Οικονομικό Πανεπιστήμιο Αθηνών

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Ανάπτυξη φίλτρου διήθησης ηλεκτρονικής αλληλογραφίας για το Mozilla Thunderbird»

Δημήτρης Μπόχτης

Επιβλέπων: Ίων Ανδρουτσόπουλος

ΑΘΗΝΑ, ΟΚΤΩΒΡΙΟΣ 2007

Περιεχόμενα

Περίληψη

Στην παρούσα εργασία αναπτύχθηκε ένα φίλτρο ανεπιθύμητης ηλεκτρονικής αλληλογραφίας (spam filter) που ενσωματώθηκε στο πρόγραμμα ανάγνωσης ηλεκτρονικής αλληλογραφίας Mozilla Thunderbird. Το φίλτρο χρησιμοποιεί μια βιβλιοθήκη λογισμικού που είχε αναπτυχθεί σε προηγούμενη εργασία, η οποία παρέχει υλοποιήσεις πολλών παραλλαγών του απλοϊκού ταξινομητή Bayes (Naïve Bayes) και μηχανισμούς εφαρμογής τους σε μηνύματα ηλεκτρονικού ταχυδρομείου. Το φίλτρο της παρούσας εργασίας υποστηρίζει, επίσης, αποδείξεις ανθρώπινης αλληλεπίδρασης (human interaction proofs – HIPs). Διατίθεται ελεύθερα ως λογισμικό ανοικτού πηγαίου κώδικα, όπως και το ίδιο το Mozilla Thunderbird. Πειράματα που διεξήχθησαν στη διάρκεια της εργασίας έδειξαν ότι το φίλτρο που αναπτύχθηκε επιτυγχάνει καλύτερα αποτελέσματα από το ενσωματωμένο φίλτρο ανεπιθύμητης αλληλογραφίας που παρέχει ήδη το Mozilla Thunderbird.

1. Εισαγωγή

Το ηλεκτρονικό ταχυδρομείο αποτελεί μια από τις σημαντικότερες υπηρεσίες που προσφέρει το διαδίκτυο. Το κυριότερο ίσως πλεονέκτημα της υπηρεσίας είναι ότι παρέχεται δωρεάν, αν εξαιρέσει κανείς τη χρέωση πρόσβασης στο διαδίκτυο. Η έλλειψη χρέωσης, όμως, είναι και η πηγή του προβλήματος της ανεπιθύμητης ηλεκτρονικής αλληλογραφίας (spam): καθότι δωρεάν, πολλοί χρησιμοποιούν το ηλεκτρονικό ταχυδρομείο για να αποστέλλουν διαφημιστικά και συνήθως ανεπιθύμητα μηνύματα σε χιλιάδες ή εκατομμύρια χρήστες. Υπολογίζεται ότι τα ανεπιθύμητα αυτά μηνύματα αποτελούν περίπου το 60% της διακινούμενης ηλεκτρονικής αλληλογραφίας.

Τα ανεπιθύμητα μηνύματα ηλεκτρονικού ταχυδρομείου σπαταλούν τους πόρους του διαδικτύου αλλά και το το χρόνο των χρηστών, αφού τους υποχρεώνουν να τα αφαιρούν χειρωνακτικά από τα εισερχόμενα μηνύματά τους. Η αντιμετώπιση του προβλήματος, που πλέον έχει λάβει διαστάσεις επιδημίας, αποτελεί στόχο πολλών ερευνητών και εταιρειών. Μια από τις πιο επιτυχημένες μεθόδους που έχουν προταθεί είναι η χρήση προγραμμάτων αυτόματης ταξινόμησης μηνυμάτων σε κατηγορίες, που χρησιμοποιούν συνήθως αλγορίθμους μηχανικής μάθησης. Στην προκειμένη περίπτωση, τα προγράμματα αυτά χρησιμοποιούνται ως φίλτρα που ταξινομούν τα εισερχόμενα μηνύματα ως επιθυμητά (ham) ή ανεπιθύμητα (spam), αφού εκπαιδευθούν σε παλαιότερα μηνύματα που έχουν ταξινομηθεί χειρωνακτικά.

Στην παρούσα εργασία κατασκευάσαμε ένα τέτοιο φίλτρο, το οποίο ενσωματώθηκε στο πρόγραμμα ανάγνωσης ηλεκτρονικής αλληλογραφίας Mozilla Thunderbird.[1] Το φίλτρο χρησιμοποιεί μια βιβλιοθήκη λογισμικού που είχε αναπτυχθεί σε προηγούμενη εργασία , η οποία παρέχει υλοποιήσεις πολλών παραλλαγών του απλοϊκού ταξινομητή Bayes (Naïve Bayes) και μηχανισμούς εφαρμογής τους σε μηνύματα ηλεκτρονικού ταχυδρομείου. Το φίλτρο της παρούσας εργασίας υποστηρίζει, επίσης, αποδείξεις ανθρώπινης αλληλεπίδρασης (human interaction proofs – HIPs), μια προσέγγιση στην οποία οι αποστολείς καλούνται να λύσουν απλούς γρίφους που απαιτούν, όμως, ανθρώπινη νοημοσύνη, προκειμένου να αποκλειστούν συστήματα αυτόματης μαζικής αποστολής ανεπιθύμητων μηνυμάτων [4]. Το φίλτρο της εργασίας, που ονομάζεται AUEB Spam Filter, διατίθεται ελεύθερα ως λογισμικό ανοικτού πηγαίου κώδικα, όπως και το ίδιο το Mozilla Thunderbird.

[1] Βλ. http://www.mozilla.com/thunderbird/ .

Στο επόμενο, το δεύτερο, κεφάλαιο της εργασίας παρουσιάζεται αρχικά η γενική μορφή των απλοϊκών ταξινομητών Bayes και ο τρόπος χρήσης τους κατά την κατάταξη μηνυμάτων. Παρέχονται επίσης παραπομπές προς εργασίες που εξηγούν τον τρόπο λειτουργίας του φίλτρου που διαθέτει ήδη το Thunderbird.[2] Στη συνέχεια παρουσιάζονται αναλυτικότερα οι παραλλαγές των απλοϊκών ταξινομητών Bayes που υποστηρίζει το φίλτρο της εργασίας, καθώς και ο πρόσθετος μηχανισμός αποδείξεων ανθρώπινης αλληλεπίδρασης που επίσης παρέχει το φίλτρο.

Στο τρίτο κεφάλαιο παρουσιάζονται τα πειράματα που διεξήχθησαν προκειμένου να συγκριθεί το φίλτρο της εργασίας με το προϋπάρχον φίλτρο του Thunderbird, καθώς και τα αποτελέσματά τους. Τέλος, στο τέταρτο κεφάλαιο συνοψίζεται η εργασία και προτείνονται πιθανές μελλοντικές προεκτάσεις της.

1.1 Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαίτερα τον επιβλέποντα καθηγητή της εργασίας μου κύριο Ίωνα Ανδρουτσόπουλο, για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου αυτή την εργασία, για την καθοδήγησή του καθ' όλη τη διάρκειά της και κυρίως για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα ενδιαφέρον αντικείμενο. Επίσης, θα ήθελα να ευχαριστήσω τον Άρη Κοσμόπουλο για την αμέριστη βοήθειά του στα πειράματα της εργασίας, αλλά και για την παροχή της βιβλιοθήκης της εργασίας του.

[2] Βλ. http://wiki.mozilla.org/Thunderbird2:JunkMail , http://lxr.mozilla.org/mozilla/source/mailnews/extensions/bayesian-spam-filter/src/ .

