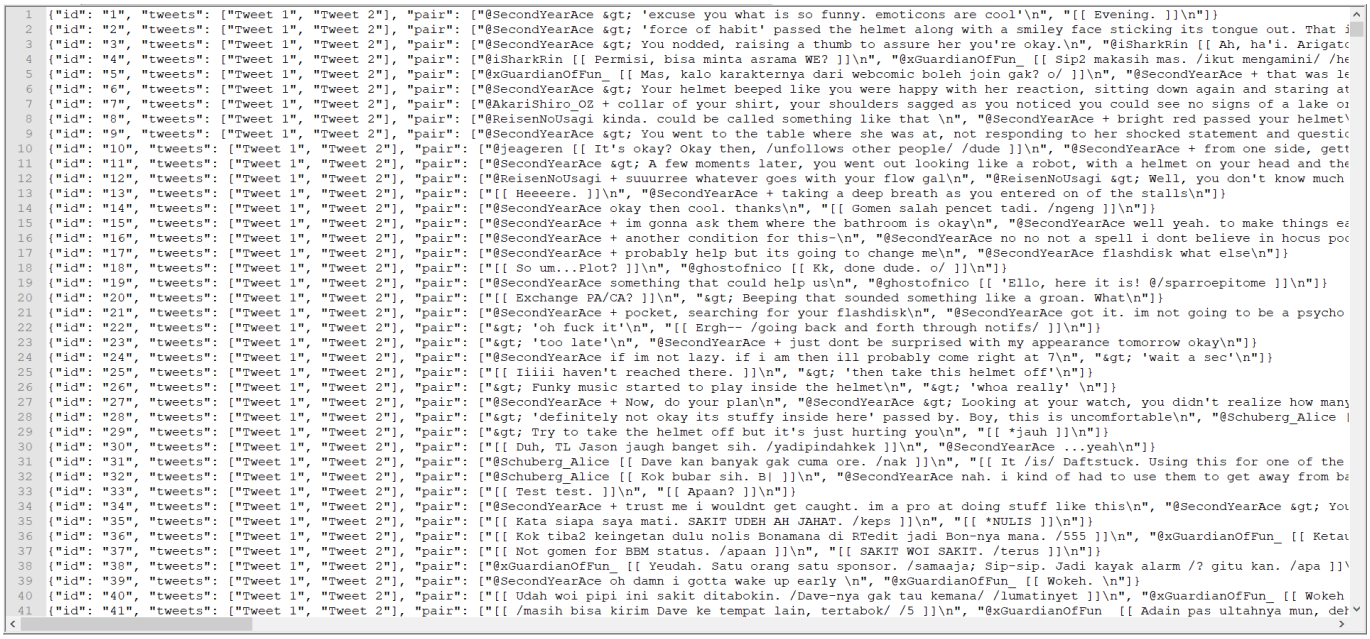
Ο αριθμός των tweets ανά χρήστη για τον υπό μελέτη μήνα, κυμαινόταν από 2.226 έως 2.788, όπως φαίνεται και στον πίνακα στην Εικόνα 19. Για το λόγο αυτό, ορίστηκε το πλήθος των tweets που θα αξιοποιηθούν από κάθε χρήστη να είναι 2.200.

Έτσι, για τη δημιουργία του τελικού dataset, επιλέχθηκαν περίπου 1.100 ζευγάρια tweets για κάθε χρήστη (δηλαδή περίπου 2.200 tweets/χρήστη). Ακολουθώντας τη μορφή του dataset του διαγωνισμού PAN2020, αποφασίστηκε να διατηρηθεί η αναλογία του πλήθους των ζευγαριών με ίδιο και διαφορετικό συγγραφέα. Ειδικότερα, δημιουργήθηκαν 33.000 ζευγάρια tweets του ίδιου συγγραφέα (~60% του πλήθους του dataset – 660 ζευγάρια/χρήστη) και 22.000 ζευγάρια tweets διαφορετικού συγγραφέα (~40% του πλήθους του dataset – ο κάθε χρήστης εμφανίζεται στατιστικά σε περίπου 440 ζευγάρια). Με άλλα λόγια, το dataset περιλαμβάνει 33.000 ζευγάρια με την ένδειξη «true» (ίδιου συγγραφέα) και 22.000 ζευγάρια με την ένδειξη “false” (διαφορετικών συγγραφέων). Ως εκ τούτου, 1.320 tweets για κάθε χρήστη προορίστηκαν για τη δημιουργία ζευγαριών ίδιου συγγραφέα (true) και περίπου 880 για τη δημιουργία ζευγαριών διαφορετικών συγγραφέων (false).

Ο τρόπος της δημιουργίας των ζευγαριών του ίδιου συγγραφέα έγινε σειριακά (το 1^{ο} tweet έγινε ζευγάρι με το 2^{ο}, το 3^{ο} tweet έγινε ζευγάρι με το 4^{ο} κ.ο.κ). Το κάθε tweet διαγράφονταν μετά την συμπερίληψή του σε κάποιο ζευγάρι, οπότε και χρησιμοποιούνταν μια και μόνο φορά.

Ο τρόπος της δημιουργίας των ζευγαριών διαφορετικών συγγραφέων πραγματοποιήθηκε με τυχαίο τρόπο, δημιουργώντας ζευγάρια tweets που είχαν διαφορετικό αναγνωριστικό χρήστη (id). Λόγω του ότι οι χρήστες δεν είχαν τον ίδιο αριθμό tweets, υπήρχε μια περίσσια tweets (12.164) τα οποία δεν αξιοποιήθηκαν για την δημιουργία κάποιου ζευγαριού δεν συμπεριλήφθηκαν στο τελικό dataset.

Έτσι, το τελικό dataset αποτελείται από 55.000 ζευγάρια tweets από τους 50 χρήστες. Τα 33.000 ήταν ζευγάρια tweets του ίδιου συγγραφέα και 22.000 ήταν ζευγάρια διαφορετικών συγγραφέων. Στην Εικόνα 21 μπορούμε να δούμε ενδεικτικά τις πρώτες γραμμές του dataset.



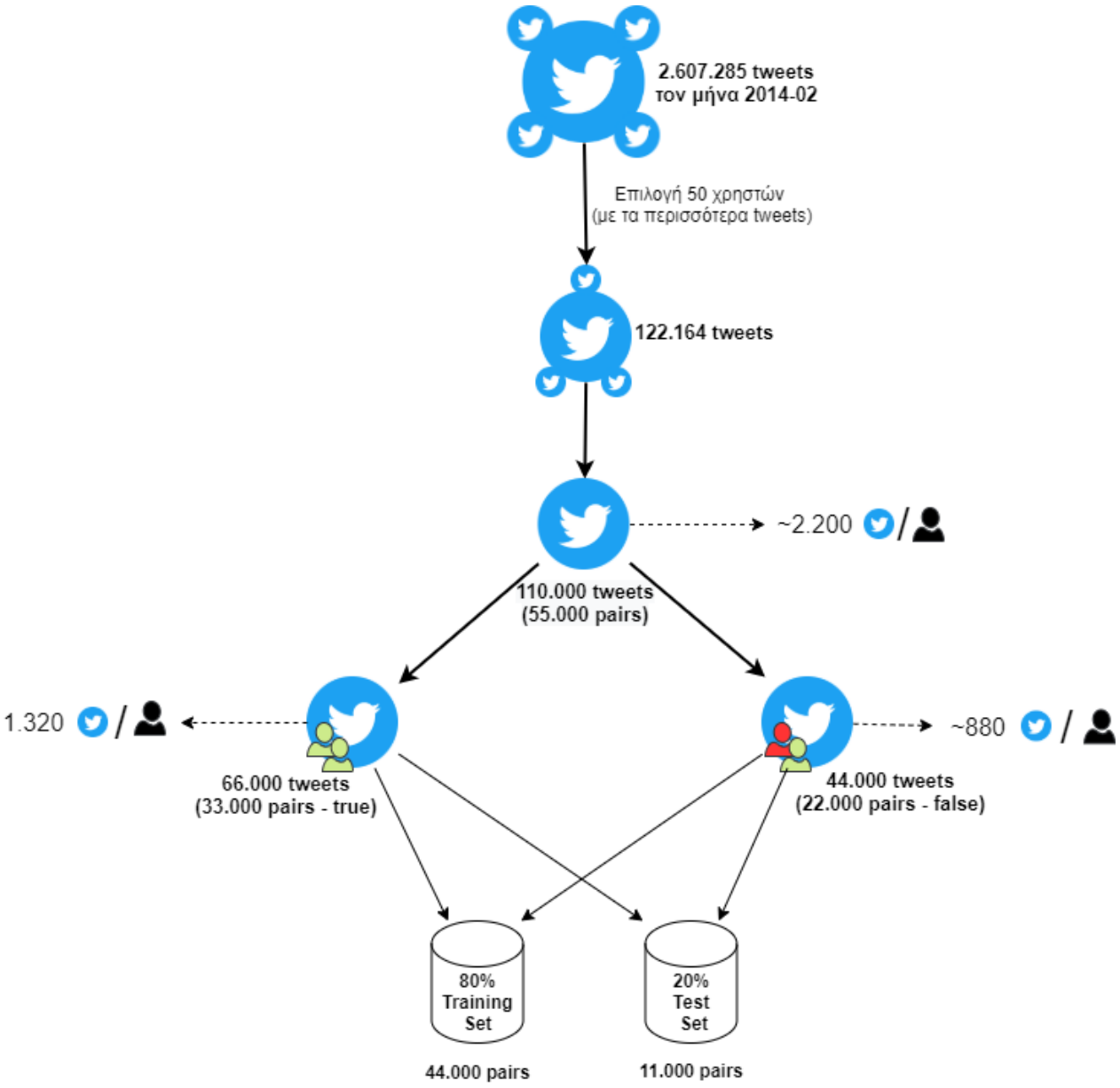
Εικόνα 21 Οι πρώτες γραμμές του τελικού dataset με τα pairs των tweets

Όπως σημειώθηκε πιο πάνω, στον διαγωνισμό PAN2020, υπάρχει κι ένα αρχείο με το όνομα truth.jsonl, όπου περιέχει το ground truth για όλα τα ζευγάρια. Έτσι, παράλληλα με την δημιουργία του dataset, δημιουργήθηκε και το αρχείο truth για το προς μελέτη dataset, όπου περιέχει το ground truth για όλα τα ζευγάρια που δημιουργήθηκαν παραπάνω (Εικόνα 21). Άρα, δημιουργήθηκε ένα αρχείο, με τις ίδιες σε πλήθος καταχωρήσεις, όπως το βασικό μας dataset, όπου είχε τα ground truth για κάθε ζευγάρι, όπως φαίνεται στην Εικόνα 22:

```
{"id": "1", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "2", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "3", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "4", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "5", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "6", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "7", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "8", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "9", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "10", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "11", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "12", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "13", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "14", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "15", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "16", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "17", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "18", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "19", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "20", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "21", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "22", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "23", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "24", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "25", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "26", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "27", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "28", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "29", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "30", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "31", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "32", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "33", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "34", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "35", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "36", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "37", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "38", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "39", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "40", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "41", "same": true, "authors": ["1124513143", "1124513143"]}
{"id": "42", "same": true, "authors": ["1124513143", "1124513143"]}
```


Εικόνα 22 Οι πρώτες γραμμές του αρχείου με τα ground truth

Η διαδικασία και τα βήματα που περιγράφηκαν παραπάνω φαίνονται διαγραμματικά στην Εικόνα 23:



Εικόνα 23 Διάγραμμα με τα βήματα δημιουργίας του dataset

Το αρχεία αυτά που δημιουργήθηκαν χρειάζονται στην τροφοδότηση του αλγορίθμου του baseline κώδικα του PAN2020. Ο baseline κώδικας του PAN2020 προσφέρει μια απλοϊκή, αλλά γρήγορη λύση σχετικά με την επαλήθευση συγγραφέα. Όλα τα έγγραφα αναπαρίστανται χρησιμοποιώντας ένα μοντέλο bag-of-character-n-grams, το οποίο είναι σταθμισμένο με TFIDF. Υπολογίζεται η ομοιότητα συνημιτόνου μεταξύ κάθε ζεύγους εγγράφων στο σύνολο δεδομένων βαθμονόμησης (calibration dataset). Τέλος, οι ομοιότητες που προκύπτουν βελτιστοποιούνται και προβάλλονται μέσω μιας απλής διαδικασίας επανακλιμάκωσης, έτσι ώστε να μπορούν να λειτουργήσουν ως ψευδο-πιθανότητες, υποδεικνύοντας την πιθανότητα

ένα ζεύγους εγγράφων να είναι ζευγάρι του ίδιου συγγραφέα. Μέσω αναζήτησης πλέγματος, καθορίζεται το βέλτιστο όριο επαλήθευσης, λαμβάνοντας υπόψη ότι ορισμένα δύσκολα προβλήματα μπορούν να μείνουν αναπάντητα.

