

**ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

ΘΕΜΑ ΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ:

Εντοπισμός ανεπιθύμητων μηνυμάτων ηλεκτρονικής αλληλογραφίας με
χρήση τεχνικών εξόρυξης πληροφορίας

Φοιτητής: Βόβολης Αλέξανδρος
ΑΜ: Π2014008

Επιβλέπων Καθηγητής: Έξαρχος Θέμης



Περίληψη

Η συγκεκριμένη πτυχιακή εργασία έχει ως στόχο την ανάλυση δεδομένων από ένα σύνολο δεδομένων με διάφορα emails και την εξόρυξη πληροφορίας με τεχνικές μηχανικής μάθησης ώστε να διαπιστωθεί πότε ένα email ανήκει στην κατηγορία spam και πότε όχι. Γίνεται γενική ανάλυση του φαινομένου spam καθώς και τον αντίκτυπο που έχει τόσο στην καθημερινή ζωή των ανθρώπων που ανταλλάσσουν μηνύματα ηλεκτρονικού ταχυδρομείου αλλά και στην παγκόσμια οικονομία.

Στην συνέχεια γίνεται μια βιβλιογραφική ανασκόπηση σε 15 μελέτες που αφορούν το θέμα του spam οι οποίες έχουν δημοσιευθεί επίσημα σε διάφορα επιστημονικά περιοδικά και συνέδρια. Γίνεται μια αρχική περίληψη με το αντικείμενο της μελέτης και στην συνέχεια αναλύεται η μεθοδολογία που χρησιμοποιήθηκε καθώς και τα αποτελέσματα τα οποία προέκυψαν για την κάθε μελέτη τα οποία παίζουν καθοριστικό ρόλο στην εξαγωγή συμπερασμάτων.

Τέλος στο πειραματικό κομμάτι της εργασίας γίνεται αρχικά μια προεπεξεργασία των δεδομένων και στην συνέχεια εφαρμόστηκαν οι αλγόριθμοι μηχανικής μάθησης ώστε να αποδειχθεί βάσει αποτελεσμάτων ποιος αλγόριθμος είχε την καλύτερη απόδοση και έκανε δηλαδή την καλύτερη πρόβλεψη.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή μου κ. Θέμη Έξαρχο για την πολύτιμη βοήθεια που μου προσέφερε στην πτυχιακή μου εργασία όπως και για την κατανόηση και υπομονή που έδειξε κατά την διάρκεια εκπόνησης της.

Πίνακας Περιεχομένων

Κεφάλαιο 1	5
1.1 Εισαγωγή.....	5
1.2 Εντοπισμός ανεπιθύμητων μηνυμάτων ηλεκτρονικής αλληλογραφίας.....	6
1.3 Χρήσιμα στατιστικά στοιχεία για τα ανεπιθύμητα μηνύματα	7
1.4 Οι πιο συνηθισμένες μορφές spam	8
1.5 Κίνδυνοι σε επιχειρήσεις και οργανισμούς	10
Κεφάλαιο 2	12
2.1 Ανάλυση Βιβλιογραφίας.....	12
Κεφάλαιο 3	27
3.1 Μηχανική μάθηση	27
3.2 Μεθοδολογία.....	27
3.3 Σύνολο Δεδομένων	28
3.4 Προεπεξεργασία δεδομένων	31
3.5 Ταξινόμηση.....	32
3.6 Αλγόριθμοι ταξινόμησης.....	33
3.7 Αποτελέσματα.....	36
Κεφάλαιο 4	41
4.1 Συμπεράσματα - Συζήτηση	41

Κεφάλαιο 1

1.1 Εισαγωγή

Ως Ανεπιθύμητη Αλληλογραφία (spam) ορίζουμε την συνήθως ανώνυμη, συνήθως μαζική αποστολή ενός ή περισσότερων μηνυμάτων ηλεκτρονικού ταχυδρομείου προς πολλαπλούς αποδέκτες. Ο πιο συχνός λόγος ύπαρξης τέτοιων μηνυμάτων είναι η δημιουργία κέρδους για τον αποστολέα ο οποίος μπορεί να είναι και ανώνυμος σε πολλές περιπτώσεις. Μαζική αποστολή μηνυμάτων σημαίνει συνήθως η αποστολή δεκάδων χιλιάδων έως και εκατομμυρίων μηνυμάτων. Ο βασικός λόγος για τον τεράστιο όγκο μηνυμάτων είναι το οικονομικό όφελος που έχουν οι αποστολείς των μηνυμάτων spam (γνωστοί ως spammers) προέρχεται από ένα πολύ μικρό ποσοστό ανταπόκρισης από τους παραλήπτες των μηνυμάτων τους.

Ανεπιθύμητη αλληλογραφία μπορεί να θεωρηθεί όμως από κάποιον και νόμιμη αλληλογραφία όπως για παράδειγμα διαφημιστικό υλικό το οποίο κάποιοι τελικοί αποδέκτες έχουν επιλέξει να λαμβάνουν από κάποιο αποστολέα. Ένα τέτοιο μήνυμα μπορεί να μοιάζει με spam, αλλά στην πραγματικότητα μπορεί να είναι νόμιμη αλληλογραφία. Με άλλα λόγια, το ίδιο μήνυμα μπορεί να χαρακτηριστεί ως spam ή ως νόμιμη αλληλογραφία ανάλογα με το αν ή όχι ο χρήστης επιλέξει να το παραλάβει. Επομένως ο όρος διαφήμιση δεν πρέπει να χρησιμοποιείται ως το αποκλειστικό κριτήριο ορισμού του spam.

Πολλά μηνύματα spam δεν είναι ούτε διαφήμιση, ούτε κάποιου είδους εμπορική πρόταση. Εκτός από την προσφορά αγαθών και υπηρεσιών, το spam μπορεί να περιλαμβάνει μεταξύ άλλων και τις ακόλουθες κατηγορίες:

- Μηνύματα πολιτικού περιεχομένου
- Οικονομικές απάτες
- Μηνύματα που προτρέπουν την προώθηση τους σε τρίτους (αλυσίδα)
- Μηνύματα που χρησιμοποιούνται για τη διάδοση κακόβουλου λογισμικού (malware)

1.2 Εντοπισμός ανεπιθύμητων μηνυμάτων ηλεκτρονικής αλληλογραφίας

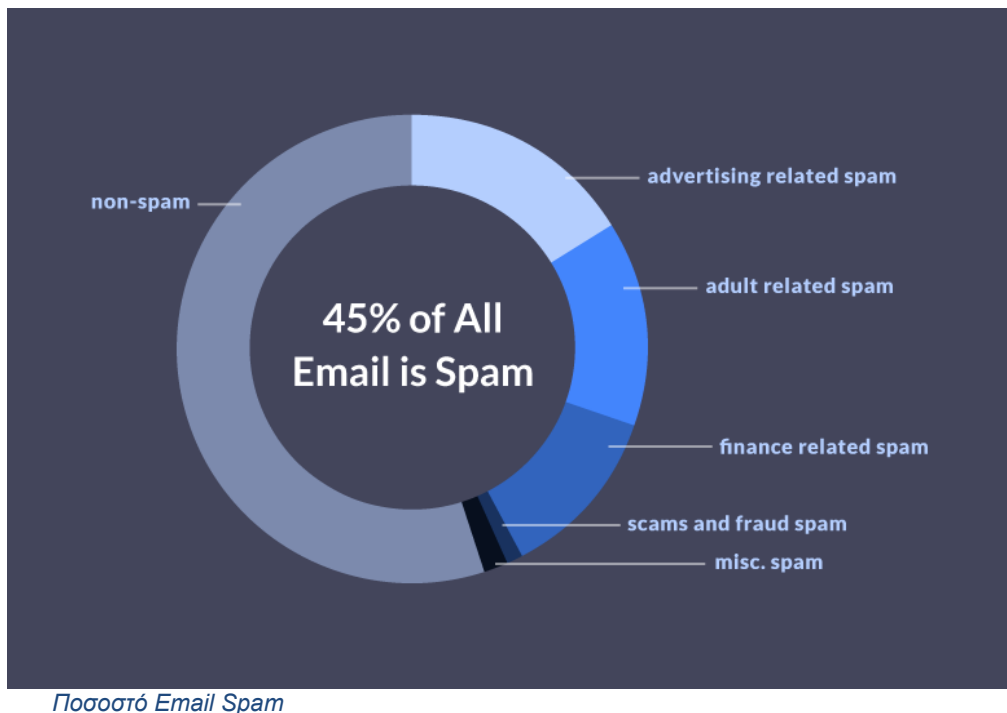
Θέμα της παρούσας πτυχιακής εργασίας είναι ο εντοπισμός ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου με χρήση τεχνικών εξόρυξης δεδομένων. Οι τεχνικές αυτές έχουν παίξει καθοριστικό ρόλο στο φιλτράρισμα των μηνυμάτων και στην αντιμετώπιση πολλών κινδύνων για τον χρήστη που έχουν προκύψει από την ραγδαία αύξηση της χρήσης των ηλεκτρονικών μέσων από την παγκόσμια κοινότητα. Εξόρυξη δεδομένων είναι η εξεύρεση μιας (ενδιαφέρουσας, αυτονόητης, μη προφανούς και πιθανόν χρήσιμης) πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων.

Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις. Ο πραγματικός στόχος της εξόρυξης δεδομένων είναι η αυτόματη ή ημιαυτόματη ανάλυση μεγάλων ποσοτήτων δεδομένων για την εξαγωγή κάποιου ενδιαφέροντος προτύπου που ήταν άγνωστο μέχρι εκείνη τη στιγμή, όπως ομάδες από εγγραφές δεδομένων (συσταδοποίηση), ασυνήθιστες εγγραφές (*anomaly detection*) και εξαρτήσεις (κανόνες συσχετίσεων). Αυτό συνήθως συμπεριλαμβάνει τη χρήση βάσης δεδομένων όπως χωρικά ευρετήρια. Αυτά τα πρότυπα ύστερα μπορούν να θεωρηθούν ως μία περιγραφή των δεδομένων εισαγωγής και να χρησιμοποιηθούν για περαιτέρω ανάλυση ή για παράδειγμα στην εκμάθηση μηχανής και στην προγνωστική ανάλυση. Για παράδειγμα, η εξόρυξη δεδομένων θα μπορούσε να προσδιορίσει πολλαπλά σύνολα στα δεδομένα, τα οποία μπορούν να χρησιμοποιηθούν μετά για να εξασφαλίσουν περισσότερο ακριβή αποτελέσματα από ένα σύστημα υποστήριξης αποφάσεων. Παρότι η συλλογή δεδομένων και η προετοιμασία δεδομένων, αλλά και η ερμηνεία των αποτελεσμάτων και εκθέσεων δεν αποτελούν μέρος της εξόρυξης δεδομένων, παρ' όλα αυτά ανήκουν στην ανακάλυψη γνώσης από βάσεις δεδομένων σαν κάποια επιπρόσθετα βήματα.

1.3 Χρήσιμα στατιστικά στοιχεία για τα ανεπιθύμητα μηνύματα

Σύμφωνα με έρευνα που διεξήχθη το 2018 προέκυψαν τα παρακάτω κρίσιμα στατιστικά στοιχεία:

- α) Σχεδόν το ½ των email που ανταλλάσσονται καθημερινά αποτελούν spam. Σε μια μέρα υπολογίζεται ότι στέλνονται περίπου 14.5 δισεκατομμύρια email.
- β) Το 36% των spam email έχουν να κάνουν με κάποια μέθοδο διαφήμισης - προώθηση
- γ) Οι spammers λαμβάνουν 1 απάντηση για κάθε 12.500.000 emails που στέλνουν.
- δ) Οι spammers κερδίζουν περίπου 7000\$ την μέρα
- ε) Τα spam emails κοστίζουν στην οικονομία περίπου 20.5 δισεκατομμύρια δολάρια κάθε χρόνο ποσό το οποίο μπορεί να αυξηθεί δραματικά
- στ) Οι χώρες οι οποίες έχουν το μεγαλύτερο πρόβλημα με τα spam email είναι οι παρακάτω: Ηνωμένες Πολιτείες Αμερικής, Κίνα, Ρωσία, Ουκρανία, Ιαπωνία, Ηνωμένο Βασίλειο, Ινδία, Hong Kong, Τουρκία, Βραζιλία
- ζ) Το 80% των spam στην Βόρεια Αμερική και στην Ευρώπη στέλνονται από 100 κοινές οργανώσεις spam
- η) 3 στις 4 εταιρείες έχουν πέσει θύμα σε απάτες phishing





2 Τρόποι απάτης με τεχνική «ψαρέματος»

1.4 Οι πιο συνηθισμένες μορφές spam

Διαφημίσεις: Αυτός είναι ένας από τους πιο συνηθισμένους τύπους ανεπιθύμητων μηνυμάτων. Σίγουρα έχετε ήδη λάβει πολλά ανεπιθύμητα email που προσφέρουν προϊόντα και υπηρεσίες, όπως χάπια απώλειας βάρους και προσφορές τένις. Σε πολλές περιπτώσεις μπορεί να είναι απάτη, αλλά η προσφορά μπορεί επίσης να είναι πραγματική.

Chain Letters: «Κάτι κακό θα σου συμβεί». Συνήθως, αλυσιδωτά γράμματα λένε συναρπαστικές και συναρπαστικές ιστορίες και σας πείθουν να περάσετε το μήνυμα υπό την ποινή του να συμβεί κάτι πολύ κακό σε εσάς. Να είστε προσεκτικοί ή θα έχετε μια κακή τύχη.

Πλαστογράφηση ηλεκτρονικού ταχυδρομείου: Η πλαστογραφία (Spoofing) μέσω ηλεκτρονικού ταχυδρομείου σχετίζεται με τις απάτες ηλεκτρονικού ψαρέματος (phishing). Συμβαίνουν όταν spammers ή phishers προσπαθούν να σας ξεγελάσουν πλαστοπροσωπώντας κάποιον που γνωρίζετε ή μια εταιρεία με την οποία έχετε σχέση. Αυτός είναι ένας από τους πιο επικίνδυνους τύπους spam.

Το ηλεκτρονικό ψάρεμα (phishing) είναι η προσπάθεια απόκτησης ευαίσθητων πληροφοριών ή δεδομένων, όπως ονόματα χρήστη, κωδικοί πρόσβασης και στοιχεία πιστωτικής κάρτας ή άλλα ευαίσθητα στοιχεία, με την πλαστοπροσωπία ως αξιόπιστου φορέα σε μια ψηφιακή επικοινωνία. Συνήθως πραγματοποιείται μέσω πλαστογράφησης email, άμεσων μηνυμάτων, και μηνυμάτων κειμένου, το ηλεκτρονικό ψάρεμα (phishing) κατευθύνει συχνά τους χρήστες να εισάγουν προσωπικά στοιχεία σε έναν ψεύτικο ιστότοπο που ταιριάζει με την εμφάνιση και την αίσθηση του νόμιμου ιστότοπου. Το ηλεκτρονικό ψάρεμα είναι ένα παράδειγμα τεχνικών κοινωνικής μηχανικής που χρησιμοποιούνται για να εξαπατήσουν τους χρήστες. Οι χρήστες δελεάζονται από επικοινωνίες που ισχυρίζονται ότι προέρχονται από αξιόπιστα μέρη, όπως ιστότοπους κοινωνικής δικτύωσης, ιστότοπους δημοπρασιών, τράπεζες, μηνύματα / μηνύματα από φίλους ή συναδέλφους / στελέχη, διαδικτυακά συστήματα πληρωμών ή διαχειριστές πληροφορικής.

Hoaxes: Σε αυτόν τον τύπο ανεπιθύμητου περιεχομένου, συμπεριλαμβάνουμε προσφορές και υποσχέσεις θαύματος, όπως, για παράδειγμα, "πλούσιος σε λιγότερο από ένα μήνα" ή "κερδίζοντας το σώμα του ονείρου σου τρώγοντας περισσότερα και ασκήσεις λιγότερο" Σε γενικές γραμμές, αυτή η τακτική χρησιμοποιείται από spammers για να σας κρατήσει την προσοχή και να σας οδηγήσει σε έναν κακόβουλο ιστότοπο.

Απάτες χρημάτων: Εδώ έχουμε ανεπιθύμητα μηνύματα με εύκολες υποσχέσεις χρημάτων, όπως το σχέδιο του πρίγκιπα της Νιγηρίας. Σε αυτήν την περίπτωση, προφανώς πρέπει μόνο να δανείσετε ένα μικρό χρηματικό ποσό για να λάβετε μια μεγάλη ανταμοιβή στο μέλλον. Το ανεπιθύμητο περιεχόμενο χρημάτων περιλαμβάνει επίσης το να ζητούνται χρήματα για παιδιά που τα έχουν ανάγκη στην Αφρική ή για οικογένειες που έχουν υποστεί απώλειες ως αποτέλεσμα φυσικής καταστροφής.

Προειδοποιήσεις Ιών: Εάν λάβετε ένα μήνυμα ηλεκτρονικού ταχυδρομείου που σας προειδοποιεί για λοίμωξη από κακόβουλο λογισμικό στις συσκευές σας, όπως ransomware ή ιού, αυτό είναι πιθανώς ανεπιθύμητο μήνυμα προειδοποίησης για κακόβουλο λογισμικό. Σε ορισμένες περιπτώσεις, οι spammers λένε ότι έχουν τη λύση στο πρόβλημά σας και ότι πρέπει απλώς να δώσετε κάποιες πληροφορίες ή να κατεβάσετε ένα συνημμένο.

Ανεπιθύμητο ερωτικό περιεχόμενο: Η αποστολή πορνογραφίας μέσω email χρησιμοποιείται ευρέως από spammers, επειδή η αγορά πορνογραφίας είναι πολύ επικερδής, αυξάνοντας το ενδιαφέρον των ανθρώπων. Φυσικά, οι spammers το γνωρίζουν και το χρησιμοποιούν για να δημιουργήσουν κακόβουλα email χρησιμοποιώντας ερωτικές εικόνες και βίντεο.

1.5 Κίνδυνοι σε επιχειρήσεις και οργανισμούς

Οι άνθρωποι αποτελούν το πιο ισχυρό στοιχείο στην επιχειρησιακή λειτουργία, όμως αναφορικά με την κυβερνοασφάλεια, μπορούν επίσης να αποτελέσουν τον πιο αδύναμο κρίκο. Πολλοί υπάλληλοι αν δεν είναι εκπαιδευμένοι, μπορεί να γίνουν επιρρεπείς σε επιθέσεις που εκμεταλλεύονται την ανθρώπινη φύση, και όχι μόνο τις τεχνικές ευπάθειες. Οι κυβερνοεγκληματίες κάνουν έρευνα σχετική με τα θύματα τους για να εντοπίσουν τους “καλύτερους” ανθρώπινους στόχους σε μία εταιρεία. Θα μελετήσουν την εταιρεία και θα επισκέπτονται συχνά την ιστοσελίδα της. Θα επισκέπτονται στα social media, όπως στο LinkedIn για να μαθαίνουν σχετικά με τους ανθρώπους που εργάζονται, τον τίτλο τους, που έχουν εργαστεί αλλού, με ποιους έχουν συνεργαστεί και ποια ενδιαφέροντα έχουν.

Τακτικές – Οι απάτες μέσω email στοχεύουν τους ανθρώπους. Οι επιθέσεις σχεδιάζονται ώστε να ξεγελούν τους υπαλλήλους, νομίζοντας ότι έλαβαν email από κάποιο υψηλόβαθμο στέλεχος της εταιρείας, ή κάποιον προμηθευτή ή συνεργάτη ή συνάδελφο. Ο αποστολέας του spoof email ζητά την εκτέλεση ενεργειών όπως τη μεταφορά χρημάτων, την παροχή φορολογικών στοιχείων, ή άλλων ευαίσθητων εταιρικών ή προσωπικών δεδομένων. Με μία γρήγορη ματιά, δε φαίνεται τίποτα ασυνήθιστο στο συγκεκριμένο email. Όμως μικρές διαφορές, όπως το όνομα του αποστολέα, η διεύθυνση του, ή το email απάντησης, μπορούν να προδώσουν τον απατεώνα. Οι κυβερνοεγκληματίες βασίζονται στο ότι ο στόχος τους δεν θα ασχοληθεί με την επαλήθευση του email.

Οι 4 πιο κοινές εκδοχές των email-απάτης είναι οι εξής:

Spoofed Name – Αυτή η εκδοχή αντιπροσωπεύει το 75% των επιθέσεων. Χρησιμοποιεί το όνομα του spoofed στελέχους στο πεδίο «from». Όμως η διεύθυνση email προέρχεται από εξωτερική υπηρεσία, όπως το Gmail, και ανήκει στον επιτιθέμενο.

Reply-To Spoofing – Αυτή η τεχνική χρησιμοποιεί το πραγματικό όνομα και email του αποστολέα που πλαστοπροσωπείται (συνήθως ο CEO). Το όνομα στο «Reply-to» χρησιμοποιεί επίσης το όνομα του πλαστοπροσωποποιημένου αποστολέα. Όμως η διεύθυνση «απάντηση σε», όπου αποστέλλονται οι απαντήσεις, ανήκει στον επιτιθέμενο.