2. Θεωρητική Περιγραφή του Συστήματος

Στο παρόν κεφάλαιο θα περιγράψουμε αρχικά τη γενική μορφή των απλοϊκών ταξινομητών Bayes και τον τρόπο χρήσης τους κατά την κατάταξη μηνυμάτων. Κατόπιν θα περιγράψουμε τον τρόπο λειτουργίας του φίλτρου που διαθέτει ήδη το Thunderbird. Στη συνέχεια θα παρουσιάσουμε αναλυτικότερα τις παραλλαγές των απλοϊκών ταξινομητών Bayes που υποστηρίζει το φίλτρο της εργασίας, καθώς και τον πρόσθετο μηχανισμό αποδείξεων ανθρώπινης αλληλεπίδρασης που επίσης παρέχει το φίλτρο.

2.1 Οι απλοϊκοί ταξινομητές Bayes ως φίλτρα μηνυμάτων

Τα φίλτρα που βασίζονται στους απλοϊκούς ταξινομητές Bayes (Naïve Bayes, NB) παριστάνουν κάθε μήνυμα ως ένα διάνυσμα $\vec{x} = (x_1, x_2, x_3, ..., x_n)$, όπου τα χαρακτηριστικά (features) x_1, … x_n αντιστοιχούν το καθένα σε μια διαφορετική λέξη. Στην απλούστερη περίπτωση, τα χαρακτηριστικά δείχνουν αν είναι οι αντίστοιχες λέξεις εμφανίζονται (τιμή 1) ή όχι (τιμή 0) στο κείμενο του μηνύματος. Εναλλακτικά, οι τιμές των χαρακτηριστικών μπορεί να είναι οι συχνότητες εμφάνισης των λέξεων στο μήνυμα, πιθανώς κανονικοποιημένες με τρόπους που θα εξετάσουμε αργότερα.

Η επιλογή των λέξεων για τις οποίες θα υπάρχουν χαρακτηριστικά στα διανύσματα μπορεί να γίνει με μεθόδους επιλογής χαρακτηριστικών (feature selection), όπως για παράδειγμα με το μέτρο του πληροφοριακού κέρδους (information gain), ή απλούστερα εισάγοντας στα διανύσματα ένα χαρακτηριστικό για κάθε λέξη που εμφανίζεται τουλάχιστον π.χ. 5 φορές στα μηνύματα εκπαίδευσης.

Χρησιμοποιώντας το θεώρημα του Bayes [2], η πιθανότητα να ανήκει ένα μήνυμα με διάνυσμα $\vec{x}$ στην κατηγορία c (όπου c η κατηγορία spam ή ham) είναι:

$$P(C = c \mid \vec{X} = \vec{x}) = \frac{P(C = c) \cdot P(\vec{X} = \vec{x} \mid C = c)}{\sum_{k \in \{ham, spam\}} P(C = k) \cdot P(\vec{X} = \vec{x} \mid C = c)}$$

Οι ταξινομητές NB κάνουν την απλοϊκή παραδοχή πως οι τιμές των χαρακτηριστικών είναι ανεξάρτητες δεδομένης της κατηγορίας, οπότε ο παραπάνω τύπος γίνεται:

$$P(C=c \mid \vec{X}=\vec{x}) = \frac{P(C=c)\cdot\prod_{i=1}^{n} P(X_i=x_i \mid C=c)}{\sum_{k\in\{ham,spam\}} P(C=k)\cdot\prod_{i=1}^{n} P(X_i=x_i \mid C=c)}$$

Τα $P(X_i|C)$ και $P(C)$ είναι δυνατόν να εκτιμηθούν εύκολα από τα δεδομένα εκπαίδευσης. Αν και στην πραγματικότητα η παραδοχή ανεξαρτησίας δεν ισχύει, οι ταξινομητές NB επιτυγχάνουν πολύ καλά αποτελέσματα.

Όλες οι μορφές των ταξινομητών NB που χρησιμοποιούνται στην παρούσα εργασία κατατάσσουν τα μηνύματα υπολογίζοντας τις $P(spam \mid \vec{x})$ και $P(ham \mid \vec{x})$, όπου $\vec{x}$ το διάνυσμα του υπό κατάταξη μηνύματος. Ακριβέστερα, υπολογίζουν το λόγο των δύο παραπάνω πιθανοτήτων. Οι παρανομαστές των δύο πιθανοτήτων είναι ίδιοι, οπότε αγνοούνται. Χρησιμοποιώντας λογαρίθμους, ο λόγος γίνεται διαφορά:

$$[\log(P(spam)) + \log(P(\vec{x} \mid spam))] - [\log(P(ham)) + \log(P(\vec{x} \mid ham))]$$

και με την παραδοχή ανεξαρτησίας:

$$[\log(P(spam)) + \sum_{i=1}^{n} \log(P(x_i \mid spam))] - [\log(P(ham)) + \sum_{i=1}^{n} \log(P(x_i \mid ham))]$$

Αν η παραπάνω διαφορά υπερβαίνει ένα κατώφλι δ, τότε το μήνυμα κατατάσσεται ως ανεπιθύμητο, διαφορετικά ως επιθυμητό.

2.2 Το προϋπάρχον φίλτρο του Thunderbird

Το προϋπάρχον φίλτρο του Thunderbird αποτελεί υλοποίηση του συστήματος SpamBayes [1] στο περιβάλλον ανάπτυξης του Mozilla.[3] Το SpamBayes ακολούθησε αρχικά, όπως και πολλά άλλα φίλτρα, την προσέγγιση του Paul Graham, η οποία είναι παρόμοια με τους ταξινομητές NB, αλλά εμπεριέχει αρκετές, αμφισβητήσιμες από μαθηματικής σκοπιάς, μετατροπές, καθώς και παραμέτρους των οποίων οι τιμές δεν είναι εύκολο να επιλεγούν.[4] Ένα πρόβλημα της προσέγγισης του Paul Graham, αλλά και πολλών από τις παραλλαγές των ταξινομητών NB, είναι πως οι πιθανότητες που επιστρέφουν παρουσιάζουν πολύ ακραίες τιμές, με αποτέλεσμα το φίλτρο να

[3] Βλ. http://spambayes.sourceforge.net/ .
[4] Βλ. http://www.paulgraham.com/spam.html .

εμφανίζεται συνήθως εξαιρετικά βέβαιο για τις αποφάσεις του, ακόμα και όταν παίρνει λάθος αποφάσεις. Προκειμένου να αντιμετωπιστούν τα προβλήματα αυτά, το SpamBayes βασίστηκε στη συνέχεια στην προσέγγιση του Gary Robinson.[5]

2.3 Το φίλτρο της εργασίας

Η ανάπτυξη του φίλτρου της εργασίας έγινε σε C++ και JavaScript, σύμφωνα με τα πρότυπα ανάπτυξης του Mozilla Development Center. Το φίλτρο χρησιμοποιεί δύο βασικές μορφές απλοϊκών ταξινομητών Bayes (NB) [2]: την πολυμεταβλητή μορφή Bernoulli (Multivariate Bernoulli NB) και την πολυωνυμική μορφή (Multinomial NB). Η πολυωνυμική μορφή NB υλοποιήθηκε σε τρεις παραλλαγές: (α) με δυαδικά χαρακτηριστικά, (β) με χαρακτηριστικά που αντιστοιχούν σε συχνότητες (term frequencies, TF) και (γ) με χαρακτηριστικά που αντιστοιχούν σε μετασχηματισμένες συχνότητες [3]. Ο μετασχηματισμός της περίπτωσης (γ) περιγράφεται παρακάτω.