Ο αλγόριθμος του baseline απαιτεί την τροφοδότηση με ένα training και ένα test set, καθώς και με τις αντίστοιχες απαντήσεις τους σχετικά με την ύπαρξη ή όχι ζευγαριών κειμένου του ίδιου συγγραφέα (ground truth). Έτσι, πραγματοποιήθηκε διαίρεση του τελικού dataset σε σύνολο εκπαίδευσης (training set) και σύνολο ελέγχου (test set). Τα ποσοστά των προηγούμενων συνόλων ήταν 80% για το σύνολο εκπαίδευσης και 20% για το σύνολο ελέγχου.

Με μια πρώτη δοκιμή του baseline κώδικα του PAN2020 παρατηρήθηκαν κάποιες τιμές “NaN” (κενές) στο αρχείο answers.jsonl που παράχθηκε από το pan20-verif-baseline. Από τα αντίστοιχα id στο αρχείο με τα ζευγάρια των tweets, παρατηρήθηκε ότι το πρόβλημα ήταν σε tweets με μικρό μήκος.

Έτσι, σε αυτό το σημείο ακολουθήθηκε μια διαδικασία αφαίρεσης των tweets με μήκος μικρότερο των 5 λέξεων. Μαζί με αυτήν την διαδικασία ακολουθήθηκε και η διαδικασία αφαίρεσης των tweets που περιείχαν μέσα το αναγνωριστικό «RT», όπου δείχνει αν το εκάστοτε tweet ήταν re-tweet.

Έπειτα δημιουργήθηκε ένα αντίστοιχο dataset με 100 χρήστες, το οποίο όπως θα δούμε παρακάτω στα Αποτελέσματα δεν είχε καλύτερα αποτελέσματα από αυτό των 50 χρηστών, γι’ αυτό και τα επόμενα πειράματα έγιναν στο dataset των 50 χρηστών.

Επιπρόσθετα, δημιουργήθηκε και ένα άλλο dataset στο οποίο αφαιρέθηκαν τα tweets που είχαν την συμβολοσειρά «RT» η οποία υποδεικνύει ότι το συγκεκριμένο tweet ήταν re-tweet. Αυτό γίνεται βασίζοντας στο γεγονός ότι ένα re-tweet δεν αφορά τον εκάστοτε χρήστη οπότε λειτουργεί αρνητικά στο μοντέλο μας. Παρ’ όλ’ αυτά τα αποτελέσματα δεν ήταν πιο ικανοποιητικά από πριν. Αυτό πιθανώς γίνεται γιατί πολλά tweets που είχαν την συμβολοσειρά «RT» μπορεί να μην αφορούν re-tweet και με αυτόν τον τρόπο το μοντέλο «χάνει» πληροφορία.

Επειδή η παρούσα διπλωματική αφορά την επαλήθευση συγγραφέα σε κοινωνικά δίκτυα και πιο συγκεκριμένα, με την χρήση των tweets, ακολουθήθηκε και μια άλλη προσέγγιση, η οποία αφορά την δημιουργία του dataset με την συνένωση *n* αριθμό tweets, για τα «γνωστά» tweets του εκάστοτε χρήστη, έναντι ενός αγνώστου. Σε αυτό το σημείο, αξίζει να σημειωθεί ότι ο τρόπος δημιουργίας των dataset έγινε με τέτοιο τρόπο όπου έχει διατηρηθεί η χρονική συνέχεια των tweets. Δηλαδή, τα «γνωστά» tweets είναι χρονικά νωρίτερα από το «άγνωστο» tweet. Άρα, με την δημιουργία αυτού του dataset, εξασφαλίζουμε πως το προς μελέτη tweet είναι μεταγενέστερο των «γνωστών» tweets. Αυτό σημαίνει, πως με αυτόν τον τρόπο, μπορεί να γίνει η παρατήρηση το «άγνωστο» tweet έχει γραφτεί ή όχι από τον εκάστοτε χρήστη.

Για να γίνουν περισσότερο κατανοητά τα παραπάνω, το dataset που προέκυψε ήταν της παρακάτω μορφής:

```
{"id": "1", "tweets": [“Tweet1”, “Tweet2”], "pair": ["Tweet1 + Tweet2 +…+Tweetn", "Tweetn+1"]}
```

Έτσι, σύμφωνα με τα παραπάνω, δημιουργήθηκαν 17 dataset, για n = 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150 και 200. Για παράδειγμα, το dataset για n = 5 είναι της μορφής:

```
{"id": "1", "tweets": ["Tweet1", "Tweet2"], "pair": ["Tweet1 + Tweet2 + Tweet3 + Tweet4 + Tweet5", "Tweet 6..."]}
```

Στη συνέχεια δοκιμάστηκε μια παραλλαγή όπου όταν υπάρχουν π.χ. 10 γνωστά tweets υπολογίζεται η απάντηση συγκρίνοντας ένα-ένα τα γνωστά με το άγνωστο και συνδυάζοντας τις 10 απαντήσεις (πλειοψηφία). Επίσης, για την περίπτωση των 100 tweets δοκιμάστηκαν 10-αδες, 20-αδες, 25-αδες και 50αδες γνωστά tweets και συγκρίθηκαν με το άγνωστο και στη συνέχεια συνδυάστηκαν οι απαντήσεις για τον υπολογισμό της τελικής απάντησης.

Η παραλλαγή αυτή προσπαθεί να λάβει υπόψη της τα tweets ξεχωριστά ή σε πολλές ομάδες (ενώ μέχρι τώρα τα λαμβάνονταν υπόψη όλα μαζί σαν μία ενιαία ομάδα). Δηλαδή, έστω ότι υπάρχουν 10 tweets: t_1, t_2, t_3, …, t_{10} και ένα άγνωστο t_u. Στα προηγούμενα πειράματα πάρθηκε η συνένωση των 10 γνωστών και συγκρινόταν με το άγνωστο. Σε αυτήν την περίπτωση συγκρίνεται το κάθε γνωστό με το άγνωστο ξεχωριστά (t_1-t_u, t_2-t_u, …, t_{10}-t_u). Από αυτές τις συγκρίσεις θα ληφθούν 10 απαντήσεις, κάποιες θα είναι θετικές (ίδιος συγγραφέας) κάποιες θα είναι αρνητικές (διαφορετικός συγγραφέας). Ο συνδυασμός αυτών των απαντήσεων θα είναι και η τελική απάντηση. Όπως αναφέρθηκε και νωρίτερα, οι απαντήσεις που εξάγονται είναι μια τιμή (score), όπου όταν είναι μεγαλύτερο του 0.5 η απάντηση είναι θετική και όταν είναι μικρότερο του 0.5 η απάντηση είναι αρνητική. Έτσι, υπάρχουν διάφοροι τρόποι για να συνδυαστούν οι απαντήσεις αυτές. Κάποιοι τρόποι από αυτούς είναι:

1. Majority voting: δηλαδή τι λέει η πλειοψηφία, π.χ. αν εμφανίζονται 8 θετικά και 2 αρνητικά αποτελέσματα, η τελική απάντηση είναι θετική.

2. Weighted voting: υπολογίζεται ο μέσος όρος των 10 scores και αυτό δείχνει την τελική απάντηση (αν είναι μεγαλύτερο του 0.5 είναι θετικό, αν είναι μικρότερο του 0.5 είναι αρνητικό).

3. Confident model selection: λαμβάνονται υπόψιν μόνο οι απαντήσεις που έχουν score είτε σχετικά μεγάλο (π.χ. >0.7) είτε σχετικά μικρό (π.χ. <0.3) και συνδυάζονται είτε με simple majority είτε με weighted majority.

Για την δημιουργία των παραπάνω datasets χρησιμοποιήθηκαν συνδυασμοί προηγούμενων κωδίκων, που βρίσκονται στα Παραρτήματα, με μικρές αλλαγές για τις ανάγκες της κάθε περίπτωσης.

Ο τρόπος που ακολουθήθηκε στην αρχή είναι ο weighted voting, δηλαδή ο υπολογισμός του μέσου όρου των απαντήσεων. Όπως θα δούμε στο κεφάλαιο Αποτελέσματα, υπήρχε ένα πρόβλημα σε αυτήν την προσέγγιση. Το πρόβλημα αυτό, αφορούσε το γεγονός ότι αν σε μια 10-άδα εμφανιζόταν 8 αποτέλεσμα που είχαν ~0.6 τότε αρκούσαν δύο μόνο αποτελέσματα 0.0, για να ρίξουν τον μέσο όρο κάτω από το 0.4. Γι' αυτόν τον λόγο, πρώτα υπολογίστηκε το πλήθος των θετικών και αρνητικών απαντήσεων και στην συνέχεια υπολογιζόταν ο μέσος όρος της κατηγορίας με το μεγαλύτερο πλήθος (οι απαντήσεις με 0.5 εξαιρούνται από την διαδικασία – αφού θεωρούνται non-decisions). Έτσι, για παράδειγμα για τις απαντήσεις στην Εικόνα 24 βγήκε τα αποτέλεσμα στην Εικόνα 25.


```
{"id": "71", "value": 0.5226826002769492}
{"id": "72", "value": 0.5225400290527278}
{"id": "73", "value": 0.518872634446571}
{"id": "74", "value": 0.5310310603566523}
{"id": "75", "value": 0.3700546837209865 4}
{"id": "76", "value": 0.356949040506576}
{"id": "77", "value": 0.520171324921289}
{"id": "78", "value": 0.5135401671452766}
{"id": "79", "value": 0.5146838779335795}
{"id": "80", "value": 0.5125226117044456}
```


Εικόνα 24 Κάποιες από τις 10-άδες των απαντήσεων, για τα 100 tweets.


```
{"id": "8", "value": 0.5195055382296864}
```


Εικόνα 25 Τελικό αποτέλεσμα απάντησης για την δεκάδα στην Εικόνα 24

7

Αποτελέσματα

Για να γίνουν κατανοητά τα αποτελέσματα πρέπει πρώτα να δούμε διάφορες μετρικές που χρησιμοποιούνται για την αξιολόγηση των μοντέλων μηχανικής μάθησης.

