Rough Set Theory Approach for Filtering Spams from boundary messages in a Chat System

Sanjiban Sekhar Roy[1], Saptarshi Charaborty[1], Swapnil Sourav[1] and Ajith Abraham[2,3]

[1]School of Computing Science and Engineering, VIT University, Vellore
{sanjibanroy09,sssaptarshii,swapnishu}@gmail.com

[2]IT4Innovations,VSB - Technical University of Ostrava, Czech Republic
[3]Machine Intelligence Research Labs (MIR Labs), Washington 98071, USA
ajith.abraham@ieee.org

Abstract—**This paper purports a refreshing spam discovery technology for chat system based on rough set theory. Nowadays, spam is very much allied with a huge chunk of data transferred through internet involving all disturbing and unsolicited contents received via different web-services such as chat systems, e-mail, forums and web logs. In this paper, we have reviewed various past research works of filtering SPAM and propose a novel filtering technique for SPAM especially for chat system with the support of classical rough set theory. Simulation results clearly indicate that our proposed method, can achieve higher accuracy in spam detection as compared to the existing strategies.**

Keywords—spam filtering;web logs; rough set

I. INTRODUCTION

Spam is a severe widespread predicament which induces troubles for approximately every computer users. This concern not only strikes regular users of the cyberspace, but too causes a enormous trouble for companies and institutions in view of the fact that it expenses a vast quantity of capital in vanished output, wasting users' moment in time and involvement of bandwidth. A lot of examines on spam points that spam tolls institutions one thousand millions of dollars annually. It is the misuse of electronic content sending arrangements to fling unwanted, begrudge vastness messages. Among all spams, electronic mail spam(email) is the majority common but this term can be valid to analogous abuses like instantaneous unsolicited messaging, spam used in mobile phone and while searching in the web [1].So, with the increase of communication technology, one of the most significant steps in the advancement of human communication is the promotion and commercialization of internet, which was developed as a packet switching network [1]. Today, there are millions of web sites sharing multimedia data (e.g. slides, photos, videos, etc.), well-known peer-to-peer networks and social networking sites. But, malware, spy ware and spamming activities are therefore also increasing at the same time. In this context, spam is a term used to define all types of unwanted commercial communication and can be categorically demarcated as an electronic message satisfying the following three conditions:

(i) The personal identity of the recipient and context of the message sent are irrelevant because the message is equally applicable to many other potential recipients.

(ii) The beneficiary has not verifiably settled intentional, unambiguous, and still-revocable authorization for it to be sent

(iii) Finally, the communication of the message provides an' inappropriate benefit' to the sender, as solely determined by the recipient [2].

Today, spam is spread through internet using variety of mediums including posting comments in a blog or in a video, e-mail, forum entries, social networks, pop-up windows, instant messaging bots, etc. Recent studies have indicated that the rapid increase of spam traffic is turning out to be the latest worrying problem delaying the trouble-free usage of the latest communication technologies [3][4]. In this context, the scientists are trying their best to come out with effective spam filtering techniques to filter the message content during message delivery time [5].

Although there subsists a few machine learning (ML) overtures for spam filtering and some of the recent works have successfully implemented well-known classifiers to the spam problem domain [6][7], but the exact application of rough set theory for spam filtering has not been thoroughly and widely dissected yet. Instinctively, the proper implementation of rule-based systems such as rough set theory seem perfectly suitable for addressing disjoint concepts like spam and ham (legitimate) classes on spam filtering [8]. A**t** this time, in this manuscript we have followed up and conveyed a variety of

times of yore investigation works on SPAM, which were carried out in spam filtering area and first time aiming specially for chat system, which is our own well-brought-up option with the support of traditional rough set theory. Imitation outcome evidently point towards our proposed method, which is laid down on rough set theory, can very effortlessly attain superior accurateness in spam detection as compared to the obtainable arrangements on the off chance of chat application[9].