2.3.1 Πολυμεταβλητή μορφή Bernoulli του απλοϊκού ταξινομητή Bayes

Στην πολυμεταβλητή μορφή Bernoulli [2], κάθε μήνυμα παριστάνεται από ένα δυαδικό διάνυσμα της μορφής $\vec{x} =< x_1, x_2, ..., x_n >$, όπου $x_i \in \{0,1\}$. Κάθε χαρακτηριστικό x_i αντιστοιχεί σε μια διαφορετική λεκτική μονάδα (token) και δείχνει αν η λεκτική μονάδα εμφανίζεται (τιμή 1) ή όχι (τιμή 0) στο μήνυμα. Υποθέτουμε ότι κάθε μήνυμα κατηγορίας c είναι το αποτέλεσμα n ανεξάρτητων δοκιμών Bernoulli, κατά τις οποίες καθορίζεται αν η λεκτική μονάδα t_i , που αντιστοιχεί στο χαρακτηριστικό x_i, εμφανίζεται ή όχι στο μήνυμα. Η υπόθεση ανεξαρτησίας δεν ισχύει στην πραγματικότητα, αλλά παρ' όλα αυτά τα αποτελέσματα του ταξινομητή είναι συχνά ικανοποιητικά.

Με τις παραπάνω υποθέσεις, η πιθανότητα $P(\vec{x} \mid c)$ (βλ. ενότητα 2.1) γίνεται:

$$P(\vec{x} \mid c) = \prod_{i=1}^{n} P(t_i \mid c)^{x_i} \cdot (1 - P(t_i \mid c))^{1-x_i}$$

[5] Βλ. http://www.linuxjournal.com/article/6467 και http://spambayes.sourceforge.net/background.html.

Οι πιθανότητες P(t|c) εκτιμώνται ως ακολούθως. Οι σταθεροί όροι στον αριθμητή και τον παρανομαστή προστίθενται για να αποφεύγονται οι μηδενικές εκτιμήσεις.

$$P(t \mid c) = \frac{1 + M_{t,c}}{2 + M_c}$$

όπου:

$M_{t,c}$: ο αριθμός των μηνυμάτων της κατηγορίας c που εμπεριέχουν την t.

M_c : ο αριθμός των μηνυμάτων της κατηγορίας c.

2.3.2 Πολυωνυμικός απλοϊκός ταξινομητής Bayes

Μορφή 1 : Με δυαδικές ιδιότητες

Σε αυτή τη μορφή, χρησιμοποιούνται πάλι δυαδικά διανύσματα, όπως ακριβώς στην προηγούμενη μορφή, αλλά οι πιθανότητες p(t | c) εκτιμώνται ως εξής:

$$P(t \mid c) = \frac{1 + N_{t,c}}{n + N_c}$$

όπου $N_{t,c}$ ο αριθμός εμφανίσεων του token t στα μηνύματα της κατηγορίας c και N_c το άθροισμα των n διαφορετικών $N_{t,c}$. Άλλη μια διαφορά εντοπίζεται στον τύπο υπολογισμού του $P(\vec{x} \mid c)$: δεν λαμβάνονται τώρα υπόψη οι λεκτικές μονάδες που απουσιάζουν από το μήνυμα, με αποτέλεσμα ο όρος $(1 - p(t_i \mid c))^{1-x_i}$ να απαλείφεται. Το πώς προκύπτουν αυτοί οι τύποι εξηγείται στην εργασία [2].

Μορφή 2: Με χαρακτηριστικά TF

Στη μορφή αυτή, οι τιμές x_i των διανυσμάτων δείχνουν πόσες φορές εμφανίζονται οι αντίστοιχες λεκτικές μονάδες στο μήνυμα. Η πιθανότητα $P(\vec{x} \mid c)$ υπολογίζεται πάλι ως $\prod_{i=1}^{n} p(t_i \mid c_s)^{x_i}$. Οι πιθανότητες P(t|c) υπολογίζονται επίσης όπως στην προηγούμενη μορφή. Και πάλι, το πώς προκύπτουν αυτοί οι τύποι εξηγείται στην εργασία [2].

Μορφή 3: Με μετασχηματισμένες ιδιότητες TF

Ο υπολογισμός του $P(\vec{x} \mid c)$ γίνεται όπως ακριβώς στην προηγούμενη μορφή. Οι τιμές x_i των διανυσμάτων, όμως, δεν δείχνουν τώρα απευθείας πόσες φορές εμφανίζονται οι αντίστοιχες λεκτικές μονάδες στο μήνυμα (term frequencies, TF), αλλά υφίστανται τον παρακάτω μετασχηματισμό [3].

Πρώτα λογαριθμίζεται το άθροισμα της τιμής TF με την μονάδα. Η μονάδα προστίθεται για να μην προκύπτουν μηδενικά ορίσματα στο λογάριθμο, ενώ ο λογάριθμος χρησιμοποιείται προκειμένου να αντιμετωπιστεί το πρόβλημα ότι οι κατανομές των λεκτικών μονάδων στα μηνύματα δεν ακολουθούν πολυωνυμική κατανομή (βλ. εργασία [3] για περισσότερες εξηγήσεις).

Έπειτα πολλαπλασιάζουμε το αποτέλεσμα με τον παράγοντα $\log \frac{\sum_k 1}{\sum_k \delta_{ki}}$, που εκφράζει την ανάστροφη συχνότητα εγγράφων (inverse document frequency, IDF) της λεκτικής μονάδας στην οποία αντιστοιχεί το χαρακτηριστικό x_i. Το k είναι δείκτης προς κάθε ένα μήνυμα εκπαίδευσης και το δ_{ki} είναι 1 αν η λεκτική μονάδα που αντιστοιχεί στο χαρακτηριστικό x_i εμφανίζεται στο μήνυμα j ή 0 αν δεν εμφανίζεται. Οι τιμές IDF χρησιμοποιούνται προκειμένου να αυξηθούν τα βάρη των πιο σπάνιων λεκτικών μονάδων [3].

Τέλος διαιρούμε το έως τώρα αποτέλεσμα με την τιμή $\sqrt{\sum_i (d_i)^2}$, όπου d_i οι μετασχηματισμένες τιμές των χαρακτηριστικών (για το συγκεκριμένο μήνυμα) που έχουν προκύψει μέχρι και το προηγούμενο βήμα.

2.3.3 Λευκές λίστες

Το φίλτρο που αναπτύχθηκε κατά τη διάρκεια της εργασίας υποστηρίζει και λευκές λίστες (white-listing). Η λευκή λίστα κάθε χρήστη περιέχει όλες τις διευθύνσεις που περιλαμβάνονται στο βιβλίο διευθύνσεών του (address book). Όταν η χρήση λευκής λίστας είναι ενεργοποιημένη, τα εισερχόμενα μηνύματα που προέρχονται από διευθύνσεις τις λίστας κατατάσσονται άμεσα ως επιθυμητά, χωρίς να εφαρμόζεται σε αυτά το φίλτρο του απλοϊκού ταξινομητή Bayes.

2.3.4 Αποδείξεις ανθρώπινης αλληλεπίδρασης

Όπως προαναφέρθηκε, το φίλτρο της εργασίας υποστηρίζει επίσης αποδείξεις ανθρώπινης αλληλεπίδρασης (Human Interactive Proofs, HIPs) [4]. Όταν είναι ενεργοποιημένος ο μηχανισμός HIPs του φίλτρου, όποτε ένα μήνυμα κατατάσσεται ως ανεπιθύμητο, το φίλτρο ζητά από τον αποστολέα να απαντήσει σε μια προκαθορισμένη εύκολη ερώτηση φυσικής γλώσσας που έχει ορίσει ο παραλήπτης (π.χ. «Ποια είναι η πρωτεύσουσα της Ελλάδας;» – ο παραλήπτης μπορεί να αλλάζει περιοδικά την ερώτησή του). Αν ο αποστολέας απαντήσει ορθά μέσα σε ένα εύλογο χρονικό διάστημα, το μήνυμα, που προηγουμένως είχε καταταγεί ως ανεπιθύμητο, κατατάσσεται ως επιθυμητό και η διεύθυνση του αποστολέα προστίθεται στη λευκή λίστα του παραλήπτη. Ο μηχανισμός αυτός βασίζεται στην υπόθεση πως οι αποστολείς ανεπιθύμητων διαφημιστικών μηνυμάτων δεν μπορούν να απαντήσουν σε διαφορετικούς ανά χρήστη γρίφους που απαιτούν ανθρώπινη νοημοσύνη, εξαιτίας του όγκου των μηνυμάτων που αποστέλλουν και άρα και του μεγάλου αριθμού γρίφων που θα δέχονται.