Spoofed Sender (with No Reply-to Address) – Αυτή η μορφή απάτης μέσω email χρησιμοποιεί το όνομα και το email του spoofed στελέχους. Όμως το μήνυμα δεν περιλαμβάνει διεύθυνση «Απάντηση σε», οπότε είναι αδύνατη η αμφίδρομη αλληλογραφία. Το μήνυμα συχνά περιλαμβάνει οδηγίες μεταφοράς χρημάτων, που καταργούν την ανάγκη για περεταίρω διευκρινίσεις.

Lookalike Domain – Σε αυτή τη μορφή απάτης μέσω email, η διεύθυνση «Από» του επιτιθέμενου έχει παρόμοια εμφάνιση με αυτή του πλαστοπροσωποποιημένου στελέχους. Το παρόμοιο domain μπορεί να διαφέρει μόνο προς ένα γράμμα από το πραγματικό. Για παράδειγμα, το legitcompany.com μπορεί να γίνει legtcompany.com

Κεφάλαιο 2

2.1 Ανάλυση Βιβλιογραφίας

Στην συνέχεια γίνεται μια βιβλιογραφική ανασκόπηση επιλεγμένων εργασιών που αφορούν το θέμα των ανεπιθύμητων μηνυμάτων και μπορούν να προσφέρουν χρήσιμες πληροφορίες και συμπεράσματα στην παρούσα εργασία.

Ξεκινώντας την ανάλυση στην πρώτη εργασία[1] αναφέρεται πως τα ανεπιθύμητα μηνύματα ηλεκτρονικού ταχυδρομείου (spam) πολλές φορές περιλαμβάνουν ιούς αλλά και λογισμικό υποκλοπής(spyware), επομένως υπάρχει μια σημαντική ανάγκη για την ανίχνευση τέτοιου είδους μηνυμάτων ηλεκτρονικού ταχυδρομείου. Κατα καιρούς έχουν προταθεί διάφορες μέθοδοι για τον εντοπισμό ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου με τεχνικές machine learning, οι οποίες υποβλήθηκαν για να μειώσουν τα μη επιθυμητά μηνύματα ηλεκτρονικού ταχυδρομείου και για τη λήψη αποτελεσμάτων υψηλής ακρίβειας για την ταξινόμηση των ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου.

Στην παρούσα εργασία που αναλύεται, υποβάλλεται μια νέα μέθοδος πρόβλεψης με βάση τη βελτιστοποίηση antlion (ALO) και την ενίσχυση που ονομάζεται ALO-Boosting για την επίλυση του προβλήματος των ανεπιθύμητων email. Το ALO είναι ένα υπολογιστικό μοντέλο που μιμείται την τεχνική θήρας των μυρμηγκοφάγων στα μυρμήγκια και χρησιμοποιήθηκε για να τροποποιήσει την πραγματική θέση του πληθυσμού στην ξεχωριστή περιοχή αναζήτησης, αποκτώντας έτσι το βέλτιστο υποσύνολο χαρακτηριστικών για την καλύτερη υποβολή ταξινόμησης με βάση την ενίσχυση του ταξινομητή. Η ενίσχυση του ταξινομητή (Boosting classifier) είναι ένας αλγόριθμος ταξινόμησης που αναφέρεται σε μια ομάδα αλγορίθμων η οποία μετατρέπει τους soft learners σε powerful learners.

Η προτεινόμενη μέθοδος αποτελείται από δύο κύρια στάδια: επιλογή χαρακτηριστικών και ταξινόμηση. Στο πρώτο στάδιο, το ALO προτάθηκε για αναζήτηση του βέλτιστου χαρακτηριστικού στην περιοχή αναζήτησης. Στο δεύτερο στάδιο, ο ταξινομητής ενίσχυσης χρησιμοποιήθηκε για την πρόβλεψη βάσει των χαρακτηριστικών που αποκτήθηκαν στο πρώτο στάδιο. . Αυτές οι μέθοδοι εφαρμόστηκαν στο σετ δεδομένων CSDMC2010 και SpamAssassin για τον εντοπισμό των ανεπιθύμητων μηνυμάτων. Τα

πειραματικά αποτελέσματα δείχνουν ότι η προτεινόμενη μέθοδος λαμβάνει χαμηλό αριθμό επιλεγμένων χαρακτηριστικών και επιτυγχάνει υψηλό βαθμό ακρίβειας ταξινόμησης. Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα ταξινόμησης 8 μεθόδων με βάση ακρίβεια, ευαισθησία, ειδικότητα, G-mean, F-measure, και αριθμό επιλεγμένων χαρακτηριστικών από το CSDMC2010 dataset.

Measure	Method							
	KNN	ALO-KNN	SVM	ALO-SVM	Bagging	ALO-Bagging	Boosting	ALO-Boosting
A	0.9534	0.9700	0.9316	0.9606	0.9515	0.9931	0.9800	0.9980
SN	0.9022	0.9190	0.9187	0.9255	0.9730	0.9840	0.9799	0.9900
SP	0.9867	0.9934	0.9691	0.9816	0.9404	0.9968	0.9735	0.9998
P	0.9036	0.9836	0.9300	0.9675	0.8944	0.9919	0.9697	1.0000
GM	0.9435	0.9555	0.9436	0.9531	0.9566	0.9904	0.9767	0.9949
FM	0.9029	0.9502	0.9243	0.9460	0.9320	0.9879	0.9748	0.9950
Features No	507	245	507	242	507	236	507	211

3 Αποτελέσματα ταξινόμησης 8 μεθόδων

Η επόμενη εργασία[2] που αναλύεται προτείνει μια υβριδική λύση ταξινομητή ανεπιθύμητης αλληλογραφίας χρησιμοποιώντας ένα μοντέλο ταξινόμησης email βάσει περιβάλλοντος ως κύριου αλγορίθμου ο οποίος συμπληρώνεται από τον υπολογισμό του κέρδους πληροφοριών για την βελτιστοποίηση της ακρίβειας της ταξινόμησης των ανεπιθύμητων μηνυμάτων. Η προτεινόμενη λύση αποτελείται από τρία στάδια: Το πρώτο αφορά την προεπεξεργασία email, το δεύτερο την εξαγωγή χαρακτηριστικών και το τρίτο την ταξινόμηση email. Μέσα από την έρευνα διαπιστώθηκε ότι το φίλτρο ανεπιθύμητων μηνυμάτων “LingerIG” είναι πολύ αποτελεσματικό στο διαχωρισμό των ανεπιθύμητων email από τα email εργασίας. Επίσης, το αποτέλεσμα του πειράματος απέδειξε ότι η ακρίβεια του φιλτραρίσματος spam είναι 100% όπως καταγράφηκε από την ομάδα προγραμματιστών στο πανεπιστήμιο του Σίδνεϊ. Η μελέτη έδειξε ότι είναι εφικτή η εφαρμογή του φίλτρου στο μοντέλο ταξινόμησης email βάσει περιβάλλοντος και το πείραμα της μελέτης απέδειξε πως μπορεί να βελτιωθεί σε σημαντικό βαθμό το

φιλτράρισμα των ανεπιθύμητων μηνυμάτων.

Η καλύτερη λύση που υπάρχει στην παρούσα φάση είναι ο αλγόριθμος εξόρυξης γραφημάτων. Ονομάζεται σύστημα ταξινόμησης email βάσει περιεχομένου και προτάθηκε από τους Wasi et al. Αποτελείται από αλγόριθμο εξόρυξης γραφημάτων και σύστημα αναγνώρισης συμβάντων. Ένα σημαντικό ελάττωμα σε αυτό το σύστημα βασίζεται στον τεράστιο αριθμό email που απαιτεί για να φτάσει την ακρίβεια του 100%.

Η ονομασία της λύσης που προτείνεται είναι Spam Filtered Email Classification Model (SFECM). Βασίζεται στο μοντέλο ταξινόμησης ηλεκτρονικού ταχυδρομείου βάσει περιεχομένου για να φιλτράρονται τα ανεπιθύμητα μηνύματα πριν από την πραγματική ταξινόμηση email. Αυτό το σύστημα αποτελείται από τρία στάδια: Προεπεξεργασία μέσω email, εξαγωγή χαρακτηριστικών και ταξινόμηση email. Το προτεινόμενο σύστημα εκτελεί το POS Tagger σε email σε στάδιο προεπεξεργασίας email για να μετατρέψει τα email κειμένου σε email χαρακτηριστικών. Στο στάδιο εξαγωγής χαρακτηριστικών, το προτεινόμενο σύστημα φιλτράρει τα ανεπιθύμητα μηνύματα από ένα σύνολο email που έχουν εισαχθεί. Στη συνέχεια, από τα φιλτραρισμένα μηνύματα εξάγονται λέξεις αποσύνδεσης, λέξεις χαιρετισμού, λέξεις-κλειδιά για να σχηματίσουν ένα γράφημα email. Σε αυτό το στάδιο, τα γραφήματα προτύπων ενημερώνονται χρησιμοποιώντας νέα γραφήματα ηλεκτρονικού ταχυδρομείου. Στη συνέχεια, τα γραφήματα προτύπων κατατάσσονται στο στάδιο ταξινόμησης email για να εκχωρηθούν και να αντιπροσωπεύουν τον σχετικό φάκελο. Στη συνέχεια, τα γραφήματα ηλεκτρονικού ταχυδρομείου αντιστοιχίζονται σε αντιπροσωπευτικά πρότυπα γραφήματα και τοποθετούνται σε φάκελο του αντιπροσωπευτικού προτύπου γραφήματος με το οποίο ταιριάζουν περισσότερο. Η συγκεκριμένη εργασία έχει αναγνωρίσει ότι η ακρίβεια 100% στην ταξινόμηση ανεπιθύμητων μηνυμάτων του συστήματος email εξακολουθεί να είναι μια ανεκπλήρωτη ανάγκη. Όσον αφορά το πείραμα, ο χρόνος επεξεργασίας μεταξύ της χρήσης του φίλτρου ανεπιθύμητου και της μη χρήσης του φίλτρου ανεπιθύμητων μηνυμάτων διαφέρει ασήμαντα. Ωστόσο, είναι σημαντικό να τονιστεί η ανάγκη μείωσης του χρόνου επεξεργασίας της ταξινόμησης ανεπιθύμητων μηνυμάτων, επειδή ο χρόνος επεξεργασίας 0,1 δευτερολέπτων είναι μια ανεκπλήρωτη ανάγκη σε αυτήν τη λύση.

Στην επόμενη εργασία[3] προς ανάλυση επισημαίνεται πως τα ανεπιθύμητα μηνύματα ηλεκτρονικού ταχυδρομείου είναι μηνύματα που αποστέλλονται μέσω του Διαδικτύου με σκοπό να βλάψουν τους χρήστες. Οι αναλυτές ανεπιθύμητων μηνυμάτων, οι χρήστες του Διαδικτύου και η κοινότητα κατά των ανεπιθύμητων μηνυμάτων χρειάζονται μια τεχνική για να προσδιορίσουν εάν τα μηνύματα ηλεκτρονικού ταχυδρομείου είναι ασφαλή ή κακόβουλα. Για την αποστολή κακόβουλων μηνυμάτων, οι spam domains χρησιμοποιούνται από spammers. Οι domains ανεπιθύμητης αλληλογραφίας είναι εκείνοι που χρησιμοποιούνται για τη διαφήμιση ψευδών υπηρεσιών και προϊόντων. Αυτό το άρθρο αντικατοπτρίζει τις διάφορες έρευνες που προωθήθηκαν σχετικά με τα spam domains χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων. Η μεθοδολογία αξιολόγησης σχετίζεται με την ανασκόπηση της βιβλιογραφίας που επιλέγει πολλές ερευνητικές ερωτήσεις για να διαλέξει όλα τα σχετικά ερευνητικά στοιχεία που απαιτούνται για την έρευνα. Ο σκοπός είναι να δείξει την εκτίμηση του ερευνητικού θέματος χρησιμοποιώντας μια αξιόπιστη, αυστηρή και ελεγχόμενη μεθοδολογία αξιοπιστίας.

Κατά τη διεξαγωγή του σταδίου αναθεώρησης:

- Αναγνωρίζονται οι πρωτογενείς μελέτες
- Ορίζονται οι αξιολογήσεις ποιότητας
- Πραγματοποιείται εξαγωγή και παρακολούθηση δεδομένων
- Τέλος, κατά την επεξεργασία του σταδίου αναθεώρησης:
- Καθορίζονται οι μέθοδοι δημοσίευσης
- Παρουσιάζονται οι αναφορές επανεξέτασης

Κατά τη διάρκεια της μελέτης πλαισιώνονται τα ακόλουθα ερευνητικά ερωτήματα:

- Θα μειωθεί ο αριθμός των ανεπιθύμητων μηνυμάτων μέσα από την αναγνώριση και ανάλυση του ανεπιθύμητου domain;
- Πώς βοηθούν οι αλγόριθμοι Εξόρυξης Δεδομένων στον εντοπισμό domain ανεπιθύμητης αλληλογραφίας;

Η παρούσα εργασία προσπαθεί να απαντήσει στα ερωτηματολόγια συνδυάζοντας διάφορους αλγόριθμους μαζί που ανιχνεύουν τον ανεπιθύμητο domain από τον οποίο προέρχονται τα spam email. Εξακολουθεί να είναι ένα μεγάλο πρόβλημα η εύρεση της προέλευσης των ανεπιθύμητων μηνυμάτων από όπου αποστέλλονται. Ακόμη, υπάρχουν πολλά εργαλεία που είναι διαθέσιμα, αλλά υπάρχουν και πολλά άλλα ζητήματα για αυτό και απαιτείται να υπάρξει βελτίωση για πιο αποτελεσματική χρήση ώστε

στο μέλλον τα ανεπιθύμητα μηνύματα ηλεκτρονικού ταχυδρομείου να μειωθούν αισθητά. Ο συσχετισμένος αλγόριθμος της εξόρυξης δεδομένων που υπήρχε στην παρούσα φάση δεν ήταν ο ιδανικός για την αντιμετώπιση προκλήσεων, επομένως απαιτείται περισσότερη εξέταση και ανάλυση.

Στην επόμενη εργασία[4] που αναλύουμε γίνεται αναφορά στο τεράστιο μέρος της επικοινωνίας, επαγγελματικής αλλά και ιδιωτικής, πραγματοποιείται με τη μορφή email. Ωστόσο, λόγω διαφημιστικών εταιρειών και ιστότοπων κοινωνικής δικτύωσης, τα περισσότερα από τα email περιέχουν ανεπιθύμητες πληροφορίες που δεν σχετίζονται με τον χρήστη. Τα ανεπιθύμητα μηνύματα ηλεκτρονικού ταχυδρομείου προκαλούν σύγχυση στους παραλήπτες ενώ πολλές φορές μπορούν να οδηγήσουν και σε οικονομικές απώλειες, επομένως υπάρχει ανάγκη να φιλτράρονται και να διαχωρίζονται από τα χρήσιμα μηνύματα ηλεκτρονικού ταχυδρομείου. Έχουν αναπτυχθεί πολλοί αλγόριθμοι και φίλτρα για τον εντοπισμό των ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου, όμως οι spammers εξελίσσονται συνεχώς και εκσυγχρονίζουν τις τεχνικές spamming τους, λόγω των οποίων τα υπάρχοντα φίλτρα καθίστανται λιγότερο αποτελεσματικά. Η μέθοδος που προτείνεται σε αυτό το άρθρο περιλαμβάνει τη δημιουργία ενός φίλτρου spam χρησιμοποιώντας δυαδικές και συνεχείς κατανομές πιθανότητας. Οι αλγόριθμοι που εφαρμόστηκαν για την κατασκευή του μοντέλου ταξινόμησης είναι οι Naive Bayes και Decision Trees.

Τα μοντέλα ταξινόμησης Naive Bayes και Decision Tree τόσο σε δυαδικές όσο και σε συνεχείς διανομές δοκιμάστηκαν σε τρία τυπικά σύνολα δεδομένων. Το πρώτο σύνολο δεδομένων συλλέχθηκε από το αποθετήριο του Κρατικού Πανεπιστημίου της Ιντιάνα, το οποίο περιείχε 5200 μηνύματα spam και μη. Το δεύτερο και τρίτο σύνολο δεδομένων λήφθηκε από το Ling-Spam dataset (περιείχε συνολικά 2893 αρχεία). Η χρήση της κατανομής Bernoulli σε αυτό το ερευνητικό έργο γίνεται για να υποδηλώσει την εμφάνιση χαρακτηριστικών που εξετάζονται σε κάθε email. Εάν ένα χαρακτηριστικό εμφανίζεται σε ένα email, ο αριθμός που σχετίζεται με αυτό το χαρακτηριστικό είναι 1 ή 0. Σε περίπτωση συνεχούς διανομής, ο συνολικός αριθμός εμφανίσεων για κάθε χαρακτηριστικό σε ένα email λαμβάνεται υπόψη για την κατασκευή των μοντέλων ταξινόμησης. Ο ταξινομητής Naive Bayes υπολογίζει πρώτα τις πιθανότητες πιθανότητας όλων των χαρακτηριστικών που υπάρχουν στο training set και στη συνέχεια χρησιμοποιεί την προηγούμενη πιθανότητα (είτε ανεπιθύμητη είτε μη ανεπιθύμητη) για

να προβλέψει την ετικέτα ενός εγγράφου. Μόλις εκπαιδευτεί το μοντέλο, ο ταξινομητής εμφανίζει δέκα πιο αντιπροσωπευτικές και λιγότερο αντιπροσωπευτικές λέξεις για ανεπιθύμητη αλληλογραφία μαζί με τις υπολογισμένες πιθανότητες.

Τα πειραματικά αποτελέσματα αποκάλυψαν ότι τα μοντέλα ταξινομητών αποδίδουν καλύτερα με την κατανομή πιθανότητας του Bernoulli παρά με τη συνεχή κατανομή πιθανότητας. Μπορεί επίσης να συναχθεί το συμπέρασμα ότι δεν υπάρχει εγγενώς ανώτερο μοντέλο ταξινόμησης μεταξύ Naive Bayes και Decision Trees. Η απόδοση των μοντέλων ταξινόμησης ποικίλλει και εξαρτάται από διάφορους παράγοντες όπως η κατανομή πιθανότητας που χρησιμοποιείται, το σύνολο δεδομένων και το σχετικό πρόβλημα. Στον αλγόριθμο Naive Bayes, ο ταξινομητής πρέπει να εκπαιδευτεί χειροκίνητα, ενώ στον Decision Tree ο ταξινομητής επιλέγει το καλύτερο χαρακτηριστικό ελέγχοντας τον πίνακα ή το αρχείο μοντέλου.

Στην επόμενη εργασία[5] αναφέρεται πως ο μεγάλος όγκος ηλεκτρονικών μηνυμάτων προκαλεί απειλή για την ασφάλεια των χρηστών. Παρά τις διάφορες προσεγγίσεις ασφαλείας, οι spammers προκαλούν μεγάλη ευπάθεια στο Διαδίκτυο. Αυτό το άρθρο ασχολείται με τις αποτελεσματικές μεθόδους χρήσης μερικών από τους δημοφιλείς αλγόριθμους για τη δημιουργία ενός μοντέλου μηχανικής μάθησης που μπορεί να ταξινομήσει εάν ένα μήνυμα αλληλογραφίας είναι ανεπιθύμητο ή μη. Για το πείραμα χρησιμοποιείται το UCI Machine Learning Repository Spambase Data. Η απόδοση πέντε σημαντικών αλγορίθμων ταξινόμησης μηχανικής μάθησης, δηλαδή οι Logistic Regression, Decision Tree, Naive Bayes, KNN και SVM αξιολογούνται προκειμένου να εκπαιδεύσουν και να δημιουργήσουν ένα αποτελεσματικό μοντέλο μηχανικής μάθησης για την ανίχνευση ανεπιθύμητων e-mail. Το εργαλείο Weka χρησιμοποιείται για την εκπαίδευση και τον έλεγχο του συνόλου δεδομένων.

Από τα αποτελέσματα που εξήχθησαν για το μοντέλο μηχανικής μάθησης, είναι προφανές ότι το decision tree ξεπερνά τους άλλους αλγόριθμους ταξινόμησης όσον αφορά όλες τις μετρήσεις απόδοσης. Αν και ο αλγόριθμος KNN(k-nearest neighbors) παράγει το ίδιο αποτέλεσμα, χρειάστηκε περισσότερος χρόνος για την κατασκευή του μοντέλου από τον αλγόριθμο decision tree. Η μελλοντική εργασία περιλαμβάνει την ενίσχυση του μοντέλου με περισσότερες παραμέτρους αξιολόγησης για ένα πιο αποτελεσματικό φιλτράρισμα ανεπιθύμητων μηνυμάτων. Στον παρακάτω πίνακα βλέπουμε την απόδοση των 5 αλγορίθμων. Η πρώτη στήλη accuracy μας λέει πόσα παραδείγματα ταξινομήθηκαν σωστά από το σύνολο των ταξινομημένων παραδειγμάτων. Η στήλη Precision υποδηλώνει τον αριθμό της σωστής πρόβλεψης των θετικών διαιρούμενη με

το άθροισμα της σωστής και της εσφαλμένης πρόβλεψης των θετικών. Αυτό δείχνει ότι όταν ένα μοντέλο προβλέπει θετικά, η ακρίβεια διασφαλίζει ότι τα στοιχεία επισημαίνονται σωστά ως θετικά. Η στήλη Recall υποδηλώνει τον αριθμό της σωστής πρόβλεψης των θετικών διαιρούμενη με τη σωστή πρόβλεψη των θετικών και την εσφαλμένη πρόβλεψη των αρνητικών. Αυτό υποδεικνύει ότι όταν ένα μοντέλο προβλέπει θετικά, η ακρίβεια διασφαλίζει ότι τα στοιχεία επισημαίνονται σωστά ως θετικά. Η βαθμολογία F1 υπολογίζεται συνδυάζοντας την ακρίβεια (acc) και την ανάκληση (recall) για να αξιολογηθεί η συνολική ακρίβεια του αλγορίθμου. Ως εκ τούτου, η χαμηλή ψευδής θετική και η χαμηλή λανθασμένη αρνητική τιμή δίνει ένα καλό μοντέλο που έχει προβλέψει το αποτέλεσμα με ακρίβεια.

