ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Αναγνώριση Ανεπιθύμητης Αλληλογραφίας με Χρήση Νευρωνικών Δικτύων

Αθανάσιος Γ. Γιαννόπουλος

Επιβλέπων: **Παναγιώτης Σταματόπουλος,** Επίκουρος Καθηγητής ΕΚΠΑ

ΑΘΗΝΑ

ΙΟΥΝΙΟΣ 2016

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Αναγνώριση Ανεπιθύμητης Αλληλογραφίας με Χρήση Νευρωνικών Δικτύων

Αθανάσιος Γ. Γιαννόπουλος
Α.Μ.: 1115201200024

ΕΠΙΒΛΕΠΩΝ: Παναγιώτης Σταματόπουλος, Επίκουρος Καθηγητής ΕΚΠΑ

ΠΕΡΙΛΗΨΗ

Στην εποχή που τα Ανεπιθύμητα (Spam) Μηνύματα κατακλύζουν κάθε διαθέσιμο “γραμματοκιβώτιο”, η ανάγκη για την αυτοματοποιημένη αναγνώριση και αντιμετώπισή τους φαίνεται επιτακτική. Στην εργασία αυτή αρχικά μελετήθηκαν οι τεχνικές Αναγνώρισης Ανεπιθύμητης Αλληλογραφίας που χρησιμοποιούνται τα τελευταία χρόνια. Συγκεκριμένα, έμφαση δόθηκε σε προσεγγίσεις που χρησιμοποιούν αλγορίθμους Μηχανικής Μάθησης. Στη συνέχεια, αναπτύχθηκε ένα φίλτρο Ανεπιθύμητης Αλληλογραφίας χρησιμοποιώντας Νευρωνικά Δίκτυα και συγκεκριμένα τον αλγόριθμο *Πολυεπίπεδου Δικτύου Αισθητήρων* (Multilayer Perceptron / MLP). Έγινε χρήση της συλλογής αλγορίθμων Μηχανικής Μάθησης και Εξόρυξης Δεδομένων Weka του Πανεπιστημίου του Waikato. Μελετώνται οι παράμετροι του Νευρωνικού Δικτύου και αποφασίζονται οι καταλληλότερες τιμές προκειμένου να μεγιστοποιηθεί η απόδοση της κατηγοριοποίησης του φίλτρου. Τόσο οι τεχνικές που χρησιμοποιήθηκαν γι’ αυτό όσο και τα πειραματικά αποτελέσματα παρουσιάζονται στην εργασία.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Αναγνώριση Ανεπιθύμητης Αλληλογραφίας

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: ηλεκτρονικό ταχυδρομείο (email), αναγνώριση ανεπιθύμιτης αλληλογραφίας, μηχανική μάθηση, νευρωνικά δίκτυα, tf-idf

ABSTRACT

In the period that Spam emails deluge every mailbox available, the need of automated recognition and filtering seems mandatory. During this project, the techniques of Email Spam Filtering used in the latest years were studied. Specifically, more attention was given to the approaches that use Machine Learning Algorithms. Subsequently, a Spam Filter was developed using Neural Networks and more specifically the *Multilayer Perceptron / MLP* algorithm. The collection of Machine Learing and Data Mining algorithms Weka developed in University of Waikato was used. The Neural Network's parameters were studied and the optimal values are decided in order to maximize classification accuracy of the Spam Filter. Both the techniques that were used and the experimental results are reported in this project.

SUBJECT AREA: Spam filtering

KEYWORDS: email, spam filtering, machine learning, neural networks, tf-idf

Στους γονείς μου, Γιώργο και Αγγελική.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον κ. Παναγιώτη Σταματόπουλο που μου πρότεινε να ασχοληθώ με αυτό το ενδιαφέρον αντικείμενο, αλλά και για την καθοδήγηση, τις προτάσεις και τις συμβουλές που μου παρείχε καθόλη τη διάρκεια εκπόνησης αυτής της εργασίας.

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ 9

2. ΤΡΕΧΟΥΣΑ ΚΑΤΑΣΤΑΣΗ 11

2.1 Δομή ενός Φίτρου Ανεπιθύμητης Αλληλογραφίας 12

2.1.1 Αναπαράσταση 12

2.2 Συλλογές Δεδομένων (Datasets) 15

2.3 Μέτρα Απόδοσης 17

2.4 Αλγόριθμοι Κατηγοριοποίησης 20

2.4.1 Naive Bayes 20

2.4.2 k-Nearest Neighbors 21

2.4.3 Support Vector Machines 22

2.4.4 Νευρωνικά Δίκτυα 23

3. ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΠΟΥ ΥΛΟΠΟΙΕΙΤΑΙ 25

4. ΠΕΡΙΒΑΛΛΟΝ ΔΙΕΞΑΓΩΓΗΣ ΠΕΙΡΑΜΑΤΩΝ 27

5. ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ 29

6. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ 42

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ 43

ΣΥΝΤΜΗΣΕΙΣ, ΑΡΚΤΙΚΟΛΕΞΑ ΚΑΙ ΑΚΡΩΝΥΜΙΑ 44

ΑΝΑΦΟΡΕΣ 45

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

Σχήμα 1: Βασική δομή ενός Φίλτρου Ανεπιθύμητης Αλληλογραφίας 15

Σχήμα 2: Βασική δομή ενός ROC γραφήματος, απεικονίζοντας 3 κατηγοριοποιητές A,B,C. 19

Σχήμα 3: Γραμμικά Διαχωρίσιμα δεδομένα του δισδιάστατου χώρου. 22

Σχήμα 4: Βασική δομή ενός Νευρώνα. 23

Σχήμα 5: Βασική δομή ενός πολυεπίπεδου Νευρωνικού Δικτύου. 24

Σχήμα 6: False Positive Rate ανάλογα με τη μετρική επιλογής χαρακτηριστικών 30

Σχήμα 7: Ακρίβεια κατηγοριοποίησης ανάλογα με τη μετρική επιλογής χαρακτηριστικών 31

Σχήμα 8: False Positive Rate ανάλογα με το πλήθος χαρακτηριστικών 32

Σχήμα 9: Ακρίβεια ανάλογα με το πλήθος χαρακτηριστικών 33

Σχήμα 10: False Positive Rate ανάλογα με την Ορμή του Δικτύου 34

Σχήμα 11: Ακρίβεια ανάλογα με την Ορμή του Δικτύου 35

Σχήμα 12: False Positive Rate ανάλογα με το Ρυθμό Εκμάθησης του Δικτύου . 36

Σχήμα 13: Ακρίβεια ανάλογα με το Ρυθμό Εκμάθησης του Δικτύου 37

Σχήμα 14: False Positive Rate ανάλογα με το πλήθος των κρυφών κόμβων του Δικτύου 38

Σχήμα 15: Ακρίβεια ανάλογα με το πλήθος των κρυφών κόμβων του Δικτύου . 39

Σχήμα 16: False Positive Rate ανάλογα με το χρόνο εκπαίδευσης του Δικτύου 40

Σχήμα 17: Ακρίβεια ανάλογα με το χρόνο εκπαίδευσης του Δικτύου 41

1. ΕΙΣΑΓΩΓΗ

Η χρήση **ηλεκτρονικού ταχυδρομείου** (**email**) έχει διαδοθεί πάρα πολύ τα τελευταία χρόνια για την επικοινωνία μεταξύ των ανθρώπων. Η χρήση email καθιστά την επικοινωνία γρήγορη, δωρεάν και χωρίς σύνορα, μπορεί να χρησιμοποιηθεί τόσο για επίσημη όσο και για ανεπίσημη αλληλογραφία, μπορεί επίσης να μεταφέρει αρχεία όπως π.χ. έγγραφα χωρίς την απώλεια ακρίβειας των τηλεομοιοτυπιών (Fax) και πρακτικά οτιδήποτε άλλο.

Μαζί με την ευκολία του email βέβαια έρχονται και ορισμένοι κίνδυνοι. Ένας από αυτούς είναι η **ανεπιθύμητη αλληλογραφία** ή αλλιώς **spamming**. Η ανεπιθύμητη αλληλογραφία, αν και όχι αποκλειστικά, συναντάται πολύ συχνά στα email. Σαν ανεπιθύμητο/αθέμιτο (spam) email εννοούμε τα email που στέλνονται μαζικά συνήθως σε χιλιάδες παραλήπτες. Ένα ηλεκτρονικό μήνυμα συγκαταλέγεται ως ανεπιθύμητο εάν η ταυτότητα του παραλήπτη και το γενικό πλαίσιο (context) του μηνύματος είναι αδιάφορα, επειδή το μήνυμα στοχεύει εξ ίσου σε πολλούς άλλους πιθανούς παραλήπτες, ή αν ο παραλήπτης δεν έχει διαπιστωμένα δώσει σαφή έγκριση να σταλεί το μήνυμα σε αυτόν [2].

Τα ανεπιθύμητα email μπορούν να βλάψουν τον παραλήπτη με πολλούς τρόπους. Εκτός του ότι μπορεί ενοχλούν από την υπερβολική διαφήμιση κάποιου προϊόντος ή υπηρεσίας, ενδέχεται να παραπέμπουν σε κακόβουλες ιστοσελίδες με σκοπό την παραπλάνηση του χρήστη όπως π.χ. ιστοσελίδες κλοπής ταυτότητας (phising sites) έτσι ώστε να αποσπάσουν από αυτόν ευαίσθητες προσωπικές πληροφορίες όπως κωδικούς καρτών ή οτιδήποτε άλλο. Ακόμα μπορεί να έχουν παραπλανητικές ή εντελώς ψεύτικες προσφορές ή και απάτες προκειμένου να ξεγελάσουν τον χρήστη και στις περισσότερες περιπτώσεις να του αποσπάσουν χρήματα, όπως τα “nigerian scam emails”. Επιπλέον μπορεί να μεταφέρουν και κακόβουλο λογισμικό όπως ιούς ή “δούρειους ίππους” (Trojan horses) προκειμένου να προσβάλουν τον υπολογιστή του παραλήπτη [3].

Εκτός από τα παραπάνω, τα ανεπιθύμητα email δημιουργούν πρόβλημα από τον υπερβολικά μεγάλο αριθμό τους στο εύρος ζώνης του δικτύου (bandwidth), που μπορεί να καταναλώσει ακόμα και το μισό [2], και στον χώρο αποθήκευσης, τόσο στον παραλήπτη όσο και στον εξυπηρετητή (mail server) που είναι υπεύθυνος για την διανομή και την αποθήκευση τους. Ακόμα μπερδεύουν τους χρήστες και σπαταλάνε το χρόνο τους αφού είναι αναγκασμένοι να τα ξεχωρίζουν από τα θεμιτά (legitimate/ ham) email [1, 3].

Οι αποστολείς ανεπιθύμητης αλληλογραφίας (spammers) μπορούν να συλλέξουν μεγάλο αριθμό από διευθύνσεις ηλεκτρονικού ταχυδρομείου χρηστών με σχετική ευκολία. Μπορούν να αντλήσουν τις διευθύνσεις αυτές από σελίδες κοινωνικής δικτύωσης, από λίστες των πελατών διάφορων υπηρεσιών ακόμα και από διαθέσιμους τηλεφωνικούς καταλόγους από ιδιώτες ή επιχειρήσεις (whitepages, yellowpages). Αυτό μπορεί να οδηγήσει τους χρήστες στο να έχουν εκατοντάδες ανεπιθύμητα email στα εισερχόμενά τους, τα οποία δεν έχουν καμία χρησιμότητα για αυτούς και είναι επικίνδυνα αν δεν δοθεί η απαραίτητη προσοχή [3].

Συγκεκριμένα υπολογίζεται ότι 14.5 δισεκατομμύρια email ημερησίως είναι ανεπιθύμητα. Αυτό σημαίνει ότι απαρτίζουν το 45% των συνολικών email που στέλνονται, δηλαδή σχεδόν ένα στα δύο email είναι αθέμιτο αν και κάποιες έρευνες βρίσκουν αυτό το ποσοστό ακόμα μεγαλύτερο έως και 73%. Τα ανειπθύμητα email στέλνονται κατά κύριο λόγο από τις ΗΠΑ ακολουθούμενες από την Κορέα. Το μεγαλύτερο ποσοστό αυτών, που υπολογίζεται γύρω στο 36%, αποτελούνται από διαφημιστικά email ενώ η δεύτερη πιο συχνή κατηγορία αθέμιτων email είναι αυτά που σχετίζονται με πορνογραφικό περιεχόμενο και απαρτίζουν το 31.7% των συνολικών αθέμιτων email και ακολουθούν αυτά με οικονομικά ζητήματα στο 26.5%. Πολύ μικρό ποσοστό των ανεπιθύμητων email, μόλις 2.5%, αποτελούν τα μηνύματα απάτης ή παραπλάνησης εκ των οποίων το 73% είναι μηνύματα κλοπής ταυτότητας γνωστά ως phising [1].

Γι’ αυτό το λόγο έχει προκύψει η ανάγκη να αντιμετωπιστεί η ανεπιθύμητη αλληλογραφία. Πολλές χώρες αναλογιζόμενες τον οικονομικό αντίκτυπο έχουν προχωρήσει σε νομοθετικές ρυθμίσεις για την αντιμετώπιση του spamming και την προστασία επιχειρήσεων και ιδιωτών. Όπως για παράδειγμα οι ΗΠΑ με το “CAN-SPAM Act of 2003” και η Ευρωπαϊκή Ένωση με το “Directive 2002/58 on Privacy and Electronic Communications” ή αλλιώς “E-Privacy Directive” [4] που απαγορεύουν τη χρήση ηλεκτρονικών διευθύνσεων για λόγους προώθησης αγαθών και υπηρεσιών εάν δεν υπάρχει εκ των προτέρων συμφωνία του παραλήπτη [2].

2. ΤΡΕΧΟΥΣΑ ΚΑΤΑΣΤΑΣΗ

Εκτός από την νομική αυτή προσέγγιση έχουν αναπτυχθεί και αυτοματοποιημένες μέθοδοι για την αντιμετώπιση του spamming και αυτό επιτυγχάνεται με τη χρήση των **φίλτρων ανεπιθύμητης αλληλογραφίας** (**spam filters**), ειδικών φίλτρων δηλαδή που έχουν σαν ρόλο να κατατάξουν κάθε εισερχόμενο μήνυμα σε μία από τις δύο κατηγορίες: {αθέμιτο, θεμιτό} ή αλλιώς **{spam, ham/legitimate}**. Το πρόβλημα αυτό δηλαδή αποτελεί ειδική περίπτωση προβλήματος εξόρυξης δεδομένων (data mining) και συγκεκριμένα της κατηγοριοποίησης κειμένου (text classification / text categorization). Στην περίπτωση που ένα εισερχόμενο μήνυμα κατηγοριοποιηθεί ως ανεπιθύμητο τότε η αντιμετώπιση του ποικίλει ανάλογα με τον τρόπο εφαρμογής του φίλτρου ανεπιθύμητης αλληλογραφίας. Για παράδειγμα, εάν αυτό λειτουργεί στον υπολογιστή του παραλήπτη (client side) τότε η πιο συνηθισμένη πρακτική είναι το αθέμιτο μήνυμα να μπαίνει σε κάποιον ειδικό φάκελο του ηλεκτρονικού του ταχυδρομείου που συχνά αναφέρεται ως spam/junk folder. Αν από την άλλη το φίλτρο λειτουργεί σε κάποιον εξυπηρετητή ηλεκτρονική αλληλογραφίας (mail server) που εξυπηρετεί πολλούς χρήστες (server side) τότε το μήνυμα μπορεί είτε να μαρκαριστεί ως αθέμιτο είτε να διαγραφεί τελείως [5].

Τα φίλτρα ανεπιθύμητης αλληλογραφίας μπορούν να χωριστούν σε δύο κατηγορίες. Σε αυτά που δεν χρησιμοποιούν μηχανική μάθηση και σε αυτά που χρησιμοποιούν. Ειδικά παλαιότερα, τα φίλτρα λειτουργούσαν χωρίς κάποιο είδος μηχανικής μάθησης και βασίζονταν κυρίως σε τεχνικές όπως μαύρες και λευκές λίστες (blacklists και whitelists) ελέγχοντας το κείμενο του μηνύματος για λέξεις κλειδιά ή διευθύνσεις IP και DNS του αποστολέα. Για παράδειγμα, αν ένα κείμενο μηνύματος περιέχει λέξεις που έχουν κατηγοριοποιηθεί ως ακατάλληλες και βρίσκονται σε μία μαύρη λίστα (blacklist) τότε το μήνυμα χαρακτηρίζεται ως spam. Επίσης αν ο αποστολέας είναι εξακριβωμένα ασφαλής και βρίσκεται σε κάποια λευκή λίστα (whitelist) τότε το μήνυμα χαρακτηρίζεται αυτόματα ως θεμιτό / ham. Οι τεχνικές αυτές δεδομένου ότι βασίζονται σε λίστες, σχετικά στάσιμες, και σε κανόνες “γραμμένους με το χέρι”, μπορούν εύκολα να παρακαμφθούν από αποστολείς ανεπιθύμητης αλληλογραφίας καθώς με την πάροδο του χρόνου οι λίστες γίνονται γνωστές και τα αθέμιτα email προσαρμόζονται. Παραδείγματος χάρη αν σε μία μαύρη λίστα υπάρχει η λέξη “Κερδίσατε” ως ακατάλληλη, μπορεί αντί για αυτή να χρησιμοποιηθεί η λέξη “Κ3ρδίσατε” με σκοπό να περάσει τον έλεγχο. Οι τεχνικές αυτές απαιτούν “χειροκίνητη” ανανέωση και υπάρχει συχνά λανθασμένη κατηγοριοποίηση θεμιτών μηνυμάτων ως αθέμιτων [3, 5].

Λόγω αυτού τα φίλτρα ανεπιθύμητης αλληλογραφίας πλέον βασίζονται σε τεχνικές μηχανικής μάθησης. Οι αλγόριθμοι που χρησιμοποιούνται σε τέτοιες τεχνικές αντί να βασίζονται εξ ολοκλήρου σε προκαθορισμένους κανόνες, λαμβάνουν υπόψιν τους σύνολα από διαθέσιμα μηνύματα ήδη κατηγοριοποιημένα $\{spam, ham\}$ και χρησιμοποιούν τη γνώση που εξήγαγαν για την κατηγοριοποίηση των νέων εισερχόμενων μηνυμάτων. Ακόμα μία χαρακτηριστική και ενδιαφέρουσα ιδιότητα των αλγορίθμων αυτών είναι η δυνατότητα τους

να βελτιώνουν τις επιδόσεις τους με την πάροδο του χρόνου αποκτώντας “εμπειρία” και είναι προσαρμόσιμοι στις αλλαγές των χαρακτηριστικών των ανεπιθύμητων email [5].

Μερικοί τέτοιοι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται για να φιλτράρουν τα ανεπιθύμητα email είναι η Bayesian κατηγοριοποίηση (Bayesian classification) και πιο συγκεκριμένα η Naive Bayesian, οι Μηχανές Διανύσματος Υποστήριξης (Support Vector Machines - SVM), ο αλγόριθμος k-Πλησιέστεροι Γείτονες (k-Nearest Neighbors - k-NN), τα Νευρωνικά Δίκτυα (Neural Networks / Artificial Neural Networks - ANN), τα Δέντρα Απόφασης (Decision Trees) και άλλοι, Μερικούς από αυτούς θα τους αναλύσουμε περισσότερο παρακάτω.