Τονίζεται ότι ο μηχανισμός αυτός βρίσκεται σε δοκιμαστικό στάδιο και δεν θα πρέπει να χρησιμοποιείται ακόμη στην πράξη. Ένα σημαντικό πρόβλημά του είναι πως οι αποστολείς ανεπιθύμητων μηνυμάτων ενδέχεται να αρχίσουν να στέλνουν μηνύματα χρησιμοποιώντας ως διευθύνσεις αποστολέων τις διευθύνσεις υπαρκτών, αθώων χρηστών του διαδικτύου. Αυτό θα έχει ως αποτέλεσμα οι γρίφοι να καταλήγουν σε άσχετους χρήστες, ουσιαστικά επιτείνοντας το πρόβλημα της ανεπιθύμητης αλληλογραφίας. Το πρόβλημα αυτό ενδέχεται να λυθεί στο μέλλον, αν διαδοθεί ευρέως η χρήση μηχανισμών πιστοποίησης αποστολέα (DKIM, SenderID).[6] Ένα άλλο πρόβλημα είναι πως στο φίλτρο της εργασίας η αποστολή των γρίφων και των απαντήσεών τους γίνεται και αυτή μέσω ηλεκτρονικού ταχυδρομείου, με αποτέλεσμα να είναι δυνατόν ένας γρίφος ή η απάντησή του να θεωρηθεί από άλλο φίλτρο ανεπιθύμητο μήνυμα και να μην παραδοθεί.

[6] Βλ. http://dkim.org/ και http://www.microsoft.com/mscorp/safety/technologies/senderid/default.mspx

3 Πειραματική Προσέγγιση – Συγκρίσεις

Σε αυτό το κεφάλαιο θα παρουσιάσουμε τα αποτελέσματα των πειραμάτων με τις διαφορετικές μορφές του απλοϊκού ταξινομητή Bayes που εκτελέσθηκαν στη διάρκεια της εργασία, καθώς και τα συμπεράσματα που προκύπτουν από αυτά.

3.1 Περιγραφή Πειραματικής Μεθόδου

Στα πειράματα της παρούσας εργασίας χρησιμοποιήθηκε το σύνολο μηνυμάτων Enron-Spam [2], που περιέχει επιθυμητά και ανεπιθύμητα μηνύματα έξι ψευδο-χρηστών. Η συλλογή αυτή χρησιμοποιήθηκε και στην εργασία [5], κάτι που καθιστά εφικτή τη σύγκριση και σύνθεση των αποτελεσμάτων αυτής της εργασίας και των προηγούμενων. Στα μηνύματα της παραπάνω συλλογής αξιολογούμε μόνο το σώμα κάθε μηνύματος, αγνοώντας συνημμένα αρχεία, ετικέτες HTML και κεφαλίδες (πλην του θέματος, που προστίθεται στο σώμα).

Σύμφωνα με την εργασία [5], τα καλύτερα αποτελέσματα επιτυγχάνονται όταν χρησιμοποιούνται όλα τα χαρακτηριστικά, για όλες τις λεκτικές μονάδες, ασχέτως του πληροφοριακού κέρδους που αποκομίζουμε από αυτά.Ακολουθούμε, επομένως, αυτή την προσέγγιση στα πειράματά μας.

Για την αναπαράσταση των αποτελεσμάτων θα χρησιμοποιηθούν οι καμπύλες ROC, στις οποίες ο κατακόρυφος άξονας παριστάνει το ποσοστό ανάκλησης ανεπιθύμητων μηνυμάτων (Spam Recall, SR), ενώ ο οριζόντιος άξονας το ποσοστό ανάκλησης επιθυμητών μηνυμάτων (Ham Recall, HR). Τα SR και HR αντιστοιχούν στους όρους ευαισθησία (sensitivity) και σαφήνεια (specificity), αντίστοιχα, που συνήθως χρησιμοποιούνται στα διαγράμματα ROC. Ακριβέστερα, στα παρακάτω διαγράμματα οι τιμές του οριζόντιου άξονα αντιστοιχούν στο 1 – HR. Τα SR και HR ορίζονται ως εξής:

$$SR = \frac{TP}{TP + FN} \quad \text{και} \quad HR = \frac{TN}{TN + FP}$$

όπου TP : true positives, δηλαδή πόσα ανεπιθύμητα μηνύματα κατετάγησαν σωστά

TN: true negatives, δηλαδή πόσα επιθυμητά μηνύματα κατετάγησαν σωστά

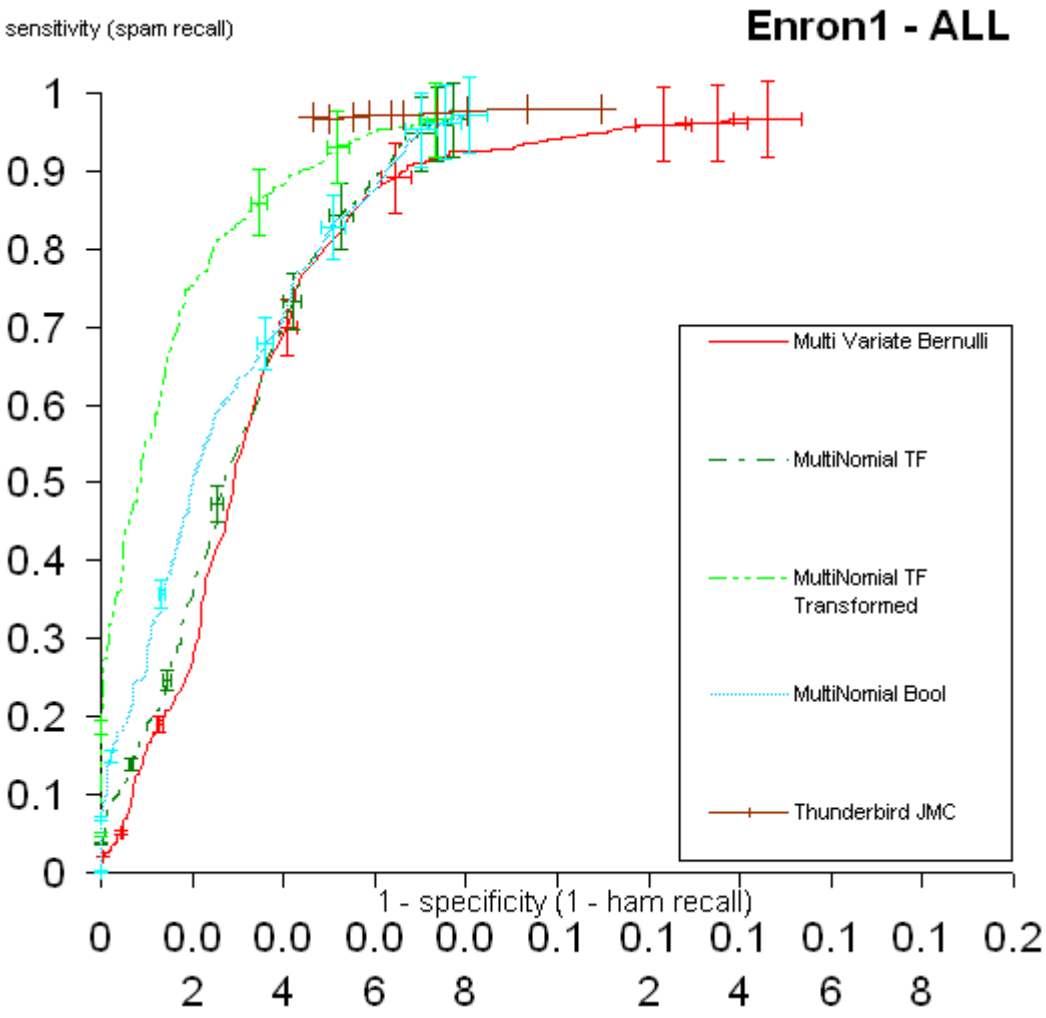
FP: false positive, αριθμός ανεπιθύμητων μηνυμάτων που κατετάγησαν λανθασμένα