7.1 Μετρικές Αξιολόγησης

Accuracy:

Το accuracy είναι η αναλογία του αριθμού των σωστών προβλέψεων προς το συνολικό αριθμό των δειγμάτων εισόδου.

$$\text{Accuracy} = \frac{Αριθμός\ σωστών\ προβλέψεων}{Συνολικός\ αριθμός\ προβλέψεων\ που\ έγιναν}$$

Το accuracy σε πολλές περιπτώσεις δεν είναι η ιδανική μετρική για να αξιολογήσουμε την αποδοτικότητα ενός μοντέλου μηχανικής μάθησης. Μια από αυτές τις περιπτώσεις είναι όταν η κατανομή της κλάσης είναι μη-ισορροπημένη (imbalanced – μία κλάση είναι πιο συχνή από άλλες).

Precision:

Την λύση της μη-ισσοροπημένης κλάσης έρχεται να λύσει το precision. Το precision ορίζεται ως:

$$Precision = \frac{Αληθώς\ θετικές\ προβλέψεις}{Αληθώς\ θετικές\ προβλέψεις + Ψευδώς\ θετικές\ προβλέψεις}$$

Recall:

Το recall είναι μια άλλη σημαντική μέτρηση, η οποία ορίζεται ως το κλάσμα των δειγμάτων από μια κατηγορία που προβλέπονται σωστά από το μοντέλο. Ορίζεται ως:

$$Recall = \frac{Αληθώς\ θετικές\ προβλέψεις}{Αληθώς\ θετικές\ προβλέψεις + Ψευδώς\ αρνητικές\ προβλέψεις}$$

F1:

Ανάλογα με την εφαρμογή, μπορεί να χρειαστεί να δωθεί μεγαλύτερη προτεραιότητα στην ανάκληση (recall) ή στην ακρίβεια (precision). Υπάρχουν όμως πολλές εφαρμογές στις οποίες τόσο η ανάκληση όσο και η ακρίβεια είναι σημαντικές. Ως εκ τούτου, έχει δημιουργηθεί ένας τρόπος να συνδυάζονται αυτά τα δύο σε μια ενιαία μέτρηση. Η βαθμολογία F1 μπορεί να ερμηνευθεί ως ο σταθμισμένος μέσος όρος της ακρίβειας (precision) και της ανάκλησης (recall), όπου το F1 φτάνει την καλύτερη του τιμή στο 1 και τη χειρότερη βαθμολογία στο 0. Η σχετική συμβολή της ακρίβειας και της ανάκλησης στη βαθμολογία F1 είναι ίση. Ο τύπος για τη βαθμολογία F1 είναι:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

Η γενικευμένη εκδοχή του F-score ορίζεται ως παρακάτω. Όπως μπορούμε να δούμε, η βαθμολογία F1 είναι ειδική περίπτωση F_b όταν b = 1.

$$F_b = (1 + b^2) \times \frac{precision \times recall}{b^2 \times precision + recall}$$

Είναι καλό να αναφερθεί ότι υπάρχει πάντα ένα αντάλλαγμα μεταξύ της ακρίβειας και της ανάκλησης ενός μοντέλου.

ROC Curve:

Η χαρακτηριστική καμπύλη του δέκτη (ROC – Receiver Operating Characteristic) είναι η γραφική παράσταση που δείχνει την απόδοση ενός δυαδικού ταξινομητή ως συνάρτηση του ορίου αποκοπής του (cut-off threshold). Ουσιαστικά δείχνει το πραγματικό θετικό ποσοστό

(TPR – True Positive Rate) έναντι του ψευδώς θετικού ποσοστού (FPR – False Positive Rate) για διάφορες τιμές κατωφλίου. Για να γίνει πιο κατανοητή η μετρική θα χρησιμοποιηθεί ένα παράδειγμα.

Ένα μοντέλο μπορεί να προβλέψει τις παρακάτω πιθανότητες για 4 δείγματα κειμένων: [0.45, 0.6, 0.7, 0.3]. Στη συνέχεια, ανάλογα με τις τιμές κατωφλίου παρακάτω, θα ληφθούν διαφορετικές ετικέτες:

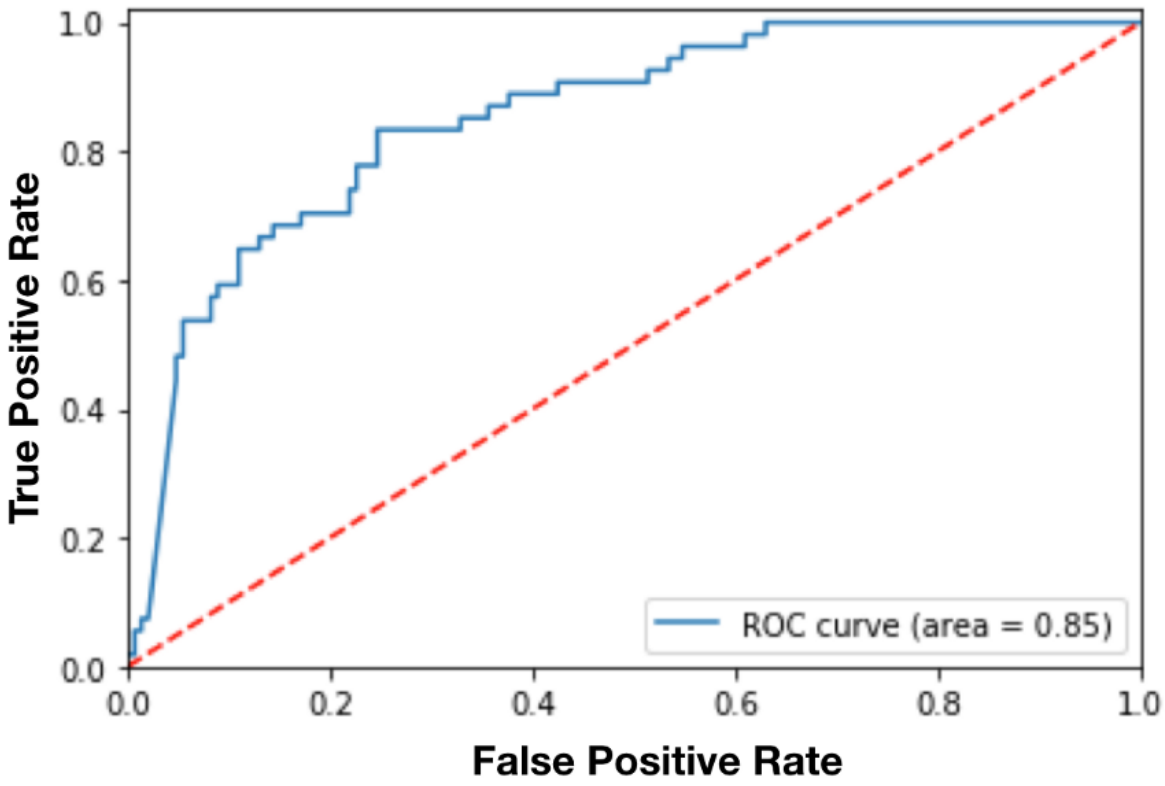
cut-off= 0.5: predicted-labels= [0,1,1,0] (default threshold)

cut-off= 0.2: predicted-labels= [1,1,1,1]

cut-off= 0.8: predicted-labels= [0,0,0,0]

Όπως φαίνεται μεταβάλλοντας τις τιμές κατωφλίου, θα εξαχθούν εντελώς διαφορετικές ετικέτες. Και ως αποτέλεσμα, καθένα από αυτά τα σενάρια θα είχε ως αποτέλεσμα διαφορετική τιμή ακριβείας και ανάκλησης (καθώς και TPR, FPR).

Η καμπύλη ROC ανακαλύπτει ουσιαστικά το TPR και το FPR για διάφορες τιμές κατωφλίου και σχεδιάζει το TPR έναντι του FPR. Ένα δείγμα της καμπύλης ROC φαίνεται στην Εικόνα 26.



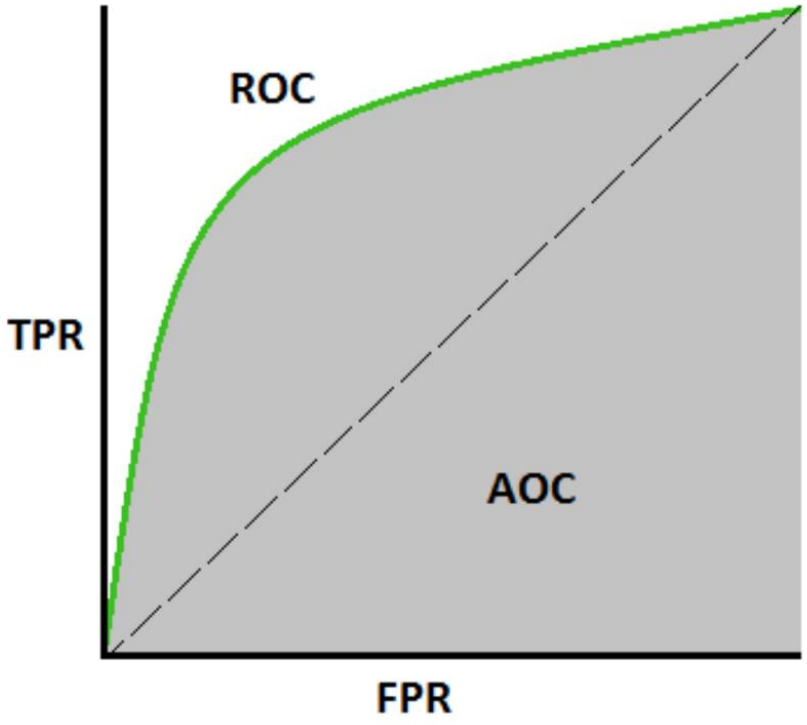
Εικόνα 26 πηγή: https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce

Όπως μπορούμε να δούμε από αυτό το παράδειγμα, όσο χαμηλότερο είναι το όριο αποκοπής στη θετική κλάση, τόσο περισσότερα δείγματα προβλέπονται ως θετική κατηγορία, δηλαδή υψηλότερο TPR (ανάκληση) και επίσης υψηλότερο FPR (που αντιστοιχεί στη δεξιά πλευρά αυτής της καμπύλης). Επομένως, υπάρχει μια αντιστάθμιση μεταξύ του πόσο υψηλή θα μπορούσε να είναι η ανάκληση έναντι του πόσο θέλουμε να περιορίσουμε το σφάλμα (FPR).

Η καμπύλη ROC είναι μια δημοφιλής μέθοδος για να εξετάσει τη συνολική απόδοση ενός μοντέλου και να επιλέξει ένα καλό όριο αποκοπής για το μοντέλο.