2.1 Δομή ενός Φίτρου Ανεπιθύμητης Αλληλογραφίας

Για την εφαρμογή των περισσότερων τέτοιων αλγορίθμων χρειάζεται αρχικά μία προεπεξεργασία στο εισερχόμενο μήνυμα. Αυτή αφορά κυρίως το Σώμα (Body) και το Θέμα (Subject), του email και συνίσταται στις εξής διαδικασίες [2, 3, 5]:

- Στοιχειοποίηση (Tokenization), δηλαδή διαχωρισμός κάθε λέξης από το σώμα του μηνύματος σαν στοιχείο (token).
- Λημματοποίηση (Lemmatization / Word Stemming), δηλαδή μετατροπή κάθε λέξης – token στην ριζική της μορφή / θέμα κάθε λέξης απαλείφοντας προθέσεις και καταλήξεις από ρήματα, μετατρέποντας τον πληθυντικό σε ενικό αριθμό στα ουσιαστικά, και χρησιμοποιώντας άλλους τέτοιους γραμματικούς κανόνες. (π.χ. “τρέχοντας” σε “τρέχω”, ή “αυτοκίνητα” σε “αυτοκίνητο”).
- Αφαίρεση κοινών λέξεων (Stop Word Removal / Stop Term Removal), δηλαδή εξάλειψη των όρων που είναι κοινοί σε όλα σχεδόν τα μηνύματα όπως άρθρα, προθέσεις και συνδέσμους π.χ. (“το”, ”και”, “για”).
- Επιλογή Χαρακτηριστικών (Feature Selection), δηλαδή επιλογή των χαρακτηριστικών, πόσα θα είναι και ποια συγκεκριμένα, από το κείμενο που θα ληφθούν υπόψιν για την κατηγοριοποίηση του μηνύματος.
- Αναπαράσταση (Representation), δηλαδή απεικόνιση του κειμένου συνήθως του μηνύματος με κάποια προαποφασισμένη δομή δεδομένων ώστε να είναι σε θέση να το διαχειριστεί ο αλγόριθμος.

2.1.1 Αναπαράσταση

Συνήθως η αναπαράσταση ενός κειμένου γίνεται μέσω ενός διανύσματος διάστασης ίσης με το πλήθος των χαρακτηριστικών που έχουν επιλεγεί. Μία από τις συνηθέστερες αναπαραστήσεις για ένα κείμενο και κατ’ επέκταση και για ένα email είναι η λεγόμενη Bag of Words (BoW) αναπαράσταση ή αλλιώς μοντέλο Χώρου Διανύσματος (Vector Space model). Με αυτόν τον τρόπο έχοντας δεδομένο ένα σύνολο / λίστα από N το πλήθος ξεχωριστούς όρους έστω $\mathrm{T} = \{t_1, t_2, \ldots, t_N\}$ που έχουν αποφασιστεί εκ των προτέρων (a priori),

και κάθε κείμενο *d* αναπαρίσταται από ένα διάνυσμα διάστασης *N* ως $\vec{x} = [x_1, x_2, \ldots, x_N]$ που έχει μία τιμή / βάρος (weight) για κάθε τέτοιο όρο. Αυτοί αποτελούν τους *N* πιο αντιπροσωπευτικούς όρους για κάθε κείμενο και τα βάρη αυτά δείχνουν πόσο συνεισφέρει κάθε όρος στην σημασιολογία του κειμένου *d*. Οι τιμές x_i των χαρακτηριστικών του διανύσματος μπορεί στην απλούστερη περίπτωση να είναι δυαδικές $\{0, 1\}$, σηματοδοτώντας την ύπαρξη ή μη του όρου t_i στο κείμενο *d* [5, 8].

Μπορεί επίσης να παίρνουν τιμές και από άλλα στατιστικά / βάρη όπως για παράδειγμα τη συχνότητα του κάθε όρου, και η τιμή $x_i = n_{t_i}$ δηλαδή, να ισούται με το πλήθος των εμφανίσεων του όρου t_i στο κείμενο *d*. Οι D. Puniškis, R. Laurutis και R. Dirmeikis στο [12], χρησιμοποιούν για χαρακτηριστικά κυρίως τη σχετική συχνότητα κάθε όρου δηλαδή $t_{i_{freq}} = 100 \cdot \frac{n_{t_i}}{N_T}\%$, όπου n_{t_i} το πλήθος των εμφανίσεων του όρου t_i στο κείμενο *d* και N_T το συνολικό πλήθος όλων των όρων στο *d*, καθώς και την αντίστοιχη σχετική συχνότητα για τους ASCII κωδικούς των πιο κοινών χαρακτήρων δηλαδή $c_{i_{freq}} = 100 \cdot \frac{n_{c_i}}{N_C}$, όπου n_{c_i} το πλήθος των εμφανίσεων του χαρακτήρα c_i στο *d* και N_C το πλήθος όλων των χαρακτήρων στο *d*. Ακόμα, πολύ συχνά επιλέγεται η χρήση του βάρους *tf-idf* (*term frequency - inverse document frequency*) ως αναπαράσταση χαρακτηριστικού. Όπου *term frequency* είναι κάποιο είδος συχνότητας (απόλυτη, σχετική, κλπ) ενός όρου t_i σε κάποιο κείμενο *d* (συνήθως $tf_{i,d} = \frac{n_{t_i}}{N_T}$) και *inverse document frequency* είναι ένα μέγεθος το οποίο καθορίζει το πόση πληροφορία προσφέρει ο κάθε όρος, δηλαδή το πόσο συχνά ή σπάνια συναντάται σε ένα πλήθος κειμένων (Συνήθως $idf_i = \log(\frac{N_D}{|\{d \in D : t_i \in d\}|})$, όπου $N_D = |D|$ το πλήθος όλων των εγγράφων). Το βάρος *tf-idf* υπολογίζεται ως το γινόμενο των δύο $tfidf_{i,d} = tf_{i,d} \cdot idf_i$. Τόσο η *term frequency* όσο και η *inverse document frequency* μπορούν να πάρουν πιο απλούς ή πιο σύνθετους τύπους ανάλογα με τις απαιτήσεις της εφαρμογής. Η χρήση του *tf-idf* βοηθάει στο να μειώνεται το βάρος των όρων που συναντώνται πολύ συχνά αλλά χωρίς να έχουν την απαραίτητη σημαντικότητα και να αυξάνεται το βάρος εκείνων που εμφανίζονται λιγότερο συχνά αλλά έχουν μεγαλύτερη σημασία.

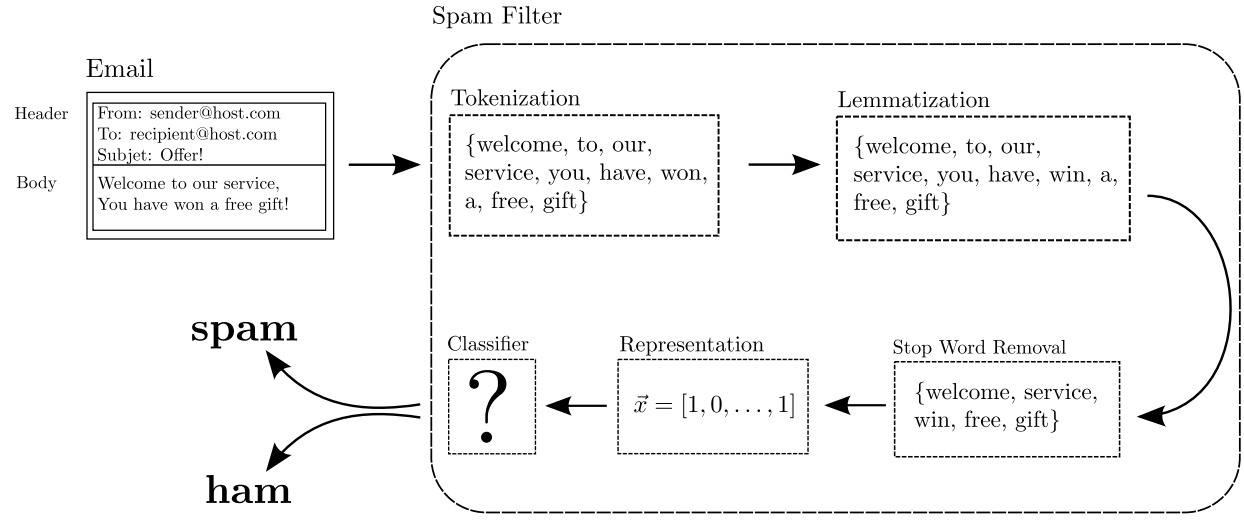
Συνήθως σε αυτές τις εφαρμογές οι διαστάσεις είναι πάρα πολύ μεγάλες, κυρίως λόγω του μεγάλου πλήθους των όρων ή των άλλων χαρακτηριστικών που εξετάζονται. Δεν είναι όμως δυνατό να κατασκευαστούν διανύσματα αναπαράστασης των εγγράφων (ή ο πίνακας όρων-εγγράφων / term by document matrix) τόσο μεγάλης διάστασης γιατί δεν θα είναι καθόλου αποδοτική η διαχείριση και η επεξεργασία τους. Γι'αυτό το λόγο γίνεται μία επιλογή ενός πλήθους N διαστάσεων / χαρακτηριστικών που θα χρησιμοποιηθούν. Συνηθίζεται να διαλέγονται τα χαρακτηριστικά με τις N μεγαλύτερες τιμές της μετρικής συνάρτησης που δείχνει τη σημαντικότητα των χαρακτηριστικών που εξετάζονται, ή ορισμένες φορές όπως στο [11] εφαρμόζονται πιο σύνθετοι αλγόριθμοι επιλογής όπως ο SVD (Singular Value Decomposition) που βασίζονται σε ιδιότητες της γραμμικής άλγεβρας.

Βέβαια, σε ορισμένες περιπτώσεις όπως για παράδειγμα στο [6] που ειδικεύεται στην αναγνώριση ενός συγκεκριμένου τύπου ανεπιθύμητων emails, αυτό των email κλοπής ταυτότητας (phising), οι T. Kathirvalavakumar, K. Kavitha και R. Palaniappan επιλέ-

γουν χαρακτηριστικά (features) για την αναπαράσταση των email που εξετάζουν που δεν βασίζονται σε συναρτήσεις της στατιστικής. Πιο συγκεκριμένα εκείνοι έχουν επιλέξει 18 χαρακτηριστικά για να αναπαραστήσουν τα email, και αυτά παίρνουν δυαδικές τιμές ανάλογα με τον αν στο εκάστοτε email ισχύει αυτό το χαρακτηριστικό. Τα χαρακτηριστικά που επιλέγουν είναι:

1. Αν χρησιμοποιείται HTML κώδικας μέσα στο email.
2. Αν το πλήθος των εικόνων στο email που λειτουργούν σαν σύνδεσμος (link) είναι μεγαλύτερος από 2.
3. Αν ο αριθμός των domain στο σώμα του email είναι μεγαλύτερος από 3.
4. Αν ο αριθμός των συνδέσμων που εμφανίζονται στο email είναι μεγαλύτερος από 3.
5. Αν ο HTML κώδικας που χρησιμοποιείται περιέχει το tag <form>
6. Αν το domain στο πεδίο From δεν είναι ίδιο με αυτό στο πεδίο Reply-To στο Header κομμάτι του email
7. Αν το μέγεθος του email είναι μικρότερο από 25 KB.
8. Αν υπάρχει ενσωματωμένος κώδικας JavaScript.
9. Αν το domain περιέχει παραπάνω από 3 τελείες (dots).
10. Αν το email περιέχει IP σαν σύνδεσμο.
11. Αν το κείμενο του email περιέχει τις λέξεις “like”, “click here”, “click”, “here” ή “login”.
12. Αν υπάρχουν περισσότερα από 3 domain στο Header κομμάτι του email.
13. Αν υπάρχει κάποιο URL που περιέχει το σύμβολο “@”.
14. Αν υπάρχει κάποιο URL με διαφορετική θύρα (port) από 80 ή 443.
15. Αν το domain κάποιου συνδέσμου στο email δεν αναδιευθύνεται σε αυτό του αποστολέα.
16. Αν το URL στο Header κομμάτι του email ξεκινάει από https:// αντί για http:// προκειμένου να κάνει τον παραλήπτη να πιστεύει ότι το email είναι θεμιτό.
17. Αν το URL στο Header κομμάτι έχει δεκαεξαδική μορφή
18. Αν το email έχει κατηγοριοποιηθεί από το φίλτρο Spam Assassin σαν ανεπιθύμητο.

Αν κάποιο από τα παραπάνω ισχύει για ένα email, τότε το αντίστοιχο χαρακτηριστικό παίρνει την τιμή 1, αλλιώς 0.



Σχήμα 1: Βασική δομή ενός Φίλτρου Ανεπιθύμητης Αλληλογραφίας

Βέβαια, στο [5] αναφέρεται ότι παρά την συχνή χρήση της Bag of Words αναπαράστασης στις εφαρμογές για αναγνώριση ανεπιθύμητης αλληλογραφίας, αυτή έχει την αδυναμία του ότι δεν ανταπεξέρχεται καλά στην πάροδο του χρόνου λόγω του φαινομένου που ονομάζεται ολίσθηση έννοιας (concept drift). Το πρόβλημα στην BoW αναπαράσταση είναι ότι τα διανύσματα που χρησιμοποιούνται για την αναπαράσταση των εγγράφων / email, είναι στατικά και δεν λαμβάνουν υπόψιν τους τις μεταβολές που μπορεί να έχουν μερικά χαρακτηριστικά / όροι στην πάροδο του χρόνου. Αυτό σημαίνει ότι εάν ένα χαρακτηριστικό μεταβληθεί ή αν πρέπει να ληφθεί υπόψιν ένα καινούριο χαρακτηριστικό, τα φίλτρα ανεπιθύμητης αλληλογραφίας που ακολουθούν αυτήν την αναπαράσταση θα πρέπει να εκπαιδευθούν ξανά από την αρχή, δηλαδή όλα τα έγγραφα να αναπαρασταθούν ξανά δεδομένων των καινούριων χαρακτηριστικών και να κατηγοριοποιηθούν εκ νέου. Η διαδικασία αυτή δεν είναι καθόλου αποδοτική εάν εφαρμόζεται συχνά.

2.2 Συλλογές Δεδομένων (Datasets)

Ένα σημαντικό στοιχείο κατά την λειτουργία των διάφορων φίλτρων ανεπιθύμητης αλληλογραφίας είναι τα δεδομένα που χρησιμοποιεί. Τα δεδομένα αυτά αφορούν τα email που έχει στη διάθεσή του κάθε φίλτρο ώστε να εκπαιδευτεί, να μάθει δηλαδή ποιά email θεωρούνται αθέμιτα και ποια όχι για να μπορεί στο μέλλον να βγάζει συμπεράσματα για καινούρια email που δεν γνωρίζει.

Οι συλλογές από email (email corpora) που χρησιμοποιούνται έχουν ομαδοποιημένα email που ανήκουν διαπιστωμένα σε μία κατηγορία, είτε θεμιτά είτε αθέμιτα. Αυτά έχουν συλλεγεί ήδη για κάποια χρονική περίοδο από διάφορες πηγές. Οι συλλογές αυτές μπορούν να είναι είτε δημόσια διαθέσιμες είτε ιδιωτικές. Κάθε συλλογή έχει τα δικά της χαρακτηριστικά ανάλογα με το από που έχουν συλλεγεί τα μηνύματα. Γι'αυτό το λόγο δεν είναι δυνατόν να χρησιμοποιηθεί οποιαδήποτε συλλογή σε οποιαδήποτε εφαρμογή / φίλ-

τρο ανεπιθύμητης αλληλογραφίας. Για παράδειγμα, αν η συλλογή των email έχει γίνει από κάποια λίστα ηλεκτρονικής αλληλογραφίας (mailing list) που συνήθως έχει συγκεκριμένη θεματολογία, τότε είναι πολύ πιθανόν να μην είναι χρήσιμη για αξιολόγηση μηνυμάτων που δεν προέρχονται από τη λίστα αυτή. Αυτό συμβαίνει, επειδή τα χαρακτηριστικά που θα έχουν αποθηκευτεί για τα θεμιτά ή αθέμιτα email δεν θα αντιπροσωπεύουν κατ' ανάγκη οποιαδήποτε άλλα θεμιτά ή αθέμιτα email αντίστοιχα.

Η διαφορετικότητα κάθε συλλογής μηνυμάτων που χρησιμοποιείται είναι και ένας λόγος για τον οποίο δεν έχει νόημα να συγκρίνονται δύο φίλτρα ανεπιθύμητης αλληλογραφίας αποκλειστικά και μόνο από τα μέτρα απόδοσης τους. Επιπλέον, μερικές εφαρμογές χρησιμοποιούν ιδιωτικές συλλογές μηνυμάτων προκειμένου να δώσουν έμφαση σε συγκεκριμένες ιδιαιτερότητες κάθε συλλογής όπως οι γλώσσες που χρησιμοποιούνται ή τα τυχόν ιδιαίτερα θεματολογικά χαρακτηριστικά τα οποίτα δεν είναι διαθέσιμα σε δημόσια διαθέσιμες συλλογές. Αυτό βέβαια μπορεί να επηρεάσει την δυνατότητα αναπαραγωγής των αποτελεσμάτων καθώς και να δυσκολέψει την σύγκριση με άλλα φίλτρα όπως αναφέρεται στο [5]. Μερικές διαθέσιμες συλλογές μηνυμάτων που χρησιμοποιούνται συχνά είναι το *CEAS 2008 Live Spam Challenge Corpus* [13] που χρησιμοποιήθηκε για την δοκιμασία Αναγώρισης Ανεπιθύμητης Αλληλογραφίας κατά τη διάρκεια του συνεδρίου CEAS 2008 (5 Αυγούστου – 8 Αυγούστου, 2008), το *2005-2007 TREC Public Spam Corpus* [14] που χρησιμοποιήθηκαν για το TREC Spam Evaluation Track, ή το *Reuters-21578* [15] που περιέχει 21578 μηνύματα διάφορων θεμάτων που έχουν κατηγοριοποιηθεί χειροκίνητα απο το προσωπικό της Carnegie Group, Inc. και της Reuters Ltd.

Σημαντικό ρόλο παίζει επίσης η χρονική περίοδος κατά την οποία έγινε η συλλογή των μηνυμάτων που θα χρησιμοποιηθούν για την εκπαίδευση του φίλτρου ανεπιθύμητης αλληλογραφίας. Όπως αναφέρεται στο [10] τα δεδομένα που χρησιμοποιόυνται για εκπαίδευση πρέπεί να ανανεώνονται κατά καιρούς καθώς αλλάζουν και τα χαρακτηριστικά των email. Τα χαρακτηριστικά των θεμιτών / ham email δεν αλλάζουν τόσο συχνά και σε τόσο μεγάλο βαθμό καθώς οι άνθρωποι σε σπάνιες περιπτώσεις αλλάζουν τον τρόπο που γράφουν μηνύματα. Το περεχόμενο των αθέμιτων / spam όμως αλλάζει συνεχώς καθώς λαμβάνει χώρα ο "αγώνας" μεταξύ των αποστολέων ανεπιθύμητης αλληλογραφίας και των δημιουργών φίλτρων. Πιο συγκεκριμένα το να χρησιμοποιούνται απαρχαιομένα δεδομένα για εκπαίδευση μειώνει την απόδοση του φίλτρου, αλλά η μείωση αυτή της απόδοσης δεν συγκρίνεται με την τεράστια μείωση που υφίσταται η απόδοση όταν τα δεδομένα που αφορούν αθέμιτα email και τα δεδομένα που αφορούν θεμιτά email έχουν διαφορετικές ηλικίες.

2.3 Μέτρα Απόδοσης

> Email is not just text; it has structure. Spam filtering is not just classification, because false positives are so much worse than false negatives that you should treat them as a different kind of error. And the source of error is not just random variation, but a live human spammer working actively to defeat your filter.
>
> *Paul Graham*
> *Better Bayesian Filtering*

Σε ότι αφορά την απόδοση του φίλτρου ανεπιθύμητης αλληλογραφίας, μια αφελής προσέγγιση θα ήταν να λαμβάνεται υπόψιν το πλήθος των μηνυμάτων που αποφασίστηκε ότι ανήκουν σε άλλη κατηγορία από αυτή που ανήκουν κανονικά (δηλαδή αυτή που θα τα κατέτασσε ένας άνθρωπος). Αυτή η προσέγγιση όμως δεν είναι σωστή διότι δεν έχει το ίδιο αντίκτυπο το να κατηγοριοποιηθεί ένα αθέμιτο email ως θεμιτό και το να κατηγοριοποιηθεί ένα θεμιτό email ως αθέμιτο. Η πρώτη περίπτωση ονομάζεται *False Negative* και η δεύτερη *False Positive*. Προφανώς το πλήθος και των δύο αυτών μέτρων θα πρέπει να είναι το ελάχιστο δυνατό αλλά η μικρή τιμή των False Positive είναι πολύ πιο σημαντική. Συγκεκριμένα οποιαδήποτε μη μηδενική τιμή δεν είναι αποδεκτή και αυτό διότι το να κατηγοριοποιηθεί ένα αθέμιτο email ως θεμιτό και να το δει ο χρήστης στα εισερχόμενά του δεν αποτελεί τόσο σοβαρό πρόβλημα (αυτό θα συνέβαινε ούτως ή άλλως χωρίς τη χρήση του φίλτρου), το να κατηγοριοποιηθεί όμως ένα θεμιτό email ως αθέμιτο μπορεί να προκαλέσει την απώλεια ενός σημαντικού για το χρήστη μηνύματος. Ειδικά σε περιπτώσεις που το μήνυμα διαγράφεται τελείως αν αναγνωριστεί ως αθέμιτο, αυτό το είδος της λάθος κατηγοριοποίησης αποτελεί ακόμα μεγαλύτερο πρόβλημα. Επιπλέον, το να έχει ακόμα και μικρή, μη μηδενική όμως τιμή το πλήθος των *False Positive* σημαίνει ότι ο χρήστης θα πρέπει να ψάχνει και στο φάκελο των spam για κάποιο τυχόν λάθος κατηγοριοποιημένο μήνυμα κάτι που ακυρώνει ουσιαστικά νόημα του φίλτρου.