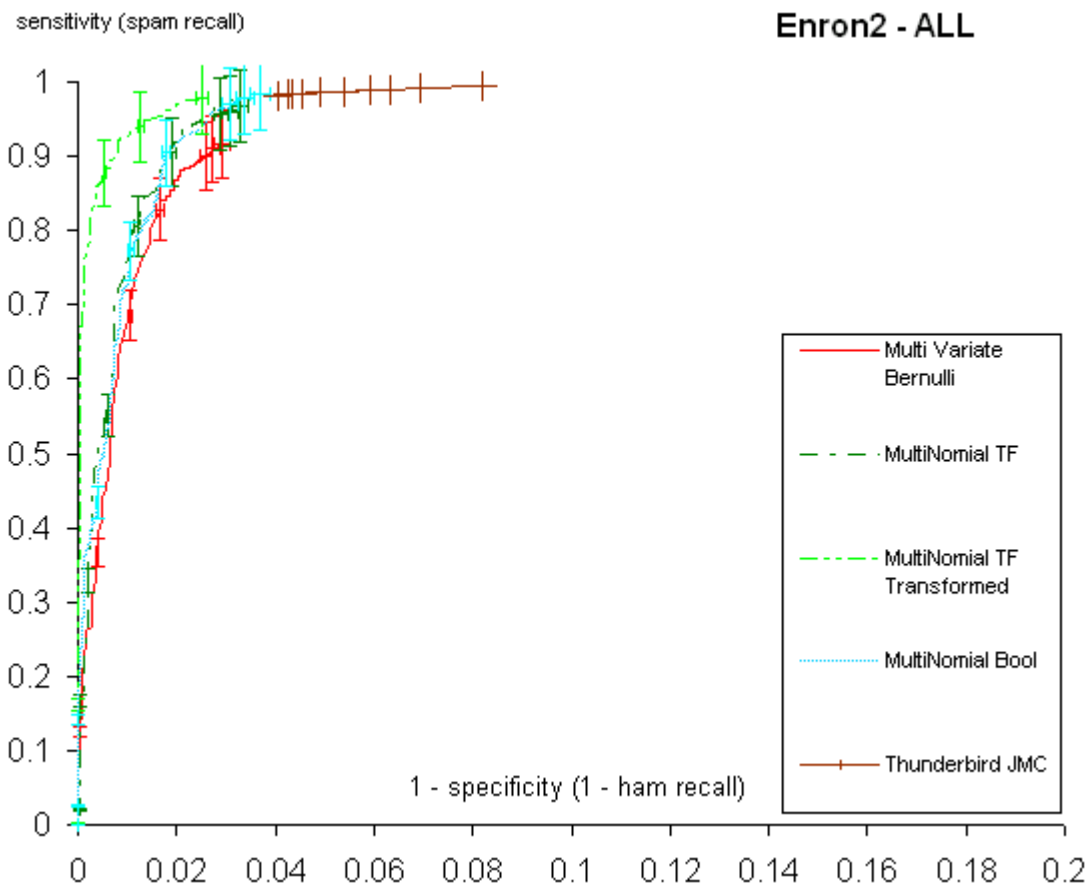
FN: false negatives, αριθμός επιθυμητών μηνυμάτων που κατετάγησαν λανθασμένα

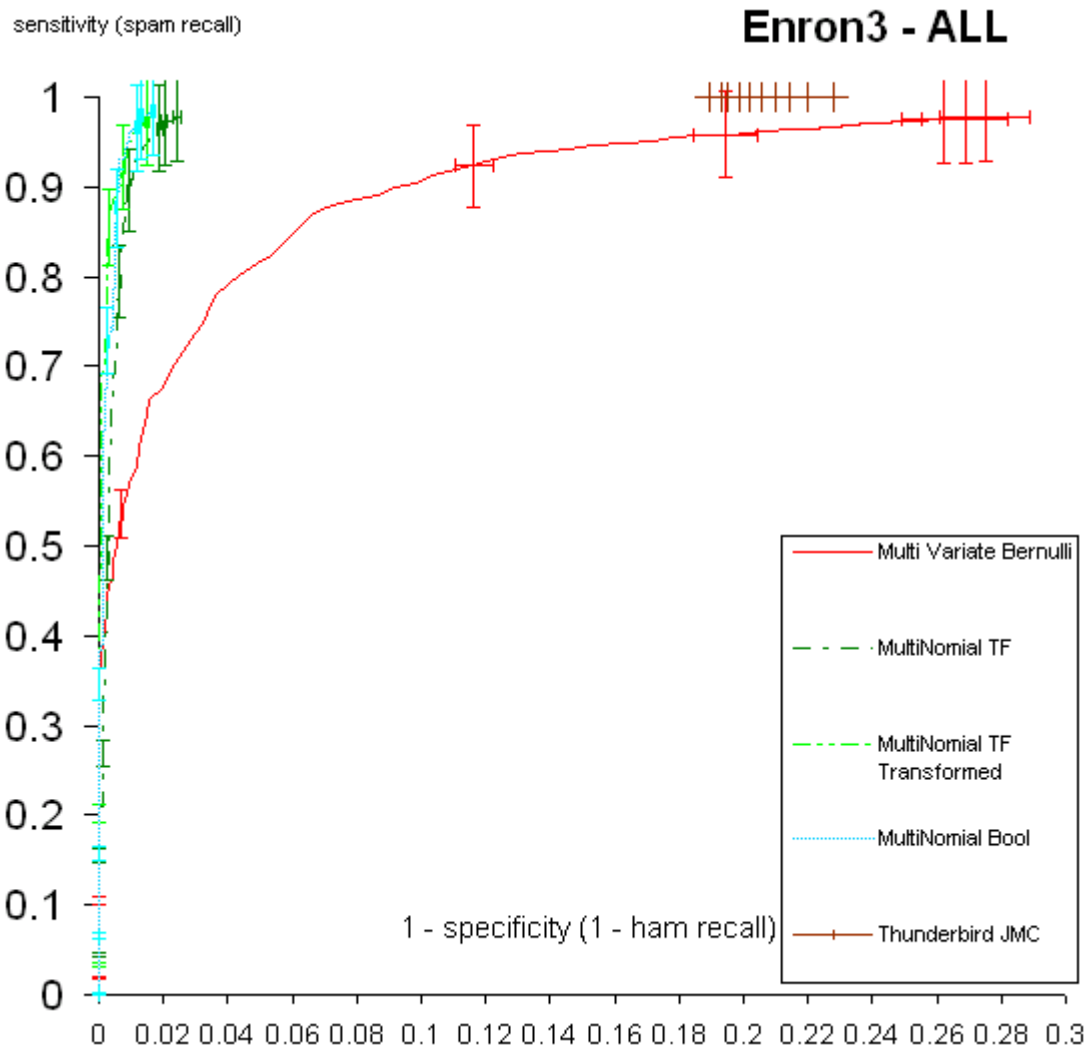
Για να δημιουργήσουμε τα σημεία (ζεύγη τιμών SR και HR) στις καμπύλες ROC, επαναλαμβάνουμε τις ταξινομήσεις των μηνυμάτων αξιολόγησης για διάφορες τιμές του κατωφλίου κατάταξης. Οι καμπύλες που βρίσκονται ψηλότερα αντιστοιχούν στα καλύτερα φίλτρα, αφού τα φίλτρα αυτά εντοπίζουν περισσότερα ανεπιθύμητα μηνύματα (υψηλότερο SR), ενώ ταυτόχρονα επιτρέπουν στον ίδιο αριθμό επιθυμητών μηνυμάτων (ίδιο HR) να περάσουν το φίλτρο. Όπως στις εργασίες [2] και [5], η αξιολόγηση γίνεται κατά δέσμες χρονικά διατεταγμένων μηνυμάτων (οι δέσμες θα μπορούσαν να αντιστοιχούν π.χ. στα εισερχόμενα μηνύματα μιας ημέρας ή εβδομάδας). Σε κάθε δέσμη μηνυμάτων αξιολόγησης, τα φίλτρα έχουν εκπαιδευθεί στα μηνύματα των προηγούμενων δεσμών, τα οποία έχουν καταταγεί χειρωνακτικά (υποτίθεται ότι οι χρήστες έχουν διορθώσει τα λάθη των φίλτρων). Στα πειράματά μας, κάθε δέσμη περιλαμβάνει 100 μηνύματα, όπως στις εργασίες [2] και [5].