AUC:

Η περιοχή κάτω από την καμπύλη (Area Under the Curve – AUC), είναι ένα συγκεντρωτικό μέτρο απόδοσης ενός δυαδικού ταξινομητή σε όλες τις πιθανές τιμές κατωφλίου (και ως εκ τούτου είναι αμετάβλητο κατωφλίου).



Εικόνα 27 πηγή: https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce

Το AUC υπολογίζει το εμβαδόν κάτω από την καμπύλη ROC και επομένως είναι μεταξύ 0 και 1. Ένας τρόπος ερμηνείας του AUC είναι η πιθανότητα το μοντέλο να κατατάσσει ένα τυχαίο θετικό παράδειγμα σε υψηλότερη θέση από ένα τυχαίο αρνητικό παράδειγμα.

c@1:

Όπως εξηγήθηκε και στο κεφάλαιο 6.1.5 το c@1 είναι μια παραλλαγή του F1-score, η οποία χρησιμοποιείται σε μοντέλα όπου είναι προτιμότερο να μην λάβουμε κάποια απάντηση, από το να λάβουμε λανθασμένη απάντηση [36]. Ο τύπος του c@1 είναι:

$$c@1 = \frac{1}{n}\left(n_{ac} + \frac{n_{ac}}{n}\, n_u\right)$$

όπου το n δηλώνει τον αριθμό των προβλημάτων, n_{ac} τον αριθμό των σωστών απαντήσεων και n_u τον αριθμό των μη-απαντήσεων. Το μέτρο c@1 λύνει το πρόβλημα «αβεβαιότητας» επιβραβεύοντας τις μη-απαντήσεις καθώς τους δίνει την ίδια ακρίβεια με τα υπόλοιπα προβλήματα.

F_0.5u:

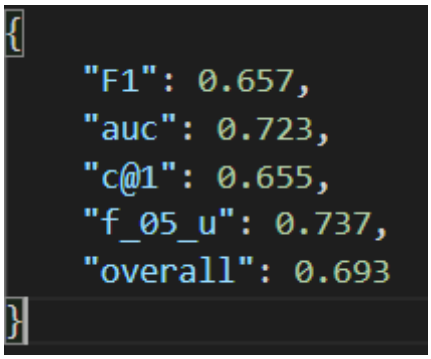
Όπως κι εδώ εξηγήθηκε στο κεφάλαιο 6.1.5 το F_0.5u είναι καινούρια μετρική όπου δίνει μεγαλύτερη έμφαση στο να αποφασίσει σωστά περιπτώσεις ίδιου συγγραφέα [37]. Ένα πρόβλημα με το c@1 είναι ότι έχει σχεδιαστεί για δυαδική ταξινόμηση με ίσα βάρη και για τις δύο κλάσεις. Αν μας ενδιαφέρει κυρίως το αν δύο κείμενα γράφτηκαν από τον ίδιο συγγραφέα, αλλά δεν χρειάζεται αξιόπιστη απόφαση σε άλλη περίπτωση, το c@1 δεν μας εξυπηρετεί καλά. Για το λόγο αυτό, έχει προταθεί ως εναλλακτική λύση το μέτρο F0.5 όπου αντιμετωπίζει τις μη-απαντήσεις ως ψευδώς αρνητικά (false negatives):

$$F\_0.5u = \frac{(1+0.5^2)n_{tp}}{(1+0.5^2)n_{tp} + 0.5^2\left(n_{fn} + n_u\right) + n_{fp}}$$

με το n_{tp} να δηλώνει τον αριθμό των αληθώς θετικών (true positives), n_{fn} τον αριθμό των ψευδώς αρνητικών (false negatives) και n_{fp} τον αριθμό των ψευδώς θετικών (false positives). Όπως και πριν, το n_u είναι ο αριθμός των αναπάντητων προβλημάτων. Η παράμετρος β = 0,5 του μέτρου F επιλέχθηκε έτσι ώστε να ζυγίζει σημαντικά μεγαλύτερη ακρίβεια (precision) από την ανάκληση (recall) χωρίς να μειώνεται εντελώς η συμβολή της. Άλλες τιμές μπορούν να χρησιμοποιηθούν ανάλογα με τις μεμονωμένες περιπτώσεις χρήσης. Αυτό το ονομάζουμε εξειδικευμένο μέτρο F_0.5u [37].

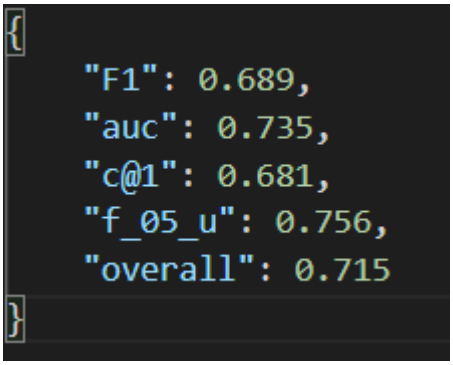
7.2 Αποτελέσματα

Το πρώτο πείραμα που δοκιμάστηκε ήταν ο baseline κώδικας του PAN2020, pan20-verif-baseline, για τα dataset που δημιουργήθηκαν με την σειρά που περιγράφηκαν στα πιο πάνω κεφάλαια. Συγκεκριμένα το πρώτο πείραμα έγινε με το dataset των 50 χρηστών πριν την προ-επεξεργασία. Τα αποτελέσματα φαίνονται στην Εικόνα 28.



Εικόνα 28 Αποτελέσματα μετρικών του dataset 50 χρηστών πριν την προ-επεξεργασία

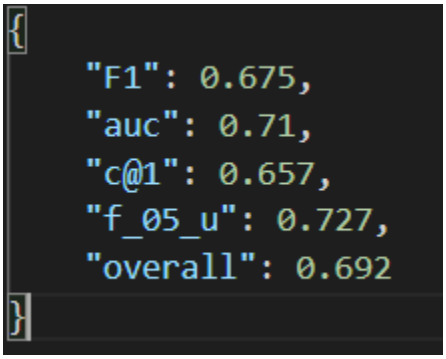
Στην συνέχεια, στην έχουμε τα αποτελέσματα των μετρικών για το dataset των 50 χρηστών μετά την προ-επεξεργασία. Η προ-επεξεργασία σε αυτό το σημείο, αφορούσε την αντικατάσταση των url's με το tag «URL» και την αντικατάσταση των ημερομηνιών με το tag «DATE».



Εικόνα 29 Αποτελέσματα μετρικών του dataset 50 χρηστών μετά την προ-επεξεργασία

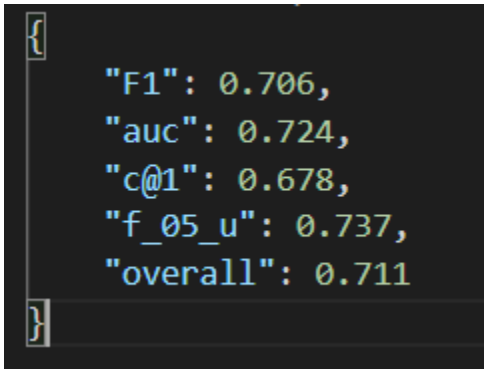
Παρατηρούμε ότι τα αποτελέσματα μετά την προ-επεξεργασία ήταν εμφανώς βελτιωμένα. Για τον λόγο αυτό τα επόμενα πειράματα αφορούν δοκιμές με το dataset μετά την προ-επεξεργασία, αφού έδειξε πολύ καλύτερα αποτελέσματα.

Έτσι, στην συνέχεια δοκιμάστηκε η ίδια μέθοδος για τους 100 χρήστες. Στην Εικόνα 30 βλέπουμε τα αποτελέσματα των μετρικών της αξιολόγησης του PAN2020.



Εικόνα 30 Αποτελέσματα μετρικών του dataset 100 χρηστών μετά την προ-επεξεργασία

Όπως αναφέρθηκε και στο κεφάλαιο Δημιουργία των dataset, η ίδια μέθοδος δοκιμάστηκε και για το dataset όπου είχαν αφαιρεθεί τα tweets με την συμβολοσειρά «RT» όπου υποδεικνύουν ότι το tweet είναι μάλλον re-tweet. Στην Εικόνα 31 βλέπουμε τα αποτελέσματα, τα οποία είναι ικανοποιητικά, αλλά ελαφρώς μειωμένα πριν την αφαίρεση αυτών των tweets.

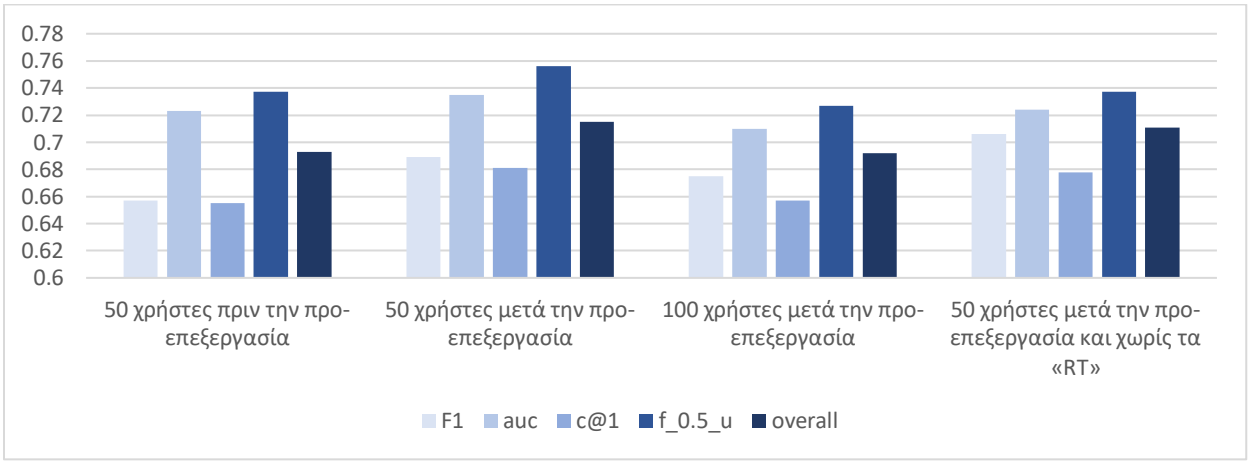


Εικόνα 31 Αποτελέσματα μετρικών του dataset 50 χρηστών μετά την προ-επεξεργασία και την αφαίρεση των tweet με το "RT"

Στην παρακάτω πίνακα βλέπουμε συγκριτικά τα αποτελέσματα έως τώρα, όπου όλα αφορούν τον baseline κώδικα από τον διαγωνισμό PAN2020:

Dataset	F1	auc	c@1	f_0.5_u	overall
50 χρήστες πριν την προ-επεξεργασία	0.657	0.723	0.655	0.737	0.693
50 χρήστες μετά την προ-επεξεργασία	0.689	**0.735**	**0.681**	**0.756**	**0.715**
100 χρήστες μετά την προ-επεξεργασία	0.675	0.71	0.657	0.727	0.692
50 χρήστες μετά την προ-επεξεργασία και χωρίς τα «RT»	**0.706**	0.724	0.678	0.737	0.711

Παρακάτω βλέπουμε διαγραμματικά τα παραπάνω αποτελέσματα:

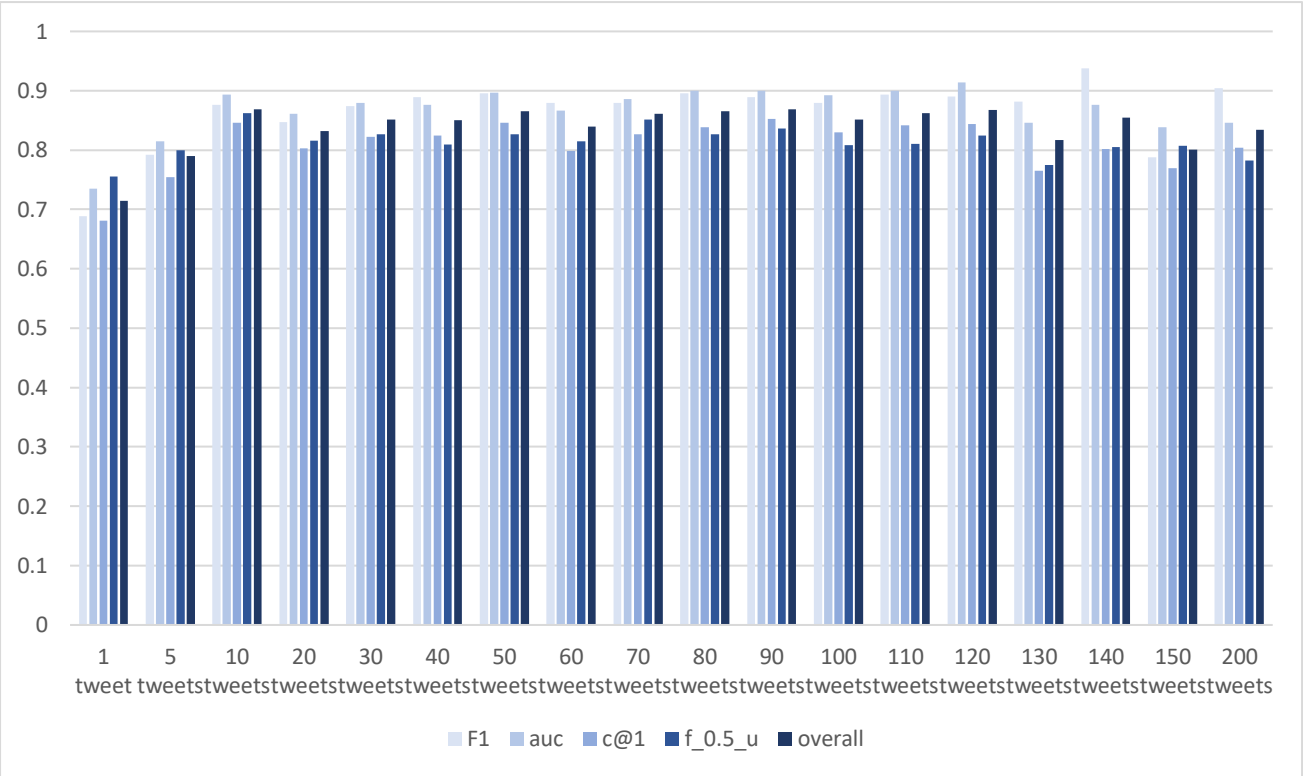


Εικόνα 32 Αποτελέσματα των dataset του παραπάνω πίνακα

Στην συνέχεια, δημιουργήθηκαν τα dataset όπως περιγράφηκαν στο κεφάλαιο Δημιουργία των dataset και παρακάτω παρουσιάζονται τα αποτελέσματα.

Dataset 50 χρηστών μετά την προ-επεξεργασία με	**F1**	**auc**	**c@1**	**f_0.5_u**	**overall**
1 tweet	0.689	0.735	0.681	0.756	0.715
5 tweets	0.792	0.815	0.754	0.8	0.79
10 tweets	0.876	0.893	0.846	**0.862**	**0.869**
20 tweets	0.847	0.861	0.803	0.816	0.832
30 tweets	0.874	0.879	0.822	0.827	0.851
40 tweets	0.889	0.876	0.824	0.809	0.85
50 tweets	0.896	0.897	0.846	0.827	0.866
60 tweets	0.88	0.867	0.799	0.815	0.84
70 tweets	0.879	0.886	0.827	0.852	0.861
80 tweets	**0.896**	0.9	0.839	0.827	0.866
90 tweets	0.889	0.9	**0.853**	0.836	**0.869**
100 tweets	0.879	0.892	0.83	0.808	0.852
110 tweets	0.894	0.9	0.842	0.811	0.862
120 tweets	0.89	**0.914**	0.844	0.825	0.868
130 tweets	0.882	0.846	0.765	0.775	0.817
140 tweets	**0.938**	0.876	0.802	0.805	0.855
150 tweets	0.788	0.839	0.77	0.807	0.801
200 tweets	0.904	0.846	0.804	0.782	0.834

Παρακάτω βλέπουμε τα παραπάνω αποτελέσματα διαγραμματικά:

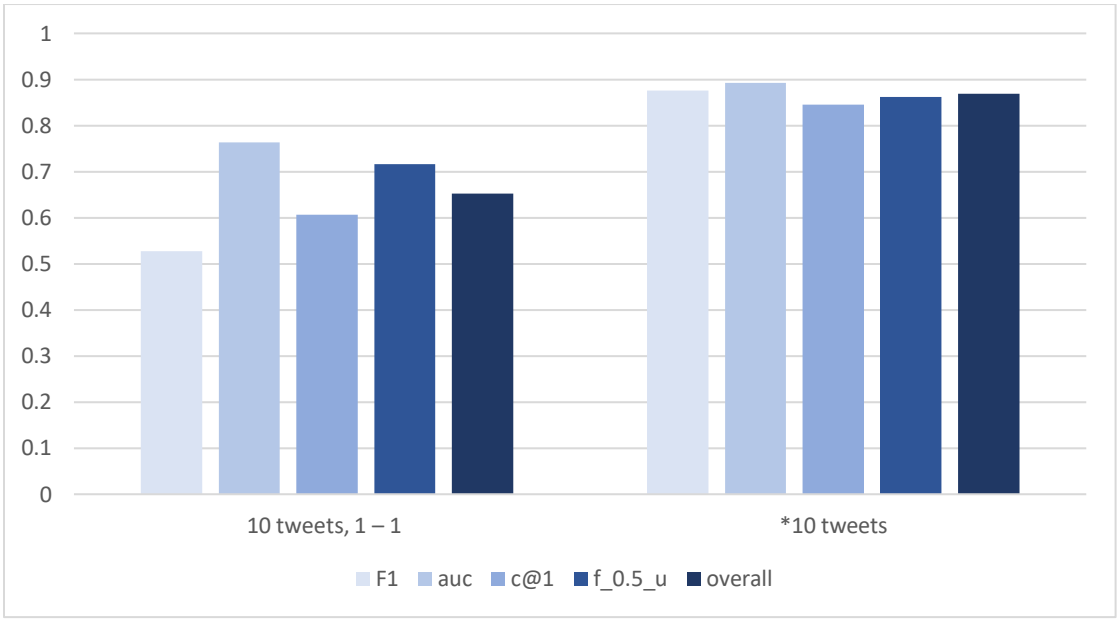


Εικόνα 33 Αποτελέσματα των dataset του παραπάνω πίνακα

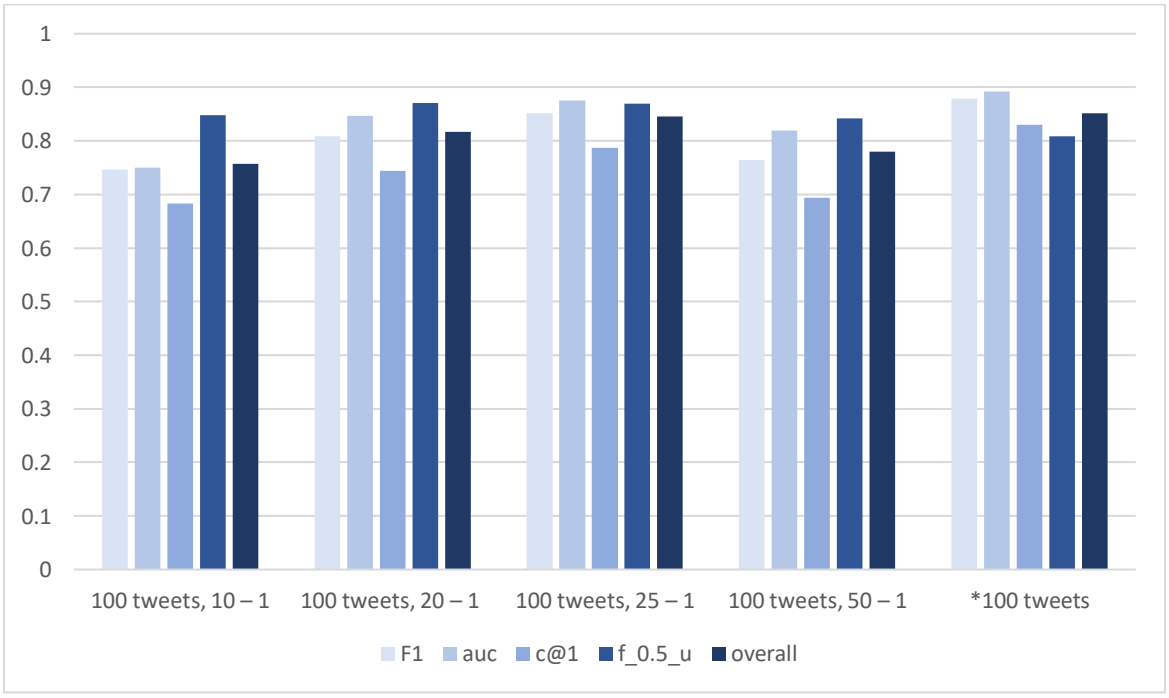
Έπειτα, θα παρουσιαστούν τα αποτελέσματα για τα dataset που περιγράφηκαν που αφορούσαν τον συνδυασμό των αποτελεσμάτων με σύγκριση των tweets ξεχωριστά. Τα αποτελέσματα αφορούν τα 10 και 100 tweets.

Dataset 50 χρηστών μετά την προ-επεξεργασία	F1	auc	c@1	f_0.5_u	overall
10 tweets, 1 – 1	0.527	0.764	0.607	0.716	0.653
***10** tweets	**0.876**	**0.893**	**0.846**	**0.862**	**0.869**
100 tweets, 10 – 1	0.747	0.75	0.683	**0.848**	0.757
100 tweets, 20 – 1	0.808	0.847	0.744	0.871	0.817
100 tweets, 25 – 1	0.851	0.875	0.787	0.87	0.846
100 tweets, 50 – 1	0.764	0.819	0.694	0.842	0.78
***100** tweets	**0.879**	**0.892**	**0.83**	0.808	**0.852**

*συμπεριλαμβάνονται και σε αυτόν τον πίνακα για σύγκριση με τα νέα αποτελέσματα.