Γι'αυτό το λόγο λοιπόν, σχεδόν σε όλα τα φίλτρα λαμβάνονται υπόψιν τα εξής μέτρα απόδοσης:

- *True Positive*, δηλαδή αθέμιτα email που έχουν αναγνωριστεί ως αθέμιτα.

$$TP = \frac{n_{spam} \rightarrow spam}{N_{spam}}$$

- *True Negative*, δηλαδή θεμιτά email που έχουν αναγνωριστεί ως θεμιτά.

$$TN = \frac{n_{ham} \rightarrow ham}{N_{ham}}$$

- *False Positive*, δηλαδή θεμιτά email που έχουν αναγνωριστεί ως αθέμιτα.

$$FP = \frac{n_{ham} \rightarrow spam}{N_{ham}}$$

- *False Negative*, δηλαδή αθέμιτα email που έχουν αναγνωριστεί ως θεμιτά.

$$FN = \frac{n_{spam} \rightarrow ham}{N_{spam}}$$

Όπου N_{spam} το συνολικό πλήθος των αθέμιτων / spam email, N_{ham} το συνολικό πλήθος των ham / θεμιτών email και $n_i \rightarrow j$, το πλήθος των email που ανήκουν στην κλάση i και αναγνωρίστικαν από το φίλτρο ότι ανήκουν στην κλάση j με $i,j \in \{spam, ham\}$. Για την ακρίβεια (accuracy) της κατηγοριοποίησης του φίλτρου χρησιμοποιείται η τιμή του πλήθους των σωστά κατηγοριοποιημένων μηνυμάτων προς το πλήθος των συνολικών μηνυμάτων. Δηλαδή, όπως και στα [11, 7, 8] αναπαρίσταται ως:

$$ACC = \frac{n_{spam} \rightarrow spam + n_{ham} \rightarrow ham}{N_{spam} + N_{ham}}$$

Ή όπως στο [6] ως:

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

Επίσης συχνά χρησιμοποιούνται τα μέτρα:

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

$$True\ Positive\ Rate = Recall = \frac{TP}{N_{spam}} \quad \text{και} \quad False\ Positive\ Rate = \frac{FP}{N_{ham}}$$

Βέβαια πολλά φίλτρα ανεπιθύμητης αλληλογραφίας χρησιμοποιούν πιο σύνθετα μέτρα για την ακρίβεια της εφαρμογής, προκειμένου η απόδοση να είναι πιο κοντά στα επιθυμητά αποτελέσματα. Ένας τρόπος εκτίμησης της απόδοσης που είναι ιδιαίτερα δημοφιλής με τους αλγορίθμους μηχανικής μάθησης λόγω του φαινομένου ότι δεν είναι ίσα τα κόστη λάθους κατηγοριοποίησης, (κόστος *False Positive* πολύ μεγαλύτερο κόστους *False Negative*), είναι το γράφημα ROC (Receiver Operating Characteristic).

Όπως αναφέρεται και στο [16], το γράφημα ROC είναι ένα γράφημα 2 διαστάσεων που απεικονίζει τη σχέση των True Positive με των False Positive κατηγοριοποιήσεων. Συγκεκριμένα, στον κατακόρυφο άξονα (Y) απεικονίζεται η αναλογία των *True Positive* (*True Positive Rate* ή *Recall*) και στον οριζόντιο άξονα (X) η αναλογία των *False Positive* (*False Positive Rate*). Ένας κατηγοριοποιητής (classifier) που αποφασίζει ότι ένα έγγραφο / email ανήκει σε μία κλάση, παράγει ένα ζευγάρι (*True Positive Rate*, *False Positive Rate*) που αντιστοιχεί σε ένα σημείο στο διάγραμμα ROC.

Σχήμα 2: Βασική δομή ενός ROC γραφήματος, απεικονίζοντας 3 κατηγοριοποιητές A,B,C.

Προφανώς, σκοπός του κάθε κατηγοριοποιητή ή του φίλτρου ανεπιθύμητης αλληλογραφίας στην προκειμένη περίπτωση είναι να βρίσκεται πάνω από την ευθεία $y = x$ αλλα φυσικά το πόσο σωστά θεωρείται ότι γίνεται η κατηγοριοποίηση εξαρτάται και από την εφαρμογή. (Π.χ. στην Αναγνώριση Ανεπιθύμητης Αλληλογραφίας είναι σχεδόν ανεπίτρεπτο το μη μηδενικό *False Positive Rate*). Το πλεονέκτημα της αναπαράστασης με ROC γραφήματα είναι η δυνατότητα που έχει ο κατηγοριοποιητής να θέτει ένα συμβιβασμό (tradeoff) ανάμεσα στα *True Positive* και στα *False Positive* όταν δεν μπορεί να πάρει απόλυτα ικανοποιητική απόδοση.

Μερικοί κατηγοριοποιητές μπορούν να αποδώσουν μια πιθανότητα στο να ανήκει το έγγραφο *d* στην κλάση *c*, ή κάποιο άλλο παρόμοιο σύστημα πόντων. Με αυτόν τον τρόπο, στα γραφήματα ROC μπορούν να σχεδιαστούν καμπύλες (ROC Curves) για τις διάφορες τιμές των εγγράφων και των αποτελεσμάτων τους. Έτσι όπως αναφέρεται και στο [5] ένα μέτρο απόδοσης που χρησιμοποιείται συχνά από αλγορίθμους μηχανικής μάθησης είναι η επιφάνεια κάτω από την καμπύλη αυτή (AUC). Η τιμή της AUC είναι ίση με την πιθανότητα ότι ο κατηγοριοποιητής θα προτιμήσει την ”*Positive*” (*spam* στην περίπτωση μας) κλάση για ένα τυχαίο ”*Positive*” δείγμα αντί για ένα τυχαίο ”*Negative*” (*ham*) δείγμα. Έτσι υψηλότερη τιμή AUC σημαίνει γενικά υψηλότερη ακρίβεια / μεγαλύτερη απόδοση του κατηγοριοποιητή. Βέβαια στα [5, 17] φαίνεται ότι οι V. Metsis, I. Androutsopoulos και G. Paliouras δεν θεωρούν ότι η χρήση της AUC ως μέτρο σύγκρισης για την απόδοση των φίλτρων ανεπιθύμητης αλληλογραφίας ενδείκνυται επειδή επηρεάζεται από περιοχές με ηψηλό *Ham Recall / False Positive Rate* που δεν έχουν ενδιαφέρον στην πράξη. Δηλαδή μπορεί ένας κατηγοριοποιητής να έχει καλύτερη απόδοση σχετικά με το *False Positive Rate* και παρόλα αυτά να έχει μικρότερη τιμή AUC. Γι’αυτό η σύγκριση μεταξύ τέτοιων φίλτρων χρησιμοποιώντας μόνο ένα μέτρο απόδοσης μπορεί να προκαλεί λανθασμένα συμπεράσματα.

2.4 Αλγόριθμοι Κατηγοριοποίησης

Το πιο σημαντικό σημαντικό ίσως κομμάτι ενός Φίλτρου Αναγνώρισης Ανεπιθύμητης Αλληλογραφίας είναι ο αλγόριθμος που θα χρησιμοποιήσει για να ταξινομήσει το εισερχόμενο ηλεκτρονικό μήνυμα σε μία από τις κλάσεις $\{spam, ham\}$ να αποφανθεί δηλαδή αν είναι θεμιτό ή αθέμιτο. Παρακάτω αναφέρονται τα βασικά στοιχεία για κάποιους από τους πιο διαδεδομένους αλγορίθμους κατηγοριοποίησης που χρησιμοποιούνται σε Φίλτρα Ανεπιθύμητης Αλλήλογραφίας.

2.4.1 Naive Bayes

Για την κατηγοριοποίηση / ταξινόμηση ενός εγγράφου σε μία κλάση, όπως και στην περίπτωση της Αναγνώρισης Ανεπιθύμητης Αλληλογραφίας που πρέπει ένα μήνυμα να κατηγοριοποιηθεί ως θεμιτό ή αθέμιτο, μπορεί να χρησιμοποιηθεί ο αλγόριθμος **Naive Bayes**. Όπως βλέπουμε και στα [5, 19], αρχικά η χρήση του προτάθηκε από τους M. Sahami, S. Dumais, D. Heckerman και E. Horvitz στο [18]. Στο Bayesian μοντέλο, σύμφωνα με το νόμο του Bayes, η πιθανότητα ότι ένα μήνυμα / έγγραφο d που αναπαρίσταται ως $\vec{x} = [x_1, x_2, \ldots, x_N]$ ανήκει σε μία κλάση $c \in \{spam, ham\}$ δίνεται από τον τύπο:

$$P(c|\vec{x}) = \frac{P(\vec{x}|c)P(c)}{P(\vec{x})} = \frac{P(\vec{x}|c)P(c)}{P(\vec{x}|spam)P(spam) + P(\vec{x}|ham)P(ham)} \tag{1}$$

Όπου $P(\vec{x}|c)$ η πιθανότητα ένα έγγραφο που έχει κατηγοριοποιηθεί στην κλάση c να αναπαρίσταται από το διάνυσμα $\vec{x}$, $P(c)$ η πιθανότητα το έγγραφο να ανήκει στην κλάση c, και $P(\vec{x})$ η εκ των προτέρων (*a priori*) πιθανότητα ένα έγγραφο d να αναπαρίσταται από το διάνυσμα $\vec{x}$. Σημαντική παραδοχή για τη χρήση του Naive Bayes κατηγοριοποιητή είναι ότι τα χαρακτηριστικά x_i, $i = 1, 2, \cdots, N$ είναι μεταξύ τους στοχαστικά ανεξάρτητα, δηλαδή η τιμή του ενός δεν μας δίνει καμία πληροφορία για την τιμή του άλλου (εξ ου και η ονομασία naive (αφελής)). Γι'αυτό το λόγο η πιθανότητα $P(\vec{x}|c)$ μπορεί να γραφτεί ως το γινόμενο των πιθανοτήτων των επιμέρους χαρακτηριστικών δηλαδή: $P(\vec{x}|c) = \prod_{i=1}^{N} P(x_i|c)$. Έτσι η (1) γίνεται:

$$P(c|\vec{x}) = \frac{\prod_{i=1}^{N} P(x_i|c)P(c)}{\prod_{i=1}^{N} P(x_i|spam)P(spam) + \prod_{i=1}^{N} P(x_i|ham)P(ham)} \tag{2}$$

Έτσι ο κανόνας της κατηγοριοποίησης είναι απλός. Αν $P(spam|\vec{x}) > P(ham|\vec{x})$, δηλαδή αν η εκ των υστέρων (*a posteriori*) πιθανότητα το έγγραφο με αναπαράσταση $\vec{x}$ να είναι αθέμιτο είναι μεγαλύτερη από την αντίστοιχη πιθανότητα να είναι θεμιτό, τότε το κατηγοριοποιούμε ως αθέμιτο, σε διαφορετική περίπτωση το κατηγοριοποιούμε ως θεμιτό. Βέβαια στην πράξη το να υπολογιστούν αυτές οι 2 πιθανότητες είναι περίπλοκο και γίνεται στη φάση της Εκπαίδευσης (Training) του κατηγοριοποιητή αφού πρώτα υπολογιστούν οι πιθανότητες $P(x_i|c)$, $i = 1, 2, \cdots, N$ για $c \in \{spam, ham\}$ από τα έγγραφα εκπαίδευσης. Και τελικά στη φάση της Κατηγοριοποίησης / Ταξινόμησης (Classification),

δεδομένου ενός εγγράφου / μηνύματος *d* υπολογίζεται το διάνυσμα αναπαράστασης $\vec{x}$ και από αυτό ανάλογα με την τιμή των παραπάνω πιθανοτήτων και τον κανόνα απόφασης, επιλέγεται η κλάση $\{spam, ham\}$ στην οποία ανήκει το *d*.

Αυτή είναι η βασική ιδέα πίσω από τον Naive Bayes κατηγοριοποιητή. Παραλλαγές αυτού χρησιμοποιούνται και σε σύγχρονα φίλτρα ανεπιθύμητης αλληλογραφίας όπως στα [17, 18] και οι διαφοροποιήσεις μπορούν να συνίστανται στο πλήθος και το είδος των χαρακτηριστικών που επιλέγονται, στο πιθανοτικό μοντέλο που υιοθετείται ή η συχνότητα και ο τρόπος της Εκπαίδευσης του κατηγοριοποιητή. Ο αλγόριθμος αυτός θεωρείται από τους πιο απλούς, έχοντας όμως ικανοποιητικά αποτελέσματα κατηγοριοποίησης και συχνά χρησιμοποιείται για σύγκριση με νέες μεθόδους.

2.4.2 k-Nearest Neighbors

Ο αλγόριθμος των k-Πλησιέστερων Γειτόνων (k-Nearest Neighbors) προϋποθέτει ότι υπάρχει η έννοια της απόστασης μεταξύ μηνυμάτων. Δηλαδή είμαστε σε θέση να αποφανθούμε πόσο ”κοντά” είναι ένα έγγραφο / μήνυμα με ένα άλλο. Γι αυτό το σκοπό μπορεί να χρησιμοποιηθεί και η ευκλείδεια απόσταση δύο διανυσμάτων για τα διανύσματα αναπαράστασης των δύο μηνυμάτων.

$$dst(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$

όπου $\vec{x} = [x_1, x_2, \cdots, x_N]$ και $\vec{y} = [y_1, y_2, \cdots, y_N]$, οι αναπαραστάσεις των δύο μηνυμάτων. Η ιδέα του αλγορίθμου αυτού είναι να κατηγοριοποιήσει ένα μήνυμα ανάλογα με την κλάση που ανήκουν τα k ”πλησιέστερα” (με τη μικρότερη απόσταση) μηνύματα από το σύνολο των μηνυμάτων εκπαίδευσης.

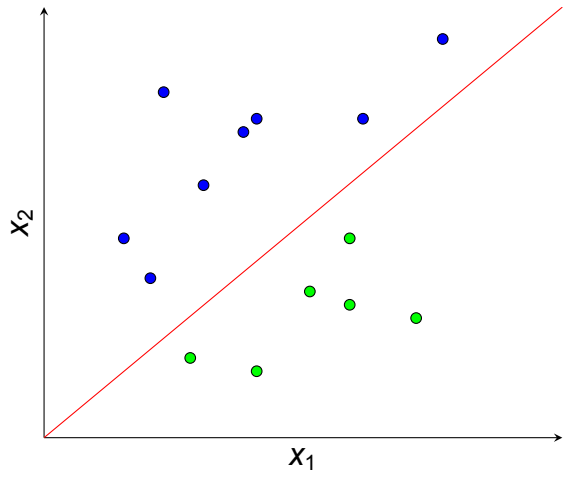
Στο [19] αναφέρεται μια μετατροπή του κανόνα κατηγοριοποίησης του αλγορίθμου προκειμένου να μειωθούν τα ποσοστά των False Positives, κάτι που είναι ζωτικής σημασίας για ένα φίλτρο ανεπιθύμητης αλληλογραφίας. Η μετατροπή αυτή συνίσταται στην εφαρμογή του *l*/*k* - *κανόνα*. Δηλαδή, αν για ένα μήνυμα *d*, *l* ή περισσότερα μηνύματα μεταξύ των k πλησιέστερων γειτόνων του ανήκουν στην κλάση *spam* (είναι αθέμιτα), τότε κατέταξε και το *d* ως αθέμιτο, αλλιώς κατέταξέ το ως θεμιτό.

Ο Αλγόριθμος των των k-Πλησιέστερων Γειτόνων απαιτεί, για κάθε μήνυμα που πρέπει να κατηγοριοποιηθεί, να υπολογιστεί η απόσταση του προς όλα τα μηνύματα εκπαίδευσης προκειμένου να επιλεχθούν τα k πλησιέστερα σε αυτό. Γι αυτό το λόγο ο αλγόριθμος αυτός έχει αρκετά μεγάλη χρονική πολυπλοκότητα [8, 19] αλλά όπως αναφέρεται και στο [8] έχει μεγαλύτερη ακρίβεια (Accuracy) συγκριτικά με 2 άλλους αλγορίθμους που δοκιμάστηκαν (Naive Bayes, Term Graph Model).

2.4.3 Support Vector Machines

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines) είναι ένας από τους πιο διαδεδομένος αλγορίθμους κατηγοριοποίησης. Ας υποθέσουμε ότι μεταξύ των δεδομένων μας (μηνύματα) που ανήκουν σε μία απο δύο κλάσεις ($\{spam, ham\}$ στην περίπτωσή μας) υπάρχει *γραμμική διαχωριστικότητα*, δηλαδή ότι αν τα δεδομένα αναπαρίστανται από σημεία στο επίπεδο, μπορούν να διαχωριστούν όλα τα στοιχεία της μίας κλάσης από όλα τα στοιχεία της άλλης κλάσης μέσω μίας γραμμής, που έχει εκατέρωθεν τα εκάστοτε δεδομένα. Η ιδέα του αλγορίθμου είναι να βρεί το μεγαλύτερο δυνατό περιθώριο μεταξύ των δεδομένων των δύο κλάσεων.

Αφού έχει μεγιστοποιηθεί το περιθώριο μεταξύ των δύο ξεχωριστών συνόλων από δεδομένα, τότε κάθε καινούριο μήνυμα που πρέπει να κατηγοριοποιηθεί αντιστοιχίζεται και αυτό στο χώρο που βρίσκονται τα δεδομένα της εκπαίδευσης και ανάλογα με τη θέση του στο χώρο αυτό αποφασίζεται η κλάση στην οποία ανήκει. Βέβαια η διαδικασία αυτή ανάγεται αμέσως και σε περισσότερες από δύο διαστάσεις, και η λειτουργία της Μηχανής Διανυσμάτων Υποστήριξης είναι να επιλέξει το βέλτιστο *υπερεπίπεδο (hyperplane)* (υποχώρος με μία διάσταση λιγότερη) κατ'αναλογία με την ευθεία στον δισδιάστατο χώρο. Βέλτιστο θεωρείται το υπερεπίπεδο με τη μεγαλύτερη απόσταση από το κοντινότερο δεδομένο εκπαίδευσης που αντιστοιχεί και σε βέλτιστη ακρίβεια του κατηγοριοποιητή. Τα δεδομένα που βρίσκονται πλησιέστερα από όλα στο υπερεπίπεδο ονομάζονται *διανύσματα υποστήριξης* από όπου προκύπτει και η ονομασία του αλγορίθμου.

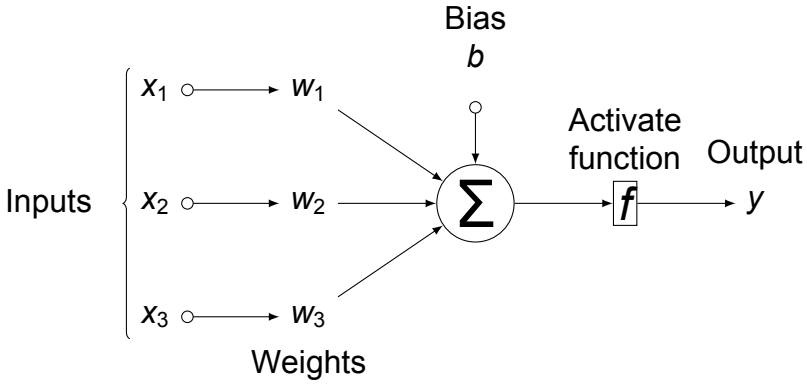


Σχήμα 3: Γραμμικά Διαχωρίσιμα δεδομένα του δισδιάστατου χώρου.