Algorithm	Evaluation Metrics			
	Accuracy	Precision	Recall	F1 measure
Logistic Regression	93.1319 %	0.931	0.931	0.931
Random Tree	99.9348 %	0.999	0.999	0.999
Naïve Bayes	79.5262 %	0.845	0.795	0.797
KNN	99.9348 %	0.999	0.999	0.999
SVM	90.7629 %	0.117	0.908	0.907

Figure 4: Απόδοση 5 Αλγορίθμων

Στην επόμενη εργασία[6], εμπνευσμένη από έναν προσαρμοστικό αλγόριθμο παρουσιάζει μια τροποποιημένη τεχνική μηχανικής μάθησης του ανθρώπινου ανοσοποιητικού συστήματος που ονομάζεται αλγόριθμος αρνητικής επιλογής (Negative Selection Algorithm). Μια τοπική διαφορική εξίσωση επιλογής (Differential Equation) δημιουργεί ανιχνευτές στη φάση τυχαίας παραγωγής ανιχνευτών του NSA(NSA-DE). Ο τοπικός συντελεστής εξώθησης (Local Outlier Factor) εφαρμόζεται ως συνάρτηση φυσικής κατάστασης για τη μεγιστοποίηση της απόστασης των παραγόμενων ανιχνευτών ανεπιθύμητων μηνυμάτων από το χώρο που δεν είναι spam. Το πρόβλημα των αλληλεπικαλυπτόμενων ανιχνευτών επιλύεται επίσης υπολογίζοντας την ελάχιστη και τη μέ-

γιστη απόσταση δύο αλληλεπικαλυπτόμενων ανιχνευτών στον χώρο των ανεπιθύμητων. Από τα πειράματα, τα αποτελέσματα δείχνουν ότι η ακρίβεια ανίχνευσης του NSA – DE είναι 83,06%, ενώ ο τυπικός αλγόριθμος αρνητικής επιλογής είναι 68,86% στους 7000 παραγόμενους ανιχνευτές. Η επικύρωση του προτεινόμενου πλαισίου πραγματοποιήθηκε με δημοσιευμένο σύνολο δεδομένων και μια σε βάθος συγκριτική μελέτη του προτεινόμενου υβριδικού μοντέλου και του τυπικού μοντέλου πραγματοποιήθηκε προκειμένου να δείξει τη βελτίωση του προτεινόμενου υβριδικού μοντέλου έναντι του μοντέλου NSA. Το στατιστικό t-test χρησιμοποιήθηκε επίσης για τον προσδιορισμό της συσχέτισης μεταξύ αλγορίθμου αρνητικής επιλογής και υβριδικής αρνητικής επιλογής. Συνολικά, η εμπειρική έκθεση έδειξε την ανωτερότητα του προτεινόμενου υβριδικού μοντέλου NSA-DE έναντι του μοντέλου NSA. Το προτεινόμενο υβριδικό σύστημα θα είναι χρήσιμο σε άλλες εφαρμογές καθώς ο αλγόριθμος αρνητικής επιλογής επιλύει μεγάλο αριθμό πολύπλοκων προβλημάτων. Με βάση τα αποτελέσματα που παρουσιάζονται, αυτή η έρευνα πρέπει να θεωρηθεί ως βελτίωση στον τομέα της υπολογιστικής νοημοσύνης.

Στην επόμενη εργασία που αναλύεται[7] αναφέρεται πως η συνεχή αύξηση του όγκου ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου έχει δημιουργήσει μια έντονη ανάγκη για την ανάπτυξη πιο αξιόπιστων και ισχυρών φίλτρων ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου. Οι πιο πρόσφατες μέθοδοι μηχανικής εκμάθησης χρησιμοποιούνται για τον επιτυχημένο εντοπισμό και φιλτράρισμα ανεπιθύμητων μηνυμάτων. Στο συγκεκριμένο άρθρο παρουσιάζεται μια συστηματική ανασκόπηση ορισμένων από τις πιο δημοφιλείς προσεγγίσεις φιλτραρίσματος ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου που βασίζονται σε μηχανική εκμάθηση όπως η τεχνική φιλτραρίσματος βάσει περιεχομένου, βάσει περιπτώσεων κ.ο.κ. Καλύπτεται η έρευνα των σημαντικών εννοιών, των προσπαθειών, της αποτελεσματικότητας και της ερευνητικής τάσης στο φιλτράρισμα ανεπιθύμητων μηνυμάτων. Επίσης εξετάζονται οι εφαρμογές τεχνικών μηχανικής μάθησης στη διαδικασία φιλτραρίσματος ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου από τους κορυφαίους παρόχους υπηρεσιών διαδικτύου (ISP) όπως τα φίλτρα ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου Gmail, Yahoo και Outlook. Ένα email αποτελείται από δύο βασικά στοιχεία που είναι η κεφαλίδα και το κύριο σώμα. Η κεφαλίδα είναι η περιοχή που διαθέτει ευρείες πληροφορίες σχετικά με το περιεχόμενο του email. Περιλαμβάνει δηλαδή το θέμα, τον αποστολέα και τον παραλήπτη. Το σώμα αποτελεί την καρδιά του μηνύματος. Μπορεί να περιλαμβάνει πληροφορίες που δεν έχουν προκαθορισμένα δεδομένα

π.χ. ιστοσελίδα, ήχο, βίντεο, αναλογικά δεδομένα. Οι κεφαλίδες επιτρέπουν στον χρήστη να βλέπει την ολόκληρη την πορεία ενός email και τον χρόνο που χρειάζεται κάθε διακομιστής για την διαχείριση της αλληλογραφίας. Οι διαθέσιμες πληροφορίες περνάνε από επεξεργασία πριν ο ταξινομητής μπορέσει να τις χρησιμοποιήσει για φιλτράρισμα.

Στην παρακάτω εικόνα απεικονίζεται μια αρχιτεκτονική διακομιστή αλληλογραφίας καθώς και πως γίνεται το φιλτράρισμα των ανεπιθύμητων μηνυμάτων.

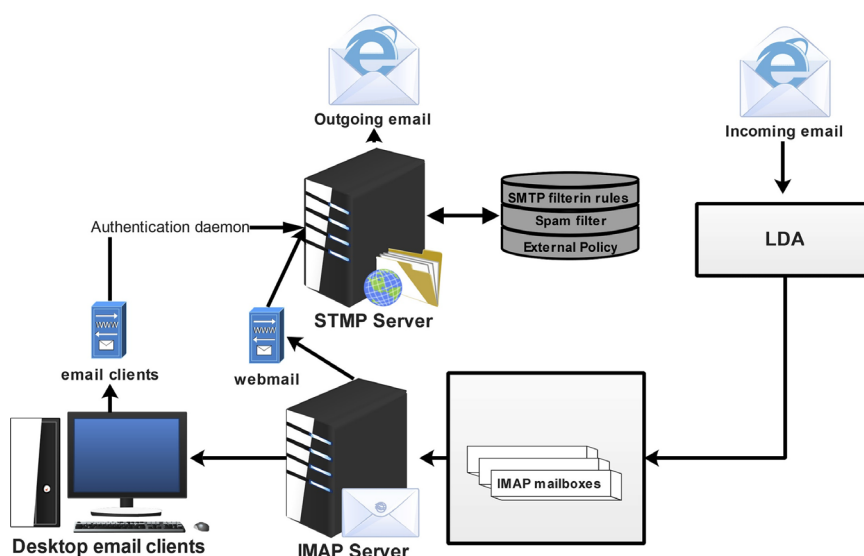


Figure 5: Αρχιτεκτονική Διακομιστή Αλληλογραφίας

Επιπλέον γίνεται αναφορά για τη γενική διαδικασία φιλτραρίσματος ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου και τις διάφορες προσπάθειες διαφορετικών ερευνητών για την καταπολέμηση των ανεπιθύμητων μηνυμάτων μέσω των τεχνικών μηχανικής εκμάθησης. Στην συνέχεια γίνεται μια σύγκριση για τα πλεονεκτήματα και τα μειονεκτήματα των υφιστάμενων προσεγγίσεων μηχανικής μάθησης και τα ανοιχτά ερευνητικά προβλήματα στο φιλτράρισμα ανεπιθύμητων μηνυμάτων. Τέλος προτείνεται η βαθιά μάθηση και η βαθιά αντίπαλη μάθηση ως μελλοντικές τεχνικές που μπορούν να χειριστούν αποτελεσματικά την απειλή των spam email.

Στην επόμενη εργασία[8] αναλύεται η σημασία των email στην σύγχρονη κοινωνία και αναφέρει το πρόβλημα που έχει δημιουργηθεί λόγω του τεράστιου όγκου

καθημερινών ανεπιθύμητων μηνυμάτων. Στην συνέχεια, εξετάζεται μια αποτελεσματική προσέγγιση φιλτραρίσματος email που βασίζεται σε σημασιολογικές μεθόδους. Η προτεινόμενη προσέγγιση χρησιμοποιεί την οντολογία του WordNet και εφαρμόζει διαφορετικές σημασιολογικές μεθόδους και μέτρα ομοιότητας για τη μείωση του τεράστιου αριθμού των εξαγόμενων χαρακτηριστικών κειμένου, με αποτέλεσμα να μειώνεται η πολυπλοκότητα χώρου και χρόνου. Επιπλέον, για να αποκτηθεί το σύνολο των ελάχιστων βέλτιστων δυνατοτήτων, η μείωση της διάστασης των χαρακτηριστικών ενσωματώθηκε χρησιμοποιώντας τεχνικές επιλογής χαρακτηριστικών, όπως η Ανάλυση Κύριων Συστατικών (PCA) και η Επιλογή Χαρακτηριστικών Συσχετίσεων (CFS). Η προτεινόμενη μέθοδος περιλαμβάνει 2 φάσεις, την κατάρτιση και τις δοκιμές. Η φάση της κατάρτισης αποτελείται από 4 κύριες ενότητες(προεπεξεργασία, στάθμιση χαρακτηριστικών, μείωση και ταξινόμηση χαρακτηριστικών). Η φάση των δοκιμών αποτελείται από την προεπεξεργασία τη στάθμιση χαρακτηριστικών και την ταξινόμηση.

Τα πειραματικά αποτελέσματα στο βασικό σημείο αναφοράς Enron Dataset έδειξαν ότι η προτεινόμενη προσέγγιση σημασιολογικού φιλτραρίσματος σε συνδυασμό με την επιλογή χαρακτηριστικών επιτυγχάνει υψηλή υπολογιστική απόδοση σε υψηλούς ρυθμούς μείωσης χώρου και χρόνου. Μια συγκριτική μελέτη για διάφορους αλγόριθμους ταξινόμησης έδειξε ότι ο αλγόριθμος Logistic Regression επιτυγχάνει την υψηλότερη ακρίβεια σε σύγκριση με τα δίκτυα Naïve Bayes, Support Vector Machine, J48, Random Forest και radial basis. Με την ενσωμάτωση της τεχνικής επιλογής χαρακτηριστικών CFS, η μέση καταγεγραμμένη ακρίβεια για όλους τους χρησιμοποιημένους αλγόριθμους είναι πάνω από 90%, με περισσότερο από 90% μείωση χαρακτηριστικών. Επιπλέον, τα πειράματα που πραγματοποιήθηκαν έδειξαν ότι η προτεινόμενη εργασία έχει πολύ σημαντική απόδοση με μεγαλύτερη ακρίβεια και λιγότερο χρόνο σε σύγκριση με άλλα σχετικά έργα.

Στην επόμενη εργασία[9] προς ανάλυση προτείνεται μια νέα μέθοδος ανίχνευσης η οποία συνδυάζει τον αλγόριθμο Artificial Bee Colony (αλγόριθμος τεχνητής αποικίας μελισσών) με ένα μοντέλο ταξινόμησης Logistic Regression. Τα αποτελέσματα που έχουν προκύψει είναι διαθέσιμα σε τρία δημόσια διαθέσιμα datasets (Enron, CSDMC2010, TurkishEmail) και δείχνουν πως το προτεινόμενο μοντέλο μπορεί να διαχειριστεί δεδομένα υψηλών διαστάσεων χάρη στις εξαιρετικά αποτελεσματικές δυνατότητες τοπικής αλλά και παγκόσμιας αναζήτησης.

Συγκρίνοντας την απόδοση του παραπάνω μοντέλου με εκείνα που χρησιμοποι-

ούν vector machine, logistic regression και naive Bayes ταξινομητές, αλλά και με γνωστές μεθόδους που έχουν αναπτυχθεί σε προηγούμενες μελέτες παρατηρείται ότι τα ποσοστά ακρίβειας ταξινόμησης είναι μεγαλύτερα στο μοντέλο που προτάθηκε σε αυτό το άρθρο και επιτυγχάνεται σημαντική πρόοδος, με ποσοστό επιτυχίας 99,25% στο σύνολο δεδομένων TurkishEmail και ποσοστό επιτυχίας 98,70% στο σύνολο δεδομένων CSDMC2010. Ο κύριος περιορισμός της προτεινόμενης μεθόδου είναι ότι έχει υψηλή υπολογιστική πολυπλοκότητα σε σύγκριση με άλλους αλγόριθμους.

Σκοπός της επόμενης εργασίας[10] είναι η διερεύνηση της χρήσης μιας σημασιολογικής ανάλυσης με σκοπό να βελτιωθεί η ακρίβεια εντοπισμού ανεπιθύμητης αλληλογραφίας. Προτείνεται μια μέθοδος βασισμένη στην ανάλυση 2 σημασιολογικών επιπέδων. Στο πρώτο επίπεδο κατηγοριοποιούνται τα e-mail με βάση τον τομέα στον οποίο ανήκουν. Για να επιτευχθεί αυτό δημιουργήθηκε ένας εποπτευόμενος ταξινομητής με επιλεγμένο λεξιλόγιο. Με αυτόν τον τρόπο είναι πολύ πιο αποτελεσματική η ανίχνευση ανεπιθύμητης αλληλογραφίας. Στο δεύτερο επίπεδο χρησιμοποιείται ένα υβριδικό σχήμα που αποτελείται από σημασιολογικά χαρακτηριστικά τα οποία καθορίζονται χειροκίνητα και εξάγονται αυτόματα για την ανίχνευση ανεπιθύμητων email ανά τομέα. Αυτές οι λειτουργίες έχουν ως σκοπό να συνοψίσουν το περιεχόμενο των email σε κατηγορίες διαχωρίζοντας τα ανεπιθύμητα μηνύματα από τα επιθυμητά. Για την επικύρωση αυτής της προσέγγισης χρησιμοποιήθηκαν εκτεταμένα πειράματα σε ένα μεγάλο αριθμό email που αποτελούνται από διαφορετικούς τομείς π.χ. υπολογιστής, ενήλικας εκπαίδευση κ.ο.κ. Στα πλαίσια αξιολόγησης δημιουργήθηκε ένα σύνολο δεδομένων δοκιμής χρησιμοποιώντας μηνύματα που έχουν συλλεχθεί από διάφορες δημόσιες πηγές και ορισμένα εξειδικευμένα φόρουμ.

Το Ling-spam corpus περιέχει μηνύματα που συλλέγονται από μια λίστα αλληλογραφίας για την επιστήμη και το επάγγελμα της γλωσσολογίας. Το σώμα περιλαμβάνει 2893 μηνύματα, εκ των οποίων τα 2412 είναι νόμιμα και τα 481 είναι ανεπιθύμητα. Όμως οι βάσεις δεδομένων Enron και Ling-spam δεν καλύπτουν όλους τους τομείς ανεπιθύμητων μηνυμάτων οπότε συλλέχθηκαν μηνύματα από εξειδικευμένα φόρουμ. Χρησιμοποιήθηκαν συνολικά 8188 email που διανέμονται σύμφωνα με το περιεχόμενό τους στις έξι προεπιλεγμένες κατηγορίες: Υγεία, Οικονομία Ενηλίκων, Εκπαίδευση, Υπολογιστής κ.ο.κ. Κάθε κατηγορία περιλαμβάνει email spam και ham(όχι spam). Για να επεκταθεί περαιτέρω η αξιολόγηση δοκιμάστηκε και σε ένα άλλο σύνολο δεδομένων το οποίο ονομάζεται CSDMC2010 SPAM. Το σύνολο δεδομένων περιέχει συνολικά 4327 μηνύματα ηλεκτρονικού ταχυδρομείου με ετικέτα όπου 1378 είναι

spam και 2949 είναι ham.

	Health	Adult	Finance	Education	Computer	Others
Ham	753	466	593	1075	808	1011
Spam	1061	563	465	254	861	278
Total	1814	1029	1058	1329	1669	1289

Figure 6: Αποτελέσματα ανα κατηγορία

Τα πειραματικά αποτελέσματα έδειξαν ότι η συγκεκριμένη προσέγγιση ξεπερνά πολλές σύγχρονες μεθόδους βασισμένες σε BoW(bag of words) και σε λανθάνουσα σημασιολογική ανάλυση.

Στην επόμενη εργασία[11] προτείνεται ένα νέο πλαίσιο το οποίο συνδυάζει τα νευρωνικά συστήματα με την ενίσχυση εκμάθησης για την ανίχνευση επιθέσεων phishing για πρώτη φορά σε online λειτουργία. Το προτεινόμενο μοντέλο έχει την δυνατότητα να προσαρμοστεί και να παράγει ένα νέο σύστημα ανίχνευσης phishing στα μηνύματα ηλεκτρονικού ταχυδρομείου το οποίο αντικατοπτρίζει τις αλλαγές στις συμπεριφορές που έχουν εξερευνηθεί πρόσφατα το οποίο επιτυγχάνεται υιοθετώντας την ιδέα της ενίσχυσης της μάθησης για τη δυναμική ενίσχυση του συστήματος με την πάροδο του χρόνου. Επίσης το προτεινόμενο μοντέλο επιλύει το πρόβλημα του περιορισμένου συνόλου δεδομένων προσθέτοντας αυτόματα περισσότερα μηνύματα ηλεκτρονικού ταχυδρομείου.

Ένας νέος αλγόριθμος προτείνεται για να διερευνήσει τυχόν νέες συμπεριφορές ηλεκτρονικού “ψαρέματος” στο νέο σύνολο δεδομένων. Μέσω αυστηρών δοκιμών χρησιμοποιώντας τα γνωστά σύνολα δεδομένων, αποδεικνύεται πως η προτεινόμενη τεχνική μπορεί να χειριστεί επιθέσεις ηλεκτρονικού ψαρέματος με υψηλά επίπεδα απόδοσης επιτυγχάνοντας υψηλή ακρίβεια, TPR και TNR στο 98.63%, 99,07% και 98.19% αντίστοιχα. Η σύγκριση με άλλες παρόμοιες τεχνικές στο ίδιο σύνολο δεδομένων δείχνει ότι το προτεινόμενο μοντέλο ξεπερνά τις υπάρχουσες μεθόδους. Τα πειράματα διεξήχθησαν χρησιμοποιώντας έναν συνδυασμό συνόλων δεδομένων από τρία διαθέσιμα στο κοινό δεδομένα, δύο από τα οποία είναι σύνολα δεδομένων email και ένα είναι σύνολο URL phishing. Παρακάτω βρίσκονται τα αποτελέσματα της

προτεινόμενης μεθόδου:

FNR	FPR	TPR	TNR	ACC	Precision	Recall	F_Measure	AUC
0.93%	1.81%	99.07%	98.19%	98.63%	98.21%	99.07%	98.64%	99.43%

Figure 7: Αποτελέσματα προτεινόμενης μεθόδου

Στην επόμενη εργασία προς ανάλυση[12], προτείνεται ένα νέο σύνολο γενικών περιγραφικών χαρακτηριστικών που εξάγονται από όλα τα στοιχεία ηλεκτρονικού ταχυδρομείου (κεφαλίδα, σώμα και συνημμένα) για βελτιωμένη ανίχνευση κακόβουλων μηνυμάτων ηλεκτρονικού ταχυδρομείου χρησιμοποιώντας μεθόδους μηχανικής εκμάθησης. Τα προτεινόμενα χαρακτηριστικά εξάγονται μόνο από το ίδιο το email. Για την διαδικασία εξαγωγής δεν απαιτείται σύνδεση στο διαδίκτυο, χρήση εξωτερικών υπηρεσιών ή άλλων εργαλείων, ικανοποιώντας έτσι τις ανάγκες των συστημάτων ανίχνευσης σε πραγματικό χρόνο. Χρησιμοποιώντας μια συλλογή από 33.142 emails που περιέχει 38,73% κακόβουλα και 61,27% καλοήγη email πραγματοποιήθηκε μια εκτενή αξιολόγηση των νέων καινοτόμων χαρακτηριστικών της προτεινόμενης μεθόδου σε σχέση με σύνολα χαρακτηριστικών που προτάθηκαν από προηγούμενες ακαδημαϊκές εργασίες. Τα αποτελέσματα της μελέτης έδειξαν ότι τα κακόβουλα email μπορούν να εντοπιστούν αποτελεσματικά κατά τη χρήση των νέων χαρακτηριστικών με αλγόριθμους μηχανικής μάθησης. Επιπλέον, αυτά τα νέα χαρακτηριστικά βελτιώνουν τον εντοπισμό κακόβουλων μηνυμάτων ηλεκτρονικού ταχυδρομείου όταν χρησιμοποιούνται σε συνδυασμό με λειτουργίες που προτείνονται από σχετική εργασία. Ο ταξινομητής Random Forest πέτυχε τα υψηλότερα ποσοστά ανίχνευσης με πραγματικό θετικό ρυθμό (TPR) 0,947 και ψευδώς θετικό ρυθμό (FPR) 0,03. Τέλος παρουσιάζεται επίσης το IDR (ενσωματωμένο ποσοστό ανίχνευσης), ένα νέο μέτρο που βοηθά στη βαθμολογηση του ορίου ενός ταξινομητή μηχανικής μάθησης, προκειμένου να επιτευχθούν τα βέλτιστα ποσοστά TP και FP.