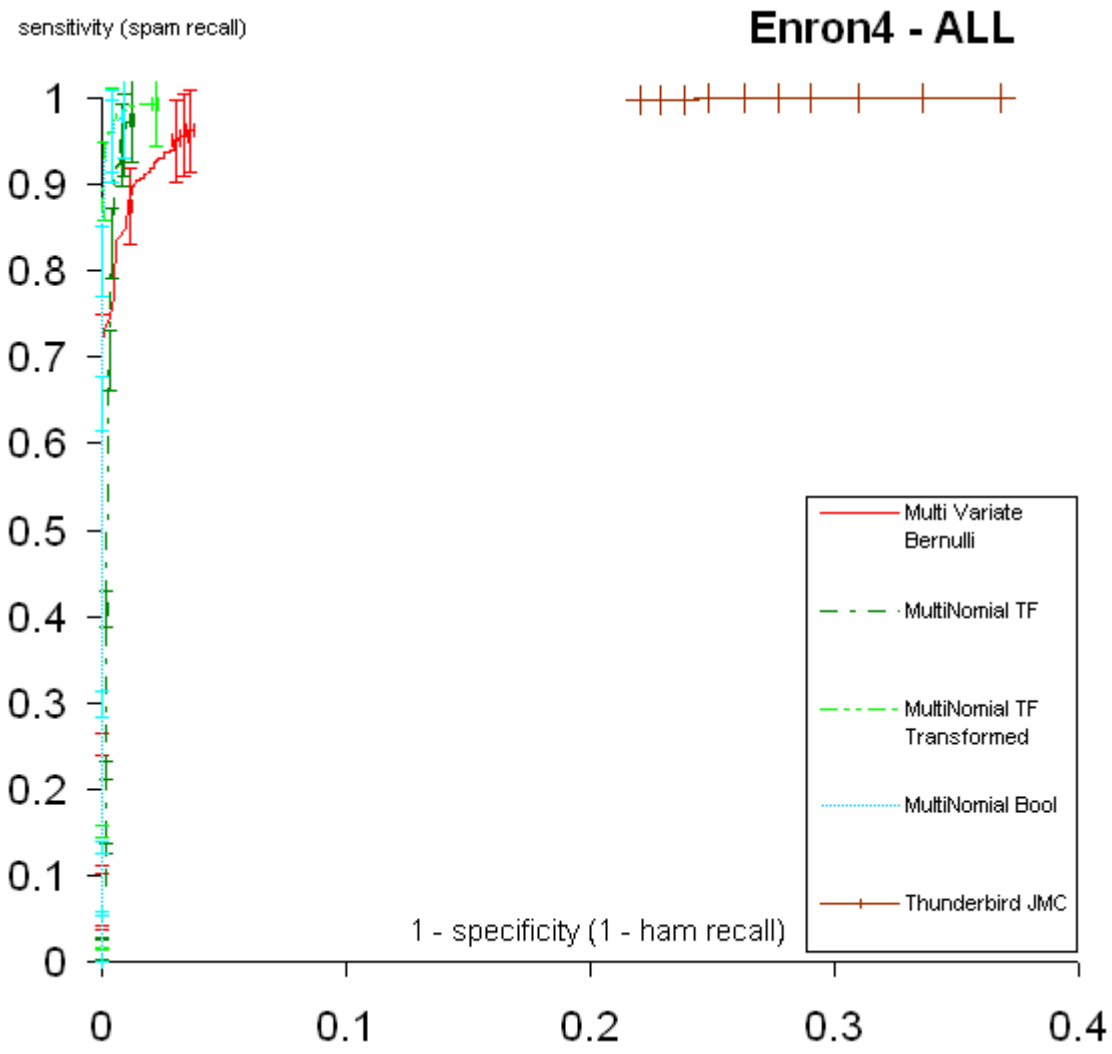
3.2 Πειραματικά αποτελέσματα ανά χρήστη

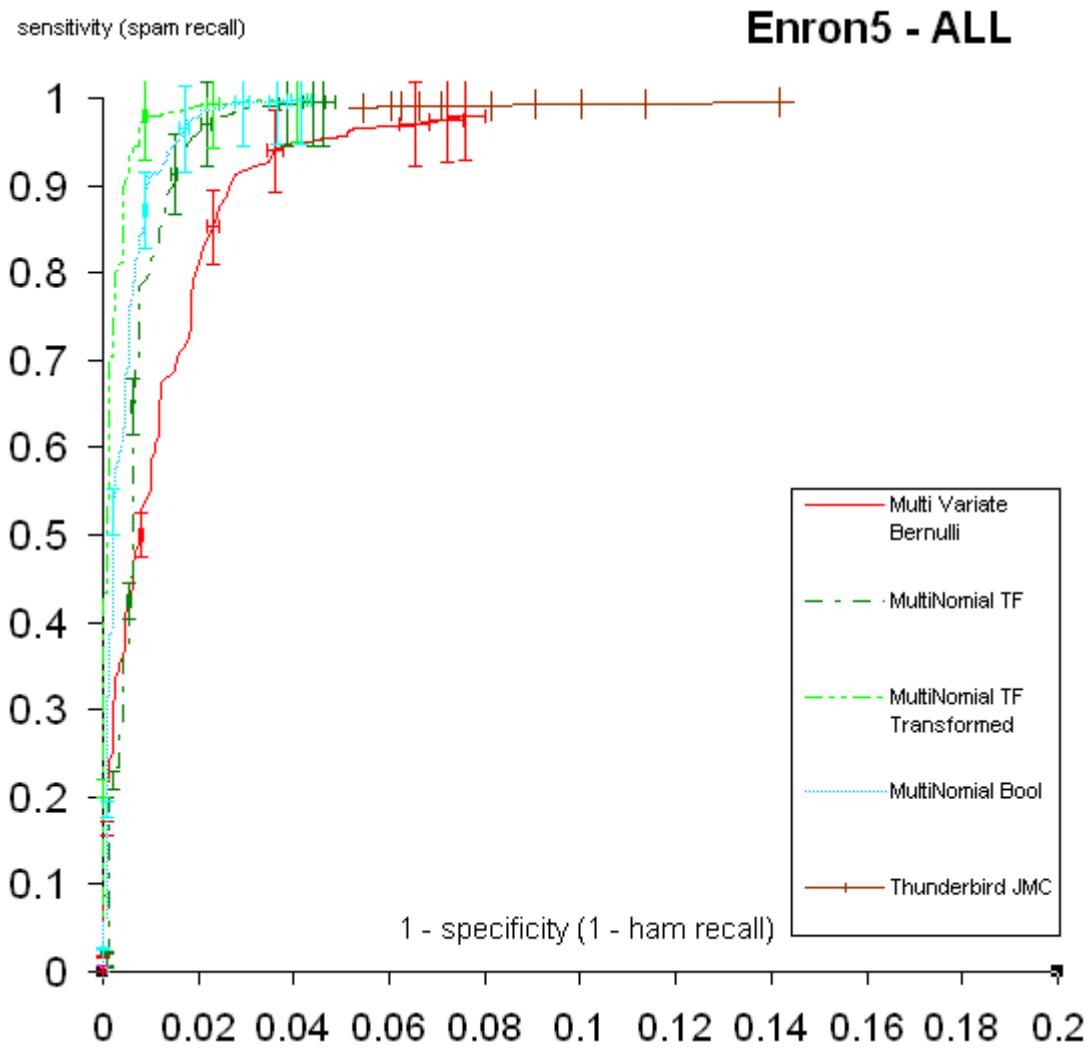
Στα παρακάτω διαγράμματα φαίνονται τα πειραματικά αποτελέσματα ανά ψευδο-χρήστη του Enron-Spam και συγκεντρωτικά για όλους τους ψευδο-χρήστες μαζί. Οι ράβδοι λάθους αντιστοιχούν σε διαστήματα εμπιστοσύνης 95%.