Όπως αναφέρεται και στο [20], στην περίπτωση που τα δεδομένα εκπαίδευσης δεν είναι γραμμικά διαχωρίσιμα, υπάρχουν τεχνικές για να εφαρμοστεί ο αλγόριθμος. Επίσης προτέρημα του αλγορίθμου των Μηχανών Διανυσμάτων Υποστήριξης είναι ότι μπορούν να αντεπεξέλθουν εύκολα σε διανύσματα μεγάλων διαστάσεων και γι'αυτό πολλές φορές δεν χρειάζεται να γίνει επιλογή των πιο αντιπροσωπευτικών χαρακτηριστικών για την αναπαράσταση, προκειμένου να μειωθεί η διάσταση των διανυσμάτων.

2.4.4 Νευρωνικά Δίκτυα

Τα Νευρωνικά Δίκτυα (Neural Networks / Artificial Neural Networks - ANN) αποτελούν σύνθετες συναρτήσεις που απαρτίζονται από μικρότερες μονάδες, τους *Νευρώνες*, και γραφικά απεικονίζονται σαν ένα *δίκτυο* από τέτοιους Νευρώνες. Ένας Νευρώνας είναι και αυτός μία συνάρτηση εμπνευσμένη από τους βιολογικούς νευρώνες και έχει N εισόδους $(x_1, x_2, \cdots, x_N)$ και συνήθως μία σταθερά b, και μία έξοδο y. Κάθε μία από τις N εισόδους έχει και ένα βάρος $(w_1, w_2, \cdots, w_N)$ που της αντιστοιχεί. Ο Νευρώνας χαρακτηρίζεται και από μία *συνάρτηση ενεργοποίησης* f και τελικά η έξοδος του καθορίζεται από την παράσταση: $y = f(\sum_{i=1}^{N} w_i x_i + b)$



Σχήμα 4: Βασική δομή ενός Νευρώνα.

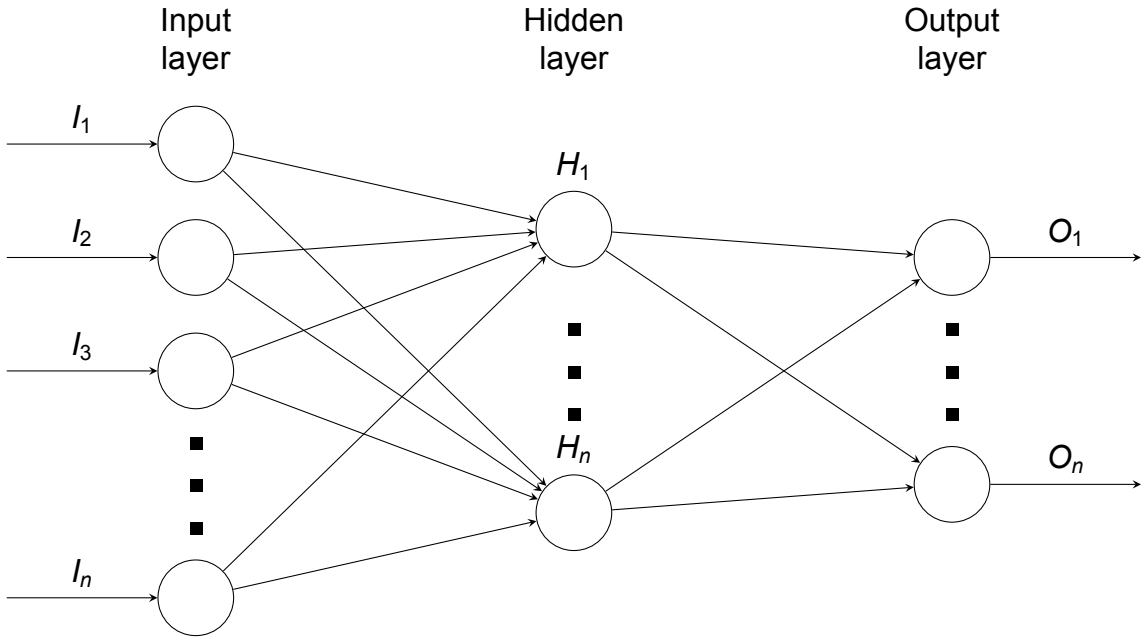
Τα Νευρωνικά Δίκτυα χρησιμοποιούνται συχνά ως αλγόριθμος κατηγοριοποίησης. Η πιο τετριμμένη περίπτωση Νευρωνικού Δικτύου είναι αυτή που έχει μόνο μία επεξεργαστική μονάδα (ένα Νευρώνα / Perceptron), και εκεί το ζητούμενο είναι η αναζήτηση μίας συνάρτησης ενεργοποίησης τέτοιας ώστε αν αυτή τροφοδοτηθεί με το διάνυσμα αναπαράστασης ενός εγγράφου / μηνύματος, τότε ανάλογα με την τιμή της, να μπορεί να αποφασιστεί η κλάση του. Στην περίπτωση του Φίλτρου Ανεπιθύμητης Αλληλογραφίας που υπάρχουν δύο κλάσεις στις οποίες μπορεί να ανήκει ένα έγγραφο ($\{spam, ham\}$), μπορούν αυτές να αντιστοιχιστούν στις τιμές $\{1, -1\}$ αντίστοιχα και η κλάση θα μπορούσε να αποφασίζεται από το πρόσημο της συνάρτησης f.

Η εκπαίδευση του Νευρώνα γίνεται επαναληπτικά και σε κάθε επανάληψη προσαρμόζονται τα βάρη $(w_1, w_2, \cdots, w_N)$ και η σταθερά b μέχρις ότου η συνάρτηση απόφασης να ταξινομεί σωστά κάθε ένα από τα δεδομένα εκπαίδευσης. Ο επαναληπτικός αυτός αλγόριθμος ξεκινά με αυθαίρετες τιμές για τα βάρη, και στη k-οστή επανάληψη επιλέγεται ένα από τα μηνύματα εκπαίδευσης που αναπαρίσταται ως $\vec{x}$ και ανήκει στην κλάση c το οποίο δνε έχει ταξινομηθεί σωστά και τα $(\vec{w}_k, b_k)$ καθορίζονται ως εξής: $\vec{w}_{k+1} = \vec{w}_k + c\vec{x}$, $b_{k+1} = b_k + c$ [19]

Βέβαια τα συνηθέστερα Νευρωνικά Δίκτυα αποτελούνται απο περισσότερους του ενός Νευρώνες και συγκεκριμένα από μερικά επίπεδα Νευρώνων (πολυεπίπεδα Νευρωνικά Δίκτυα / Multilayer Perceptron) ένα από τα οποία είναι το επίπεδο *Εισόδου*, ένα το

επίπεδο *Εξόδου* και τα ενδιάμεσα ονομάζονται *Κρυφά* επίπεδα. Οι έξοδοι καθενός από τα επίπεδα τροφοδοτεί την είσοδο του επόμενου. Βέβαια το πλήθος των Νευρώνων σε κάθε επίπεδο και η δομή του Δικτύου δεν είναι τετριμμένα, και η εκπαίδευση του πολυεπίπεδου Δικτύου είναι μια πιο σύνθετη και χρονοβόρα διαδικασία από αυτή του ενός Νευρώνα και σκοπό έχει την ελαχιστοποίηση της *συνάρτησης σφάλματος*.

Νευρωνικά Δίκτυα ως αλγόριθμος κατηγοριοποίησης σε Φίλτρο Ανεπιθύμητης Αλληλογραφίας χρησιμοποιείται αρκετά συχνά όπως στα [2, 11] με ικανοποιητικά αποτελέσματα. Από την άλλη οι D. Puniškis, R. Laurutis και R. Dirmeikis στο [12] αναφέρουν πως δεν συνιστούν την αποκλειστική χρήση Νευρωνικών Δικτύων για αντιμετώπιση Ανεπιθύμητης Αλληλογραφίας λόγω του χαμηλού αλλά μη μηδενικού ποσοστού *False Positive*. Βέβαια οι T. Kathirvalavakumar, K. Kavitha και R. Palaniappan στο [6] που προσπαθούν να αναγνωρίσουν email κλοπής ταυτότητας (phising) και χρησιμοποιούν κάποια ειδικά χαρακτηριστικά για την αναπαράστασή των μηνυμάτων αναφέρουν μηδενικό ποσοστό *False Positive*.



Σχήμα 5: Βασική δομή ενός πολυεπίπεδου Νευρωνικού Δικτύου.

3. ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΠΟΥ ΥΛΟΠΟΙΕΙΤΑΙ

Στο πλαίσιο αυτής της εργασίας κατασκευάστηκε ένα Φίλτρο Ανεπιθύμητης Αλληλογραφίας. Ο αλγόριθμος που χρησιμοποιήθηκε για την κατηγοριοποίηση των ηλεκτρονικών μηνυμάτων σε μία εκ των κατηγοριών *{ham, spam}* δηλαδή *{θεμιτά, αθέμιτα}* είναι τα Νευρωνικά Δίκτυα. Πιο συγκεκριμένα, χρησιμοποιήθηκε ο αλγόριθμος *Πολυεπίπεδου Δικτύου Αισθητήρων* (*Multilayer Perceptron / MLP*).

Ένα Πολυεπίπεδο Δίκτυο Αισθητήρων είναι ένα *προς τα εμπρός τροφοδοτούμενο* (*feedforward*) Νευρωνικό Δίκτυο. Το MLP αποτελείται από επίπεδα νευρώνων. Στην περίπτωση της αναγνώρισης ανεπιθύμητης αλληλογραφίας, που είναι ένα είδος κατηγοριοποίησης κειμένου, το επίπεδο *Εισόδου* αποτελείται από όσους Νευρώνες όσα και τα χαρακτηριστικά που χρησιμοποιούνται για την αναπαράσταση των μηνυμάτων. Το επίπεδο *Εξόδου* αποτελείται από δύο κόμβους που ο καθένας αντιπροσωπεύει μία κλάση, που στην συγκεκριμένη περίπτωση είναι μία εκ των *{ham, spam}*. Ενδιάμεσα υπάρχουν ένα η περισσότερα κρυφά επίπεδα Νευρώνων. Για την εκπαίδευση του δικτύου χρησιμοποιείται ο αλγόριθμος της *προς τα πίσω διάδοσης* (*backpropagation*) μεταβάλλοντας τα βάρη έως ότου η *συνάρτηση σφάλματος* φτάσει σε κάποια προαποφασισμένη τιμή, ή μετά από κάποιο προαποφασισμένο πλήθος επαναλήψεων / εποχών (*epochs*).

Τα ηλεκτρονικά μηνύματα έχουν αναπαρασταθεί ως ένα διάνυσμα, διάστασης ίσης με το πλήθος των χαρακτηριστικών που επιλέχτηκαν. Οι τιμές που παίρνει κάθε συντεταγμένη x_i του διανύσματος αναπαράστασης είναι δυαδικές $\{0, 1\}$ και σηματοδοτούν την ύπαρξη ή την απουσία του χαρακτηριστικού t_i στο μήνυμα. Οι τιμές αυτές είναι και οι τιμές εισόδου στους κόμβους του Νευρωνικού Δικτύου.

Στη φάση της προεπεξεργασίας των μηνυμάτων εφαρμόστηκε αρχικά η απομόνωση του σώματος (Body) του κάθε μηνύματος από το οποίο και θα αποσπάσουμε όλη την πληροφορία που θα χρησιμοποιήσουμε. Στη συνέχεια τα μηνύματα πέρασαν από τη διαδικασία της Στοιχειοποίησης (Tokenization) προκειμένου να διαχωριστούν όλες οι λέξεις μέσα στο κείμενο σαν στοιχεία (tokens) και να κρατηθούν μόνο αυτά. Δηλαδή, τώρα κάθε μήνυμα αντικαθίσταται από ένα σύνολο από τις λέξεις που περιέχει.

Στη συνέχεια εφαρμόζεται η τεχνική της Λημματοποίησης (Lemmatization), δηλαδή κάθε λέξη που υπάρχει σε ένα μήνυμα μετατρέπεται στη ριζική της μορφή. Με αυτόν τον τρόπο όλες οι ομόρριζες λέξεις ταυτίζονται πλέον στον ίδιο όρο. Από αυτούς τους όρους αφαιρούνται εκείνοι οι οποίοι είναι πολύ κοινοί σχεδόν σε κάθε μήνυμα (Stop Word Removal) και δεν προσφέρουν καμία ουσιαστική πληροφορία για την κατηγορία που ανήκει το κείμενο αυτό.

Για την επιλογή των πιο αντιπροσωπευτικών χαρακτηριστικών για κάθε μήνυμα ούτως ώστε να γίνει η αναπαράσταση χρησιμοποιούνται οι όροι με την μεγαλύτερη τιμή μιας μετρικής (*C-tf-idf / Corpus-tf-idf*). Η συγκεκριμένη μετρική αποτελεί μια παραλλαγή της *tf-idf* και πειραματικά έδειξε ότι έχει καλύτερα αποτελέσματα κατηγοριοποίησης από άλλες

που δοκιμάστηκαν. Τα πειραματικά αποτελέσματα είναι εμφανή στην αντίστοιχη ενότητα. Η *C-tf-idf* υπολογίζει μια μόνο τιμή για κάθε όρο σε αντίθεση με την *tf-idf* που υπολογίζει για κάθε όρο όσες τιμές όσο και το πλήθος των εγγράφων που εμφανίζεται ο όρος αυτός. Η μετρική αυτή χρησιμοποιεί το άθροισμα των *tf* (*term frequency*) συχνοτήτων του όρου σε όλα τα έγγραφα που εμφανίζεται. Συγκεκριμένα η τιμή ενός όρου t_i της μετρικής υπολογίζεται ως εξής:

$$Ctfidf_i = (\sum_{d \in D_i} tf_{i,d}) \cdot idf_i$$

Όπου D_i το σύνολο των εγγράφων που εμφανίζεται ο όρος t_i, και $tf_{i,d}$ και idf_i, όπως αυτά έχουν οριστεί και προηγουμένως.

Για να πραγματοποιηθεί αυτό, υπολογίζεται για κάθε μοναδικό όρο που υπάρχει στο σύνολο των μηνυμάτων η *C-tf-idf* τιμή του, κατατάσσονται με φθίνουσα σειρά και στη συνέχεια επιλέγονται οι πρώτοι *N* από αυτούς. Σαν χαρακτηριστικό μηνύματος έχουμε θεωρήσει την ύπαρξη ή μη αυτών των όρων στο μήνυμα, και έτσι δημιουργείται για κάθε μήνυμα το διάνυσμα αναπαράστασης του.

4. ΠΕΡΙΒΑΛΛΟΝ ΔΙΕΞΑΓΩΓΗΣ ΠΕΙΡΑΜΑΤΩΝ

Για την υλοποίηση του Φίλτρου Ανεπιθύμητης Αλληλογραφίας χρησιμοποιήθηκε η προγραμματιστική βιβλιοθήκη της πλατφόρμας Weka [22] από το πανεπιστήμιο του Waikato. Το περιβάλλον Weka προσφέρει, μεταξύ άλλων, μια συλλογή από αλγορίθμους μηχανικής μάθησης υλοποιημένους σε Java. Χρησιμοποιήθηκε λοιπόν ο αλγόριθμος κατηγοριοποίησης MultilayerPerceptron από το package weka.classifiers.functions. Ένα παράδειγμα δημιουργίας ενός αντικειμένου τύπου MultilayerPerceptron στην Java, και εισαγωγή τιμών σε κάποιες παραμέτρους του φαίνεται παρακάτω:

```
1 MultilayerPerceptron mlp = new MultilayerPerceptron();
2
3 mlp.setLearningRate(0.1);
4 mlp.setMomentum(0.2);
5 mlp.setTrainingTime(700);
6 mlp.setHiddenLayers("6");
```

Listing 1: Παράδειγμα αρχικοποίησης ενός MultilayerPerceptron σε Java

Για εκπαίδευση του MultilayerPerceptron διαβάζεται ένα .arff αρχείο μέσω του Loader που προσφέρει το Weka. Διαβάζονται τα δεδομένα εκπαίδευσης από το αρχείο και ορίζεται ως κλάση ($\{spam, ham\}$) το τελευταίο χαρακτηριστικό (data.numAttributes() - 1) κάθε δεδομένου / μηνύματος.

```
1 loader = new ArffLoader();
2 loader.setFile(new File(trainFile));
3 Instances data = loader.getDataSet();
4 data.setClassIndex(data.numAttributes() - 1);
```

Listing 2: Παράδειγμα διαβάσματος δεδομένων από .arff αρχείο

Για την παραγωγή του .arff αρχείου, αρχικά απομονώνεται το Σώμα (Body) από κάθε μήνυμα των αρχείων εκπαίδευσης. Για αυτό χρησιμοποιείται το package javax.mail που προσφέρει δυνατότητες για προγραμματιστική διαχείριση ηλεκτρονικών μηνυμάτων (email). Δημιουργούνται έτσι και αποθηκεύονται καινούρια αρχεία που περιέχουν μόνο το Σώμα κάθε μηνύματος.

Στη συνέχεια, γίνεται μια επεξεργασία πάνω στα αρχεία με το περιεχόμενο προκειμένου να πραγματοποιηθεί η τεχνική της Λημματοποίησης (Lemmatization). Για αυτό χρησιμοποιήθηκε το εργαλείο *morpha* [23]. Το εργαλείο αυτό μετατρέπει κάθε λέξη στη ριζική της μορφή. Η διαφορά του *morpha* με άλλα εργαλεία που πραγματοποιούν Λημματοποίηση είναι ότι το *morpha* αντιστοιχίζει κάθε λέξη του κειμένου σε κάποια υπάρχουσα λέξη, αντί να κάνει απλή αφαίρεση των καταλήξεων (Stemming). Για παράδειγμα οι λέξεις *having* και *have* θα αντιστοιχιστούν και οι δύο στον όρο *have* ενώ με άλλα εργαλεία θα αντιστοιχίζονταν στον όρο *hav*. Το προτέρημα αυτής της τεχνικής είναι ότι το αποτέλεσμα της Λημματοποίησης είναι πιο αναγνώσιμο και κατανοήσημο σε περίπτωση που κάποιος

θέλει να το περιεργαστεί. Δημιουργούνται λοιπόν αρχεία (.lemma) που περιέχουν μόνο τις ριζικές μορφές των λέξεων των αρχείων εκπαίδευσης.

```
for f in $FILES
do
    morpha -uc < $f > $f.lemma
done
```

Listing 3: bash script για τη Λημματοποίηση των αρχείων

Πάνω στα *.lemma* αρχεία πραγματοποιείται η αφαίρεση των πολύ κοινών λέξεων (Stop Word Removal). Το σύνολο με τις λέξεις της αγγλικής γλώσσας που χρησιμοποιήθηκε είναι το ακόλουθο:

{a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, i, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your}

Για τις λέξεις όρους που απομένουν υπολογίζεται η τιμή της μετρικής *C-tf-idf* που περιγράφηκε παραπάνω και κατατάσονται σε φθίνουσα σειρά με βάση αυτή. Ανάλογα με την ύπαρξη ή όχι των *N* πρώτων όρων αυτής της κατάταξης σε ένα μήνυμα, δημιουργείται ένα διάνυσμα με τις τιμές 0, 1 στη θέση κάθε όρου. Από αυτά τα διανύσματα δημιουργείται το .arff αρχείο προκειμένου να δοθεί σαν αρχείο εκπαίδευσης στο MultilayerPerceptron της βιβλιοθήκης *Weka*.

Για το 10-fold Cross Validation των πειραμάτων χρησιμοποιείται η crossValidateModel() όπως φαίνεται και παρακάτω:

```
Evaluation eval = new Evaluation(data);
eval.crossValidateModel(mlp, data, folds, new Random(1));
```

Listing 4: Παράδειγμα 10-fold Cross Validation

Στη συνέχεια αντλούνται τα αποτελέσματα από τον πίνακα σύγχυσης (confusion matrix) και γράφονται σε κάποιο log file, προκειμένου να γίνει η σύγκριση μεταξύ διάφορων εκτελέσεων.

5. ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Πραγματοποιήθηκε μια πειραματική ανάλυση προκειμένου να βρεθούν οι κατάλληλες τιμές όσον αφορά το πλήθος των χαρακτηριστικών που επιλέγονται για την αναπαράσταση, αλλά και τις παραμέτρους του Νευρωνικού Δικτύου.

Η συλλογή ηλεκτρονικών μηνυμάτων που χρησιμοποιήθηκε για την εκπαίδευση του Νευρωνικού Δικτύου είναι μια δημόσια συλλογή του *SpamAssasin* [21]. Από αυτή χρησιμοποιήθηκαν 8853 μηνύματα, εκ των οποίων 6453 είναι θεμιτά και 2400 αθέμιτα. Συγκεκριμένα χρησιμοποιήθηκαν τα αρχεία των καταλόγων *easy_ham, easy_ham_2, spam* και *spam_2*.

Για την αξιολόγηση της απόδοσης του Φίλτρου χρησιμοποιήθηκε *10-Fold Cross Validation*. Η τεχνική του *Cross Validation* χωρίζει τα δεδομένα σε 90% δεδομένα εκπαίδευσης και 10% δεδομένα επαλήθευσης. Με *10-Fold Cross Validation* εννοούμε ότι η λειτουργία αυτή γίνεται 10 φορές, αλλά την κάθε φορά επιλέγεται διαφορετικό τμήμα των δεδομένων για εκπαίδευση ή επαλήθευση από τις προηγούμενες. Στο τέλος σαν αποτέλεσμα της απόδοσης του Κατηγοριοποιητή που εξετάζεται, παίρνουμε τον μέσο όρο των αποτελεσμάτων των 10 γύρων (folds) επαλήθευσης.

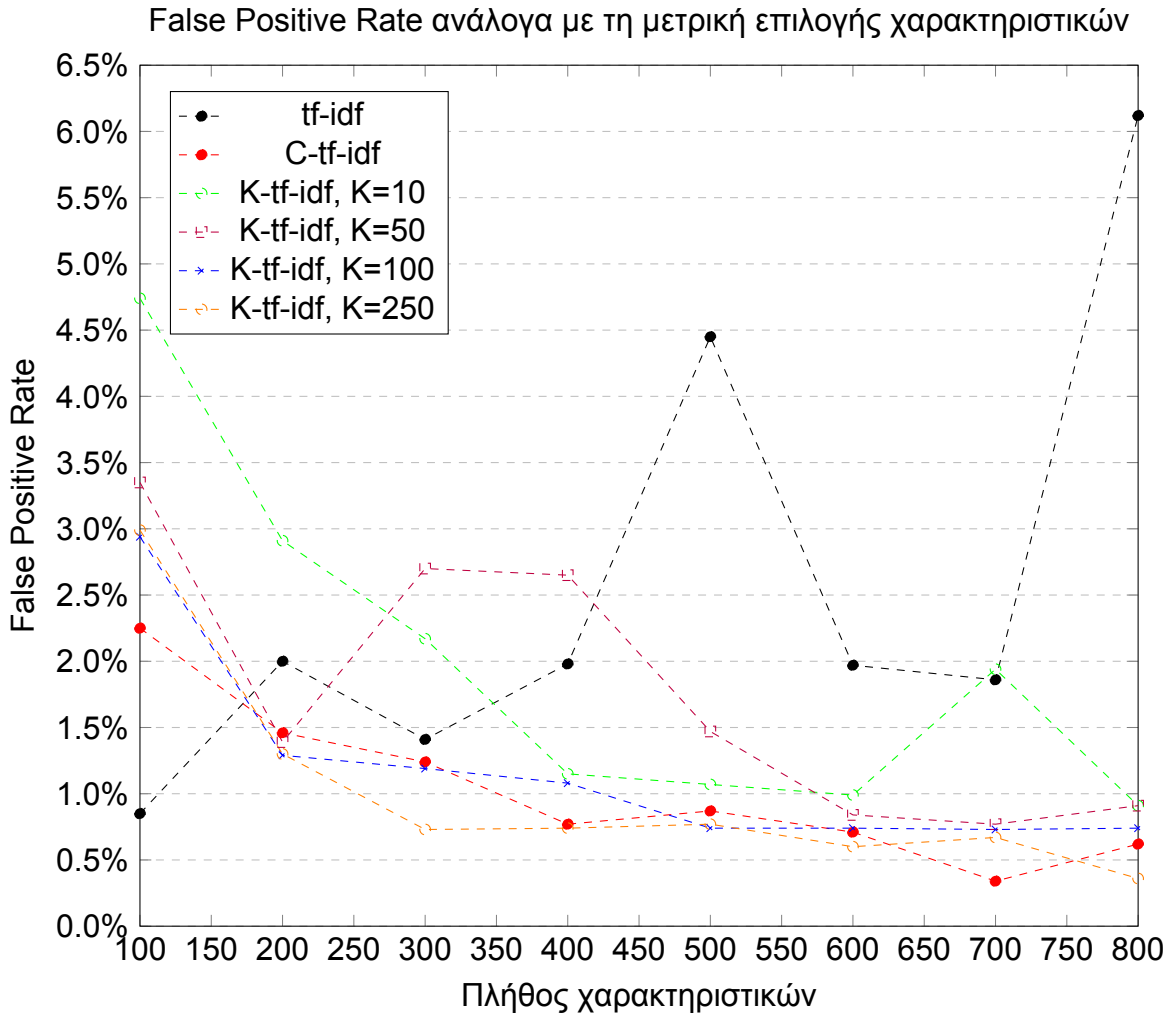
Σημαντικό είναι επίσης ο διαχωρισμός των δεδομένων σε δεδομένα εκπαίδευσης και επαλήθευσης να γίνεται με τέτοιο τρόπο ώστε να διατηρείται η αναλογία των μηνυμάτων που ανήκουν σε κάθε κλάση, σε όλους τους γύρους επαλήθευσης. Σαν μέτρο απόδοσης χρησιμοποιούνται τα ποσοστά των *True Positive*, *True Negative*, *False Positive*, *False Negative*, όπως αυα έχουν περιγραφεί προηγουμένως, αλλά και ένα γενικότερο μέτρο ακρίβειας $Acuracy = \frac{n_{spam \to spam} + n_{ham \to ham}}{N_{spam} + N_{ham}}$. Προτίστως δίνεται έμφαση στο μικρό ποσοστό *False Positive* και δευτερευόντως στο μεγάλο ποσοστό *Accuracy*.

Η σημαντικότητα του χαμηλού (ιδανικά μηδενικού) ποσοστού *False Positive* έχει αναφερθεί και παραπάνω. Πρωταρχικό μέλημα κατά την ανάπτυξη ενός φίλτρου Ανεπιθύμητης Αλληλογραφίας πρέπει να είναι να μην αποκλείονται θεμιτά μηνύματα στην προσπάθεια αντιμετώπισης των αθέμιτων. Το φραγμένο θεμιτό μήνυμα μπορεί να είναι μεγάλης σημαντικότητας και σε κάθε περίπτωση δεν είναι στην ευχέρεια του φίλτρου (ή του σχεδιαστή) να το κρίνει. Γι αυτό το λόγο προτεραιότητα έχει να κατηγοριοποιηθούν το λιγότερο δυνατόν θεμιτά μηνύματα ως αθέμιτα.

Προκειμένου να αποφασιστούν οι σημαντικότεροι όροι στη συλλογή δεδομένων που χρησιμοποιούμε για να γίνει η αναπαράσταση των μηνυμάτων με βάση αυτούς, δοκιμάστηκαν κάποιες μετρικές. Οι μετρικές αυτές δίνουν σε κάθε όρο μία τιμή που αντιστοιχεί στη σημαντικότητά του. Μία προσέγγιση ήταν να χρησιμοποιηθούν οι *N* όροι με τις μεγαλύτερες *tf-idf* τιμές. Μία δεύτερη προσέγγιση ήταν να χρησιμοποιηθούν οι *N* όροι με τους μεγαλύτερους μέσους όρους των *K* υψηλότερων *tf-idf* τιμών για αυτόν τον όρο (η τιμή του *K* είναι μεταβλητή και δοκιμάζεται). Τέλος μία τρίτη προσέγγιση ήταν να επιλεγούν οι *N* όροι με τη μεγαλύτερη τιμή της μετρικής *C-tf-idf* / *Corpus-tf-idf* όπως αυτή ορίστικε

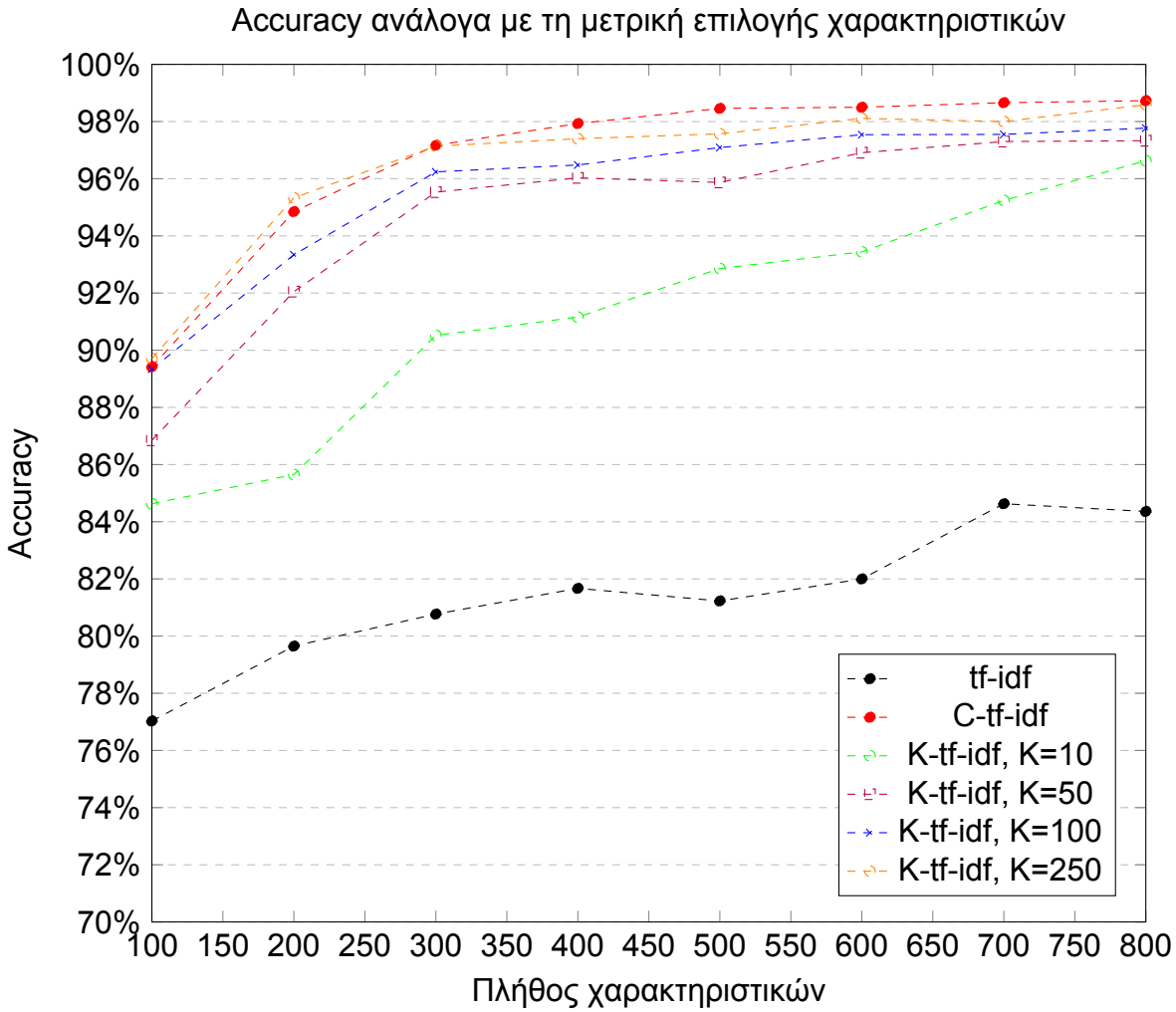
προηγουμένως.

Παρακάτω φαίνεται η απόδοση της κατηγοριοποίησης ανάλογα με τη μετρική που χρησιμοποιείται για την επιλογή των χαρακτηριστικών για διάφορες τιμές του πλήθους των χαρακτηριστικών. Σαν *tf-idf* παρουσιάζεται η πρώτη προσέγγιση που αναφέρθηκε, σαν *K-tf-idf* η δεύτερη προσέγγιση με διευκρίνιση για την τιμή του *K*, και ως *C-tf-idf* παρουσιάζεται η τρίτη προσέγγιση, με χρήση δηλαδή της μετρικής που αναφέρθηκε παραπάνω.



Σχήμα 6: False Positive Rate ανάλογα με τη μετρική επιλογής χαρακτηριστικών

Από το διάγραμμα του ποσοστού των *False Positive* παρατηρούμε ότι με χρήση της μετρικής *tf-idf* για την επιλογή των χαρακτηριστικών, υπάρχουν μεγάλα ποσοστά *False Positive Rate*, κάτι που πρέπει να αποφεύγεται στην Αναγνώριση Ανεπιθύμητης Αλληλογραφίας. Ακόμα βλέπουμε ότι στη γενική περίπτωση, η μετρική *K-tf-idf* έχει καλύτερα αποτελέσματα όσο μεγαλώνει η τιμή του *K*, και τέλος, η χαμηλότερη τιμή που παρατηρήθηκε ήταν με τη χρήση της μετρικής *C-tf-idf*. Ακολουθεί και το διάγραμμα της Ακρίβειας της κατηγοριοποίησης όπου η σύγκριση των μετρικών είναι πιο ξεκάθαρη.



Σχήμα 7: Ακρίβεια κατηγοριοποίησης ανάλογα με τη μετρική επιλογής χαρακτηριστικών

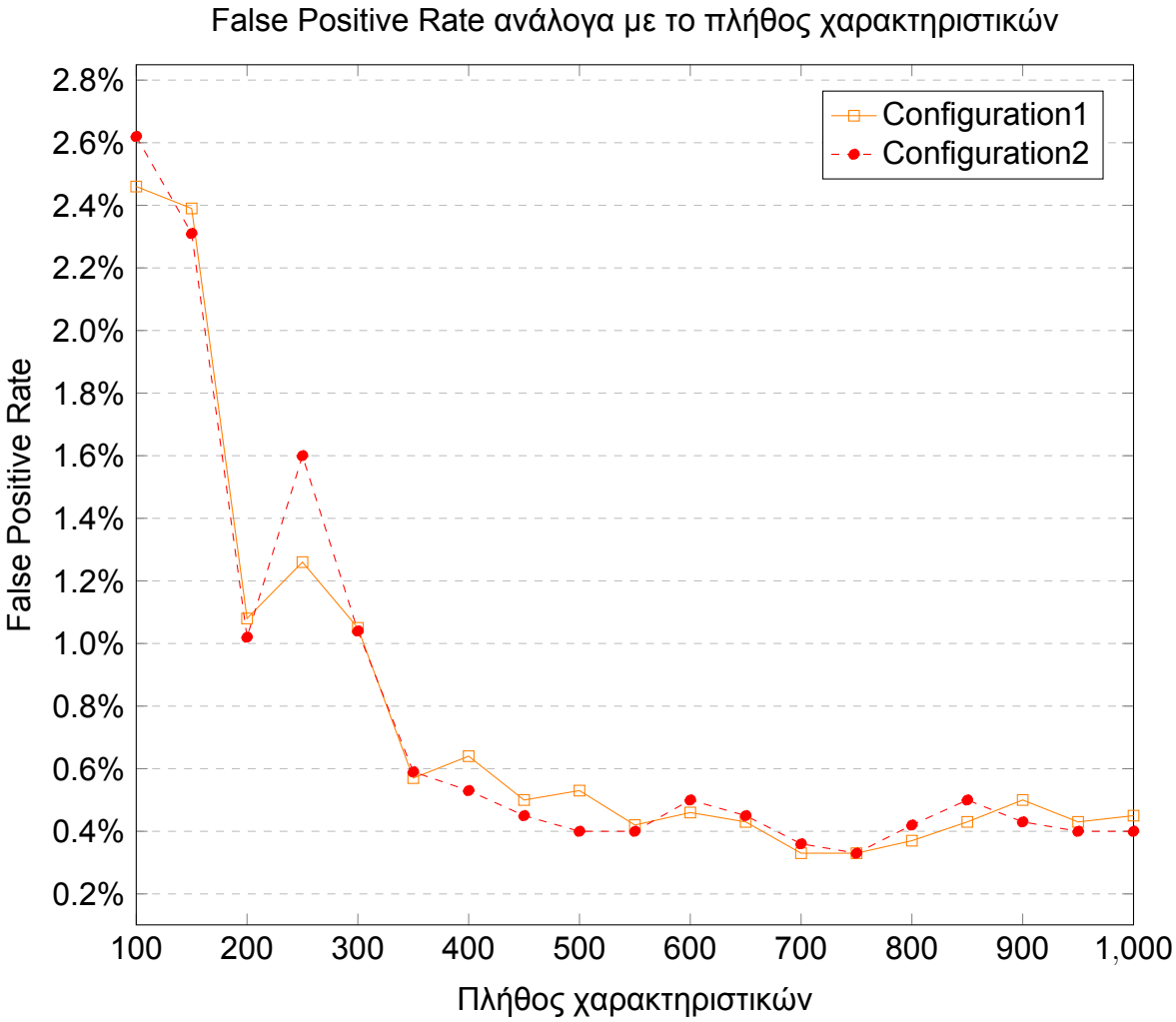
Στο διάγραμμα της Ακρίβειας της κατηγοριοποίησης παρατηρούμε ότι με τη χρήση της μετρικής *tf-idf* είχαμε το χαμηλότερο ποσοστό Ακρίβειας. Ακόμα, φαίνεται ξεκάθαρα ότι η μετρική *K-tf-idf* έχει συνεχώς καλύτερα αποτελέσματα όσο αυξάνεται η τιμή του *K*. Τέλος, βλέπουμε ότι με την χρήση της *C-tf-idf* για επιλογή χαρακτηριστικών, παίρνουμε σχεδόν σε όλες τις περιπτώσεις μεγαλύτερη Ακρίβεια κατηγοριοποίησης.

Αξίζει να σημειωθεί ότι με τη χρήση της μετρικής *K-tf-idf* για επιλογή χαρακτηριστικών, δεν μπορούν να ληφθούν υπ' όψιν όροι που εμφανίζονται σε λιγότερα από *K* έγγραφα (έχουν δηλαδή $df_i < K$). Έτσι για σχετικά μεγάλες τιμές του *K* υπάρχει περιορισμός στο πλήθος των χαρακτηριστικών που μπορεί να χρησιμοποιηθούν.

Λαμβάνοντας υπ' όψιν τα παραπάνω, επιλέγουμε για τη συνέχεια των πειραμάτων να χρησιμοποιήσουμε την μετρική *C-tf-idf* για την επιλογή των χαρακτηριστικών αναπαράστασης των μηνυμάτων / εγγράφων.