Στην επόμενη εργασία[13] απεικονίζεται μια έρευνα από διαφορετικές μεθόδους φιλτραρίσματος e-mail με βάση τεχνικές machine learning όπως ο ταξινομητής

Naive Bayes, K-nearest Neighbor, Baeyes Additive Regression, KNN tree κοκ. Επίσης παρουσιάζεται η ταξινόμηση, αξιολόγηση και σύγκριση διαφορετικών συστημάτων φιλτραρίσματος ανεπιθύμητων μηνυμάτων και συνοψίζεται το συνολικό σενάριο σχετικά με το ποσοστό ακριβείας διαφορετικών προσεγγίσεων. Με την πάροδο του χρόνου, εξακολουθεί να είναι η ταξινόμηση ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου ένας από τους σημαντικότερους τομείς έρευνας για τη γεφύρωση των κενών. Μπορεί πολλές μέθοδοι να έχουν αποδειχθεί αποτελεσματικές αλλά υπάρχουν πολλά κενά στα συστήματα φιλτραρίσματος spam γεγονός το οποίο ανησυχεί την επιστημονική κοινότητα που προσπαθεί συνεχώς να εξελίξει και να βελτιώσει την διαδικασία φιλτραρίσματος ανεπιθύμητων μηνυμάτων.

Στον παρακάτω πίνακα αναγράφονται μερικές από τις υπάρχουσες προσεγγίσεις ταξινόμησης ανεπιθύμητων μηνυμάτων με χρήση τεχνικών μηχανικής μάθησης:

Sr. No.	Author	Algorithms	Corpus or Datasets	Accuracy/ Performance
1	Mohammed et al.	Naive Bayes, SVM, KNN, Decision Tree, Rules	Email-1431	85.96% Accuracy Achieved
2	Subramaniam et al.	Naive Bayesian	Collection of spam emails from Google's Gmail Account	96.00% Accuracy Achieved
3	Sharma et al.	Various Machine Learning Algorithms Adaptions	SPAMBASE	94.28% Accuracy Achieved
4	Banday et al.	Naive Bayes, K-Nearest Neighbor, SVM, classification Bayes Additive Regression Tree	Real life data set	96.69% Accuracy Achieved
5	Awad et al.	Naive Bayes, SVM, k-Nearest Neighbor, Artificial Neural Networks, Rough Sets	Spam Assassin	99.46% Accuracy Achieved
6	Chhabra et al.	Nonlinear SVM classifier.	Enron dataset	For Dataset 3, spam: real, the ratio is 1:3, for satisfactory Recall and Precision Values
7	Tretyakov	Bayesian classification, k-NN, ANNs, SVMs	PU1 corpus	94.4% Accuracy Achieved
8	Shahi et al.	Naive Bayes, SVM	Nepali SMS	92.74% Accuracy Achieved
9	Kaul et al.	SVM	Sample emails	90% ~ 95% Accuracy Achieved
10	Suganya et al.	Rule Baseed Method	Online Social Networks (OSNs) user post	Excellence Accuracy for Given Datasets

Figure 8: Ταξινόμηση spam με τεχνικές μηχανικής μάθησης

Το αυτόματο φιλτράρισμα ανεπιθύμητων μηνυμάτων είναι ένας πολλά υποσχόμενος και άξιος ερευνητικά τομέας όπου έχουν αναφερθεί εκτεταμένα έργα για την καταλογογράφηση των ανεπιθύμητων e-mail αλλά καμία από τις μεθοδολογίες δεν εγγυάται ολοκληρωμένες λύσεις. Λόγω της ταχείας επέκτασης των ψηφιακών δεδομένων, η ανακάλυψη γνώσεων και η εξόρυξη δεδομένων έχουν προκαλέσει μεγάλη προσοχή με μια άμεση ανάγκη να μετατραπούν αυτά τα δεδομένα σε χρήσιμες πληροφο-

ρίες και γνώσεις. Στην επόμενη εργασία που αναλύεται[14] οι συγγραφείς έχουν επικεντρωθεί στον τρόπο με τον οποίο οι επικοινωνίες μέσω e-mail επηρεάζονται από τα ανεπιθύμητα μηνύματα και επικεντρώνονται σε διάφορες τεχνικές εξόρυξης δεδομένων που βασίζονται στην ταξινόμηση σε ένα σύνολο δεδομένων ανεπιθύμητης αλληλογραφίας για τον προσδιορισμό των ανεπιθύμητων μηνυμάτων και επιθυμητών, για την ανάλυση της απόδοσης όλων των ταξινομητών και τον εντοπισμό των καλύτερων εξ αυτών όσον αφορά προς την εκτέλεση. Για την εκτέλεση του σκοπού της εργασίας, ένα εργαλείο εξόρυξης δεδομένων ανοιχτού κώδικα WEKA έχει χρησιμοποιήσει για να διερευνήσει την ανάλυση απόδοσης των διαφόρων ταξινομητών και, τέλος, ο υπερ-θετικός ταξινομητής έχει εντοπίσει την ταξινόμηση των ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου και έχει αναπτύξει το μοντέλο ροής γνώσεων.

Η επόμενη εργασία που αναλύσαμε[15] προτείνει ένα μοντέλο εντοπισμού ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου που βασίζεται σε ένα σύνολο χαρακτηριστικών βασισμένα στο περιεχόμενο, στα συναισθήματα, στη σημασιολογία, σε χαρακτηριστικά χρηστών και λεξικών spam. Στόχος της συγκεκριμένης έρευνας είναι η εκμετάλλευση του ρόλου των χαρακτηριστικών συναισθημάτων μαζί με άλλα προτεινόμενα χαρακτηριστικά για την αξιολόγηση της ακρίβειας ταξινόμησης των αλγορίθμων μηχανικής μάθησης για την ανίχνευση ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου. Οι υπάρχουσες μελέτες εκμεταλλεύονται κατά κύριο λόγο την προσέγγιση της μηχανικής χαρακτηριστικών βάσει περιεχομένου. Ωστόσο, θεωρείται περιορισμένος αριθμός χαρακτηριστικών. Σε αυτό το πλαίσιο, αυτή η ερευνητική μελέτη προτείνει ένα κεντρικό χαρακτηριστικό πλαίσιο (FSEDM) βασισμένο σε υπάρχοντα και νέα χαρακτηριστικά του συνόλου δεδομένων email, τα οποία εξάγονται μετά την προεπεξεργασία. Στη συνέχεια, εφαρμόζονται ποικίλες εποπτευόμενες τεχνικές μάθησης στις προτεινόμενες δυνατότητες σε συνδυασμό με τεχνικές επιλογής χαρακτηριστικών, όπως information gain, gain ratio και Relief-F για την κατάταξη των πιο σημαντικών χαρακτηριστικών και την ταξινόμηση των μηνυμάτων ηλεκτρονικού ταχυδρομείου σε spam ή ham (όχι spam). Η ανάλυση και τα πειραματικά αποτελέσματα έδειξαν ότι το προτεινόμενο μοντέλο με ανάλυση συναισθημάτων είναι ανταγωνιστική προσέγγιση για την ανίχνευση ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου. Χρησιμοποιώντας το προτεινόμενο μοντέλο, το βαθύ νευρωνικό δίκτυο που εφαρμόστηκε με χαρακτηριστικά συναισθήματος ξεπέρασε τους άλλους ταξινομητές όσον αφορά την ακρίβεια ταξινόμησης έως και 97,2%.

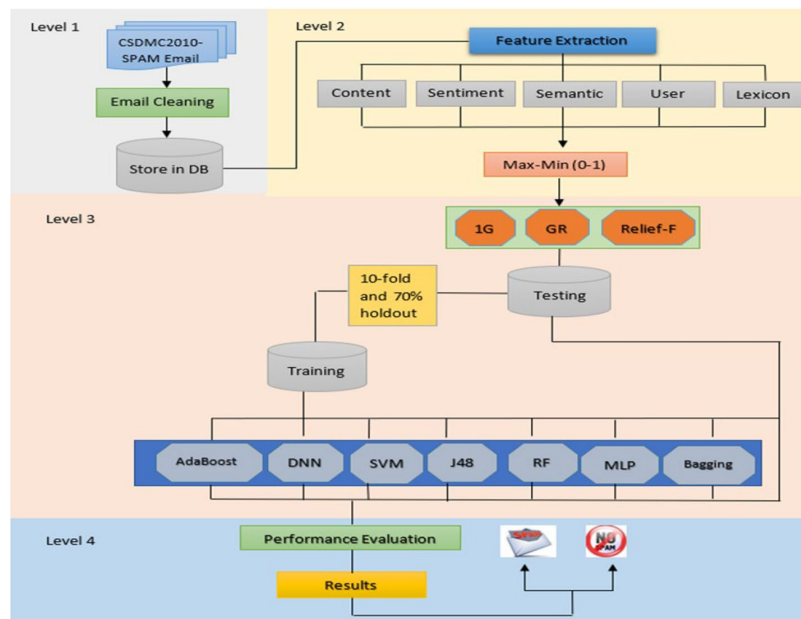


Figure 9: Προτεινόμενο μοντέλο

Κεφάλαιο 3

3.1 Μηχανική μάθηση

Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασισμένες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Η Μηχανική μάθηση εφαρμόζεται σε μια σειρά από υπολογιστικές εργασίες, όπου τόσο ο σχεδιασμός όσο και ο ρητός προγραμματισμός των αλγορίθμων είναι ανέφικτος. Το πιο χαρακτηριστικό παράδειγμα εφαρμογής μηχανικής μάθησης είναι τα φίλτρα spam στο ηλεκτρονικό ταχυδρομείο που τυχαίνει να ταυτίζεται και με την συγκεκριμένη έρευνα

3.2 Μεθοδολογία

Το εργαλείο το οποίο χρησιμοποιήθηκε στην συγκεκριμένη εργασία για την

διεξαγωγή συμπερασμάτων είναι το WEKA. Το WEKA είναι ένα από τα πιο διαδεδομένα λογισμικά Εξόρυξης Δεδομένων. Έχει χρησιμοποιηθεί σε μεγάλο αριθμό επιστημονικών εργασιών, και αρκετά βιβλία Εξόρυξης Δεδομένων αναφέρονται σε αυτό. Η μεγάλη δημοφιλία του οφείλεται στα ειδικά χαρακτηριστικά του και στις δυνατότητες που προσφέρει.

Αναλυτικότερα το WEKA:

- Περιέχει αρκετά μεγάλη ποικιλία μεθόδων για κατηγοριοποίηση, παλινδρόμηση, ανάλυση συστάδων, και κανόνες συσχέτισης. Επίσης, παρέχει δυνατότητες για προεπεξεργασία των δεδομένων, καθώς και εργαλεία οπτικοποίησης.
- Είναι λογισμικό ανοικτού κώδικα. Αυτό σημαίνει ότι ο πηγαίος κώδικας είναι δημόσιως διαθέσιμος. Χρήστες με γνώσεις προγραμματισμού μπορούν να τροποποιούν και να εξελίσσουν τους αλγόριθμους.
- Είναι γραμμένο σε γλώσσα Java, γεγονός που το καθιστά ικανό να εγκαθίσταται σε διαφορετικές πλατφόρμες υλικού και λογισμικού.
- Διαθέτει γραφικό περιβάλλον εργασίας. Στο διαδίκτυο υπάρχει διαθέσιμη μεγάλη ποικιλία βιβλιοθηκών για μηχανική μάθηση και εξόρυξη δεδομένων. Ωστόσο, η χρήση τους απαιτεί τη συγγραφή κώδικα. Αντιθέτως, το γραφικό περιβάλλον του WEKA επιτρέπει τη χρήση του λογισμικού από τελικούς χρήστες, οι οποίοι δεν διαθέτουν γνώσεις προγραμματισμού.

Στην παρούσα έρευνα χρησιμοποιήθηκαν διάφορες τεχνικές μηχανικής μάθησης ώστε να βρεθούν τα σημαντικότερα χαρακτηριστικά που οδηγούν στην πρόβλεψη της σωστής κατηγορίας για το αν ένα email είναι spam η ham και να βρεθεί ο αλγόριθμος ο οποίος πετυχαίνει τα ακριβέστερα αποτελέσματα στην κατηγοριοποίηση των μηνυμάτων ηλεκτρονικού ταχυδρομείου. Οι ταξινομητές οι οποίοι χρησιμοποιήθηκαν για την διεξαγωγή των αποτελεσμάτων είναι οι J48, JRip, Random Forest, IBK και Naive Bayes. Η απόδοση των ταξινομητών αξιολογήθηκε με βάση το ποσοστό ακριβείας (precision), ευαισθησίας(recall), και ορθότητας(accuracy).

3.3 Σύνολο Δεδομένων

Το σύνολο δεδομένων το οποίο επιλέχτηκε προς ανάλυση είναι το Spambase

Data Set από το UCI Machine Learning Repository το οποίο περιλαμβάνει μια συλλογή από ανεπιθύμητα μηνύματα ηλεκτρονικού ταχυδρομείου. Το συγκεκριμένο σύνολο δεδομένων έχει διάφορους τύπους μεταβλητών, και τα 57 χαρακτηριστικά του έχουν real και integer τιμές.

Πιο αναλυτικά το dataset έχει:

- 48 χαρακτηριστικά με τιμές συνεχείς πραγματικούς $[0,100]$ του τύπου `word_freq_WORD` το οποίο αφορά το ποσοστό των συγκεκριμένων λέξεων μέσα στο e-mail.
- 6 χαρακτηριστικά με τιμές συνεχείς πραγματικούς $[0,100]$ του τύπου `char_freq_CHAR` το οποίο αναφέρει το ποσοστό των χαρακτήρων μέσα στο e-mail.
- 1 χαρακτηριστικό με τιμή συνεχή πραγματικό $[1,...]$ του τύπου `capital_run_length_average` το οποίο περιγράφει τον μέσο όρο των συνεχόμενων κεφαλαίων χαρακτήρων στη σειρά μέσα στο e-mail.
- 1 χαρακτηριστικό με τιμή συνεχή πραγματικό $[1,...]$ του τύπου `capital_run_length_longest` το οποίο περιγράφει το μήκος των μεγαλύτερων συνεχόμενων κεφαλαίων χαρακτήρων στη σειρά μέσα στο e-mail.
- 1 χαρακτηριστικό με τιμή συνεχή πραγματικό του τύπου `capital_run_length_total` το οποίο περιγράφει το σύνολο των κεφαλαίων χαρακτήρων μέσα στο e-mail.
- 1 χαρακτηριστικό με όνομαστική τιμή του τύπου `spam` το οποίο δηλώνει αν το email έχει χαρακτηρηστεί ως spam ή όχι.

Στον παρακάτω πίνακα μπορούμε να δούμε το σύνολο των χαρακτηριστικών μας

Table 1: Attributes

ATTRIBUTE	MIN	MAX	MEAN	STDDEV
#word_freq_make	0	4,54	0,10	0,30
#word_freq_address	0	14,28	0,21	1,29
#word_freq_all	0	5,10	0,28	0,50

#word_freq_3d	0	42,81	0,06	1,39
#word_freq_our	0	10	0,31	0,67
#word_freq_over	0	5,88	0,09	0,27
#word_freq_remove	0	7,27	0,11	0,39
#word_freq_internet	0	11,1	0,10	0,40
#word_freq_order	0	5,26	0,09	0,27
#word_freq_mail	0	18,18	0,24	0,64
#word_freq_receive	0	2,61	0,06	0,20
#word_freq_will	0	9,67	0,54	0,86
#word_freq_people	0	5,55	0,09	0,30
#word_freq_report	0	10	0,06	0,33
#word_freq_addresses	0	4,41	0,04	0,25
#word_freq_free	0	20	0,24	0,82
#word_freq_business	0	7,14	0,14	0,44
#word_freq_email	0	9,09	0,18	0,53
#word_freq_you	0	18,75	1,66	1,77
#word_freq_credit	0	18,18	0,08	0,51
#word_freq_your	0	11,11	0,81	1,20
#word_freq_font	0	17,10	0,12	1,02
#word_freq_000	0	5,45	0,10	0,35
#word_freq_money	0	12,5	0,09	0,44
#word_freq_hp	0	20,83	0,55	1,67
#word_freq_hpl	0	16,66	0,26	0,88
#word_freq_george:	0	33,33	0,76	3,36
#word_freq_650:	0	9,09	0,12	0,53
#word_freq_lab:	0	14,28	0,09	0,59
#word_freq_labs	0	5,88	0,10	0,45
#word_freq_telnet	0	12,50	0,06	0,40
#word_freq_857	0	4,76	0,04	0,32
#word_freq_data	0	8,18	0,09	0,55
#word_freq_415	0	4,76	0,04	0,32
#word_freq_85	0	20	0,10	0,53
#word_freq_technology	0	7,69	0,09	0,40
#word_freq_1999	0	6,89	0,13	0,42
#word_freq_parts	0	8,33	0,01	0,22
#word_freq_pm	0	11,11	0,07	0,43
#word_freq_direct	0	4,76	0,06	0,35
#word_freq_cs	0	7,14	0,04	0,36
#word_freq_meeting	0	14,28	0,13	0,76
#word_freq_original	0	3,57	0,04	0,22
#word_freq_project	0	20	0,07	0,62
#word_freq_re	0	21,42	0,30	1,01
#word_freq_edu	0	22,05	0,18	0,91
#word_freq_table	0	2,17	0,01	0,07
#word_freq_conference	0	10	0,03	0,28
#char_freq_;	0	4,38	0,03	0,24

#char_freq_(	0	9,75	0,13	0,27
#char_freq_[	0	4,081	0,017	0,10
#char_freq_!	0	32,47	0,26	0,81
#char_freq_\$	0	6,00	0,07	0,24
#char_freq_#	0	19,82	0,04	0,42
#capital_run_length_average	1	1102,5	5,19	31,72
#capital_run_length_longest	1	9989	52,17	194,89
#capital_run_length_total	1	15841	283,28	606,34
#isspam	COUNT	WEIGHT		
0	2788	2788		
1	1813	1813		

3.4 Προεπεξεργασία δεδομένων

Το στάδιο της προ-επεξεργασίας των δεδομένων περιλαμβάνει την επιλογή χαρακτηριστικών, το μετασχηματισμό των δεδομένων και την εκκαθάριση των δεδομένων. Προκειμένου να πειραματιστούμε με τα δεδομένα, αφού εισήχθησαν στο περιβάλλον του Weka επιλέχθηκαν κάποιες τεχνικές επιλογής χαρακτηριστικών.

Η πρώτη τεχνική που χρησιμοποιήθηκε ονομάζεται “RemoveWithValues” καθώς παρατηρήθηκε πως σχεδόν σε όλα τα χαρακτηριστικά πρέπει να μειωθεί το εύρος των τιμών ώστε να έχουμε ορθότερα και πιο ακριβή αποτελέσματα.

Στην συνέχεια η επόμενη τεχνική που χρησιμοποιήθηκε ονομάζεται “AttributeSelection”. Τα διαθέσιμα δεδομένα περιλαμβάνουν πολλά χαρακτηριστικά(στήλες). Ωστόσο, για μια συγκεκριμένη εργασία εξόρυξης, δεν είναι χρήσιμα όλα αυτά τα χαρακτηριστικά. Πολλά χαρακτηριστικά περιέχουν πληροφορίες που δεν σχετίζονται με το αντικείμενο της ανάλυσης. Επίσης, μπορεί να καταγράφονται διαφορετικές εκδοχές της ίδιας πληροφορίας σε διάφορα χαρακτηριστικά. Η εργασία επιλογής αυτών των δεδομένων αυτή είναι γνωστή ως επιλογή χαρακτηριστικών.