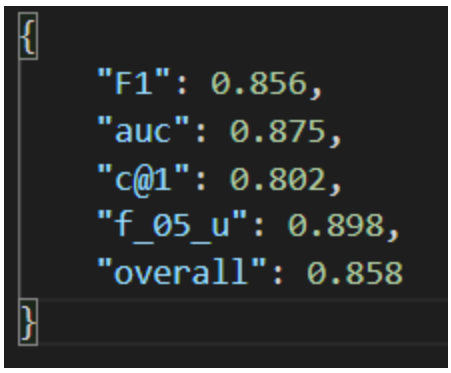


Εικόνα 34 Σύγκριση 10 ενωμένων tweets με συνδυασμό ζευγών ανά 10



Εικόνα 35 Σύγκριση 100 ενωμένων tweets με συνδυασμό ενωμένων n-άδων ανά k

Τέλος, έγινε ένα τελευταίο πείραμα που αφορούσε τα 100 tweets, όπου συγκρίθηκαν 90 συνενωμένα «γνωστά» tweets με 1. Αυτό σημαίνει ότι υπήρχαν αλληλοεπικαλυπτόμενα tweets σε κάποιες 90-άδες. Παρακάτω βλέπουμε τα αποτελέσματα αυτού του πειράματος:



Εικόνα 36 Αποτελέσματα μετρικών του dataset 50 χρηστών των 100 tweets, σύγκριση 90 – 1

8

Συμπεράσματα

Όπως παρουσιάστηκε στο προηγούμενο κεφάλαιο, τα αποτελέσματα είναι αρκετά ικανοποιητικά, καθώς η εφαρμογή των αλγορίθμων που αρχικά εφαρμόστηκαν σε κείμενα μεγάλης έκτασης στο PAN2020, αποδίδει ικανοποιητικά αποτελέσματα και σε κείμενα πολύ μικρότερης έκτασης, όπως αυτά των αναρτήσεων στο Twitter (μέγιστος αριθμός χαρακτήρων 140).

Πιο συγκεκριμένα, η προ-επεξεργασία των δεδομένων δείχνει να έχει αρκετά μεγάλη σημασία στα αποτελέσματα που θα πάρουμε. Με την προ-επεξεργασία που έγινε (αντικατάσταση διευθύνσεων και ημερομηνιών) το μοντέλο έδειξε ήδη πιο ενθαρρυντικά αποτελέσματα. Η αφαίρεση των tweets που περιείχαν το RT ως πιθανή ένδειξη για retweet, δεν παρουσίασε καλύτερα αποτελέσματα γιατί «αφαίρεσε» αρκετή πληροφορία από το μοντέλο μας (τα tweets με την ένδειξη RT, δεν είναι απαραίτητα retweets).

Επίσης, όταν τροποποιείται η εκπαίδευση και ο έλεγχος με απλά ζεύγη (1 – 1 tweet) και δημιουργούνται ζευγάρια με μεγαλύτερο πλήθος χαρακτήρων του ενός tweet (10 – 1, 20 – 1, … 200 – 1), η ακρίβεια αυξάνεται. Πιο αναλυτικά, η συνένωση πολλών tweets προοδευτικά και η σύγκρισή τους με ένα φαίνεται να βελτιώνεται και δείχνει να «σταθεροποιείται» στην μέγιστη επίδοση, κοντά στα 100 συνενωμένα tweets. Αυτό μπορεί πιθανότατα να εξηγηθεί στην βάση του ότι ο αλγόριθμος έχει «περισσότερη» πληροφορία για τον ένα χρήστη ώστε να το συγκρίνει με το κείμενο του δεύτερου χρήστη και να αποφανθεί αν πρόκειται για τον ίδιο συγγραφέα ή όχι.

Εν συνεχεία, στα πειράματα όπου δημιουργούνται n-άδες ενωμένων tweets (π.χ. δημιουργία 10-άδων, 20-άδων, 25-άδων και 50-άδων για τα 100 tweets) και συγκρίνονται με το ίδιο «άγνωστο» tweet και με την χρήση ensembles (όπου για κάθε n απαντήσεις υπολογίζεται μια νέα απάντηση – major voting σε συνδυασμό με weighted voting), όπως περιγράφηκε στο 6.2, το μοντέλο δεν δείχνει να βελτιώνεται, γιατί το διαθέσιμο μήκος του γνωστού κειμένου είναι μικρότερο. Αυτό σημαίνει πως η μέθοδος, μάλλον βελτιώνεται όταν το διαθέσιμο μήκος των

γνωστών κειμένων για το training set είναι σχετικά μεγάλο. Αυτό μπορεί να επιβεβαιωθεί με το τελευταίο πείραμα που έγινε σχετικά με τα 100 tweets που γίνεται σύγκριση 90 «γνωστών» με 1 «άγνωστο». Τα αποτελέσματα αυτού του πειράματος ήταν καλύτερα απ' ότι το πείραμα με τα 100 συνενωμένα tweets.

Επιπρόσθετα, σύμφωνα με την δημιουργία των datasets και γνωρίζοντας ότι τα tweets είναι με χρονολογική σειρά, το μοντέλο χάνει την υψηλή απόδοση με τα tweets που είναι χρονολογικά πιο παλιά.

Αν λάβουμε υπόψη όλα τα παραπάνω μπορούμε να συμπεράνουμε ότι η επαλήθευση συγγραφέα θα μπορούσε σίγουρα να συμβάλλει στην μετρίαση του προβλήματος των παραβιασμένων λογαριασμών στα κοινωνικά δίκτυα.

9

Επίλογος – Προοπτικές

Οι παραβιασμένοι λογαριασμοί στα κοινωνικά δίκτυα μπορούν να αποτελέσουν το μέσο επίθεσης κακόβουλων χρηστών και να αποβούν καταστροφικοί για άτομα ή και ομάδες ατόμων. Η ανίχνευση των παραβιασμένων λογαριασμών στα κοινωνικά δίκτυα, όπως παρουσιάστηκε, είναι πολύ σημαντική για την πρόληψη κακόβουλων ενεργειών από κακόβουλους χρήστες. Το ερευνητικό πεδίο της επαλήθευσης συγγραφέα φαίνεται να είναι ένα εργαλείο που μπορεί να βοηθήσει στην αντιμετώπιση του προβλήματος της ανίχνευσης των παραβιασμένων λογαριασμών στα κοινωνικά δίκτυα. Σύμφωνα με τα πειράματα που έγιναν και τα συμπεράσματα, τα αποτελέσματα είναι πολύ ενθαρρυντικά. Ακόμη, αξίζει να σημειωθεί η ρεαλιστικότητα των πειραμάτων που διεξήχθησαν. Με τον τρόπο που δημιουργήθηκαν τα dataset και την τυχαία επιλογή των χρηστών, το πρόβλημα της επαλήθευσης συγγραφέα γίνεται ακόμη πιο ρεαλιστικό και δύσκολο. Επίσης, τα χαρακτηριστικά που συνήθως εμφανίζονται σε ένα tweet παραβιασμένου λογαριασμού μπορεί να περιέχουν ασυνήθιστη γλώσσα, συνδέσμους που οδηγούν σε phishing ιστοσελίδες και γενικότερα μια ασυνήθιστη μορφή. Οι επιθέσεις αυτές κατά κύριο λόγο γίνονται μαζικά και άρα, ο επιτιθέμενος δεν προσπαθεί να μιμηθεί τον τρόπο γραφής ή συμπεριφοράς του έγκυρου χρήστη, με αποτέλεσμα οι μέθοδοι της επαλήθευσης συγγραφέα να μπορούν να εφαρμοστούν για την ανίχνευσή τους. Παρακάτω θα παρουσιαστούν μερικές προοπτικές σχετικά με την παρούσα εργασία.

Αρχικά, θα μπορούσε να γίνουν περαιτέρω ενέργειες και πειράματα σχετικά με την προ-επεξεργασία των tweets, ανάλογα με τα features που θα χρησιμοποιηθούν για την εκάστοτε μέθοδο. Κάποιες από αυτές τις ενέργειες θα μπορούσε να είναι η αφαίρεση των emoticons, να επιλεγούν tweets μόνο στα αγγλικά (στην παρούσα εργασία έχουν επιλεγεί tweets με λατινικούς χαρακτήρες οπότε είναι πιθανό να υπάρχουν κάποια tweets σε γλώσσα εκτός των αγγλικών με λατινικούς χαρακτήρες), τα hashtag να αντικατασταθούν με το meta tag @HASH και τα usernames να αντικατασταθούν με το meta tag @USER κ.α.

Όπως είδαμε για την παρούσα διπλωματική χρησιμοποιήθηκε ο baseline κώδικας του PAN2020, οπότε μια προοπτική είναι η ανάπτυξη άλλων αλγορίθμων και μεθόδων που θα μπορούν να χρησιμοποιηθούν με τα datasets που δημιουργήθηκαν στην παρούσα εργασία. Θα μπορούσαν ακόμα να δοκιμαστούν και υπάρχουσες μέθοδοι που προορίζονταν για επαλήθευση συγγραφέα ή ακόμα και να δοκιμαστούν νέες εφαρμογές όπως αλγόριθμοι βαθιάς μάθησης (deep learning).

Ακόμη, ο υπολογισμός με την μέθοδο που ακολουθήθηκε με τα ensembles έχει πολλούς τρόπους που μπορεί να γίνει. Ως εκ τούτου, θα μπορούσαν να δοκιμαστούν διάφοροι τρόποι συνδυασμών των απαντήσεων για κάθε ζευγάρια (π.χ. Confident model selection).

Τέλος, είναι γνωστό πως αυτές οι μέθοδοι έχουν καλύτερα αποτελέσματα όταν τα κείμενά μας έχουν μια θεματική συνάφεια. Θα μπορούσαν να δημιουργηθούν και να εξεταστούν τα αποτελέσματα αν τα datasets που θα δημιουργηθούν έχουν μεταξύ τους μια θεματική συνάφεια (π.χ. χρησιμοποιώντας tweets με συγκεκριμένα hashtags).