Στη συνέχεια έγιναν πειραματικές δοκιμές προκειμένου να αποφασιστεί το πλήθος των χαρακτηριστικών για την αναπαράσταση των μηνυμάτων που επιφέρει τα καλύτερα αποτελέσματα Με δύο αρχικές διαμορφώσεις του νευρωνικού δικτύου χρησιμοποιήθηκε διαφορετικό πλήθος χαρακτηριστικών για αναπαράσταση των μηνυμάτων και παρατηρή-

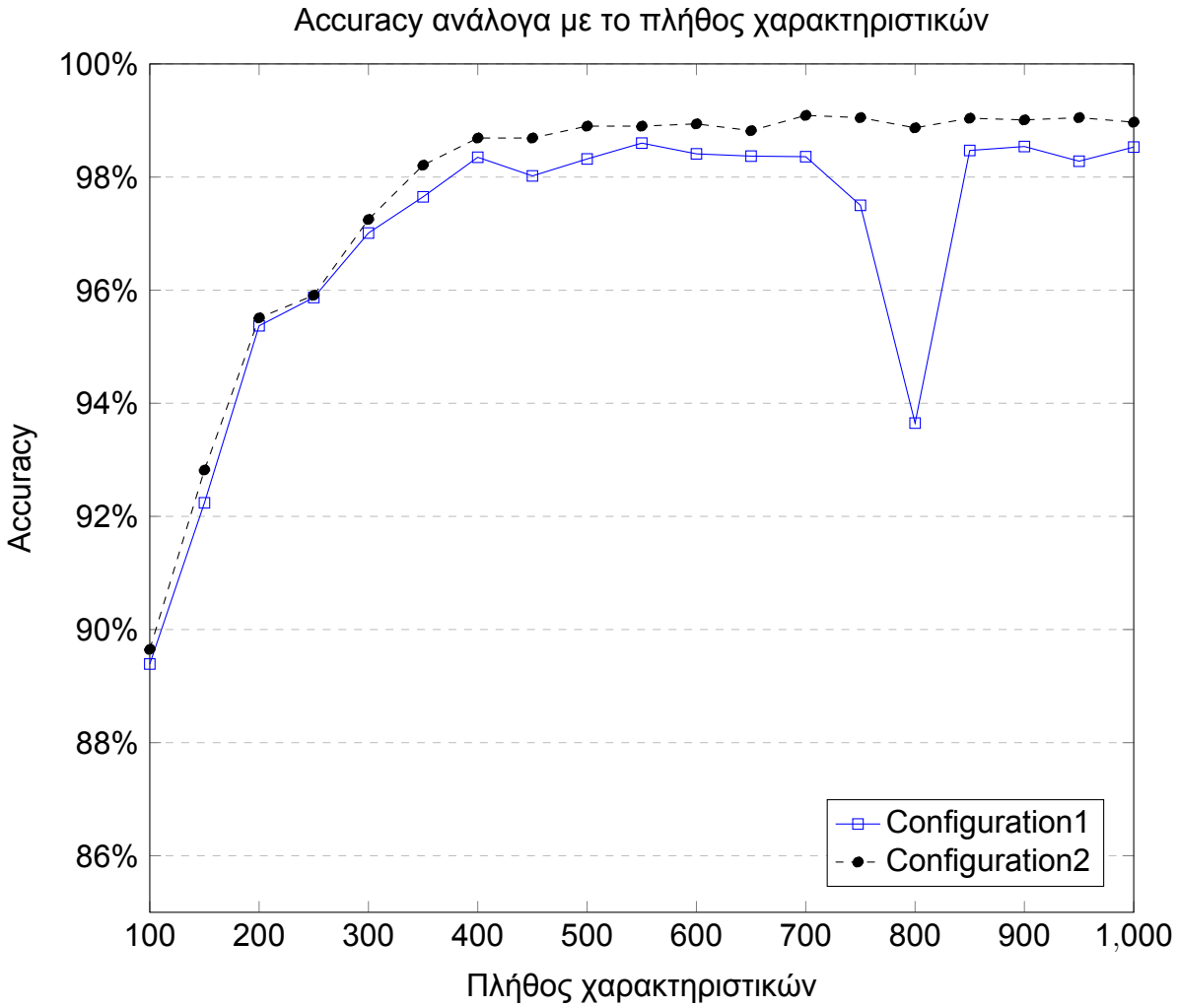
θηκαν τα μέτρα της απόδοσης της Κατηγοριοποίησης. Τα αποτελέσματα των δοκιμών αυτών φαίνονται παρακάτω.



Σχήμα 8: False Positive Rate ανάλογα με το πλήθος χαρακτηριστικών

Παρατηρούμε ότι για 700 και 750 χαρακτηριστικά, δηλαδή αναπαράσταση των μηνυμάτων με διανύσματα διάστασης 700 και 750 αντίστοιχα, έχουμε το χαμηλότερο ποσοστό *False Positive* και για τις δύο διαμορφώσεις του δικτύου. Συγκεκριμένα για την πρώτη αυτά είναι 0.33% ή 21 μηνύματα και 700 και για 750 χαρακτηριστικά ενώ για τη δεύτερη είναι 0.36% ή 23 μηνύματα για επιλογή 700 χαρακτηριστικών και 0.33% ή 21 μηνύματα για επιλογή 750 χαρακτηριστικών. Μεταξύ δύο επιλογών με ίδιο αποτέλεσμα γενικά καλύτερη φαίνεται η επιλογή των λιγότερων χαρακτηριστικών διότι οδηγεί σε διανύσματα αναπαράστασης μικρότερης διάστασης και κατ' επέκταση σε καλύτερη χρονική απόδοση του φίλτρου στη διάρκεια της εκπαίδευσης.

Παρακάτω βλέπουμε τα αντίστοιχα αποτελέσματα των πειραμάτων για την τιμή της γενικότερης ακρίβειας (*Accuracy*) του κατηγοριοποιητή και όπως φαίνεται η ακρίβεια με επιλογή 700 χαρακτηριστικών απέχει πολύ λίγο από την μέγιστη ακρίβεια που εμφανίζεται στην πρώτη διαμόρφωση του δικτύου, και έχει την μέγιστη τιμή για τη δεύτερη.



Σχήμα 9: Ακρίβεια ανάλογα με το πλήθος χαρακτηριστικών

Πιο συγκεκριμένα, με την πρώτη διαμόρφωση του δικτύου η μέγιστη τιμή για το *Accuracy* ήταν 98.6% για 550 χαρακτηριστικά και αυτή των 700 χαρακτηριστικών ήταν 98.36%, ενώ για 750 χαρακτηριστικά είναι 97.5%. Στη δεύτερη διαμόρφωση η ακρίβεια για τα 700 χαρακτηριστικά είναι η μέγιστη που εμφανίστηκε και είναι αυτή του 99.09% ενώ για 750 χαρακτηριστικά είναι 99.05%.

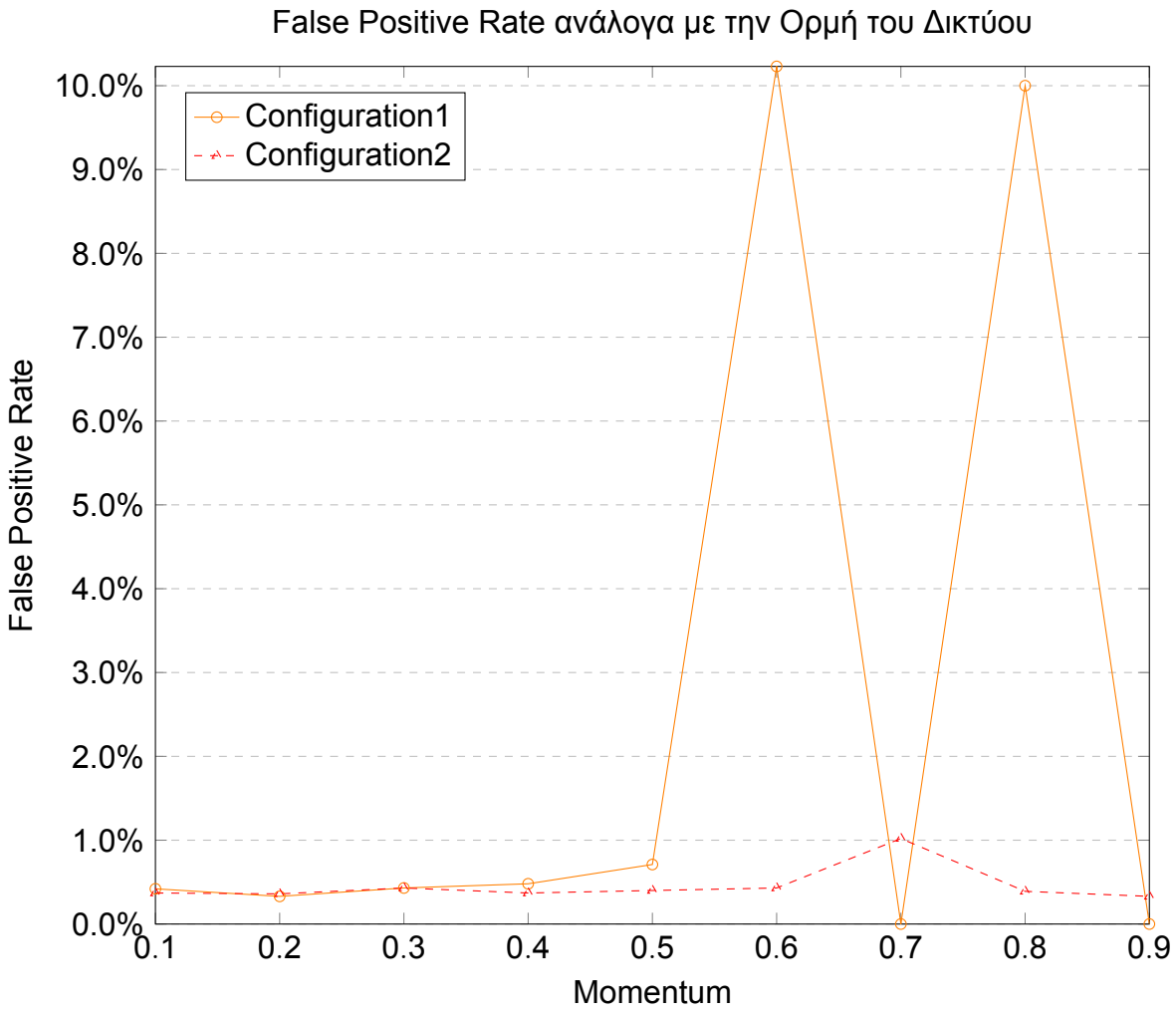
Για αυτό το λόγο αν συμπεριλάβουμε τα αποτελέσματα του ποσοστού των *False Positive* και αυτά της ακρίβειας καταλήγουμε ότι θα χρησιμοποιηθούν 700 χαρακτηριστικά για την αναπαράσταση των μηνυμάτων, δηλαδή επιλέγονται οι 700 όροι μη την μεγαλύτερη *tf-idf* τιμή.

Σε ένα Νευρωνικό Δίκτυο υπάρχουν ορισμένοι παράγοντες που παραμετροποιούν τον αλγόριθμο και συνεπώς και την απόδοσή του στην κατηγοριοποίηση των μηνυμάτων. Οι παράμετροι αυτοί είναι η *"Ορμή"* (*Momentum*) του δικτύου, ο *Ρυθμός Εκμάθησης* (*Learning Rate*), η *Αρχιτεκτονική* του, δηλαδή η διάταξη των επιπέδων των κόμβων (πλήθος κρυφών επιπέδων και κόμβων σε αυτά), αλλά και ο χρόνος εκπαίδευσης με τα κατηγοριοποιημένα δεδομένα. Τα προηγούμενα πειράματα έγιναν με δύο αρχικές διαμορφώσεις του Νευρωνικού Δικτύου με σταθερές τιμές σε αυτές τις παραμέτρους και μοναδική

μεταβολή στο πλήθος των χαρακτηριστικών.

Στη συνέχεια, χρησιμοποιώντας 700 χαρακτηριστικά για την αναπαράσταση των μνημάτων, εφαρμόζουμε αντίστοιχα πειράματα για να προσεγγίσουμε τις τιμές των παραμέτρων του Νευρωνικού Δικτύου που θα δώσουν τα καλύτερα αποτελέσματα στην κατηγοριοποίηση των μηνυμάτων.

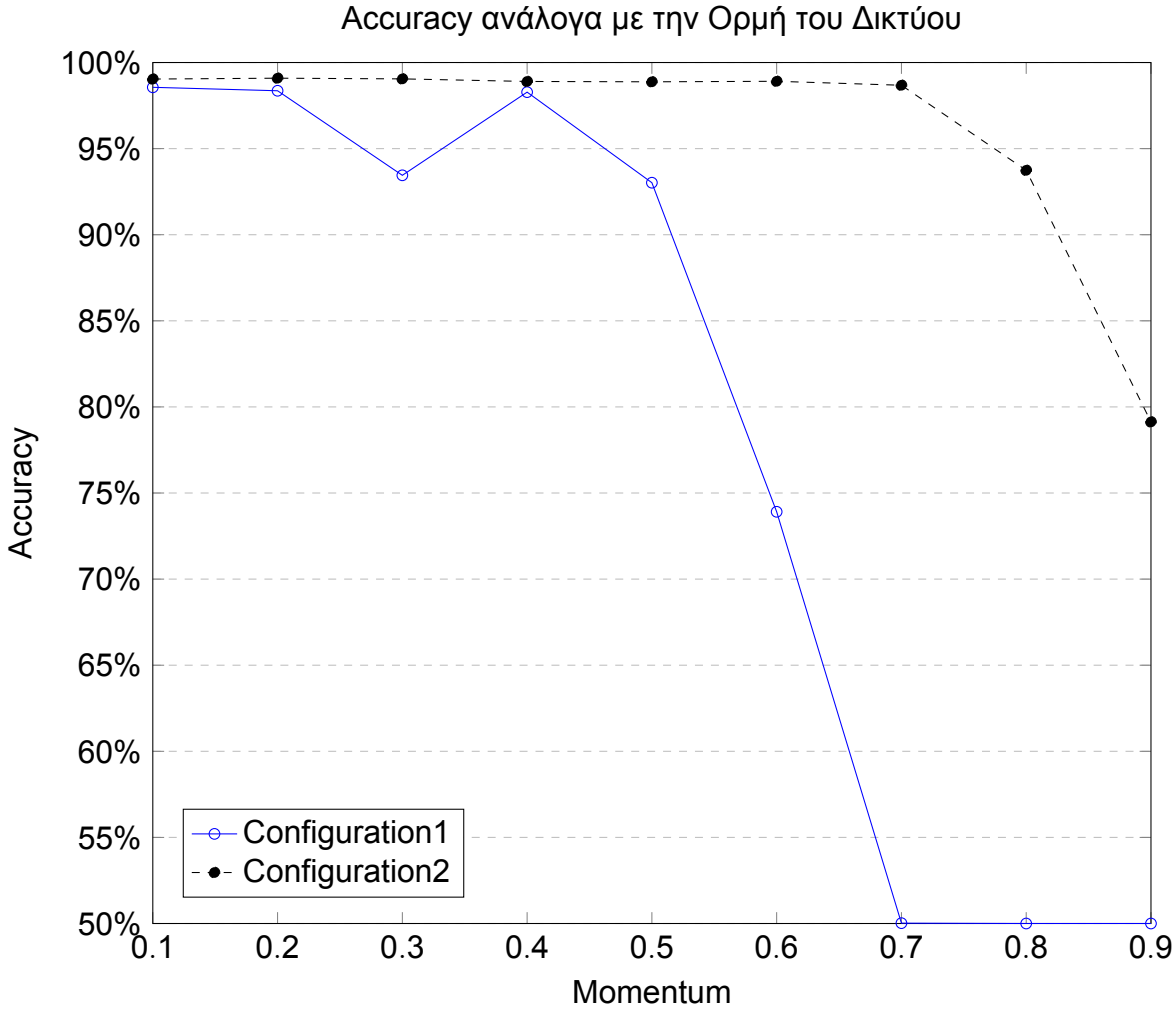
Αρχικά, έχουμε σταθερές τιμές σε όλες τις παραμέτρους εκτός της *"Ορμής"* (*Momentum*) την οποία και μεταβάλουμε. Χρησιμοποιούμε πάλι 2 διαμορφώσεις του δικτύου (που διαφέρουν στο *Ρυθμό Εκμάθησης*) και καταγράφουμε τα ποσοστά των *False Positive* και την ακρίβεια της κατηγοριοποίησης για τις δύο αυτές διαμορφώσεις. Χρησιμοποιούνται δύο διαμορφώσεις αντί για μία για να υπάρχει μία πιο καθαρή εικόνα των αποτελεσμάτων. Επίσης μια εξαντλητική αναζήτηση για κάθε τιμή κάθε παραμέτρου είναι πρακτικά σχεδόν αδύνατη με δίκτυο τέτοιων διαστάσεων.



Σχήμα 10: False Positive Rate ανάλογα με την Ορμή του Δικτύου

Από το γράφημα του ποσοστού των *False Positive* φαίνεται ότι στην πρώτη διαμόρφωση του δικτύου η κατηγοριοποίηση γίνεται πολύ ασταθής για τιμές της *Ορμής* μεγαλύτερες από 0.5. Μέχρι αυτό το σημείο όμως, το μικρότερο ποσοστό είναι 0.33 ή 21 μηνύματα και επιτυγχάνεται για *Ορμή* 0.2. Στη δεύτερη διαμόρφωση δεν είναι εμφανές

παρόμοιο φαινόμενο από την καμπύλη του ποσοστού των *False Positive* για μεγάλες τιμές της *Ορμής* και οι ελάχιστες τιμές του επιτυγχάνονται για *Ορμή* 0.2 και 0.9 και είναι 0.36 και 0.33 αντίστοιχα.



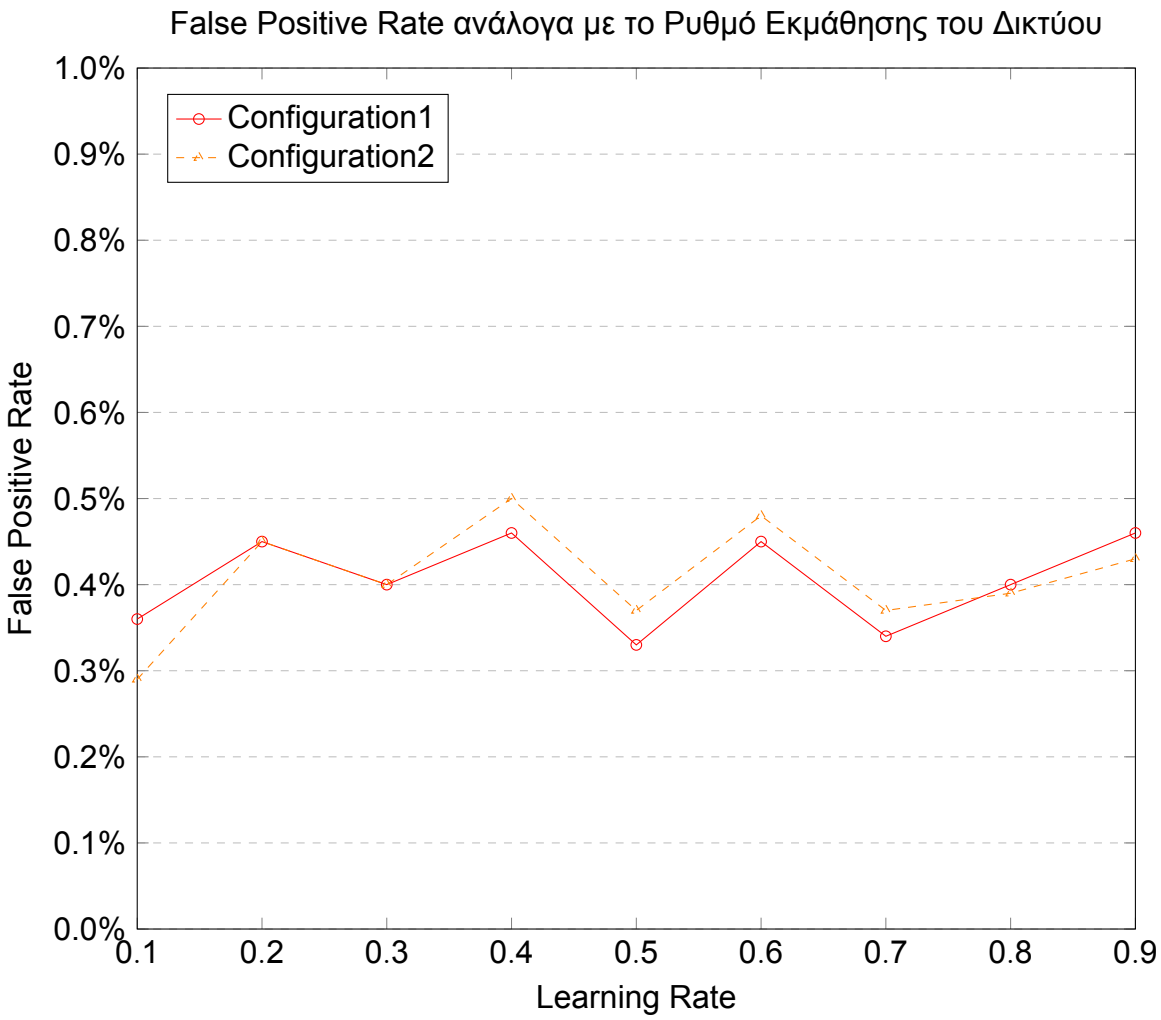
Σχήμα 11: Ακρίβεια ανάλογα με την Ορμή του Δικτύου

Τα πράγματα γίνονται πιο ξεκάθαρα όταν παρουσιάζεται η ακρίβεια της κατηγοριοποίησης. Εδώ βλέπουμε ότι και στις δύο διαμορφώσεις του δικτύου η ακρίβεια όντως φθίνει με την αύξηση της *Ορμής*. Αυτό συμβαίνει πιο νωρίς και με πολύ πιο γρήγορο τρόπο στην πρώτη διαμόρφωση αλλά πιο ήπια και στις δύο. Στην δεύτερη διαμόρφωση η μέγιστη ακρίβεια επιτυγχάνεται για *Ορμή* 0.2, που είναι και η δεύτερη μέγιστη τιμή της πρώτης διαμόρφωσης.