Μια άλλη τεχνική που χρησιμοποιήθηκε κατά την επεξεργασία δεδομένων ονομάζεται “SpreadSubsample” κατά την οποία παράγεται ένα υποσύνολο ενός συνόλου δεδομένων. Το αρχικό σύνολο δεδομένων πρέπει να χωράει εξ ολοκλήρου στη μνήμη. Αυτό το φίλτρο επιτρέπει να καθοριστεί η μέγιστη "εξάπλωση" μεταξύ της πιο σπάνιας

και της πιο κοινής κατηγορίας.

Αφού έχουμε περιορίσει το εύρος των τιμών των δεδομένων μας με την τεχνική “RemoveWithValues” εφαρμόζουμε την τεχνική(attribute selection) στο σύνολο των δεδομένων μας και παρατηρούμε πως από τα 58 χαρακτηριστικά που υπήρχαν συνολικά μόνο τα 17 παρέμειναν, που σημαίνει πως αυτά τα 17 περιέχουν τις σημαντικές πληροφορίες και δεδομένα που χρειαζόμαστε για να συνεχίσουμε το πείραμα.

Παρακάτω παρατίθεται μια οπτικοποίηση των χαρακτηριστικών μετά από την εφαρμογή των 2 παραπάνω τεχνικών.



Figure 10 Οπτικοποίηση χαρακτηριστικών

3.5 Ταξινόμηση

Στην συνέχεια προχωράμε στο στάδιο της ταξινόμησης με σκοπό να δούμε και

να συγκρίνουμε τα αποτελέσματα των αλγορίθμων των οποίων χρησιμοποιήσαμε. Συνολικά τρέξαμε τα δεδομένα 4 φορές. Την πρώτη φορά χωρίς καμία τεχνική επιλογής χαρακτηριστικών. Στην συνέχεια μία φορά με τεχνική “attributeSelection”. Την 3^η φορά με τεχνική “SpreadSubsample”, και τέλος με τεχνικές “attributeSelection” και “SpreadSubsample” μαζί. Οι αλγόριθμοι ταξινόμησης οι οποίοι χρησιμοποιήθηκαν είναι οι J48, JRip, Random Forest, Ibk και Naive Bayes. Η απόδοση των ταξινομητών αξιολογήθηκε με βάση το ποσοστό ακριβείας (precision), ευαισθησίας(recall), και ορθότητας(accuracy).

Η μέθοδος η οποία χρησιμοποιήθηκε από τους αλγορίθμους ταξινόμησης είναι η διασταυρούμενη επικύρωση 10 τμημάτων (10 Fold Cross Validation). Στη μέθοδο επικύρωσης 10 τμημάτων (10 fold cross validation) το σύνολο δεδομένων διαιρείται σε 10 υποσύνολα. Κάθε υποσύνολο περιέχει διαφορετικές παρατηρήσεις. Η επιλογή των υποσυνόλων είναι τυχαία. Ένα από τα υποσύνολα χρησιμοποιείται ως σύνολο επικύρωσης και τα υπόλοιπα εννέα συνενώνονται και δημιουργούν το σύνολο εκπαίδευσης. Το μοντέλο εκπαιδεύεται χρησιμοποιώντας το σύνολο εκπαίδευσης και δοκιμάζεται έναντι του συνόλου επικύρωσης. Η διαδικασία επαναλαμβάνεται δέκα φορές, κάθε φορά χρησιμοποιώντας ένα διαφορετικό σύνολο ως σύνολο επικύρωσης και τα υπόλοιπα εννέα ως σύνολο εκπαίδευσης. Στο τέλος υπολογίζεται η μέση επίδοση του μοντέλου. Η μέθοδος μπορεί να διαφοροποιηθεί ως προς το πλήθος των τμημάτων. Γενικότερα ονομάζεται μέθοδος επικύρωσης k τμημάτων, όπου k συμβολίζει τον αριθμό των δημιουργημένων υποσυνόλων και των επαναλήψεων.

3.6 Αλγόριθμοι ταξινόμησης

Δέντρα Απόφασης με τον C4.5(J48)

Ο J48 ανήκει στην κατηγορία των δέντρων αποφάσεων. Ένα δέντρο αποφάσεων αποτελείται από κόμβους αποφάσεων και κόμβους φύλλων, όπου ο κόμβος αποφάσεων καθορίζει μια δοκιμή για ένα από τα χαρακτηριστικά και ένας κόμβος φύλλων αντιπροσωπεύει την τιμή κλάσης. Κάθε μονοπάτι από τον ριζικό κόμβο προς τον κόμβο των φύλλων είναι κανόνας. Το σφάλμα ταξινόμησης είναι ο βασικός παράγοντας

απόδοσης για το δέντρο αποφάσεων. Στο δέντρο αποφάσεων όταν οι κατηγορίες κλάσεων είναι περισσότερες, η ακρίβεια ταξινόμησης μειώνεται σημαντικά. Πρώτα, ελέγχεται αν όλες οι περιπτώσεις ανήκουν στην ίδια κλάση. Τότε το δέντρο είναι φύλλο και φέρει την ετικέτα με αυτήν την κατηγορία. Στη συνέχεια για κάθε χαρακτηριστικό, υπολογίζονται οι πληροφορίες και το κέρδος πληροφοριών. Βρίσκεται το καλύτερο χαρακτηριστικό διαχωρισμού. Εξετάζει το κέρδος πληροφοριών (information gain – διαφορά στην εντροπία) που προκύπτει από την επιλογή ενός χαρακτηριστικού για το διαχωρισμό των δεδομένων. Το χαρακτηριστικό με το υψηλότερο ομαλοποιημένο κέρδος είναι αυτό που χρησιμοποιείται για να ληφθεί μια απόφαση. Ο αλγόριθμος επαναλαμβάνεται για μικρότερες υπολίστες δεδομένων[16].

JRip:

Ο αλγόριθμος αυτός περνά από 4 στάδια: ανάπτυξη κανόνα, κλάδεμα, βελτιστοποίηση και επιλογή. Τα δεδομένα εκπαίδευσης χωρίζονται σε ένα αναπτυσσόμενο σετ και ένα σετ κλαδέματος. Πρώτον, σχηματίζεται ένα αρχικό σύνολο κανόνων, χρησιμοποιώντας κάποια ευρετική μέθοδο. Αυτό το υπερβολικά μεγάλο σύνολο κανόνων απλουστεύεται στη συνέχεια επανειλημμένα εφαρμόζοντας από ένα σύνολο τελεστών κλαδέματος. Σε κάθε στάδιο απλοποίησης, ο επιλεγμένος τελεστής κλαδέματος είναι αυτός που αποδίδει τη μεγαλύτερη μείωση σφάλματος στο σετ κλάδεσης. Η απλοποίηση τελειώνει όταν εφαρμόζοντας οποιονδήποτε τελεστή κλαδέματος θα αυξήσει το σφάλμα στο κλάδεμα[17].

Naive Bayes:

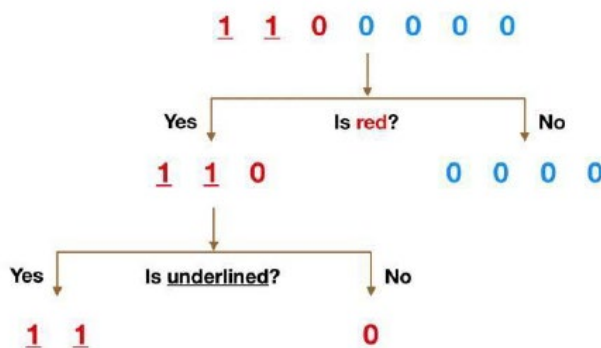
Ο Naïve Bayes για να κατηγοριοποιήσει τα δεδομένα του, χρησιμοποιεί τη θεωρία των πιθανοτήτων. Υπολογίζει την πιθανότητα κάποιο από τα δεδομένα να ανήκει σε μία κατηγορία. Βασίζεται στο θεώρημα του Bayes και κατ' επέκταση, κάνει χρήση δεσμευμένων πιθανοτήτων.

$$P(b|a) = \frac{P(a|b) \times P(b)}{P(a)}$$

Όπου $P(a)$ είναι η πιθανότητα του γεγονότος a και $P(a|b)$ είναι η δεσμευμένη πιθανότητα του a δεδομένου του γεγονότος b [18]

Random Forest:

Για να γίνει κατανοητός ο τρόπος λειτουργίας του αλγορίθμου αυτού πρέπει να εξηγήσουμε πώς λειτουργούν τα δέντρα απόφασης, καθώς είναι τα δομικά στοιχεία του Random Forest



11 Δένδρο απόφασης

Ας υποθέσουμε ότι το σύνολο δεδομένων μας αποτελείται από τους αριθμούς στο πάνω μέρος του σχήματος. Έχουμε δύο 1 και πέντε 0 (1 και 0 είναι οι κλάσεις) και επιθυμούμε να διαχωρίσουμε τις κλάσεις χρησιμοποιώντας τα χαρακτηριστικά τους. Τα χαρακτηριστικά είναι το χρώμα (κόκκινο έναντι μπλε) και αν το δείγμα είναι υπογραμμισμένο ή όχι. Το χρώμα μοιάζει ένα αρκετά προφανές χαρακτηριστικό για τον διαχωρισμό αλλά ένα από τα 0 είναι μπλε. Έτσι μπορούμε να χρησιμοποιήσουμε την ερώτηση, "Είναι κόκκινο;" για να χωρίσουμε τον πρώτο κόμβο. Ο κόμβος σε ένα δέντρο μοιάζει ως το σημείο όπου η διαδρομή χωρίζεται σε δύο - δείγματα που πληρούν τα κριτήρια τοποθετούνται κάτω από το κλαδί Ναι και αυτά που δεν πληρούν κάτω από το Όχι. Το κλαδί Όχι (τα μπλέ) είναι όλα 0 τώρα, έτσι έχουμε τελειώσει, αλλά το κλαδί Ναι μπορεί ακόμα να χωριστεί περαιτέρω. Τώρα μπορεί να χρησιμοποιηθεί το δεύτερο χαρακτηριστικό και να γίνει η ερώτηση "Είναι υπογραμμισμένο;" ώστε να γίνει ένα δεύτερο split. Τα δύο 1 που είναι υπογραμμισμένα πηγαίνουν κάτω από το δευτερεύον τμήμα Yes και το 0 που δεν είναι υπογραμμισμένο πηγαίνει κάτω από το σωστό υποσύστημα. Στην πραγματική ζωή τα δεδομένα δεν είναι τόσο καθαρά, αλλά η λογική που χρησιμοποιεί ένα δέντρο αποφάσεων παραμένει η ίδια.

Σε κάθε κόμβο, θα ρωτήσει -Ποιο χαρακτηριστικό θα μου επιτρέψει να χωρίσω τα δείγματα κατά τρόπο ώστε οι ομάδες που προκύπτουν να είναι όσο το δυνατόν πιο διαφορετικές μεταξύ τους (και τα μέλη κάθε προκύπτουσας υποομάδας να είναι όσο το

δυνατόν πιο παρόμοια μεταξύ τους). Ο Random Forest αποτελείται από μεγάλο αριθμό μεμονωμένων δέντρων απόφασης που λειτουργούν ως σύνολο. Κάθε μεμονωμένο δέντρο ξετυλίζει μια πρόβλεψη κλάσης ή αλλιώς δίνει μια κατηγοριοποίηση και η κλάση με τις περισσότερες ψήφους γίνεται η πρόβλεψη του μοντέλου μας. [19]

3.7 Αποτελέσματα

Στον παρακάτω πίνακα καταγράφονται τα αποτελέσματα των αλγορίθμων ταξινόμησης χωρίς την χρήση μεθόδων επιλογής χαρακτηριστικών για το σύνολο όλων των χαρακτηριστικών.

Στην συνέχεια ακολουθεί διάγραμμα το οποίο μας δείχνει το ποσοστό ορθότητας ανά αλγόριθμο ταξινόμησης.

Table 2: Ταξινόμηση χωρίς filters

	PRECISION	RECALL	ACCURACY
J48	91.60 %	92.50 %	91.00 %
	90.30 %	89.00 %	
RANDOM FOREST	94.90 %	96.60 %	95.20 %
	95.50 %	93.40 %	
NAÏVE BAYES	93.80 %	68.40 %	79.70 %
	69.90 %	94.20 %	
IBK	91.70 %	92.70 %	91.20 %
	90.50 %	89.30 %	
JRIP	92.70 %	92.90 %	91.90 %
	90.80 %	90.60 %	

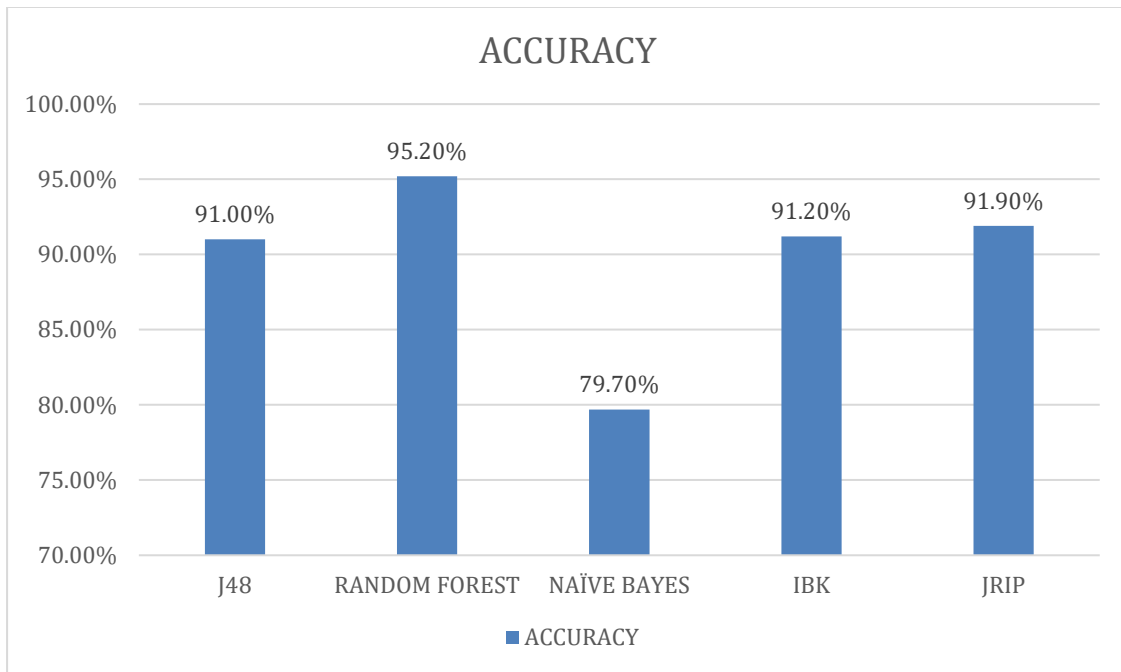


Figure 11: Accuracy Diagram χωρίς filters

Στον παρακάτω πίνακα καταγράφονται τα αποτελέσματα των αλγορίθμων ταξινόμησης μετά από την χρήση της μεθόδου επιλογής χαρακτηριστικών “AttributeSelection”.

Στην συνέχεια ακολουθεί διάγραμμα το οποίο μας δείχνει το ποσοστό ορθότητας ανά αλγόριθμο ταξινόμησης.

Table 3: Ταξινόμηση με Feature Selection

	PRECISION	RECALL	ACCURACY
J48	91.80 %	93.10 %	91.50 %
	91.00 %	89.30 %	
RANDOM FOR- EST	94.30 %	94.40 %	93.60 %
	92.80 %	92.60 %	
NAÏVE BAYES	91.90 %	91.90 %	90.90 %
	89.60 %	89.60 %	
IBK	92.50 %	90.80 %	90.70%
	88.50 %	90.60 %	
JRIP	91.70 %	93.70 %	91.70 %
	91.70 %	89.00 %	

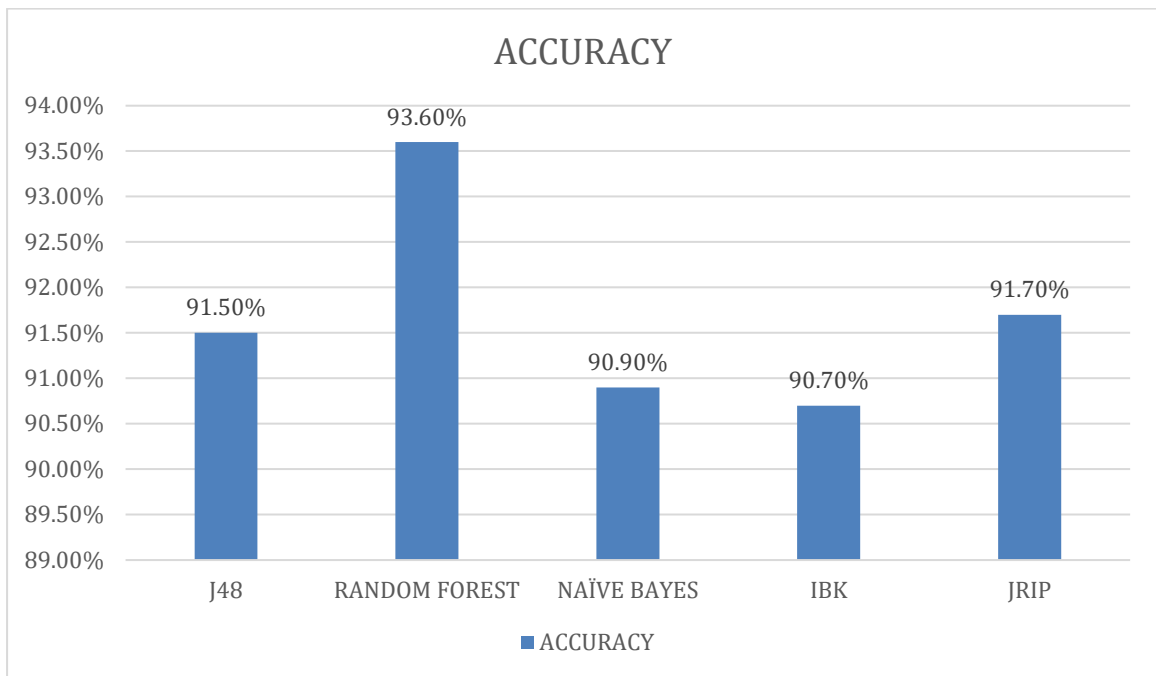


Figure 12: Accuracy Diagram with Feature Selection

Στον παρακάτω πίνακα καταγράφονται τα αποτελέσματα των αλγορίθμων ταξινόμησης μετά από την χρήση της μεθόδου επιλογής χαρακτηριστικών “ClassBalancing”.

Στην συνέχεια ακολουθεί διάγραμμα το οποίο μας δείχνει το ποσοστό ορθότητας ανά αλγόριθμο ταξινόμησης.

Table 4: Ταξινόμηση με Class Balancing

	PRECISION	RECALL	ACCURACY
J48	90.90 %	91.50 %	91.20 %
	91.40 %	90.90 %	
RANDOM FOR- EST	94.30 %	95.80 %	95.00 %
	95.70 %	94.20 %	
NAÏVE BAYES	92.10 %	68.60 %	81.40 %
	75.00 %	94.10 %	
IBK	90.00 %	91.30 %	90.60 %
	91.20 %	89.90 %	
JRIP	92.50 %	92.00 %	92.30 %
	92.10 %	92.50 %	

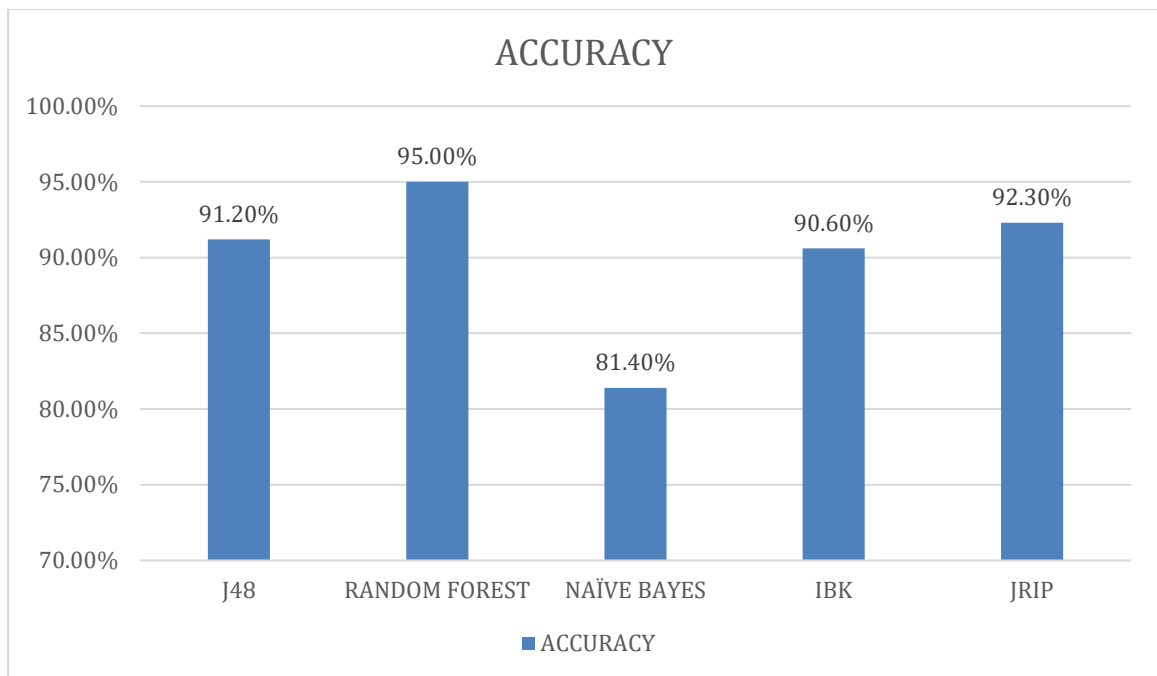


Figure 13: Accuracy Diagram with Class Balancing

Στον παρακάτω πίνακα καταγράφονται τα αποτελέσματα των αλγορίθμων ταξινόμησης μετά από την χρήση των μεθόδων επιλογής χαρακτηριστικών “AttributeSelection” και “ClassBalancing”.