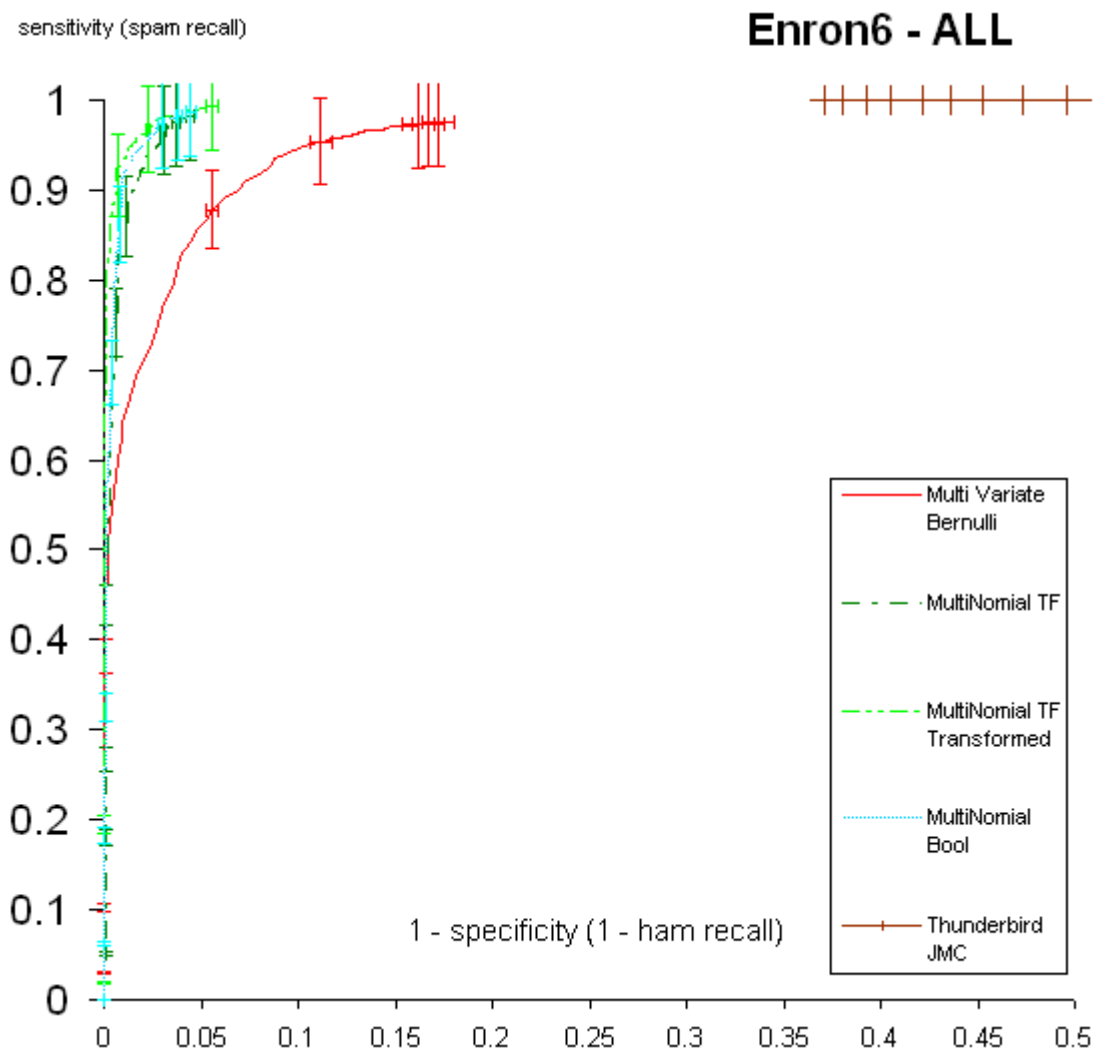



Enron2 - ALL
sensitivity (spam recall)
1 - specificity (1 - ham recall)
Multi Variate Bernulli
MultiNomial TF
MultiNomial TF Transformed
MultiNomial Bool
Thunderbird JMC
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2


Enron3 - ALL
sensitivity (spam recall)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16 0.18 0.2 0.22 0.24 0.26 0.28 0.3
1 - specificity (1 - ham recall)
Multi Variate Bernulli
MultiNomial TF
MultiNomial TF Transformed
MultiNomial Bool
Thunderbird JMC