Επομένως βλέποντας τόσο τις τιμές του ποσοστού των *False Positive* όσο και τη γενικότερη ακρίβεια της κατηγοριοποίησης του φίλτρου, φαίνεται ότι οι τιμές της *Ορμής* που είχαν μικρό ποσοστό *False Positive*, είχαν στην πραγματικότητα πολύ μεγάλο ποσοστό *False Negative*, κάτι που ρίχνει κατακόρυφα την ακρίβεια. Έτσι, καταλήγουμε ότι 0.2 είναι μία κατάλληλη τιμή για την *Ορμή* του Νευρωνικού Δικτύου.

Με σταθερές τιμές στις άλλες παραμέτρους εκτός του *Ρυθμού Εκμάθησης* παρατηρούμε το ποσοστό των *False Positive* και την Ακρίβεια του δικτύου. Χρησιμοποιούνται

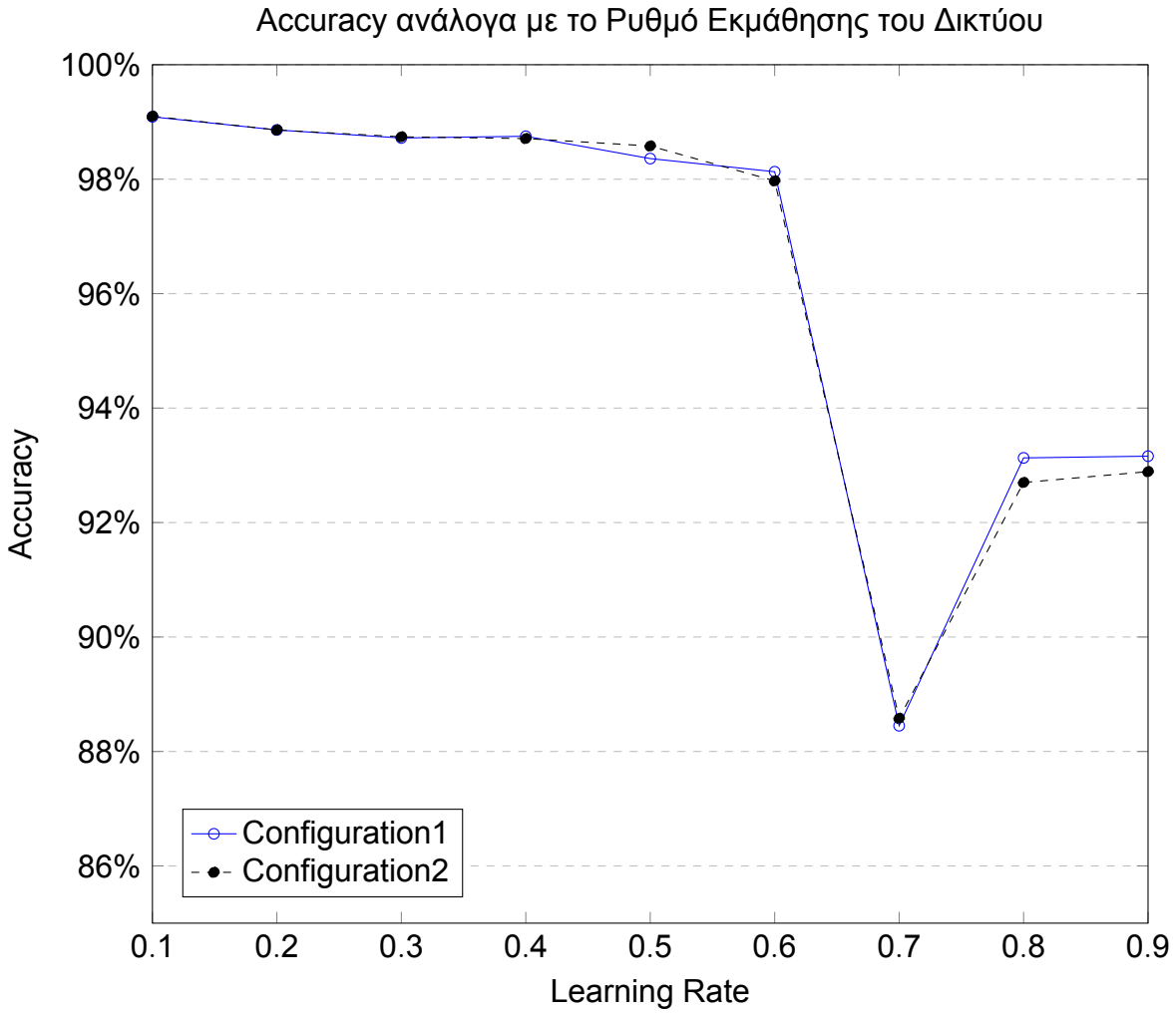
700 χαρακτηριστικά για την αναπαράσταση κάθε μηνύματος, και η ορμή του δικτύου έχει τεθεί σε 0.2. Οι δύο διαμορφώσεις (Configuration1, Configuration2) που δοκιμάζονται διαφέρουν στο χρόνο εκπαίδευσης του δικτύου.



Σχήμα 12: False Positive Rate ανάλογα με το Ρυθμό Εκμάθησης του Δικτύου

Βλέπουμε ότι το ποσοστό των *False Positive* κυμαίνεται μεταξύ 0.3% και 0.5% με την μικρότερη τιμή που παρατηρήθηκε να είναι 0.33% για *Ρυθμό Εκμάθησης* ίσο με 0.5 στην πρώτη διαμόρφωση. Στη δεύτερη διαμόρφωση παρατηρείται πολύ παρόμοια διακύμανση με μικρότερη βέβαια ελάχιστη τιμή, ίση με 0.29% για *Ρυθμό Εκμάθησης* ίσο με 0.1. Βέβαια από τις παραπάνω καμπύλες φαίνεται ότι το ποσοστό των *False Positive* δεν έχει σταθερή πορεία όσο αυξάνεται ο *Ρυθμός Εκμάθησης*. Επιπλέον εφόσον οι διαφορές μεταξύ των τοπικών ελαχίστων είναι μικρές δεν μπορούμε με σιγουριά να επιλέξουμε την καλύτερη τιμή του *Ρυθμού Εκμάθησης* καθώς τα αντίστοιχα ποσοστά μπορεί να μεταβληθούν λόγω άλλων παραγόντων, όπως ο χρόνος εκπαίδευσης ή η *Αρχιτεκτονική* του δικτύου.

Καλύτερη εικόνα παίρνουμε από το γράφημα της ακρίβειας της κατηγοριοποίησης, όπου και φαίνεται μια σταθερή (πτωτική) πορεία της ακρίβειας με την αύξηση του *Ρυθμού Εκμάθησης* του Νευρωνικού Δικτύου και για τις δύο διαμορφώσεις του.



Σχήμα 13: Ακρίβεια ανάλογα με το Ρυθμό Εκμάθησης του Δικτύου

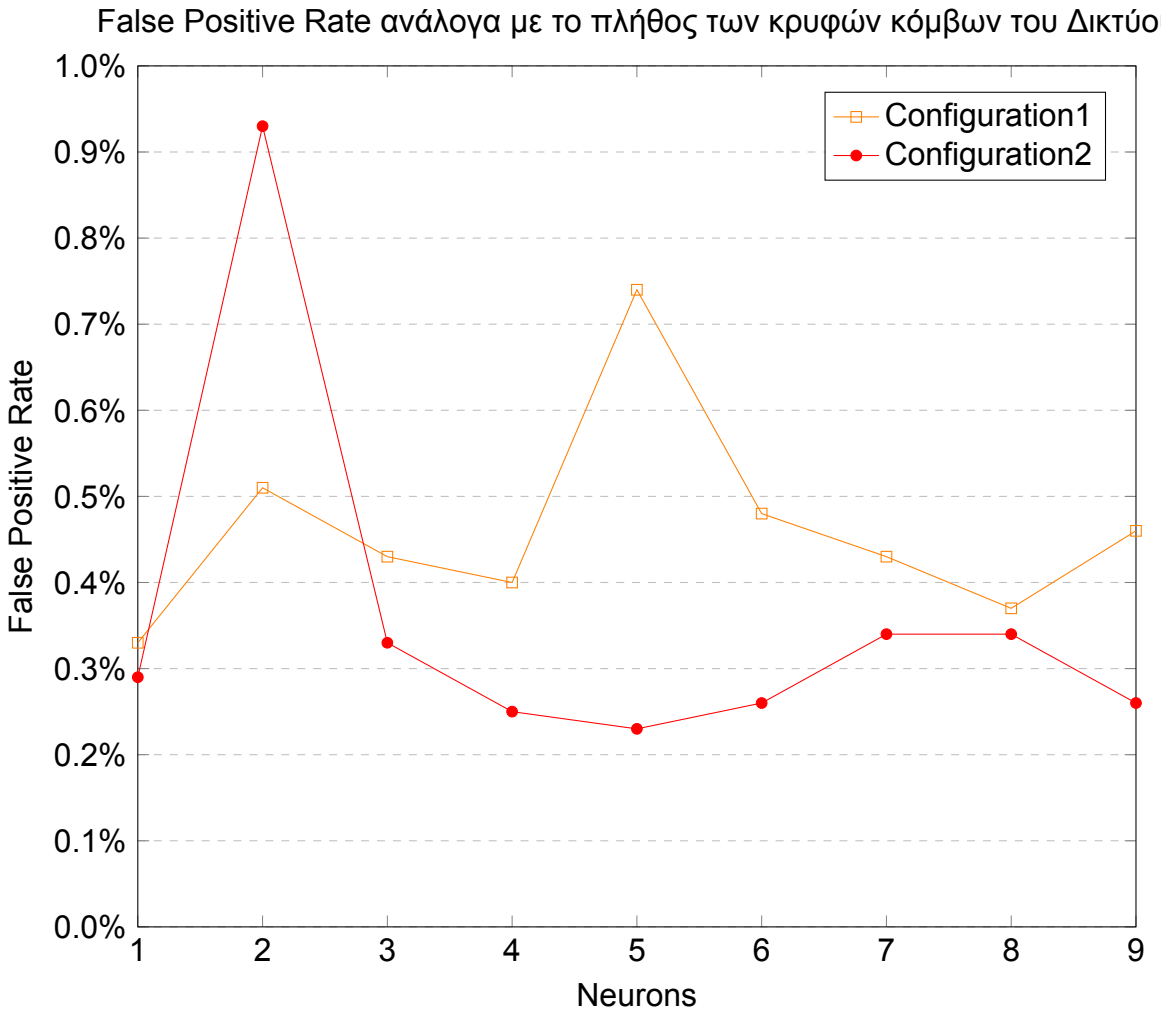
Βλέπουμε ότι η ακρίβεια του δικτύου είναι σχεδόν ίδια και για τις δύο διαμορφώσεις και πιο συγκεκριμένα φθίνει συνεχώς όσο αυξάνεται ο *Ρυθμός Εκμάθησης*. Μετά την τιμή του 0.6 μάλιστα έχει πολύ απότομη πτώση. Η μέγιστη τιμή της ακρίβειας παρατηρείται για *Ρυθμό Εκμάθησης* ίσο με 0.1. Για αυτήν την τιμή, το ποσοστό των *False Positive* που παρατηρήθηκε είναι 0.36% ή αλλιώς 23 μηνύματα, που είναι πολύ κοντά στο ελάχιστο του 0.33% ή 21 μηνύματα για πρώτη περίπτωση, και 0.29% ή 19 μηνύματα που ήταν και το ελάχιστο στη δεύτερη περίπτωση.

Έτσι φαίνεται ότι η επιλογή της τιμής 0.1 για *Ρυθμό Εκμάθησης* του δικτύου μπορεί να είναι καλύτερη από αυτή του 0.5 από άποψη επιδόσεων της τελικής κατηγοριοποίησης. Παρακάτω θα εξετάσουμε και τις δύο περιπτώσεις για να αποφασίσουμε για την καλύτερη επιλογή.

Μία ακόμα παράμετρος του Νευρωνικού Δικτύου είναι η *Αρχιτεκτονική* του. Μέχρι τώρα οι δοκιμές που έχουν γίνει αφορούν Νευρωνικό Δίκτυο με 1 κρυφό επίπεδο, αποτελούμενο από 1 κρυφό κόμβο / Νευρώνα. Στη συνέχεια έγιναν δοκιμές για διαφορετικό πλήθος Νευρώνων στο κρυφό επίπεδο του δικτύου. Το πλήθος των Νευρώνων που δοκιμάστηκε είναι σχετικά μικρό και συγκεκριμένα στο εύρος $[1, 9]$ καθώς ο χρόνος εκτέλεσης

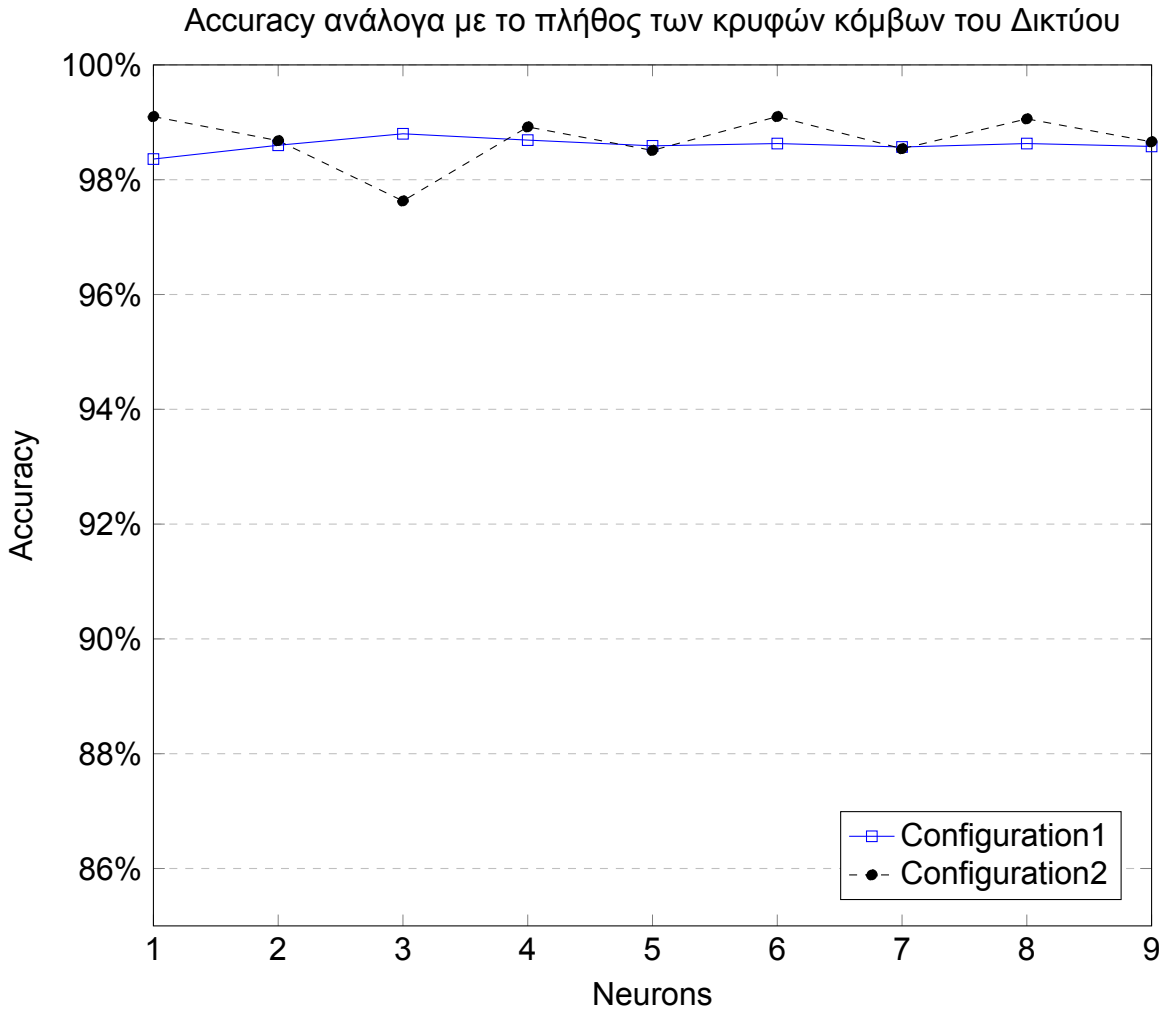
του *10-Fold Cross Validation* για τις διαστάσεις αναπαράστασης που έχουν χρησιμοποιηθεί αυξάνεται πάρα πολύ με την αύξηση των κρυφών κόμβων του δικτύου.

Στα διαγράμματα ου ακολουθούν παρουσιάζεται το ποσοστό των *False Positive* και η Ακρίβεια της ταξινόμησης του δικτύου με σταθερές παραμέτρους εκτός από το πλήθος των Νευρώνων στο κρυφό επίπεδο. Οι δύο διαμορφώσεις (Configuration1, Configuration2) που ακολουθούν διαφέρουν στο *Ρυθμό Εκμάθησης* και έχουν τις τιμές 0.5 και 0.1 αντίστοιχα.



Σχήμα 14: False Positive Rate ανάλογα με το πλήθος των κρυφών κόμβων του Δικτύου

Βλέπουμε πως σχεδόν για κάθε πλήθος κόμβων στο κρυφό επίπεδο του δικτύου χαμηλότερο ποσοστό *False Positive* έχει η δεύτερη διαμόρφωση (Configuration2). Πιο συγκεκριμένα, το χαμηλότερο ποσοστό που παρατηρείται είναι 0.23% ή 15 μηνύματα για 5 Νευρώνες στο κρυφό επίπεδο.