Στην συνέχεια ακολουθεί διάγραμμα το οποίο μας δείχνει το ποσοστό ορθότητας ανά αλγόριθμο ταξινόμησης.

Table 5: Ταξινόμηση με Feature Selection & Class Balancing

ALGORITHM	PRECISION	RECALL	ACCURACY
J48	90.80 %	92.50 %	91.60 %
	92.30 %	90.60 %	
RANDOM FOR-EST	93.20 %	94.00 %	93.60 %
	93.90 %	93.20 %	
NAÏVE BAYES	89.80 %	90.90 %	90.30 %
	90.80 %	89.60 %	
IBK	90.90 %	90.40 %	90.70 %
	90.50 %	90.90 %	
JRIP	90.80 %	91.40%	91.00 %
	91.30 %	90.70 %	

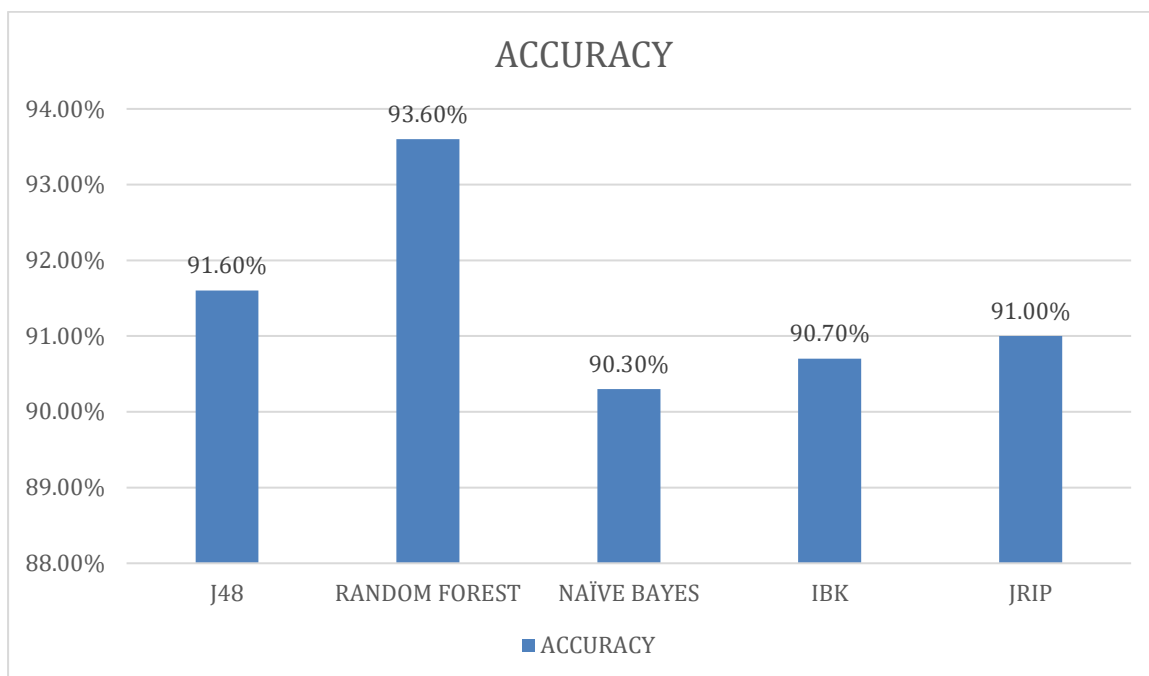


Figure 14: Accuracy Diagram with Feature selection & Class Balancing

Ενδεικτικό παράδειγμα 2 τυχαίων δέντρων από τον αλγόριθμο Random Forest

Appendix 1

Ενδεικτικό παράδειγμα JRip με χρήση μόνο των σημαντικών χαρακτηριστικών

Appendix 2

Κεφάλαιο 4

4.1 Συμπεράσματα - Συζήτηση

Οι τεχνικές μηχανικής μάθησης έχουν φέρει επανάσταση στον τρόπο ανάλυσης των δεδομένων καθώς και στην διαχείριση τεράστιου όγκου πληροφορίας. Με αυτό τον τρόπο η εξαγωγή συμπερασμάτων και πληροφορίας είναι πιο σύντομη αλλά και πιο ακριβής από ποτέ. Στην συγκεκριμένη εργασία αναλύθηκε η απόδοση 5 διαφορετικών αλγορίθμων οι οποίοι έδιναν διαφορετικά αποτελέσματα. Εφαρμόστηκαν διάφορες τεχνικές ώστε να βρεθούν τα κύρια χαρακτηριστικά που μπορούν να συνεισφέρουν στην εξαγωγή χρήσιμης και αξιόπιστης πληροφορίας. Συγκρίνοντας τους 5 αλγορίθμους ως προς την ακρίβεια, την ορθότητα και την ευαισθησία συμπεραίνουμε πως ο αλγόριθμος Random Forest αποδίδει καλύτερα από τους υπόλοιπους με καλύτερα ποσοστά ειδικά στην περίπτωση όπου δεν χρησιμοποιήθηκαν καθόλου τεχνικές επιλογής χαρακτηριστικών εκεί πετυχαίνει το υψηλότερο ποσοστά του. Πιο συγκεκριμένα ο αλγόριθμος Random Forest πετυχαίνει ποσοστό ορθότητας 95,20%, ανάκλησης 94,90% και ακρίβειας 96,60%. Στην δεύτερη θέση βρίσκεται ο JRip με ποσοστό ορθότητας 91,70%, ανάκλησης 92,90% και ακρίβειας 92,70% και στην τρίτη ο Ibk με ποσοστό ορθότητας 91,20%, ανάκλησης 92,70% και ακρίβειας 91,70%. Στις περιπτώσεις που εφαρμόσαμε τεχνικές επιλογής χαρακτηριστικών παραμένει ο Random Forest στην πρώτη θέση και παρατηρούμε πως ο JRip με τον J48 βρίσκονται πάρα πολύ κοντά.

REFERENCES

- [1] Amany A.Naem, Neveen I.Ghali,Afaf A.Saleh “Antlion optimization and boosting classifier for spam email detection”
<https://www.sciencedirect.com/science/article/pii/S2314728818300746>
- [2] M. K. Chae, Abeer Alsadoon, P.W.C. Prasad, Sasikumaran Sreedharan
“Spam filtering email classification (SFECM) using gain and graph mining algorithm” <https://ieeexplore.ieee.org/document/7905294>
- [3] Kavita Patel, Sanjay Kumar Dubey, “To recognize and analyze spam domains from spam emails by data mining” <https://ieeexplore.ieee.org/document/7725016>
- [4] Shubhi Shrivastava, R. Anju,“Spam mail detection through data mining techniques” <https://ieeexplore.ieee.org/document/8324021>
- [5] S. Nandhini , Jeen Marseline K.S. , “Performance Evaluation of Machine Learning Algorithms for Email Spam Detection” <https://ieeexplore.ieee.org/document/9077835>
- [6] Ismaila Idris, Ali Selamat, Sigeru Omatu, “Hybrid email spam detection model with negative selection algorithm and differential evolution”
<https://www.sciencedirect.com/science/article/abs/pii/S0952197613002352>
- [7] Emmanuel Gbenga Dada, Joseph Stephen Bassi, Haruna Chiroma, Shafi'i Muhammad Abdulhamid, Adebayo Olusola Adetunmbi, Opeyemi Emmanuel Ajibuwa
“Machine learning for email spam filtering: review, approaches and open research problems” <https://www.sciencedirect.com/science/article/pii/S2405844018353404>
- [8] Eman M.Bahgat,Sherine Rady, Walaa Gad, Ibrahim F.Moawad
“Efficient email classification approach based on semantic methods”
<https://www.sciencedirect.com/science/article/pii/S2090447918300455>

[9] Bilge KaganDedeturk, BahriyeAkay, “Spam filtering using a logistic regression model trained by an artificial bee colony algorithm”

<https://www.sciencedirect.com/science/article/pii/S1568494620301691>

[10] Nadjate Saidani, Kamel Adi, Mohand Saïd Allili , “A semantic-based classification approach for an enhanced spam detection”

<https://www.sciencedirect.com/science/article/pii/S0167404820300043>

[11] Sami Smadi, Nauman Aslam, LiZhang “Detection of online phishing email using dynamic evolving neural network based on reinforcement learning”

<https://www.sciencedirect.com/science/article/pii/S0167923618300010>

[12] Aviad Cohen,Nir Nissim, Yuval Elovici, “Novel set of general descriptive features for enhanced detection of malicious emails using machine learning methods”

<https://www.sciencedirect.com/science/article/pii/S0957417418303312>

[13] Hanif Bhuiyan, Akm Ashiquzzaman, Tamanna Islam Juthi, Suzit Biswas & Jinat Ara, “A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques “ <https://core.ac.uk/download/pdf/231154948.pdf>

[14] Soumyabrata Saha, Suparna DasGupta, Suman Kumar Das, “Spam Mail Detection Using Data Mining: A Comparative Analysis”

https://link.springer.com/chapter/10.1007/978-981-13-1921-1_56

[15] Ammara Zamir, Hikmat Ullah Khan, Waqar Mehmood, Tassawar Iqbal, Abubakker Usman Akram, “A feature-centric spam email detection model using diverse supervised machine learning algorithms”

<https://www.emerald.com/insight/content/doi/10.1108/EL-07-2019-0181/full/html>

[16] [17] [18] [19] Επιχειρηματική Ευφυΐα & Εξόρυξη Δεδομένων (Ευστάθιος Γ. Κύρκος)

Appendix 1

RandomTree

```
word_freq_credit < 0.1
| capital_run_length_longest < 9.5
| | word_freq_order < 0.34
| | | word_freq_email < 0.3
| | | | word_freq_your < 1.09
| | | | | word_freq_make < 0.06
| | | | | word_freq_hp < 0.7
| | | | | | word_freq_remove < 0.23
| | | | | | word_freq_free < 1.41
| | | | | | word_freq_will < 2.16
| | | | | | | char_freq_ < 0.02
| | | | | | | | word_freq_1999 < 0.14
| | | | | | | | word_freq_will < 2.05
| | | | | | | | | word_freq_re < 0.36
| | | | | | | | | word_freq_you < 1.11
| | | | | | | | | | word_freq_george < 0.53
| | | | | | | | | | word_freq_re < 0.14
| | | | | | | | | | | word_freq_will < 0.24
| | | | | | | | | | | word_freq_hp < 0.48
| | | | | | | | | | | char_freq_(< 0.71
| | | | | | | | | | | | char_freq_(< 0.01
| | | | | | | | | | | | | char_freq_! < 1.49
| | | | | | | | | | | | | | word_freq_project < 0.36
| | | | | | | | | | | | | | capital_run_length_total < 36.5
| | | | | | | | | | | | | | capital_run_length_longest < 4.5
| | | | | | | | | | | | | | | char_freq_! < 0.11
| | | | | | | | | | | | | | | capital_run_length_longest < 3.5
| | | | | | | | | | | | | | | capital_run_length_average < 1.06
| | | | | | | | | | | | | | | | capital_run_length_total < 6.5 : 0 (49/0)
| | | | | | | | | | | | | | | | capital_run_length_total >= 6.5
| | | | | | | | | | | | | | | | | capital_run_length_total < 8 : 1 (1/0)
| | | | | | | | | | | | | | | | | capital_run_length_total >= 8 : 0 (1/0)
| | | | | | | | | | | | | | | | | capital_run_length_average >= 1.06 : 0 (21/0)
| | | | | | | | | | | | | | | | | capital_run_length_longest >= 3.5
| | | | | | | | | | | | | | | | | word_freq_people < 0.81
| | | | | | | | | | | | | | | | | capital_run_length_average < 1.41 : 1 (1/0)
| | | | | | | | | | | | | | | | | capital_run_length_average >= 1.41 : 0 (12/0)
| | | | | | | | | | | | | | | | | word_freq_people >= 0.81 : 0 (3/0)
| | | | | | | | | | | | | | | | | | char_freq_! >= 0.11 : 0 (2/0)
| | | | | | | | | | | | | | | | | | capital_run_length_longest >= 4.5 : 0 (26/0)
| | | | | | | | | | | | | | | | | | capital_run_length_total >= 36.5 : 1 (1/0)
| | | | | | | | | | | | | | | | | | word_freq_project >= 0.36 : 0 (2/0)
| | | | | | | | | | | | | | | | | | char_freq_! >= 1.49
| | | | | | | | | | | | | | | | | | capital_run_length_longest < 2.5 : 1 (2/0)
| | | | | | | | | | | | | | | | | | capital_run_length_longest >= 2.5 : 0 (1/0)
| | | | | | | | | | | | | | | | | | char_freq_(>= 0.01 : 0 (17/0)
| | | | | | | | | | | | | | | | | | char_freq_(>= 0.71
| | | | | | | | | | | | | | | | | | | capital_run_length_total < 19.5 : 0 (4/0)
| | | | | | | | | | | | | | | | | | | capital_run_length_total >= 19.5 : 1 (2/0)
| | | | | | | | | | | | | | | | | | | word_freq_hp >= 0.48 : 1 (1/0)
| | | | | | | | | | | | | | | | | | | word_freq_will >= 0.24 : 0 (10/0)
| | | | | | | | | | | | | | | | | | | word_freq_re >= 0.14 : 0 (2/0)
| | | | | | | | | | | | | | | | | | | word_freq_george >= 0.53 : 0 (43/0)
| | | | | | | | | | | | | | | | | | | word_freq_you >= 1.11
| | | | | | | | | | | | | | | | | | | word_freq_free < 0.3
| | | | | | | | | | | | | | | | | | | | capital_run_length_total < 11.5
| | | | | | | | | | | | | | | | | | | | capital_run_length_total < 10.5
| | | | | | | | | | | | | | | | | | | | | char_freq_! < 0.33
| | | | | | | | | | | | | | | | | | | | | capital_run_length_longest < 1.5
| | | | | | | | | | | | | | | | | | | | | word_freq_edu < 1.39
| | | | | | | | | | | | | | | | | | | | | word_freq_george < 4.54
```

```

| word_freq_you < 5.15 : 1 (2/0)
| word_freq_you >= 5.15 : 0 (4/0)
| word_freq_george >= 4.54 : 0 (2/0)
| word_freq_edu >= 1.39 : 0 (1/0)
| capital_run_length_longest >= 1.5 : 0 (6/0)
| char_freq_! >= 0.33 : 0 (1/0)
| capital_run_length_total >= 10.5
| capital_run_length_average < 1.19 : 0 (1/0)
| capital_run_length_average >= 1.19 : 1 (3/0)
| capital_run_length_total >= 11.5
| word_freq_all < 0.18
| capital_run_length_total < 16.5 : 0 (21/0)
| capital_run_length_total >= 16.5
| word_freq_you < 1.29
| char_freq_! < 0.1
| word_freq_meeting < 0.6 : 1 (4/0)
| word_freq_meeting >= 0.6 : 0 (1/0)
| char_freq_! >= 0.1 : 0 (1/0)
| word_freq_you >= 1.29
| word_freq_re < 0.18
| capital_run_length_longest < 8.5 : 0 (31/0)
| capital_run_length_longest >= 8.5
| word_freq_george < 0.83
| capital_run_length_average < 2.29 : 0 (2/0)
| capital_run_length_average >= 2.29 : 1 (1/0)
| word_freq_george >= 0.83 : 0 (1/0)
| word_freq_re >= 0.18 : 1 (1/0)
| word_freq_all >= 0.18 : 0 (10/0)
| word_freq_free >= 0.3 : 0 (3/0)
| word_freq_re >= 0.36 : 0 (67/0)
| word_freq_will >= 2.05 : 1 (1/0)
| word_freq_1999 >= 0.14 : 0 (31/0)
| char_freq_ ; >= 0.02 : 0 (30/0)
| word_freq_will >= 2.16 : 0 (26/0)
| word_freq_free >= 1.41
| word_freq_re < 0.81
| word_freq_meeting < 0.93
| word_freq_free < 1.65 : 1 (3/0)
| word_freq_free >= 1.65 : 0 (1/0)
| word_freq_meeting >= 0.93 : 0 (1/0)
| word_freq_re >= 0.81 : 0 (2/0)
| word_freq_remove >= 0.23
| word_freq_free < 0.29 : 0 (2/0)
| word_freq_free >= 0.29 : 1 (3/0)
| word_freq_hp >= 0.7 : 0 (58/0)
| word_freq_make >= 0.06 : 0 (68/0)
| word_freq_your >= 1.09
| capital_run_length_total < 13.5
| word_freq_receive < 0.48
| word_freq_re < 0.79
| word_freq_all < 0.4
| word_freq_address < 0.81
| capital_run_length_average < 1.06 : 1 (3/0)
| capital_run_length_average >= 1.06
| word_freq_your < 4.75 : 0 (1/0)
| word_freq_your >= 4.75 : 1 (2/0)
| word_freq_address >= 0.81 : 1 (1/0)
| word_freq_all >= 0.4 : 0 (4/0)
| word_freq_re >= 0.79 : 0 (8/0)
| word_freq_receive >= 0.48 : 1 (3/0)
| capital_run_length_total >= 13.5
| word_freq_order < 0.11
| char_freq_$ < 0.02
| capital_run_length_total < 30.5
| word_freq_people < 0.7
| word_freq_re < 0.33
| word_freq_mail < 1.16 : 0 (21/0)

```



```

| | | word_freq_internet >= 0.03 : 0 (5/0)
| | | word_freq_hpl >= 0.07 : 0 (18/0)
| | | word_freq_000 >= 0.34
| | | | word_freq_business < 0.24 : 1 (8/0)
| | | | word_freq_business >= 0.24 : 0 (2/0)
| | | word_freq_1999 >= 0.3 : 0 (55/0)
| | | word_freq_85 >= 0.03 : 0 (53/0)
| | | word_freq_you >= 3.97
| | | | word_freq_people < 0.48 : 1 (6/0)
| | | | word_freq_people >= 0.48 : 0 (1/0)
| | | word_freq_our >= 1.67 : 1 (6/0)
| | char_freq_$ >= 0.1
| | | word_freq_000 < 0.19
| | | | word_freq_1999 < 1.15
| | | | | capital_run_length_average < 3.5 : 0 (8/0)
| | | | | capital_run_length_average >= 3.5 : 1 (11/0)
| | | | word_freq_1999 >= 1.15 : 0 (3/0)
| | | | word_freq_000 >= 0.19 : 1 (10/0)
| | word_freq_will >= 0.73
| | | word_freq_you < 2.08
| | | | word_freq_make < 0.02
| | | | | word_freq_1999 < 0.04
| | | | | | word_freq_hp < 0.09
| | | | | | | capital_run_length_longest < 33.5 : 0 (19/0)
| | | | | | | capital_run_length_longest >= 33.5
| | | | | | | | capital_run_length_longest < 62
| | | | | | | | | word_freq_re < 0.17 : 1 (1/0)
| | | | | | | | | word_freq_re >= 0.17 : 0 (2/0)
| | | | | | | | | capital_run_length_longest >= 62 : 1 (3/0)
| | | | | | | word_freq_hp >= 0.09 : 0 (46/0)
| | | | | word_freq_1999 >= 0.04 : 0 (56/0)
| | | | word_freq_make >= 0.02 : 0 (51/0)
| | | word_freq_you >= 2.08
| | | | word_freq_meeting < 0.41
| | | | | word_freq_receive < 0.41
| | | | | | word_freq_re < 0.43
| | | | | | | word_freq_remove < 0.5
| | | | | | | | capital_run_length_longest < 10.5 : 0 (2/0)
| | | | | | | | capital_run_length_longest >= 10.5
| | | | | | | | | word_freq_our < 0.95 : 0 (1/0)
| | | | | | | | | word_freq_our >= 0.95 : 1 (2/0)
| | | | | | | word_freq_remove >= 0.5 : 1 (3/0)
| | | | | | word_freq_re >= 0.43 : 0 (4/0)
| | | | | word_freq_receive >= 0.41 : 1 (2/0)
| | | | word_freq_meeting >= 0.41 : 0 (4/0)
| | word_freq_free >= 0.11
| | | word_freq_business < 0.14
| | | | word_freq_hp < 0.24
| | | | | word_freq_make < 0.03
| | | | | | char_freq_( < 0.11
| | | | | | | word_freq_edu < 0.22
| | | | | | | | word_freq_mail < 0.24
| | | | | | | | | word_freq_free < 0.14
| | | | | | | | | | word_freq_all < 0.44 : 1 (1/0)
| | | | | | | | | | word_freq_all >= 0.44 : 0 (2/0)
| | | | | | | | | word_freq_free >= 0.14 : 1 (12/0)
| | | | | | | | word_freq_mail >= 0.24 : 1 (4/0)
| | | | | | word_freq_edu >= 0.22 : 1 (5/0)
| | | | char_freq_( >= 0.11
| | | | | word_freq_will < 0.15
| | | | | | capital_run_length_average < 3.18
| | | | | | | word_freq_your < 0.29 : 0 (2/0)
| | | | | | | word_freq_your >= 0.29 : 1 (1/0)
| | | | | | | capital_run_length_average >= 3.18 : 0 (16/0)
| | | | | word_freq_will >= 0.15
| | | | | | word_freq_email < 0.63
| | | | | | | word_freq_will < 0.74