II. RELATED WORKS

The majority anti-spam pokes into a accurate filter to categorize mails. Numerous datum or machine learning feelers is used. Chouchoulas [10] anticipated a rough setundercoated technique for text assortment to trickle out spam. Zhao projected a rough set-based representation to sort out emails into three families - spam, no-spam, and suspicious, instead of two divisions. Zhao [11] again anticipated on decision-theoretic rough set for filtration of spam mails. The truth is that rough set theory is appropriate for cataloging data with ambiguity and uncertainty with the intention, which springs up, from inexactness, conflict ness or imperfect information. Plentiful resolutions have been suggested to trounce the spam crisis. Amongst the anticipated techniques, much attention has given on the machine learning proficiencies in spam filtering. Among those techniques, rule learning [12] [14], decision trees [14], Naıve Bayes [13, 17], SVM [15, 16] or blending of dissimilar learners [10]. Support vector machine (SVM) has beenemployed by Drucker et al. [15] for assorting e-mails based on their stuffing and equated its functioning with Ripper, boosting decision trees. Rocchio et al. has shown that spam filtering depending upon the textual substance of e-mail can be looked at as an individual case of text categorization, i.e. whether it is HAM or spam [16]. One of the frequently used methods known as Bayesian classification technique designed to filter out spams contrived by Sahami [17]. Naïve Bayes (NB) was proposed by Androutsopoulos et al. [13][17], where they exhibited the consequence of dissimilar number of lineaments and training-set applied on the filter's performance.In NB-based approaches, token information is collected in a vector of attributes denoting the target message. Due to several classification criterion and representational issues, there are various techniques available, which are based on Naïve Bayes including

(i) Multinomial NB,

(ii) Multivariate Bernoulli NB,

(iii) Multinomial NB with term frequency attributes,

(iv) Multinomial NB with Boolean attributes,

(v) Flexible Bayes

The specific implementation of NB from Spam AssassinIn order to reduce the difficultiescontinuouslycaused by spam on companies, individuals and Internet Service Providers, scientists are coming up with effective complementary alternatives such as:

(i) Schemes for domain authentication, this includes the support for both labeling the authorized servers to transfer messages from different Internet domain and authenticate the clients who are sending the messages,

(ii) Collaborative approaches, which is developed to share the necessary information about spam messages through networks so that detecting a spam message becomes easy, and

(iii) Machine learning algorithms using the concepts of content-based techniques

There are other Machine Learning techniques, combining approaches and domain authentication schemes arealso interesting research field because spam filtering can be dealt with at multiple stages in the network, from theMTA (Message Transfer Agent) to the MRA (Message receiver Agent).Apart from these, artificial neural networks (ANN), SVM, k-nearest neighbor (KNN), artificial immune systems (AIS), case-based reasoning (CBR) systems and boosting strategies also can deal with spam filtering.

Among the different available boosting techniques, Adaboost can be successfully used to filter spam e-mails as reported by the work of Carreras and Marquez [18]. Finally, CBR systems are used to collect data from the previous problems and the solutions (stored in the case base) with the goal of solving new situations by somewhat tweaking previous approaches [19]. One of the major disadvantage of the existing system is that only topical terms like 'free' or 'Viagra' or 'sex' is considered as spam content and due to this old methodology in spam filtering, very often the legitimate messages containing these terms are blocked as spam. This problem occurs in chat messages more frequently compared to email messages because of the simpler and small content size of chat message. Moreover, adaptive schemes are not strong enough to filter the spams spread via these new innovative ways because the best method for spam filtering should be self-evolving which should come up with new techniques to manipulate the classification rules according to the situation. In this paperwe have first time applied classical rough set technique for filtering spam especially for chat system. As far as we know, no one has reported classical rough set approach for detection of SPAM in chat system.

A. Proposed Scheme

In a chat system, messages can be broadly classified in three categories- ham, uncertain and, spam. As the existing spam filtering methods are not suitable for the messages belonging to uncertain category, in those cases the challenge response method is used to further categorize the uncertain messages into ham and spam category. Statistics show that a very large portion of the spam messages is machine generated. That's

why challenge response method where human authentication is required will be very helpful in spam detection. In the figure below, we can see the major stake holders- message center (central server), the message sender and message receiver. In our methodology we are suggesting a framework where only the message center should only be responsible to check the message content in a chat system, because