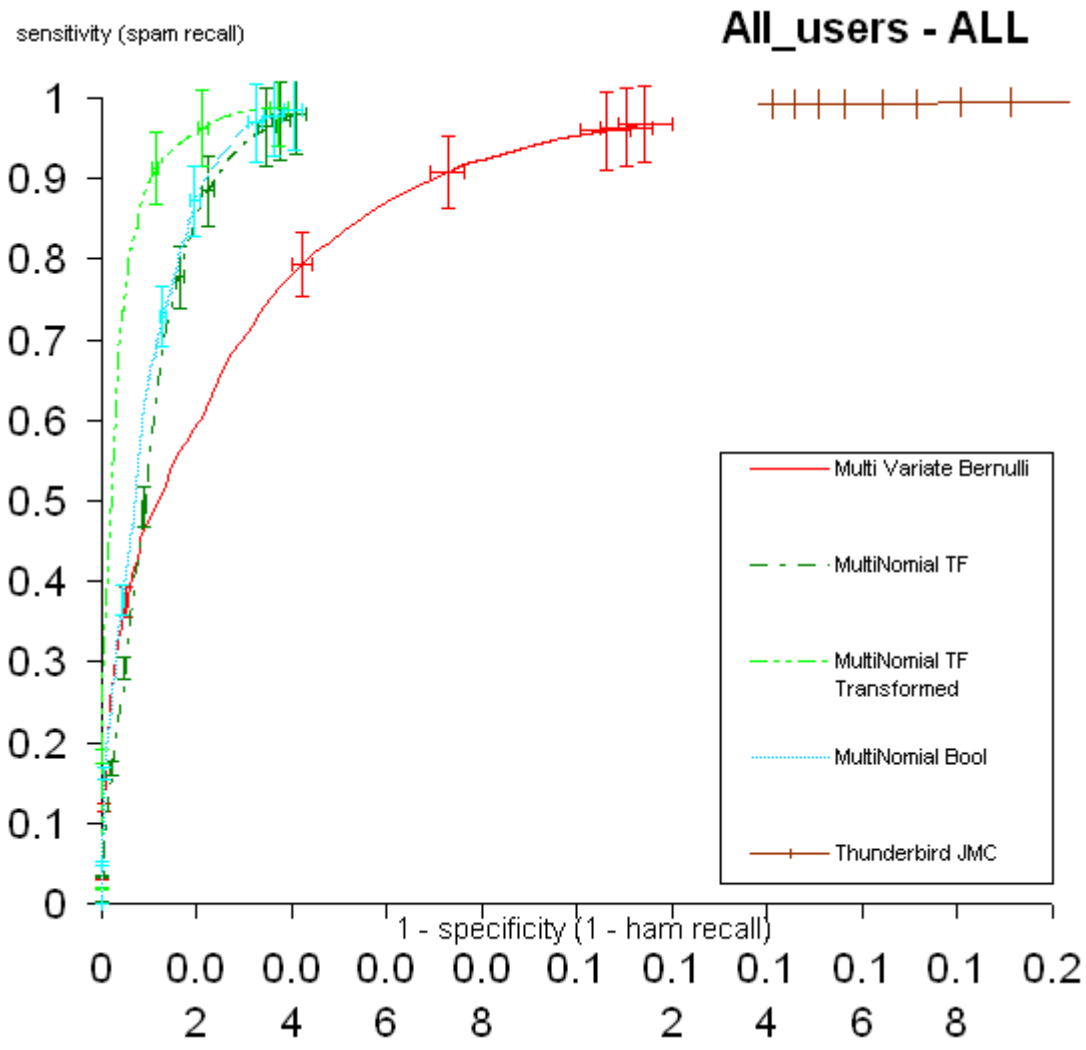

Enron4 - ALL
sensitivity (spam recall)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
1 - specificity (1 - ham recall)
Multi Variate Bernulli
MultiNomial TF
MultiNomial TF Transformed
MultiNomial Bool
Thunderbird JMC


Enron5 - ALL
sensitivity (spam recall)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2
1 - specificity (1 - ham recall)
Multi Variate Bernulli
MultiNomial TF
MultiNomial TF Transformed
MultiNomial Bool
Thunderbird JMC



Παρατηρούμε ότι σε γενικές γραμμές η πολυωνυμική μορφή με μετασχηματισμένα χαρακτηριστικά TF του απλοϊκού ταξινομητή Bayes υπερτερεί στην περιοχή υψηλού HR (1 – HR τείνει στο μηδέν) που κυρίως μας ενδιαφέρει (θέλουμε να μην χάνουμε επιθυμητά μηνύματα), ιδιαίτερα στους ψευδο-χρήστες Enron1, Enron2 και Enron5. Στους υπόλοιπους τρεις ψευδο-χρήστες, οι διαφορές από τις άλλες δύο μορφές του πολυωνυμικού απλοϊκού ταξινομητή Bayes είναι δυσδιάκριτες, αλλά και πάλι η πολυωνυμική μορφή με μετασχηματισμένα χαρακτηριστικά TF είναι μεταξύ των κορυφαίων. Η πολυμεταβλητή μορφή Bernoulli είναι εμφανώς η χειρότερη. Τα αποτελέσματα αυτά συμφωνούν με εκείνα των εργασιών [2] και [5]. Το προϋπάρχον φίλτρο του Thunderbird επιτυγχάνει υψηλές τιμές SR, αλλά δεν καταφέρνει να πλησιάσει τιμές του HR κοντά στο 1, περιοχή που κυρίως μας ενδιαφέρει, για καμιά τιμή του κατωφλίου.

Στο παρακάτω διάγραμμα φαίνονται τα αποτελέσματα συγκεντρωτικά, για όλους τους χρήστες μαζί. Είναι και πάλι εμφανής η υπεροχή της πολυωνυμικής μορφής με μετασχηματισμένα χαρακτηριστικά TF.



4 Επίλογος

Στην παρούσα εργασία αναπτύχθηκε ένα φίλτρο ανεπιθύμητης ηλεκτρονικής αλληλογραφίας, που ενσωματώθηκε στο Mozilla Thunderbird. Διατίθεται ελεύθερα ως λογισμικό ανοικτού πηγαίου κώδικα, όπως και το ίδιο το Mozilla Thunderbird. Χρησιμοποιεί μια βιβλιοθήκη λογισμικού που είχε αναπτυχθεί σε προηγούμενη εργασία, η οποία παρέχει υλοποιήσεις πολλών παραλλαγών του απλοϊκού ταξινομητή Bayes. Το φίλτρο της εργασίας χρησιμοποιεί τέσσερις μορφές: την πολυμεταβλητή μορφή Bernoulli, την πολυωνυμική με χαρακτηριστικά TF, την πολυωνυμική με μετασχηματισμένα χαρακτηριστικά TF και την πολυωνυμική με δυαδικά χαρακτηριστικά. Το φίλτρο της παρούσας εργασίας υποστηρίζει, επίσης, αποδείξεις ανθρώπινης αλληλεπίδρασης (HIPs), μηχανισμό που βρίσκεται, όμως, σε δοκιμαστικό στάδιο και δεν αξιολογήθηκε πειραματικά. Πειράματα που διεξήχθησαν στη διάρκεια της εργασίας έδειξαν ότι, για τους σκοπούς της διήθησης ανεπιθύμητης αλληλογραφίας, η καλύτερη μορφή του απλοϊκού ταξινομητή Bayes είναι η πολυωνυμική με μετασχηματισμένα χαρακτηριστικά TF. Με αυτή τη μορφή, το φίλτρο που αναπτύχθηκε επιτυγχάνει καλύτερα αποτελέσματα από το προϋπάρχον φίλτρο του Mozilla Thunderbird.

Μελλοντικά θα ήταν σκόπιμο να αξιολογηθεί πειραματικά και ο μηχανισμός αποδείξεων ανθρώπινης αλληλεπίδρασης, αφού συνδυασθεί με μηχανισμούς πιστοποίησης αποστολέα (DKIM ή SenderID) και αντιμετωπισθούν τα προβλήματα που παρουσιάζει επί του παρόντος. Επίσης, θα ήταν ιδιαίτερα ενδιαφέρον να αξιολογηθεί το φίλτρο της εργασίας στην πράξη, από πραγματικούς χρήστες.

5 Βιβλιογραφία

[1] T.A Meyer και B Whateley, «SpamBayes: Effective open-source, Bayesian based, email classification system». Πρακτικά του *1st Conference on Email and Anti-Spam* (CEAS 2004), Mountain View, CA, ΗΠΑ, 2004..

[2] Β. Μέτσης, Ι. Ανδρουτσόπουλος και Γ. Παλιούρας, «Spam Filtering with Naive Bayes -- Which Naive Bayes?». Πρακτικά του *3rd Conference on Email and Anti-Spam* (CEAS 2006), Mountain View, CA, ΗΠΑ, 2006.

[3] J.D.M. Rennie, L. Shih, J. Teevan και D.R. Karger, «Tackling the Poor Assumptions of Naive Bayes Text Classifiers». Πρακτικά του *20th International Conference on Machine Learning*, Washington DC, 2003.

[4] Δ.Κ. Βασιλάκης, Ι. Ανδρουτσόπουλος και Ε.Φ. Μαγείρου, «A Game-Theoretic Investigation of the Effect of Human Interactive Proofs on Spam E-mail». Πρακτικά του *4th Conference on Email and Anti-Spam* (CEAS 2007), Mountain View, CA, ΗΠΑ, 2007.

[5] Α. Κοσμόπουλος, «*Διήθηση ανεπιθύμητης ηλεκτρονικής αλληλογραφίας με διάφορες μορφές του απλοϊκού ταξινομητή Bayes και διαμοιρασμό φίλτρων μεταξύ χρηστών*», μεταπτυχιακή διπλωματική εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2007.