```

```

| | | | | | | | | | word_freq_people < 0.15 : 0 (1/0)
| | | | | | | | | | word_freq_people >= 0.15 : 1 (3/0)
| | | | | | | | | | word_freq_will >= 0.74 : 1 (11/0)
| | | | | | | | | | word_freq_email >= 0.63 : 0 (1/0)
| | | | | | | | | | word_freq_make >= 0.03 : 0 (9/0)
| | | | | | | | | | word_freq_hp >= 0.24 : 0 (22/0)
| | | | | | | | | | word_freq_business >= 0.14
| | | | | | | | | | word_freq_hpl < 0.26
| | | | | | | | | | word_freq_your < 0.18
| | | | | | | | | | word_freq_free < 0.2
| | | | | | | | | | word_freq_over < 0.18 : 0 (1/0)
| | | | | | | | | | word_freq_over >= 0.18 : 1 (1/0)
| | | | | | | | | | word_freq_free >= 0.2 : 0 (2/0)
| | | | | | | | | | word_freq_your >= 0.18 : 1 (24/0)
| | | | | | | | | | word_freq_hpl >= 0.26 : 0 (6/0)
| | | | | | | | | | char_freq_! >= 0.18
| | | | | | | | | | word_freq_remove < 0.03
| | | | | | | | | | capital_run_length_average < 3.38
| | | | | | | | | | word_freq_edu < 0.07
| | | | | | | | | | word_freq_hpl < 0.34
| | | | | | | | | | word_freq_money < 0.14
| | | | | | | | | | char_freq_# < 0.02
| | | | | | | | | | word_freq_all < 1.75
| | | | | | | | | | word_freq_pm < 0.17
| | | | | | | | | | capital_run_length_total < 48 : 0 (2/0)
| | | | | | | | | | capital_run_length_total >= 48
| | | | | | | | | | word_freq_1999 < 0.1
| | | | | | | | | | word_freq_mail < 0.11
| | | | | | | | | | word_freq_receive < 0.41
| | | | | | | | | | word_freq_you < 0.34 : 0 (2/0)
| | | | | | | | | | word_freq_you >= 0.34
| | | | | | | | | | capital_run_length_average < 1.59 : 0 (2/0)
| | | | | | | | | | capital_run_length_average >= 1.59
| | | | | | | | | | word_freq_will < 0.21
| | | | | | | | | | word_freq_internet < 0.32
| | | | | | | | | | word_freq_over < 0.21
| | | | | | | | | | word_freq_your < 0.34
| | | | | | | | | | word_freq_meeting < 0.89 : 1 (6/0)
| | | | | | | | | | word_freq_meeting >= 0.89 : 0 (1/0)
| | | | | | | | | | word_freq_your >= 0.34 : 1 (4/0)
| | | | | | | | | | word_freq_over >= 0.21 : 1 (1/0)
| | | | | | | | | | word_freq_internet >= 0.32 : 1 (5/0)
| | | | | | | | | | word_freq_will >= 0.21 : 1 (7/0)
| | | | | | | | | | word_freq_receive >= 0.41 : 0 (1/0)
| | | | | | | | | | word_freq_mail >= 0.11 : 1 (10/0)
| | | | | | | | | | word_freq_1999 >= 0.1 : 0 (2/0)
| | | | | | | | | | word_freq_pm >= 0.17 : 0 (5/0)
| | | | | | | | | | word_freq_all >= 1.75 : 0 (4/0)
| | | | | | | | | | char_freq_# >= 0.02
| | | | | | | | | | word_freq_3d < 0.3 : 0 (7/0)
| | | | | | | | | | word_freq_3d >= 0.3 : 1 (1/0)
| | | | | | | | | | word_freq_money >= 0.14 : 1 (11/0)
| | | | | | | | | | word_freq_hpl >= 0.34 : 0 (5/0)
| | | | | | | | | | word_freq_edu >= 0.07 : 0 (23/0)
| | | | | | | | | | capital_run_length_average >= 3.38
| | | | | | | | | | word_freq_pm < 0.33
| | | | | | | | | | word_freq_mail < 1.36
| | | | | | | | | | word_freq_will < 0.13 : 1 (48/0)
| | | | | | | | | | word_freq_will >= 0.13
| | | | | | | | | | word_freq_receive < 0.71
| | | | | | | | | | word_freq_over < 0.08
| | | | | | | | | | word_freq_you < 3.47
| | | | | | | | | | word_freq_our < 0.17 : 0 (2/0)
| | | | | | | | | | word_freq_our >= 0.17 : 1 (2/0)
| | | | | | | | | | word_freq_you >= 3.47 : 1 (3/0)
| | | | | | | | | | word_freq_over >= 0.08 : 1 (12/0)
| | | | | | | | | | word_freq_receive >= 0.71 : 0 (1/0)

```

```

| | | | | word_freq_mail >= 1.36 : 0 (4/0)
| | | | | word_freq_pm >= 0.33 : 0 (2/0)
| | | | | word_freq_remove >= 0.03 : 1 (67/0)
| | | | | word_freq_george >= 0.01 : 0 (238/0)
| | | | | word_freq_your >= 0.92
| | | | | char_freq_$ < 0.14
| | | | | word_freq_hp < 0.12
| | | | | word_freq_all < 0.01
| | | | | word_freq_address < 1.33
| | | | | char_freq_! < 0.02
| | | | | word_freq_will < 1.29
| | | | | word_freq_our < 0.31
| | | | | word_freq_edu < 0.3
| | | | | word_freq_make < 0.17
| | | | | word_freq_your < 2.13
| | | | | word_freq_remove < 0.27 : 0 (5/0)
| | | | | word_freq_remove >= 0.27
| | | | | word_freq_mail < 0.34 : 1 (3/0)
| | | | | word_freq_mail >= 0.34 : 0 (2/0)
| | | | | word_freq_your >= 2.13 : 0 (7/0)
| | | | | word_freq_make >= 0.17 : 1 (2/0)
| | | | | word_freq_edu >= 0.3 : 0 (6/0)
| | | | | word_freq_our >= 0.31 : 1 (7/0)
| | | | | word_freq_will >= 1.29 : 1 (3/0)
| | | | | char_freq_! >= 0.02
| | | | | word_freq_business < 0.11
| | | | | word_freq_remove < 0.27
| | | | | capital_run_length_average < 4.58
| | | | | word_freq_will < 0.24
| | | | | word_freq_re < 1.25
| | | | | capital_run_length_average < 4.34
| | | | | capital_run_length_total < 136.5 : 1 (19/0)
| | | | | capital_run_length_total >= 136.5
| | | | | word_freq_free < 0.13 : 1 (1/0)
| | | | | word_freq_free >= 0.13 : 0 (1/0)
| | | | | capital_run_length_average >= 4.34 : 0 (2/0)
| | | | | word_freq_re >= 1.25 : 0 (3/0)
| | | | | word_freq_will >= 0.24
| | | | | word_freq_000 < 0.24
| | | | | word_freq_mail < 0.8
| | | | | word_freq_will < 2.4
| | | | | word_freq_our < 1.71 : 0 (7/0)
| | | | | word_freq_our >= 1.71 : 1 (1/0)
| | | | | word_freq_will >= 2.4 : 1 (2/0)
| | | | | word_freq_mail >= 0.8 : 1 (1/0)
| | | | | word_freq_000 >= 0.24
| | | | | word_freq_make < 0.72 : 0 (4/0)
| | | | | word_freq_make >= 0.72 : 1 (2/0)
| | | | | capital_run_length_average >= 4.58 : 1 (12/0)
| | | | | word_freq_remove >= 0.27 : 1 (21/0)
| | | | | word_freq_business >= 0.11 : 1 (24/0)
| | | | | word_freq_address >= 1.33 : 0 (4/0)
| | | | | word_freq_all >= 0.01
| | | | | char_freq_! < 0.16
| | | | | capital_run_length_total < 71 : 0 (7/0)
| | | | | capital_run_length_total >= 71
| | | | | capital_run_length_total < 495.5
| | | | | word_freq_edu < 0.17
| | | | | word_freq_will < 0.33
| | | | | word_freq_your < 1.21 : 1 (8/0)
| | | | | word_freq_your >= 1.21
| | | | | word_freq_email < 0.99
| | | | | word_freq_order < 0.27 : 0 (10/0)
| | | | | word_freq_order >= 0.27 : 1 (3/0)
| | | | | word_freq_email >= 0.99 : 1 (3/0)
| | | | | word_freq_will >= 0.33
| | | | | char_freq_ ; < 0.05 : 1 (31/0)

```

```

| | | | | | | | | | char_freq_ >= 0.05
| | | | | | | | | | word_freq_address < 0.09 : 0 (1/0)
| | | | | | | | | | word_freq_address >= 0.09 : 1 (2/0)
| | | | | | | | | | word_freq_edu >= 0.17 : 0 (2/0)
| | | | | | | | | | capital_run_length_total >= 495.5 : 1 (20/0)
| | | | | | | | | | char_freq_! >= 0.16
| | | | | | | | | | word_freq_meeting < 0.1
| | | | | | | | | | char_freq_# < 0.08 : 1 (127/0)
| | | | | | | | | | char_freq_# >= 0.08
| | | | | | | | | | word_freq_mail < 0.9 : 1 (6/0)
| | | | | | | | | | word_freq_mail >= 0.9
| | | | | | | | | | word_freq_all < 0.38 : 1 (1/0)
| | | | | | | | | | word_freq_all >= 0.38 : 0 (1/0)
| | | | | | | | | | word_freq_meeting >= 0.1 : 0 (3/0)
| | | | | | | | | | word_freq_hp >= 0.12
| | | | | | | | | | word_freq_hpl < 0.96 : 0 (51/0)
| | | | | | | | | | word_freq_hpl >= 0.96
| | | | | | | | | | word_freq_85 < 0.09
| | | | | | | | | | word_freq_all < 0.28 : 1 (1/0)
| | | | | | | | | | word_freq_all >= 0.28 : 0 (3/0)
| | | | | | | | | | word_freq_85 >= 0.09 : 0 (14/0)
| | | | | | | | | | char_freq_$ >= 0.14
| | | | | | | | | | word_freq_free < 0.01
| | | | | | | | | | word_freq_edu < 0.63
| | | | | | | | | | word_freq_will < 1.93 : 1 (82/0)
| | | | | | | | | | word_freq_will >= 1.93 : 0 (1/0)
| | | | | | | | | | word_freq_edu >= 0.63 : 0 (1/0)
| | | | | | | | | | word_freq_free >= 0.01 : 1 (208/0)
word_freq_credit >= 0.1
| word_freq_your < 0.09
| | word_freq_project < 0.42
| | | char_freq_( < 0.15
| | | | word_freq_business < 0.46 : 0 (6/0)
| | | | word_freq_business >= 0.46 : 1 (1/0)
| | | | char_freq_( >= 0.15 : 1 (4/0)
| | | | word_freq_project >= 0.42 : 0 (4/0)
word_freq_your >= 0.09
| char_freq_! < 0.14
| | word_freq_internet < 0.03
| | | word_freq_will < 0.77
| | | | capital_run_length_longest < 14.5 : 1 (4/0)
| | | | capital_run_length_longest >= 14.5
| | | | | char_freq_( < 0.08
| | | | | | word_freq_business < 0.24 : 1 (5/0)
| | | | | | word_freq_business >= 0.24 : 0 (3/0)
| | | | | | char_freq_( >= 0.08 : 0 (9/0)
| | | | | | word_freq_will >= 0.77
| | | | | | word_freq_email < 0.28 : 1 (10/0)
| | | | | | word_freq_email >= 0.28 : 0 (1/0)
| | | | | | word_freq_internet >= 0.03 : 1 (23/0)
| | | | | | char_freq_! >= 0.14 : 1 (241/0)

```

Size of the tree : 515

RandomTree

```

=====

char_freq_! < 0.12
| capital_run_length_average < 4.49
| | char_freq_$ < 0.09
| | | word_freq_remove < 0.07
| | | | word_freq_free < 0.03
| | | | | word_freq_re < 0.01
| | | | | | word_freq_your < 4.79
| | | | | | word_freq_hpl < 0.04
| | | | | | capital_run_length_average < 2.36

```

```

word_freq_george < 0.05
word_freq_mail < 0.66
capital_run_length_longest < 30
word_freq_people < 0.16
word_freq_direct < 0.04
word_freq_our < 2.3
word_freq_edu < 2.25
word_freq_edu < 0.07
word_freq_money < 0.15
word_freq_technology < 0.82
capital_run_length_total < 19.5
capital_run_length_total < 16.5
word_freq_hp < 0.91
word_freq_email < 0.4
word_freq_you < 4.55
word_freq_your < 0.27
word_freq_will < 0.84
capital_run_length_longest < 1.5
word_freq_technology < 0.39
char_freq_ ; < 0.16
capital_run_length_total < 6.5 : 0 (52/0)
capital_run_length_total >= 6.5 : 1 (1/0)
char_freq_ ; >= 0.16 : 0 (1/0)
word_freq_technology >= 0.39 : 0 (2/0)
capital_run_length_longest >= 1.5 : 0 (32/0)
word_freq_will >= 0.84 : 0 (2/0)
word_freq_your >= 0.27 : 0 (3/0)
word_freq_you >= 4.55
word_freq_you < 4.78 : 1 (1/0)
word_freq_you >= 4.78 : 0 (5/0)
word_freq_email >= 0.4 : 0 (5/0)
word_freq_hp >= 0.91 : 0 (3/0)
capital_run_length_total >= 16.5
capital_run_length_average < 1.72
capital_run_length_total < 18.5
word_freq_will < 0.57
capital_run_length_average < 1.48
capital_run_length_total < 17.5 : 1 (1/0)
capital_run_length_total >= 17.5 : 0 (1/0)
capital_run_length_average >= 1.48 : 0 (2/0)
word_freq_will >= 0.57 : 1 (2/0)
capital_run_length_total >= 18.5 : 1 (4/0)
capital_run_length_average >= 1.72 : 0 (5/0)
capital_run_length_total >= 19.5
capital_run_length_longest < 3.5
word_freq_meeting < 0.4
word_freq_will < 0.28
word_freq_your < 0.1
capital_run_length_longest < 2 : 0 (1/0)
capital_run_length_longest >= 2 : 1 (1/0)
word_freq_your >= 0.1 : 0 (2/0)
word_freq_will >= 0.28 : 0 (3/0)
word_freq_meeting >= 0.4 : 0 (7/0)
capital_run_length_longest >= 3.5
char_freq_ ( < 0.04
word_freq_mail < 0.1 : 0 (63/0)
word_freq_mail >= 0.1 : 1 (1/0)
char_freq_ ( >= 0.04 : 0 (79/0)
word_freq_technology >= 0.82 : 1 (2/0)
word_freq_money >= 0.15 : 1 (2/0)
word_freq_edu >= 0.07 : 0 (55/0)
word_freq_edu >= 2.25
word_freq_report < 0.41 : 0 (1/0)
word_freq_report >= 0.41 : 1 (2/0)
word_freq_our >= 2.3 : 1 (1/0)
word_freq_direct >= 0.04
word_freq_you < 0.13 : 0 (2/0)

```

```

| | | | | word_freq_you >= 0.13 : 1 (3/0)
| | | | | word_freq_people >= 0.16 : 0 (26/0)
| | | | | capital_run_length_longest >= 30
| | | | | word_freq_mail < 0.04
| | | | | word_freq_will < 0.2 : 1 (1/0)
| | | | | word_freq_will >= 0.2
| | | | | word_freq_internet < 0.17 : 1 (1/0)
| | | | | word_freq_internet >= 0.17
| | | | | word_freq_over < 0.03 : 0 (1/0)
| | | | | word_freq_over >= 0.03 : 0 (5/2)
| | | | | word_freq_mail >= 0.04 : 0 (4/0)
| | | | | word_freq_mail >= 0.66
| | | | | capital_run_length_longest < 9.5 : 0 (5/0)
| | | | | capital_run_length_longest >= 9.5
| | | | | word_freq_will < 0.18
| | | | | char_freq_(< 0.21 : 1 (3/0)
| | | | | char_freq_( >= 0.21 : 0 (1/0)
| | | | | word_freq_will >= 0.18 : 1 (2/0)
| | | | | word_freq_george >= 0.05 : 0 (74/0)
| | | | | capital_run_length_average >= 2.36
| | | | | word_freq_lab < 0.01
| | | | | word_freq_1999 < 0.04
| | | | | word_freq_george < 0.04
| | | | | capital_run_length_average < 2.48
| | | | | word_freq_order < 0.16
| | | | | char_freq_(< 0.05
| | | | | capital_run_length_average < 2.37 : 1 (2/0)
| | | | | capital_run_length_average >= 2.37 : 0 (2/0)
| | | | | char_freq_( >= 0.05 : 1 (6/0)
| | | | | word_freq_order >= 0.16 : 1 (1/0)
| | | | | capital_run_length_average >= 2.48
| | | | | word_freq_email < 0.12
| | | | | capital_run_length_average < 4.21
| | | | | word_freq_make < 0.2
| | | | | word_freq_receive < 0.19
| | | | | capital_run_length_total < 99
| | | | | word_freq_over < 0.4
| | | | | word_freq_meeting < 0.39
| | | | | word_freq_address < 0.96
| | | | | word_freq_people < 0.28
| | | | | word_freq_business < 0.52
| | | | | char_freq_# < 0.1
| | | | | word_freq_edu < 0.38
| | | | | char_freq_(< 0.67
| | | | | word_freq_mail < 0.73
| | | | | word_freq_will < 0.96
| | | | | capital_run_length_longest < 13.5 : 0 (10/0)
| | | | | capital_run_length_longest >= 13.5 : 1 (2/0)
| | | | | word_freq_will >= 0.96 : 0 (2/0)
| | | | | word_freq_mail >= 0.73 : 1 (2/0)
| | | | | char_freq_( >= 0.67 : 0 (3/0)
| | | | | word_freq_edu >= 0.38 : 0 (3/0)
| | | | | char_freq_# >= 0.1 : 0 (3/0)
| | | | | word_freq_business >= 0.52 : 0 (1/0)
| | | | | word_freq_people >= 0.28 : 0 (3/0)
| | | | | word_freq_address >= 0.96 : 1 (1/0)
| | | | | word_freq_meeting >= 0.39 : 0 (2/0)
| | | | | word_freq_over >= 0.4 : 1 (2/0)
| | | | | capital_run_length_total >= 99
| | | | | char_freq_(< 0.12
| | | | | word_freq_make < 0.09
| | | | | word_freq_technology < 0.08
| | | | | capital_run_length_total < 122.5 : 1 (1/0)
| | | | | capital_run_length_total >= 122.5 : 0 (2/0)
| | | | | word_freq_technology >= 0.08 : 0 (4/0)
| | | | | word_freq_make >= 0.09 : 1 (1/0)
| | | | | char_freq_( >= 0.12

```

```

| word_freq_font < 0.1
| capital_run_length_total < 298.5 : 1 (9/0)
| capital_run_length_total >= 298.5 : 0 (2/0)
| word_freq_font >= 0.1 : 1 (7/0)
| word_freq_receive >= 0.19 : 0 (3/0)
| word_freq_make >= 0.2 : 0 (8/0)
| capital_run_length_average >= 4.21 : 0 (8/0)
| word_freq_email >= 0.12 : 0 (5/0)
| word_freq_george >= 0.04 : 0 (24/0)
| word_freq_1999 >= 0.04 : 0 (25/0)
| word_freq_lab >= 0.01 : 0 (17/0)
| word_freq_hpl >= 0.04 : 0 (235/0)
| word_freq_your >= 4.79 : 1 (3/0)
word_freq_re >= 0.01
| word_freq_re < 0.76 : 0 (204/0)
| word_freq_re >= 0.76
| word_freq_all < 0.13
| word_freq_1999 < 0.28
| word_freq_you < 2.95 : 0 (54/0)
| word_freq_you >= 2.95
| word_freq_will < 1.79 : 0 (39/0)
| word_freq_will >= 1.79
| word_freq_edu < 0.76 : 0 (4/0)
| word_freq_edu >= 0.76
| capital_run_length_average < 1.19 : 0 (3/0)
| capital_run_length_average >= 1.19 : 1 (1/0)
| word_freq_1999 >= 0.28 : 0 (27/0)
| word_freq_all >= 0.13 : 0 (39/0)
word_freq_free >= 0.03
| word_freq_people < 0.82
| capital_run_length_longest < 9.5
| word_freq_address < 0.38
| word_freq_your < 1.4 : 0 (27/0)
| word_freq_your >= 1.4
| word_freq_you < 0.81 : 1 (1/0)
| word_freq_you >= 0.81 : 0 (3/0)
| word_freq_address >= 0.38 : 1 (1/0)
capital_run_length_longest >= 9.5
| word_freq_hpl < 0.12
| char_freq_[] < 0
| word_freq_our < 0.01
| word_freq_free < 0.52
| capital_run_length_longest < 17 : 1 (4/0)
| capital_run_length_longest >= 17
| word_freq_will < 0.91
| capital_run_length_longest < 61.5
| capital_run_length_longest < 53.5
| char_freq_(< 0.04 : 0 (4/0)
| char_freq_( >= 0.04
| word_freq_conference < 0.05 : 1 (3/0)
| word_freq_conference >= 0.05 : 0 (2/0)
| capital_run_length_longest >= 53.5 : 1 (3/0)
| capital_run_length_longest >= 61.5 : 0 (3/0)
| word_freq_will >= 0.91 : 0 (9/0)
| word_freq_free >= 0.52 : 0 (15/0)
word_freq_our >= 0.01
| word_freq_project < 0.01
| word_freq_make < 0.57
| word_freq_internet < 0.2
| word_freq_over < 0.23 : 1 (28/0)
| word_freq_over >= 0.23
| word_freq_mail < 0.22 : 1 (12/0)
| word_freq_mail >= 0.22 : 0 (1/0)
| word_freq_internet >= 0.2
| word_freq_direct < 0.17 : 0 (1/0)
| word_freq_direct >= 0.17 : 1 (1/0)
word_freq_make >= 0.57 : 0 (1/0)