Παραρτήματα

```python
import pandas as pd
import os
from pathlib import Path

user_ids = []
num_of_tweets = []

# directory που είναι τα αρχεία
directory = '2015-micro-text-authorship-attribution-dataset'

# Για κάθε αρχείο μέσα στον φάκελο
for filename in os.listdir(directory):
    f = os.path.join(directory, filename)
# αν το όνομα του αρχείου είναι ακέραιος (δηλαδή κάποιος χρήστης)
    if Path(f).stem.isdigit():
# να βάλει το id του χρήστη στην λίστα user_ids
        user_ids.append(Path(f).stem)
        f_read = open(f, "r", encoding='utf-8')
# Διαβάζω τον αριθμό των tweets του χρήστη από την τελευταία γραμμή μέσα σε κάθε αρχείο
        num_of_tweets.append(f_read.readlines()[-1].split()[0])
        f_read.close()


data = {'user ID': user_ids,
        'No. of tweets': num_of_tweets}
my_dataframe = pd.DataFrame.from_dict(data)
# pd.set_option('display.max_rows', None)

# print(my_dataframe)

# Γράφω το dictionary που έχω τα τους χρήστες με τον αριθμό των tweets τους, σε ένα csv.
my_dataframe.to_csv(r'2015-micro-text-authorship-attribution-dataset/dataframe.csv')

```

Παράρτημα 1 Κώδικας για τη δημιουργία ενός csv με το id των χρηστών και τον αριθμό τον tweets αντίστοιχα.

```python
import pandas as pd
import os
from pathlib import Path
from shutil import copy2, copyfile

# Διαβάζω το excel με τους πρώτους 50 χρήστες
data_file = pd.read_excel('E:\\thesis\\my_dataset\\2015-micro-text-authorship-attribution-dataset\\users_50.xlsx')

# Οι 50 χρήστες με τα περισσότερα tweets
users_50 = data_file['users'].tolist()
users_50 = [str(x) for x in users_50]

# directory που είναι τα αρχεία
directory = 'E:\\thesis\\my_dataset\\2015-micro-text-authorship-attribution-dataset'

# Δημιουργία ενός φακέλου "users_50" με τους πρώτους 50 χρήστες
for filename in os.listdir(directory):
    f = os.path.join(directory, filename)
    if Path(f).stem.isdigit():
        if users_50.count(Path(f).stem) > 0 :
            copy2(f, directory + '\\' + 'users_50' + '\\' + Path(f).stem + '.dat')

```

Παράρτημα 2 Κώδικας για τη δημιουργία φακέλου "users_50" μόνο με τους πρώτους 50 χρήστες.

```python
import pandas as pd
import os
from pathlib import Path
from shutil import copy2, copyfile

# Διαβάζω το excel με τους πρώτους 100 χρήστες
data_file = pd.read_excel('E:\\thesis\\my_dataset\\2015-micro-text-authorship-attribution-dataset\\users_100.xlsx')

# Οι 100 χρήστες με τα περισσότερα tweets
users_100 = data_file['users'].tolist()
users_100 = [str(x) for x in users_100]

# directory που είναι τα αρχεία
directory = 'E:\\thesis\\my_dataset\\2015-micro-text-authorship-attribution-dataset'

# Δημιουργία ενός φακέλου "users_100" με τους πρώτους 100 χρήστες
for filename in os.listdir(directory):
    f = os.path.join(directory, filename)
    if Path(f).stem.isdigit():
        if users_100.count(Path(f).stem) > 0 :
            copy2(f, directory + '\\' + 'users_100' + '\\' + Path(f).stem + '.dat')

```

Παράρτημα 3 Κώδικας για τη δημιουργία φακέλου "users_100" μόνο με τους πρώτους 100 χρήστες.

```python
import pandas as pd
import os
from pathlib import Path
import re

dates_columns = []

directory = 'E:\\thesis\\my_dataset\\2015-micro-text-authorship-attribution-dataset'

# Ονόματα των στηλών (2007-01, 2007-02,...,2015-12)
for i in range(2007, 2015):
    for j in range(1, 13):
        if j < 10:
            dates_columns.append(str(i)+'-'+'0'+str(j))
        else:
            dates_columns.append(str(i)+'-'+str(j))

# δημιουργία του dataframe
df = pd.DataFrame(0, index=[], columns=dates_columns)

# για κάθε χρήστη
for filename in os.listdir(directory):
    f = os.path.join(directory, filename)
    if Path(f).stem.isdigit():
        f_read = open(f, "r", encoding='utf-8')
        userfile = f_read.readlines()
# βάζω το id του εκάστοτε χρήστη σαν νέα γραμμή
        df = df.append(pd.DataFrame(0, index=[Path(f).stem], columns=df.columns))

# βλέπω σε ποιες γραμμές είναι οι ημερομηνίες
        for i in range(len(userfile)):
            dateRegex = re.compile('(\d\d\d\d-\d\d)-(\d\d)')
            mo = dateRegex.search(userfile[i])
            result = re.search(dateRegex, userfile[i])

# αν είναι ημερομηνία βλέπω ποια ημερομηνία είναι και αυξάνω κατά 1 στην συγκεκριμένη ημερομηνία
            if result:
                if(int(mo.group(1).split('-')[0]) > 2000 and int(mo.group(1).split('-')[0]) < 2021):
                    for column in df.columns:
                        if mo.group(1) == column:
                            df.at[Path(f).stem,column]+=1
        f_read.close()

# γράφω το dataframe σε ένα csv
df.to_csv(r'2015-micro-text-authorship-attribution-dataset/dates_details.csv')
```

Παράρτημα 4 Κώδικας για δημιουργία πίνακα με τους χρήστες και πόσα tweets έχουν για κάθε ημερομηνία.

```python
import json
import pandas as pd
import os
from pathlib import Path
import re
import string


data = {'User': [], 'Tweet': []}
df = pd.DataFrame.from_dict(data)

directory = 'E:\\thesis\\my_dataset\\2015-micro-text-authorship-attribution-dataset\\users_50'

# για κάθε χρήστη
for filename in os.listdir(directory):
    f = os.path.join(directory, filename)
    if Path(f).stem.isdigit():
        user_id = Path(f).stem
        f_read = open(f, "r", encoding='utf-8')
        userfile = f_read.readlines()

        for i in range(len(userfile)):
            dateRegex = re.compile('(\d\d\d\d)-(\d\d)-(\d\d)')
            mo = dateRegex.search(userfile[i])
            result = re.search(dateRegex, userfile[i])
# αν είναι ημερομηνία που θέλουμε
            if result:
                if(int(mo.group(1).split('-')[0]) > 2000 and int(mo.group(1).split('-')[0]) < 2021):
                    if(int(mo.group(1).split('-')[0]) == 2014 and int(mo.group(2).split('-')[0]) == 2):

                        retweetRegex = re.compile('(?:^|\W)RT(?:$|\W)', re.IGNORECASE)
                        mo2 = retweetRegex.search(userfile[i+1])
                        result2 = re.search(retweetRegex, userfile[i+1])

# Αν δεν είναι retweet να το γράψει (αν δεν έχει το RT ή rt μέσα στο string)
                        if not result2:
# Να έχει περισσότερες από 5 λέξεις
                            res = sum([j.strip(string.punctuation).isalpha() for j in userfile[i+1].split()])
                            if res >= 5:
                                new_row = {'User': user_id, 'Tweet': userfile[i+1]}
# τότε το γράφει
                                df = df.append(new_row, ignore_index=True)
    f_read.close()

# το dataframe σε csv
df.to_csv(r'E:/thesis/my_dataset/2015-micro-text-authorship-attribution-dataset/users_50/users_50_tweets_without_RT_moreThanFiveWords.csv')
```

Παράρτημα 5 Κώδικας για τη δημιουργία του csv με τους χρήστες και όλα τους τα tweets για τον ενδιαφερόμενο μήνα (2014-02).

```python
import json
import pandas as pd
import random

dataframe_tweets = pd.read_csv('users_50_tweets.csv')
dataframe_number_tweets = pd.read_csv('users_50.csv')

id = 1
start = 0
trueValue = True
falseValue = False

# για κάθε χρήστη
for i in range(50):
# 660 ζευγάρια για κάθε χρήστη
    for j in range(660):
# παίρνω το πρώτο tweet
        tweet1 = dataframe_tweets.loc[start, 'Tweet']
# παίρνω το id του χρήστη
        author1 = dataframe_tweets.loc[start, 'User']
# αφαιρώ την γραμμή για το tweet που έχω πάρει
        dataframe_tweets = dataframe_tweets.drop(dataframe_tweets.index[start]).reset_index(drop=True)
#αντίστοιχα για το δεύτερο
        tweet2 = dataframe_tweets.loc[start, 'Tweet']
        author2 = dataframe_tweets.loc[start, 'User']
        dataframe_tweets = dataframe_tweets.drop(dataframe_tweets.index[start]).reset_index(drop=True)

        dataframe_number_tweets.at[i, 'Number'] = dataframe_number_tweets.at[i, 'Number'] - 2
# γράφω τις γραμμές στα τελικά dataset
        row = {
            "id": str(id),
            "tweets": ["Tweet 1", "Tweet 2"],
            "pair": [tweet1, tweet2]
            }

        with open('data.jsonl', 'a') as outfile:
            json.dump(row, outfile)
            outfile.write('\n')

```

Παράρτημα 6 Κώδικας για τη δημιουργία των τελικών dataset (1).*

```
            row2 = {
                "id": str(id),
                "same": trueValue,
                "authors": [str(author1), str(author2)]
                }

            with open('data_truth.jsonl', 'a') as outfile2:
                json.dump(row2, outfile2)
                outfile2.write('\n')

            id+=1

        start += dataframe_number_tweets.at[i, 'Number']

    random_int = random.randint(0, len(dataframe_tweets)-1)
    random_int2 = random.randint(0, len(dataframe_tweets)-1)

    # 22000 ζευγάρια για τα false (διαφορετικοί χρήστες)
    for i in range(22000):
    # δημιουργία ζευγαριού 2 tweet τυχαία και στην συνέχεια τα αφαιρώ από τον πίνακα
        while (dataframe_tweets.at[random_int, 'User'] == dataframe_tweets.at[random_int2, 'User']):
            random_int = random.randint(0, len(dataframe_tweets)-1)
            random_int2 = random.randint(0, len(dataframe_tweets)-1)

        tweet1 = dataframe_tweets.loc[random_int, 'Tweet']
        author1 = dataframe_tweets.loc[random_int, 'User']
        dataframe_tweets = dataframe_tweets.drop(random_int)

        tweet2 = dataframe_tweets.loc[random_int2, 'Tweet']
        author2 = dataframe_tweets.loc[random_int2, 'User']
        dataframe_tweets = dataframe_tweets.drop(random_int2)

        dataframe_tweets = dataframe_tweets.reset_index(drop=True)

```

Παράρτημα 7 Κώδικας για τη δημιουργία των τελικών dataset (2).*

```
74  # γράφω τις γραμμές στα τελικά dataset
75      row = {
76              "id": str(id),
77              "tweets": ["Tweet 1", "Tweet 2"],
78              "pair": [tweet1, tweet2]
79          }
80
81      with open('data.jsonl', 'a') as outfile:
82          json.dump(row, outfile)
83          outfile.write('\n')
84
85      row2 = {
86              "id": str(id),
87              "same": falseValue,
88              "authors": [str(author1), str(author2)]
89          }
90
91      with open('data_truth.jsonl', 'a') as outfile2:
92          json.dump(row2, outfile2)
93          outfile2.write('\n')
94
95
96      random_int = random.randint(0, len(dataframe_tweets)-1)
97      random_int2 = random.randint(0, len(dataframe_tweets)-1)
98
99      id+=1
100
```

Παράρτημα 8 Κώδικας για τη δημιουργία των τελικών dataset (3).*

```python
import json
import pandas as pd
import random

with open('data_after_process.jsonl', 'r') as infile:

    for i in range(26400):
        with open('data_train_after_process.jsonl', 'a') as outfile:
            outfile.write(infile.readline())

    for i in range(6600):
        with open('data_test_after_process.jsonl', 'a') as outfile2:
            outfile2.write(infile.readline())

    for i in range(17600):
        with open('data_train_after_process.jsonl', 'a') as outfile:
            outfile.write(infile.readline())

    for i in range(4400):
        with open('data_test_after_process.jsonl', 'a') as outfile2:
            outfile2.write(infile.readline())

#για το truth
with open('data_truth_after_process.jsonl', 'r') as infile:

    for i in range(26400):
        with open('data_train_truth_after_process.jsonl', 'a') as outfile:
            outfile.write(infile.readline())

    for i in range(6600):
        with open('data_test_truth_after_process.jsonl', 'a') as outfile2:
            outfile2.write(infile.readline())

    for i in range(17600):
        with open('data_train_truth_after_process.jsonl', 'a') as outfile:
            outfile.write(infile.readline())

    for i in range(4400):
        with open('data_test_truth_after_process.jsonl', 'a') as outfile2:
            outfile2.write(infile.readline())
```