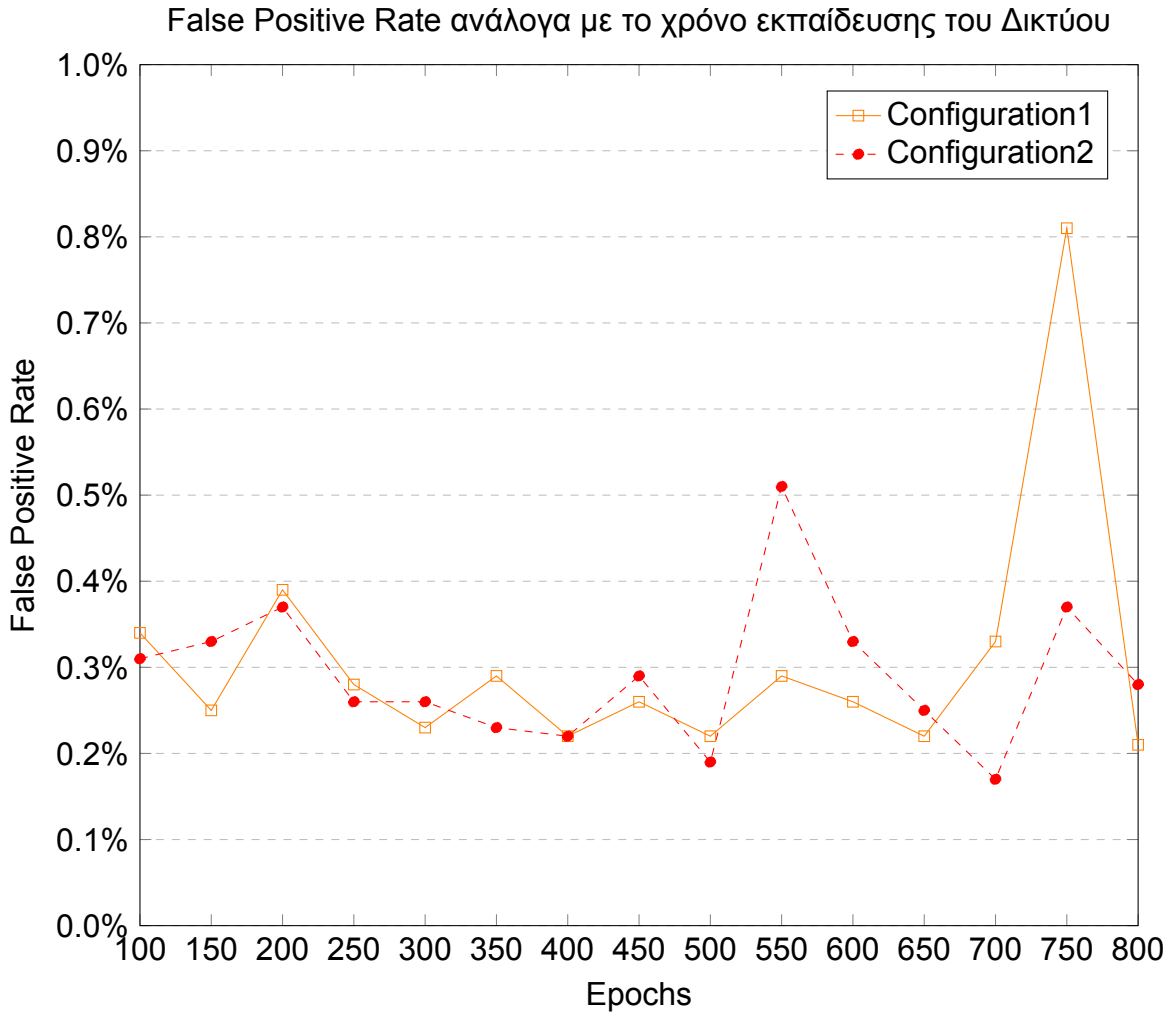


Σχήμα 15: Ακρίβεια ανάλογα με το πλήθος των κρυφών κόμβων του Δικτύου

Από την καμπύλη της ακρίβειας βλέπουμε μικρές διαφορές ανάλογα με το πλήθος των κόμβων στο κρυφό επίπεδο του δικτύου για τη δεύτερη διαμόρφωση, και ακόμα μικρότερες (σχεδόν σταθερή Ακρίβεια) για την πρώτη διαμόρφωση. Η μεγαλύτερη τιμή που παρατηρήθηκε είναι 99.1% και επιτεύχθηκε για 6 Νευρώνες στο κρυφό επίπεδο. Η αντίστοιχη τιμή για 5 Νευρώνες είναι ελαφρώς μικρότερη και ίση με 98.51%

Επομένως, 5 κόμβοι στο κρυφό επίπεδο του Νευρωνικού Δικτύου έδωσαν το μικρότερο ποσοστό *False Positive* ενώ 6 κόμβοι έδωσαν τη μεγαλύτερη ακρίβεια. Γι' αυτό το λόγο στις επόμενες μετρήσεις που αφορούν το χρόνο εκπαίδευσης του δικτύου, θα χρησιμοποιηθούν μία διαμόρφωση με 5 κρυφούς κόμβους και μία με 6, στο μοναδικό κρυφό επίπεδο, προκειμένου να αποφανθούμε την καταλληλότερη.

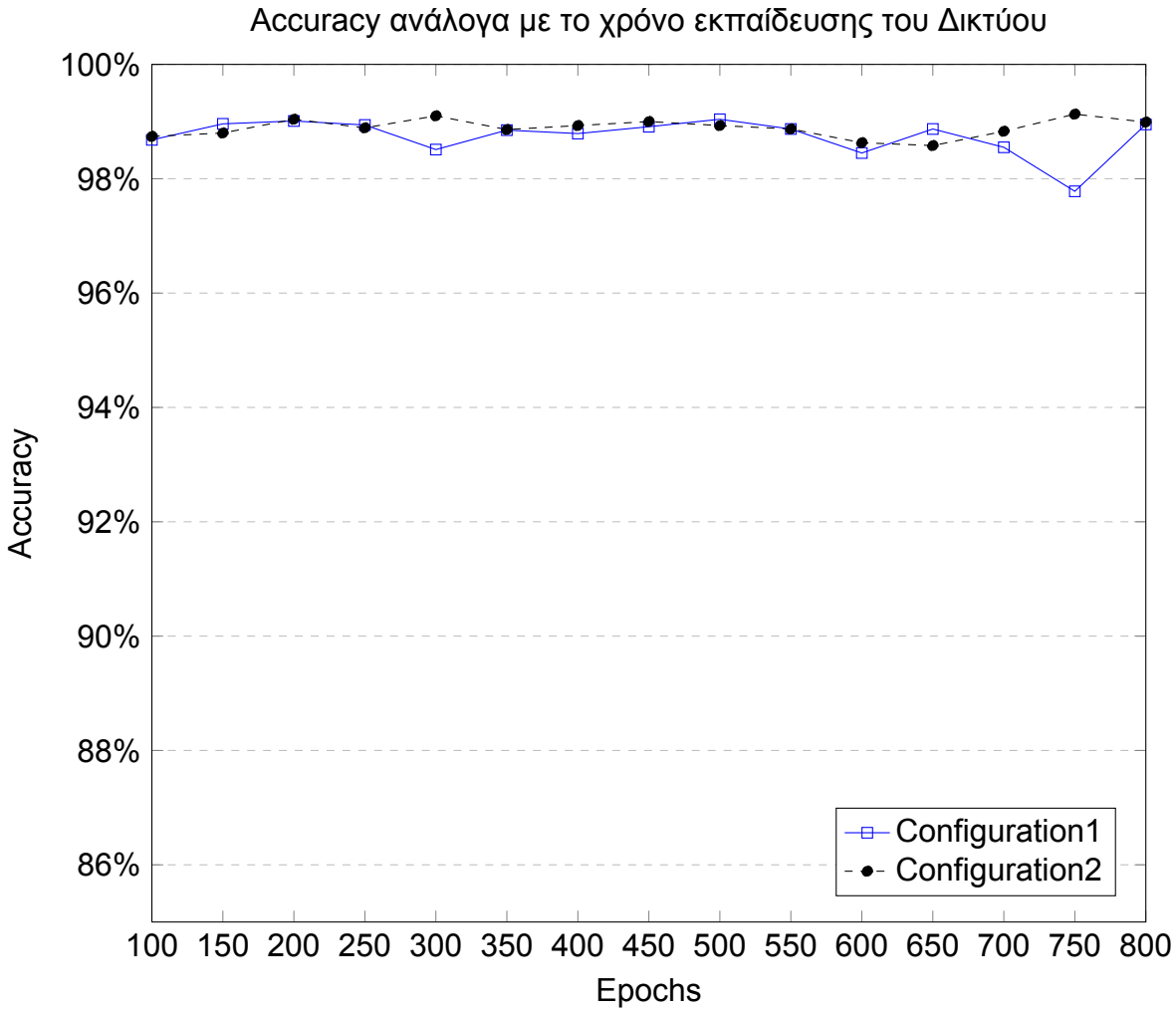
Παρακάτω φαίνονται τα αποτελέσματα της κατηγοριοποίησης του Νευρωνικού Δικτύου με σταθερές τιμές στις παραμέτρους που αναφέρθηκαν προηγουμένως και διαφορετικό πλήθος *εποχών* εκπαίδευσης. Σαν Configuration1 εννοείται η διαμόρφωση με 5 Νευρώνες και σαν Configuration2 η διαμόρφωση με 6 Νευρώνες στο κρυφό δίκτυο.



Σχήμα 16: False Positive Rate ανάλογα με το χρόνο εκπαίδευσης του Δικτύου

Από την καμπύλη του ποσοστού των *False Positive* ανάλογα με το χρόνο εκπαίδευσης του δικτύου βλέπουμε ότι οι δύο διαμορφώσεις δεν διαφέρουν πάρα πολύ στις περισσότερες των περιπτώσεων, και το μικρότερο ποσοστό που παρατηρείται ανήκει στην (κόκκινη) καμπύλη της δεύτερης διαμόρφωσης του δικτύου (Configuration2) αυτή δηλαδή που έχει 6 Νευρώνες στο κρυφό επίπεδο. Το ελάχιστο αυτό ποσοστό είναι ίσο με 0.17% ή αλλιώς 11 μηνύματα και επιτυγχάνεται για χρόνο εκπαίδευσης του δικτύου ίσο με 700 *εποχές*.

Στο γράφημα της Ακρίβειας της ταξινόμησης που φαίνεται παρακάτω, τα αποτελέσματα μοιάζουν ακόμα περισσότερο. Παρόλα αυτά και σε αυτήν την περίπτωση η μεγαλύτερη τιμή για την Ακρίβεια παρατηρείται για τη δεύτερη διαμόρφωση του δικτύου (Configuration2). Η μέγιστη αυτή τιμή είναι ίση με 99.13% και επιτυγχάνεται για 750 *εποχές* εκπαίδευσης του δικτύου. Η Ακρίβεια που επιτυγχάνεται για εκπαίδευση 700 *εποχών* που δίνουν το μικρότερο ποσοστό *False Positive* είναι ίση με 98.83% που δεν απέχει πολύ από τη μέγιστη των 750 *εποχών*.



Σχήμα 17: Ακρίβεια ανάλογα με το χρόνο εκπαίδευσης του Δικτύου

Σύμφωνα με τις παραπάνω μετρήσεις λοιπόν, η διαμόρφωση του Νευρωνικού Δικτύου που πετυχαίνει τα καλύτερα αποτελέσματα ταξινόμησης από αυτές που δοκιμάστηκαν, φαίνεται να είναι το δίκτυο με 1 κρυφό επίπεδο αποτελούμενο από 6 Νευρώνες, *Ρυθμό Εκμάθησης* ίσο με 0.1 και *Ορμή* ίση με 0.2.

Εφόσον δίνουμε πρωταρχική σημασία στο χαμηλό ποσοστό *False Positive* επιλέγουμε ως καλύτερη επιλογή την εκπαίδευση του Δικτύου για 700 *εποχές* αντί για 750, θυσιάζοντας ένα μικρό ποσοστό της συνολικής ακρίβειας της κατηγοριοποίησης. Τα αποτελέσματα του *10-Fold Cross Validation* για τη συλλογή δεδομένων του *Spam Assasin* φαίνονται παρακάτω:

Σωστά κατηγοριοποιημένα μηνύματα:	8790 ή 99.29%
Λάθος κατηγοριοποιημένα μηνύματα:	63 ή 0.71%
True Positive:	2348 ή 97.83%
True Negative:	6442 ή 99.83%
False Positive:	11 ή 0.17%
False Negative:	52 ή 2.17%
Ακρίβεια:	98.83%

6. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Από την παρούσα εργασία είναι φανερή η σημαντικότητα ενός Φίλτρου Ανεπιθύμητης Αλληλογραφίας. Επίσης, είναι φανερή η σημαντικότητα της προφύλαξης των θεμιτών μηνυμάτων και κατ' επέκταση η ανάγκη για χαμηλά ποσοστά *False Positive*. Αφού όμως τα ποσοστά αυτά ρεαλιστικά δεν μπορεί να είναι μηδενικά, θα πρέπει πρώτον να είναι όσο το δυνατόν πιο κοντά στο μηδέν και δεύτερον να μην διαγράφεται τελείως ένα μήνυμα αν κατηγοριοποιηθεί ως ανεπιθύμητο. Αντί αυτού θα πρέπει να φέρει την αντίστοιχη ένδειξη ή να περιέχεται σε κατάλληλο "φάκελο" στο γραμματοκιβώτιο του χρήστη προκειμένου αυτός να είναι σε θέση να το αξιολογήσει σε περίπτωση που το θέλει.

Σαν μελλοντική επέκταση θα είχε ενδιαφέρον η ανάδραση (feedback) ενός χρήστη του Φίλτρου Ανεπιθύμητης Αλληλογραφίας για τυχόν σφάλματα στην ταξινόμηση των εισερχόμενών του. Με αυτόν τον τρόπο θα μπορούσε το Φίλτρο να ξαναεκπαιδευτεί με τα νέα δεδομένα που θα παρείχε ο χρήστης. Κάτι τέτοιο θα είχε σαν αποτέλεσμα ένα Φίλτρο που θα ικανοποιούσε τις ανάγκες κάθε χρήστη ξεχωριστά, και θα είχε για τον καθένα καλύτερη απόδοση.

Ακόμα, όπως αναφέρεται και στο [19] θα είχε ενδιαφέρον ο συνδυασμός δύο ταξινομητών για την τελική κατηγοριοποίηση ενός μηνύματος. Συγκεκριμένα, με την προϋπόθεση ότι έχουμε δύο ταξινομητές με **πολύ χαμηλό** ποσοστό *False Positive* εκπαιδευμένους στα ίδια δεδομένα, θα κατηγοριοποιούμε ένα μήνυμα ως ανεπιθύμητο όταν τουλάχιστον ένας από τους δύο το κατηγοριοποιήσει σαν τέτοιο. Μόνο στην περίπτωση που και οι δύο το κατηγοριοποιήσουν ως θεμιτό, διαλέγουμε αυτό το αποτέλεσμα.

Εκ πρώτης όψης φαίνεται πως κάτι τέτοιο θα αυξήσει το ποσοστό των *False Positive* του τελικού κατηγοριοποιητή, παρόλα αυτά, έχουμε θεωρήσει ότι και οι δύο ταξινομητές έχουν πολύ χαμηλό ποσοστό *False Positive*. Έτσι, στην περίπτωση που κατατάξουν ένα μήνυμα σε διαφορετικές κλάσεις, δηλαδή ο ένας έχει κατηγοριοποιήσει το μήνυμα ως θεμιτό και ο άλλος ως αθέμιτο, είναι πιο πιθανό να έχει κάνει λάθος ο πρώτος. Έτσι και ο τελικός ταξινομητής (ένωση των δύο) θα έχει και αυτός χαμηλό ποσοστό *False Positive*. Ακόμα είναι δυνατόν να συνδυαστούν και τρεις ταξινομητές, με τον τρίτο να αποφασίζει το δίλημμα στην περίπτωση που οι άλλοι δύο ταξινομούν σε διαφορετικές κλάσεις.

Τέλος, αξίζει να σημειωθεί πως αν και ευρέως χρησιμοποιούμενη, η BoW (Bag of Words) αναπαράσταση των μηνυμάτων, δεν ευνοεί την ανανέωση του μοντέλου. Αν καινούριοι όροι πρέπει να ληφθούν υπ όψιν ή κάποιοι που λαμβάνονται ήδη σταμάτησαν να δίνουν χρήσιμη πληροφορία για τα μηνύματα, τότε πρέπει ο ταξινομητής να εκπαιδευτεί από την αρχή, κάτι που δεν είναι ιδιαίτερα αποδοτικό. Θα είχε ενδιαφέρον μία μέθοδος αναπαράστασης που να επιτρέπει την προσθήκη η αφαίρεση χαρακτηριστικών.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

ηλεκτρονικό ταχυδρομείο	email
ανεπιθύμητο μήνυμα	spam
θεμιτό μήνυμα	ham
δούρειος ίππος	trojan horse
Bayesian κατηγοριοποίηση	Bayesian classification
Μηχανές Διανύσματος Υποστήριξης	Support Vector Machines
k-Πλησιέστεροι Γείτονες	k-Nearest Neighbors
Δέντρα Απόφασης	Decision Trees
επιλογή χαρακτηριστικών	feature selection
θύρα	port
Συλλογές Δεδομένων	Datasets
πίνακας όρων-εγγράφων	term by document matrix
συλλογές email	email corpora
υπερεπίπεδο	hyperplane

ΣΥΝΤΜΗΣΕΙΣ, ΑΡΚΤΙΚΟΛΕΞΑ ΚΑΙ ΑΚΡΩΝΥΜΙΑ

URL	Uniform Resource Locator
IP	Internet Protocol
DNS	Domain Name System
NN	Neural Networks
ANN	Artificial Neural Networks
SVM	Support Vector Machines
k-NN	k-Nearest Neighbors
tf	term frequency
idf	inverse document frequency
tf-idf	term frequency-inverse document frequency
SVD	Singular Value Decomposition
FP	False Positive
TP	True Positive
FN	False Negative
TN	True Negative
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve

ΑΝΑΦΟΡΕΣ

[1] "Spam Statistics and Facts" [Online]. Available: http://www.spamlaws.com/spam-stats.html.

[2] A. Nossier, K. Nagati, I. Taj-eddi "Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks" *IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013.*,

[3] D. Arora, G. Rehani "A Study of Various Spam Filtering Techinques" *International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 2, Issue 1, March 2015, pp. 135-137.*

[4] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector *(Directive on privacy and electronic communications).* [Online]. Availabe: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:en:HTML.

[5] T. S. Guzella, W. M. Caminhas "A Review Of machine learning approaches to Spam filtering" *Expert Systems with Applications 36 (2009) pp. 10206–10222.*

[6] T. Kathirvalavakumar, K. Kavitha, R. Palaniappan "Efficient Harmful Email identification Using Neural Network" *British Journal of Mathematics & Computer Science 7(1): pp. 58-67, 2015, Article no.BJMCS.2015.102.*

[7] Z. Ma, R. Yan, D. Yuan, L. Liu "An Imbalanced Spam Mail Filtering Method" *International Journal of Multimedia and Ubiquitous Engineering Vol. 10, No. 3 (2015), pp. 119-126.*

[8] V. Bijalwan, V. Kumar, P. Kumari, J. Pascual "KNN based Machine Learning Approach for Text and Document Mining" *International Journal of Database Theory and Application Vol.7, No.1 (2014), pp. 61-70.*

[9] T. A. Almeida, A. Yakami "Facing the spammers: A very effective approach to avoid junk emails" *Expert Systems with Applications 39 (2012) pp. 6557-6561.*

[10] J. M. M da Cruz, G. V. Cormack "Using old Spam and Ham Samples to Train Email Filters" *CEAS 2009 Sixth Conference on Email and Anti-Spam.*

[11] B. Yu, D. Zhu "Combining neural networks and semantic feature space for email classification" *Expert Systems with Applications 22 (2009) pp. 376-381.*

[12] D. Puniškis, R. Laurutis, R. Dirmeikis "An Artificial Neural Nets for Spam email Recognition" *ELEKTRONIKA IR ELEKTROTECHNIKA 2006. Nr. 5(69).*

[13] "CEAS 2008 Live Spam Challenge Corpus" *CEAS 2008 Spam Filter Challenge.* August 5th – August 8th, 2008 [Online]. Availabe: http://plg.uwaterloo.ca/~gvcormac/ceascorpus/.

[14] "TREC Public Spam Corpus" *TREC Spam Evaluation Track.* [Online]. Availabe: 2005: http://plg.uwaterloo.ca/~gvcormac/treccorpus/about.html
2006: http://plg.uwaterloo.ca/~gvcormac/treccorpus06/
2007: http://plg.uwaterloo.ca/~gvcormac/treccorpus07/.

[15] "Reuters-21578 Text Categorization Collection" [Online]. Availabe: http://www.daviddlewis.com/resources/testcollections/reuters21578/.

[16] T. Fawcett "An introduction to ROC analysis" *Pattern Recognition Letters 27 (2006) pp. 861–874.*

[17] V. Metsis, I. Androutsopoulos, G. Paliouras "Spam Filtering with Naive Bayes – Which Naive Bayes?" *CEAS 2006 Third Conference on Email and Anti- Spam.*

[18] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz "A Bayesian Approach to Filtering Junk E-Mail" *Learning for Text Categorization: Papers from the 1998 Workshop, Madison, Wisconsin, AAAI Technical Report WS-98-05, (1998).*

[19] K. Tretyakov "Machine Learning Techniques in Spam Filtering" *Data Mining Problem-oriented Seminar, MTAT.03.177, May 2004, pp. 60-79.*

[20] F. Sebastiani "Machine Learning in Automated Text Categorization" *ACM Computing Surveys (CSUR) Volume 34 Issue 1, March 2002 pp 1-47.*

[21] *Apache SpamAssasin™* Homepage: http://spamassassin.apache.org/

Public Corpus: http://spamassassin.apache.org/publiccorpus/

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten (2009); ”The WEKA Data Mining Software: An Update” *SIGKDD Explorations, Volume 11, Issue 1*

[23] Minnen, G., J. Carroll, D. Pearc ”Applied morphological processing of English” *Natural Language Engineering, 7(3). 207-223*