```

```

| | | word_freq_project >= 0.01 : 0 (5/0)
| | | char_freq_[] >= 0 : 0 (7/0)
| | | word_freq_hpl >= 0.12
| | | | word_freq_will < 0.14 : 1 (3/0)
| | | | word_freq_will >= 0.14 : 0 (40/0)
| | | word_freq_people >= 0.82 : 0 (15/0)
| word_freq_remove >= 0.07
| | word_freq_george < 0.15
| | | word_freq_data < 0.2
| | | | word_freq_you < 3.63
| | | | word_freq_email < 0.04
| | | | word_freq_free < 0.19
| | | | | word_freq_conference < 0.1 : 1 (13/0)
| | | | | word_freq_conference >= 0.1 : 0 (1/0)
| | | | | word_freq_free >= 0.19 : 1 (10/0)
| | | | | word_freq_email >= 0.04 : 1 (35/0)
| | | | word_freq_you >= 3.63
| | | | | word_freq_business < 0.38 : 1 (3/0)
| | | | | word_freq_business >= 0.38 : 0 (1/0)
| | | word_freq_data >= 0.2
| | | | word_freq_hpl < 0.23
| | | | | word_freq_remove < 0.2 : 1 (1/0)
| | | | | word_freq_remove >= 0.2 : 0 (9/0)
| | | | | word_freq_hpl >= 0.23 : 1 (5/0)
| | | word_freq_george >= 0.15 : 0 (12/0)
| char_freq_$ >= 0.09
| | word_freq_all < 0.13
| | | char_freq_! < 0.03
| | | | word_freq_meeting < 0.04
| | | | | word_freq_your < 0.92
| | | | | word_freq_you < 0.4
| | | | | | word_freq_business < 0.33
| | | | | | capital_run_length_longest < 15 : 0 (3/0)
| | | | | | capital_run_length_longest >= 15 : 1 (3/0)
| | | | | | word_freq_business >= 0.33 : 0 (5/0)
| | | | | | word_freq_you >= 0.4 : 0 (20/0)
| | | | | word_freq_your >= 0.92
| | | | | | capital_run_length_average < 3.14 : 0 (3/0)
| | | | | | capital_run_length_average >= 3.14 : 1 (4/0)
| | | | word_freq_meeting >= 0.04 : 0 (6/0)
| char_freq_! >= 0.03
| | | word_freq_email < 0.3
| | | | char_freq_( < 0.04
| | | | | word_freq_will < 0.73 : 1 (3/0)
| | | | | word_freq_will >= 0.73 : 0 (1/0)
| | | | | char_freq_( >= 0.04 : 1 (4/0)
| | | | | word_freq_email >= 0.3 : 1 (4/0)
| word_freq_all >= 0.13
| | char_freq_ ; < 0.05
| | | word_freq_business < 0.22
| | | | word_freq_business < 0.16
| | | | | word_freq_your < 1.15
| | | | | | word_freq_money < 0.63 : 1 (11/0)
| | | | | | word_freq_money >= 0.63 : 0 (3/0)
| | | | | | word_freq_your >= 1.15 : 1 (20/0)
| | | | | | word_freq_business >= 0.16 : 0 (1/0)
| | | | | word_freq_business >= 0.22 : 1 (60/0)
| char_freq_ ; >= 0.05
| | | word_freq_your < 0.55 : 1 (2/0)
| | | | word_freq_your >= 0.55
| | | | | word_freq_mail < 0.14 : 0 (4/2)
| | | | | word_freq_mail >= 0.14 : 0 (1/0)
capital_run_length_average >= 4.49
| | word_freq_address < 0.33
| | | word_freq_1999 < 0.43
| | | | word_freq_hpl < 0.21
| | | | | word_freq_report < 0.33

```

```

| | | | | word_freq_money < 0.03
| | | | | word_freq_1999 < 0.15
| | | | | | word_freq_make < 0.09
| | | | | | word_freq_650 < 0.37
| | | | | | | word_freq_will < 0.78
| | | | | | | word_freq_address < 0.13
| | | | | | | word_freq_our < 0.55
| | | | | | | word_freq_addresses < 0.12
| | | | | | | word_freq_mail < 0.04
| | | | | | | word_freq_all < 0.12
| | | | | | | word_freq_free < 0.39
| | | | | | | | word_freq_000 < 0.39
| | | | | | | | word_freq_hp < 0.28
| | | | | | | | word_freq_free < 0.18
| | | | | | | | | capital_run_length_longest < 16 : 0 (1/0)
| | | | | | | | | capital_run_length_longest >= 16 : 1 (5/0)
| | | | | | | | | word_freq_free >= 0.18 : 1 (2/0)
| | | | | | | | | word_freq_hp >= 0.28 : 0 (1/0)
| | | | | | | | | word_freq_000 >= 0.39 : 1 (2/0)
| | | | | | | | | word_freq_free >= 0.39 : 0 (1/0)
| | | | | | | | | word_freq_all >= 0.12 : 0 (2/0)
| | | | | | | | | word_freq_mail >= 0.04 : 0 (6/0)
| | | | | | | | | word_freq_addresses >= 0.12 : 1 (1/0)
| | | | | | | | | word_freq_our >= 0.55 : 1 (3/0)
| | | | | | | | | word_freq_address >= 0.13 : 0 (3/0)
| | | | | | | | | word_freq_will >= 0.78 : 1 (4/0)
| | | | | | | | | word_freq_650 >= 0.37 : 0 (3/0)
| | | | | | | | | word_freq_make >= 0.09 : 1 (8/0)
| | | | | | | | | word_freq_1999 >= 0.15 : 1 (4/0)
| | | | | | | | | word_freq_money >= 0.03
| | | | | | | | | word_freq_free < 0.96 : 1 (20/0)
| | | | | | | | | word_freq_free >= 0.96 : 0 (2/0)
| | | | | | | | | word_freq_report >= 0.33 : 0 (3/0)
| | | | | | | | | word_freq_hpl >= 0.21 : 0 (8/0)
| | | | | | | | | word_freq_1999 >= 0.43
| | | | | | | | | word_freq_1999 < 1.94 : 0 (17/0)
| | | | | | | | | word_freq_1999 >= 1.94
| | | | | | | | | word_freq_data < 0.39 : 1 (2/0)
| | | | | | | | | word_freq_data >= 0.39 : 0 (1/0)
| | | | | | | | | word_freq_address >= 0.33 : 1 (14/0)
| | | | | char_freq_! >= 0.12
| | | | | | word_freq_business < 0.04
| | | | | | capital_run_length_total < 39.5
| | | | | | | word_freq_all < 1.52
| | | | | | | word_freq_remove < 0.47
| | | | | | | word_freq_hp < 0.49
| | | | | | | word_freq_our < 0.32
| | | | | | | | capital_run_length_average < 2.26 : 0 (110/0)
| | | | | | | | capital_run_length_average >= 2.26
| | | | | | | | | char_freq_! < 0.63
| | | | | | | | | word_freq_conference < 0.37 : 1 (1/0)
| | | | | | | | | word_freq_conference >= 0.37 : 0 (2/0)
| | | | | | | | | char_freq_! >= 0.63 : 0 (4/0)
| | | | | | | | | word_freq_our >= 0.32 : 0 (10/0)
| | | | | | | | | word_freq_hp >= 0.49
| | | | | | | | | capital_run_length_average < 1.29 : 0 (5/0)
| | | | | | | | | capital_run_length_average >= 1.29
| | | | | | | | | word_freq_all < 0.53 : 0 (5/0)
| | | | | | | | | word_freq_all >= 0.53 : 1 (3/0)
| | | | | | | | | word_freq_remove >= 0.47 : 1 (9/0)
| | | | | | | | | word_freq_all >= 1.52
| | | | | | | | | capital_run_length_average < 1.86 : 0 (5/0)
| | | | | | | | | capital_run_length_average >= 1.86 : 1 (8/0)
| | | | | | | | | capital_run_length_total >= 39.5
| | | | | | | | | capital_run_length_longest < 18.5
| | | | | | | | | word_freq_re < 0.04
| | | | | | | | | word_freq_free < 0.13

```

```

| | | | | word_freq_will < 0.72
| | | | | | word_freq_hpl < 0.34
| | | | | | | capital_run_length_total < 104.5
| | | | | | | | word_freq_will < 0.43
| | | | | | | | | char_freq_! < 0.36
| | | | | | | | | | char_freq_! < 0.21
| | | | | | | | | | | word_freq_edu < 0.88
| | | | | | | | | | | | char_freq_# < 0.02 : 1 (2/0)
| | | | | | | | | | | | char_freq_# >= 0.02 : 0 (1/0)
| | | | | | | | | | | | word_freq_edu >= 0.88 : 0 (2/0)
| | | | | | | | | | | | char_freq_! >= 0.21 : 0 (9/0)
| | | | | | | | | | | | char_freq_! >= 0.36
| | | | | | | | | | | | capital_run_length_total < 65 : 1 (13/0)
| | | | | | | | | | | | capital_run_length_total >= 65
| | | | | | | | | | | | | char_freq_! < 0.61 : 1 (1/0)
| | | | | | | | | | | | | char_freq_! >= 0.61 : 0 (1/0)
| | | | | | | | | | | | | word_freq_will >= 0.43 : 1 (6/0)
| | | | | | | | | | | | | capital_run_length_total >= 104.5 : 1 (7/0)
| | | | | | | | | | | | | word_freq_hpl >= 0.34 : 0 (4/0)
| | | | | | | | | | | | | word_freq_will >= 0.72
| | | | | | | | | | | | | char_freq_ ; < 0.07
| | | | | | | | | | | | | | char_freq_( < 0.05
| | | | | | | | | | | | | | | char_freq_! < 0.3 : 0 (5/0)
| | | | | | | | | | | | | | | char_freq_! >= 0.3 : 1 (2/0)
| | | | | | | | | | | | | | | char_freq_( >= 0.05 : 0 (8/0)
| | | | | | | | | | | | | | | char_freq_ ; >= 0.07 : 1 (1/0)
| | | | | | | | | | | | | word_freq_free >= 0.13
| | | | | | | | | | | | | | word_freq_order < 0.51 : 1 (53/0)
| | | | | | | | | | | | | | word_freq_order >= 0.51 : 0 (1/0)
| | | | | | | | | | | | | word_freq_re >= 0.04
| | | | | | | | | | | | | word_freq_edu < 0.07
| | | | | | | | | | | | | | word_freq_labs < 0.25
| | | | | | | | | | | | | | | char_freq_$ < 0.01
| | | | | | | | | | | | | | | | word_freq_order < 0.37
| | | | | | | | | | | | | | | | | word_freq_your < 0.85
| | | | | | | | | | | | | | | | | | word_freq_re < 0.61
| | | | | | | | | | | | | | | | | | | word_freq_all < 0.47 : 0 (2/0)
| | | | | | | | | | | | | | | | | | | word_freq_all >= 0.47
| | | | | | | | | | | | | | | | | | | | word_freq_our < 1.28 : 1 (1/0)
| | | | | | | | | | | | | | | | | | | | word_freq_our >= 1.28 : 0 (1/0)
| | | | | | | | | | | | | | | | | | | | word_freq_re >= 0.61 : 0 (9/0)
| | | | | | | | | | | | | | | | | | | | word_freq_your >= 0.85
| | | | | | | | | | | | | | | | | | | | | capital_run_length_longest < 13 : 1 (10/0)
| | | | | | | | | | | | | | | | | | | | | capital_run_length_longest >= 13
| | | | | | | | | | | | | | | | | | | | | | word_freq_you < 4.07 : 0 (2/0)
| | | | | | | | | | | | | | | | | | | | | | word_freq_you >= 4.07 : 1 (1/0)
| | | | | | | | | | | | | | | | | | | | | | word_freq_order >= 0.37 : 0 (4/0)
| | | | | | | | | | | | | | | | | | | | | | char_freq_$ >= 0.01
| | | | | | | | | | | | | | | | | | | | | | | word_freq_our < 0.12
| | | | | | | | | | | | | | | | | | | | | | | | word_freq_re < 2.07 : 1 (12/0)
| | | | | | | | | | | | | | | | | | | | | | | | word_freq_re >= 2.07 : 0 (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | word_freq_our >= 0.12 : 1 (7/0)
| | | | | | | | | | | | | | | | | | | | | | | | word_freq_labs >= 0.25 : 0 (8/0)
| | | | | | | | | | | | | | | | | | | | | | | | word_freq_edu >= 0.07 : 0 (18/0)
| | | | | | | | | | | | | | | | | | | | | | | | capital_run_length_longest >= 18.5
| | | | | | | | | | | | | | | | | | | | | | | | capital_run_length_longest < 70.5
| | | | | | | | | | | | | | | | | | | | | | | | capital_run_length_average < 2.93
| | | | | | | | | | | | | | | | | | | | | | | | | word_freq_george < 0.35
| | | | | | | | | | | | | | | | | | | | | | | | | word_freq_original < 0.1
| | | | | | | | | | | | | | | | | | | | | | | | | word_freq_people < 0.39
| | | | | | | | | | | | | | | | | | | | | | | | | | char_freq_! < 0.48
| | | | | | | | | | | | | | | | | | | | | | | | | | | word_freq_your < 0.53
| | | | | | | | | | | | | | | | | | | | | | | | | | | | word_freq_free < 0.81 : 0 (5/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | word_freq_free >= 0.81 : 1 (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | word_freq_your >= 0.53
| | | | | | | | | | | | | | | | | | | | | | | | | | | | word_freq_your < 1.22 : 1 (11/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | word_freq_your >= 1.22

```

```

| | | | | | | | | | word_freq_re < 0.12 : 1 (8/0)
| | | | | | | | | | word_freq_re >= 0.12 : 0 (5/0)
| | | | | | | | | | char_freq_! >= 0.48 : 1 (38/0)
| | | | | | | | | | word_freq_people >= 0.39
| | | | | | | | | | word_freq_000 < 0.49
| | | | | | | | | | word_freq_edu < 0.2
| | | | | | | | | | word_freq_mail < 0.04 : 0 (2/0)
| | | | | | | | | | word_freq_mail >= 0.04 : 1 (1/0)
| | | | | | | | | | word_freq_edu >= 0.2 : 0 (6/0)
| | | | | | | | | | word_freq_000 >= 0.49 : 1 (5/0)
| | | | | | | | | | word_freq_original >= 0.1 : 0 (3/0)
| | | | | | | | | | word_freq_george >= 0.35 : 0 (10/0)
| | | | | | | | | | capital_run_length_average >= 2.93
| | | | | | | | | | capital_run_length_longest < 69.5
| | | | | | | | | | word_freq_edu < 0.43
| | | | | | | | | | word_freq_hp < 0.23
| | | | | | | | | | capital_run_length_average < 3.98 : 1 (79/0)
| | | | | | | | | | capital_run_length_average >= 3.98
| | | | | | | | | | capital_run_length_total < 142.5
| | | | | | | | | | word_freq_will < 0.24
| | | | | | | | | | char_freq_! < 0.35
| | | | | | | | | | char_freq_$ < 0.27
| | | | | | | | | | word_freq_email < 0.36 : 1 (6/0)
| | | | | | | | | | word_freq_email >= 0.36
| | | | | | | | | | word_freq_your < 2.24
| | | | | | | | | | word_freq_email < 0.77 : 0 (1/0)
| | | | | | | | | | word_freq_email >= 0.77
| | | | | | | | | | capital_run_length_average < 4.21 : 1 (1/0)
| | | | | | | | | | capital_run_length_average >= 4.21 : 0 (1/0)
| | | | | | | | | | word_freq_your >= 2.24 : 1 (4/0)
| | | | | | | | | | char_freq_$ >= 0.27 : 0 (1/0)
| | | | | | | | | | char_freq_! >= 0.35 : 1 (21/0)
| | | | | | | | | | word_freq_will >= 0.24
| | | | | | | | | | capital_run_length_total < 118 : 1 (1/0)
| | | | | | | | | | capital_run_length_total >= 118 : 0 (2/0)
| | | | | | | | | | capital_run_length_total >= 142.5 : 1 (65/0)
| | | | | | | | | | word_freq_hp >= 0.23
| | | | | | | | | | word_freq_re < 1
| | | | | | | | | | capital_run_length_total < 272 : 0 (4/0)
| | | | | | | | | | capital_run_length_total >= 272 : 1 (3/0)
| | | | | | | | | | word_freq_re >= 1 : 1 (3/0)
| | | | | | | | | | word_freq_edu >= 0.43 : 0 (6/0)
| | | | | | | | | | capital_run_length_longest >= 69.5
| | | | | | | | | | char_freq_(< 0.11 : 1 (1/0)
| | | | | | | | | | char_freq_(>= 0.11 : 0 (6/0)
| | | | | | | | | | capital_run_length_longest >= 70.5
| | | | | | | | | | word_freq_you < 1.31
| | | | | | | | | | word_freq_pm < 0.43
| | | | | | | | | | word_freq_our < 0.07
| | | | | | | | | | char_freq_! < 0.16 : 0 (1/0)
| | | | | | | | | | char_freq_! >= 0.16
| | | | | | | | | | word_freq_re < 0.31 : 1 (16/0)
| | | | | | | | | | word_freq_re >= 0.31 : 0 (1/0)
| | | | | | | | | | word_freq_our >= 0.07 : 1 (30/0)
| | | | | | | | | | word_freq_pm >= 0.43 : 0 (1/0)
| | | | | | | | | | word_freq_you >= 1.31 : 1 (100/0)
| | | | | | | | | | word_freq_business >= 0.04
| | | | | | | | | | word_freq_650 < 0.2
| | | | | | | | | | word_freq_addresses < 0.01
| | | | | | | | | | word_freq_mail < 1.31
| | | | | | | | | | word_freq_project < 0.37
| | | | | | | | | | word_freq_email < 0.57
| | | | | | | | | | capital_run_length_average < 1.94
| | | | | | | | | | word_freq_email < 0.48
| | | | | | | | | | char_freq_(< 0.06 : 1 (9/0)
| | | | | | | | | | char_freq_(>= 0.06
| | | | | | | | | | word_freq_business < 0.59 : 0 (2/0)

```


Appendix 2

JRIP rules:

=====

(char_freq_! >= 0.085) and (capital_run_length_average >= 2.232) and (char_freq_\$ >= 0.013) => is-spam=1 (607.0/14.0)
(word_freq_remove >= 0.05) and (char_freq_! >= 0.05) => isspace=1 (207.0/4.0)
(word_freq_free >= 0.14) and (word_freq_our >= 0.29) and (word_freq_your >= 0.54) and
(word_freq_george <= 0) => isspace=1 (136.0/5.0)
(char_freq_\$ >= 0.062) and (word_freq_hp <= 0.05) and (word_freq_your >= 0.63) and
(word_freq_money >= 0.05) => isspace=1 (54.0/3.0)
(word_freq_hp <= 0.11) and (char_freq_\$ >= 0.015) and (word_freq_000 >= 0.1) and (word_freq_edu
<= 0) and (word_freq_your <= 2.13) => isspace=1 (33.0/2.0)
(word_freq_your >= 0.56) and (char_freq_! >= 0.464) => isspace=1 (63.0/13.0)
(word_freq_hp <= 0) and (capital_run_length_average >= 2.22) and (word_freq_edu <= 0) and
(word_freq_george <= 0) and (word_freq_our >= 0.2) => isspace=1 (58.0/6.0)
(word_freq_free >= 0.4) and (char_freq_! >= 0.109) and (word_freq_your <= 0.43) => isspace=1
(38.0/6.0)
(word_freq_hp <= 0) and (char_freq_! >= 0.225) and (capital_run_length_average >= 3.965) => is-spam=1 (27.0/5.0)
(word_freq_hp <= 0.11) and (word_freq_remove >= 0.13) => isspace=1 (32.0/9.0)
=> isspace=0 (1747.0/126.0)

Number of Rules : 11