Παράρτημα 9 Κώδικας για τον χωρισμό του dataset σε train set και test set.*

```python
from os import replace
import pandas as pd
import re

# μέθοδος για να βρω όπου υπάρχει url με την βοήθεια regex
def findUrl(string):
    regex = r"(?i)\b((?:https?://|www\d{0,3}[.]|[a-z0-9.\-]+[.][a-z]{2,4}
    url = re.findall(regex, string)
    return [x[0] for x in url]

df = pd.read_csv ('users_50_tweets.csv')

for i in range(len(df)) :
    tweet = df.loc[i, "Tweet"]

    # αντικαθιστώ http://example.something - > URL
    list_with_urls = findUrl(tweet)
    for j in list_with_urls:
        tweet = tweet.replace(j, "URL")
        df.at[i, 'Tweet'] = tweet

    # αντικαθιστώ ημερομηνίες -> DAT
    # της μορφής 2014-02-02 και 02-02-2014 (μόνο τέτοια μορφή βρέθηκε)
    tweet = re.sub(r"\d{2}-\w{2}-\d{4}", "DATE", tweet)
    tweet = re.sub(r"\d{4}-\w{2}-\d{2}", "DATE", tweet)
    df.at[i, 'Tweet'] = tweet

# ξαναγράφω τα tweet μετά την προ-επεξεργασία
df.to_csv(r'users_50_tweets_after_process.csv', index=False)

```

Παράρτημα 10 Κώδικας για την προ-επεξεργασία με τα URL και τις ημερομηνίες.

```python
import json
import numpy as np
import statistics

list_with_values = []

# Παίρνω τις απαντήσεις από το αρχείο answers.jsonl
with open("answers.jsonl", "r") as input:
    for line in input:
        convertedDict = json.loads(line)
        value = convertedDict["value"]
        list_with_values.append(value)

# Δημιουργώ νέα λίστα με τις n-άδες
composite_list = [list_with_values[x:x+10] for x in range(0, len(list_with_values),10)]
list_with_average = []

```

Παράρτημα 11 Κώδικας για τον υπολογισμό των νέων απαντήσεων και δημιουργία του αρχείου με αυτές (1).*

```python
# Για κάθε n-άδα υπολογίζω το πλήθος των περισσότερων (0.5> ή 0.5<) και μετά το μ.ο. αυτών
for i in range(len(composite_list)):
    more_than = 0
    less_than = 0
    equal_to = 0
    more_than_list = []
    less_than_list = []
    equal_to_list = []

    for j in (composite_list[i]):
        if j > 0.5:
            more_than += 1
            more_than_list.append(j)
        elif j < 0.5:
            less_than += 1
            less_than_list.append(j)
        else:
            equal_to += 1
            equal_to_list.append(j)

    if more_than > less_than:
        list_with_average.append(statistics.mean(more_than_list))
    elif more_than < less_than:
        list_with_average.append(statistics.mean(less_than_list))
    elif  more_than == less_than:
        if not more_than_list and not less_than_list:
            list_with_average.append(0.5)
        else:
            difference = (statistics.mean(more_than_list) + statistics.mean(less_than_list) ) / 2
            list_with_average.append(difference)

```

Παράρτημα 12 Κώδικας για τον υπολογισμό των νέων απαντήσεων και δημιουργία του αρχείου με αυτές (2).*

```
49  # Γράφω τα αποτελέσματα σε νέο αρχείο
50  id = 1
51  for i in range(len(list_with_average)):
52
53      value = list_with_average[i]
54      row = { "id": str(id), "value": value}
55
56      with open("answers_new.jsonl", "a") as outfile:
57          json.dump(row, outfile)
58          outfile.write('\n')
59      id+=1
```

Παράρτημα 13 Κώδικας για τον υπολογισμό των νέων απαντήσεων και δημιουργία του αρχείου με αυτές (3).*

*οι κώδικες τροποποιήθηκαν σε συγκεκριμένα σημεία, ανάλογα με ποιο dataset επρόκειτο να δημιουργηθεί.

Βιβλιογραφία

[1] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, “COMPA: Detecting Compromised Accounts on Social Networks.”

[2] E. Zangerle and G. Specht, “‘Sorry, i was hacked’ a classification of compromised Twitter accounts,” *Proc. ACM Symp. Appl. Comput.*, pp. 587–593, 2014, doi: 10.1145/2554850.2554894.

[3] C. Grier, K. Thomas, V. Paxson, and M. Zhang, “@spam: The Underground on 140 Characters or Less * General Terms.”

[4] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, “Detecting and characterizing social spam campaigns,” *Proc. ACM SIGCOMM Internet Meas. Conf. IMC*, pp. 35–47, 2010, doi: 10.1145/1879141.1879147.

[5] R. Kaur, S. Singh, and H. Kumar, “Authorship Analysis of Online Social Media Content,” *Lect. Notes Networks Syst.*, vol. 46, pp. 539–549, 2019, doi: 10.1007/978-981-13-1217-5_52.

[6] “Reinforcement Learning: Bottom-up Programming for Ethical Machines. Marten Kaas.” https://giorgivachnadze.medium.com/reinforcement-learning-bottom-up-programming-for-ethical-machines-marten-kaas-ca383612c778.

[7] M. J. Zaki and W. M. Jr., *Εξόρυξη και ανάλυση δεδομένων*. Κλειδάριθμος, 2017.

[8] A. Tan and A. Tan, “Text Mining: The state of the art and the challenges,” *Proc. PAKDD 1999 Work. Knowl. DISOCOVERY FROM Adv. DATABASES*, pp. 65--70, 1999.

[9] “What is Text Mining? | IBM.” .

[10] C. D. Manning, P. Raghavan, and H. Schutze, *Εισαγωγή στην ανάκτηση πληροφοριών*. Εκδόσεις Κλειδάριθμος, 2012.

[11] G. G. Chowdhury, “Natural language processing,” *Annu. Rev. Inf. Sci. Technol.*, vol. 37, no. 1, pp. 51–89, Jan. 2003, doi: 10.1002/ARIS.1440370103.

[12] V. Gurusamy and S. Kannan, “Preprocessing Techniques for Text Mining,” *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2014.

[13] Charu C. Aggarwal and ChengXiang Zhai, *Mining Text Data*. Springer Science+Business Media, 2012.

[14] Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, and Η. Σακελλαρίου, *Τεχνητή Νοημοσύνη*, 4η Έκδοση. Εκδόσεις Πανεπιστημίου Μακεδονίας, 2020.

[15] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep

Learning Based Text Classification: A Comprehensive Review," vol. 1, no. 1, pp. 1–43, 2020, doi: 10.1145/3439726.

[16] S. ELMANARELBOUANANI and I. KASSOU, "Authorship Analysis Studies: A Survey," *Int. J. Comput. Appl.*, vol. 86, no. 12, pp. 22–29, Jan. 2014, doi: 10.5120/15038-3384.

[17] "Authorship Analysis as a Text Classification or Clustering Problem | by Nabanita Roy | Towards Data Science." https://www.datasciencecentral.com/profiles/blogs/introduction-to-authorship-analysis-as-a-text-classification.

[18] "PAN." .

[19] E. Stamatatos, "Authorship Verification: A Review of Recent Advances."

[20] G. D. Miner, J. Elder, and R. A. Nisbet, "Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications," *Pract. Text Min. Stat. Anal. Non-structured Text Data Appl.*, 2012, doi: 10.1016/C2010-0-66188-8.

[21] R. Kaur, S. Singh, and H. Kumar, "Rise of spam and compromised accounts in online social networks: A state-of-the-art review of different combating approaches," *Journal of Network and Computer Applications*. 2018, doi: 10.1016/j.jnca.2018.03.015.

[22] C. Lumezanu and N. Feamster, "Observing common spam in Twitter and email," *Proc. ACM SIGCOMM Internet Meas. Conf. IMC*, pp. 461–466, 2012, doi: 10.1145/2398776.2398824.

[23] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@Spam: The underground on 140 characters or less," *Proc. ACM Conf. Comput. Commun. Secur.*, pp. 27–37, 2010, doi: 10.1145/1866307.1866311.

[24] J. Song, S. Lee, and J. Kim, "Spam Filtering in Twitter Using Sender-Receiver Relationship," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6961 LNCS, pp. 301–317, 2011, doi: 10.1007/978-3-642-23644-0_16.

[25] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: Social honeypots + machine learning," *SIGIR 2010 Proc. - 33rd Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 435–442, 2010, doi: 10.1145/1835449.1835522.

[26] J. Martinez-Romo and L. Araujo, "Detecting malicious tweets in trending topics using a statistical analysis of language," *Expert Syst. Appl.*, vol. 40, no. 8, pp. 2992–3000, Jun. 2013, doi: 10.1016/J.ESWA.2012.12.015.

[27] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting Spammers on Twitter."

[28] I. Santos, I. Miñambres-Marcos, C. Laorden, P. Galán-García, A. Santamaría-Ibirika, and P. G. Bringas, "Twitter Content-Based Spam Filtering," *Adv. Intell. Syst. Comput.*, vol. 239, pp. 449–458, 2014, doi: 10.1007/978-3-319-01854-6_46.

[29] Z. Zhang and K. Wang, “A trust model for multimedia social networks,” *Soc. Netw. Anal. Min. 2012 34*, vol. 3, no. 4, pp. 969–979, Jul. 2012, doi: 10.1007/S13278-012-0078-4.

[30] S. Barbon, R. A. Igawa, and B. Bogaz Zarpelão, “Authorship verification applied to detection of compromised accounts on online social networks: A continuous approach,” *Multimed. Tools Appl.*, 2017, doi: 10.1007/s11042-016-3899-8.

[31] M. L. Brocardo, I. Traore, S. Saad, and I. Woungang, “Authorship verification for short messages using stylometry,” 2013, doi: 10.1109/CITS.2013.6705711.

[32] R. Igawa, A. Almeida, B. Zarpelão, and S. Jr, “Recognition of Compromised Accounts on Twitter,” 2019, doi: 10.5753/sbsi.2015.5885.

[33] J. S. Li, L. C. Chen, J. V. Monaco, P. Singh, and C. C. Tappert, “A comparison of classifiers and features for authorship authentication of social networking messages,” 2017, doi: 10.1002/cpe.3918.

[34] “PAN @ CLEF 2020 - Authorship Verification.” .

[35] W. J. S. Anderson Rocha, C. W. F. T. Cavalcante, A. Theophilo, B. Shen, A. R. B. Carvalho, and E. Stamatatos, “Authorship Attribution for Social Media Forensics,” *IEEE Trans. Inf. FORENSICS Secur.*, pp. 133–150, 2016, doi: 10.1201/9781003049357-9.

[36] A. Peñas, P. Peñas, and A. Rodrigo, “A Simple Measure to Assess Non-response,” pp. 1415–1424.

[37] J. Bevendorff, B. Stein, M. Hagen, and M. Potthast, “Generalizing Unmasking for Short Texts,” pp. 654–659, Accessed: Sep. 02, 2021. [Online]. Available: https://github.com/webis-de/NAACL-19.