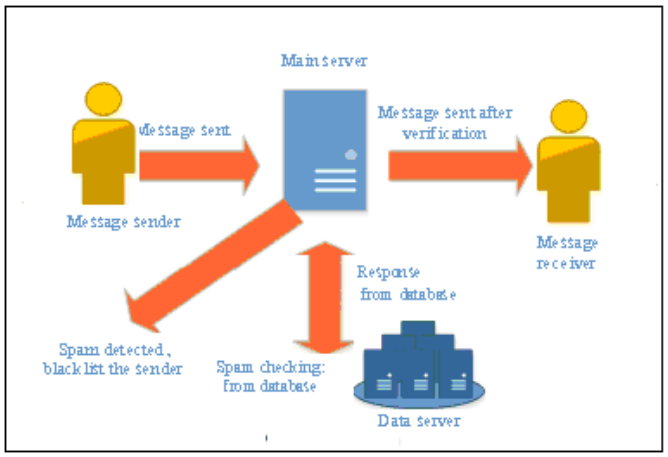

Fig 1: Working principle of the proposed chat system

1) This method will reduce the heavy traffic exhaustion and then the message should be forwarded to the recipient

2) This method will enable the message center to collect huge amount of data and this collected data can be very useful in further classification of message contents in future and the spam filtering algorithm can become self-evolving

3) Lastly, it will be a very tedious job to operate homogenous anti-spam filtering software in all the client system, so deploying the algorithm in the server sounds better idea to do the filtering.

B. Advantages of implementing Rough sets detection for SPAM detection

The rough set theory proposed by Pawlak [20] is an attempt to provide a formal framework for the automated transformation of data into knowledge. It is a mathematical tool, which deals with uncertainty [21]. It is based on the idea that any inexact concept (e.g. denoted by a class label) can be approximated from below and from above using an indiscernibility relationship. One of the major feature of the RS philosophy is the need to discover redundancy and dependencies between features [20]. The main advantages of applying rough set theory for spam filtering are

(i) It provides efficient and less time consuming algorithms to extract hidden patterns in data,

(ii) It can easily recognize those relationships which can't be easily found by traditional statistical methods,

(iii) It allows the use of both quantitative and qualitative data,

(iv) It can evaluate the minimal sets of data required for classification tasks,

(v) It can find out the significance of data and

(vi) It can also generatea set of decision rules from the given data set. This last property of the rough set will be exploited in this research paper.

III. A GENERAL VIEW OF ROUGH SET

Rough set theory proposed by Z Pawlak is popular mathematical tool to deal with uncertain data. Here in our SPAM data are the uncertain information. Pawlak in the year 1982 proposed this tool, which is a great help for scientists working on artificial intelligence, data analysis and data mining [26].

A. Data Table and indicernibility relation

In rough set theory [20] the information regarding all the objects is represented in a tabular form. In this data Table, the different attributes are denoted by different columns, and separate rows denote the distinct objects (actions). Each cell in the table denotes a qualitative or quantitative [21] evaluation of the object placed in that row associated with the specific attribute present in the corresponding column. Each data table consists of 4-tuples $S = (U, V, Q, F)$, where Urefers to a finite set of objects, $Q = \{q_1, q_2, \dots\dots\dots\dots, q_n\}$refers a finite set of attributes,$V_q$refers the domain of the attribute q, and $V = \bigcup_{q \in Q} V_qandf : U \times Q \rightarrow V$denotes a total function where$f(x, q) \in V_q$for every $q \in Q, x \in U$. So, each object xof U described by a vector (string) represented as$Des_q(x) = [f(x, q_1), f(x, q_2), \dots\dots\dots, f(x, q_n)]$ is called the description of x in terms of the evaluations of the attributes from$Q$. Every non-empty subset of attributes $P$is associated by an indiscernibility relation on U, denoted byI_P :

$$I_P = \{(x, y) \in U \times U : f(x, q) = f(y, q);\ \forall q \in P\}$$

The objects x and y can be said P-indiscernible if and only if $(x, y) \in I_P$. Certainly, this type of indiscernibility relation can also be termed as an equivalence relation.

B . Upper Approximations, lower approximation and boundary region of rough sets

Let us assume that Sbe a data table and $X$is a non-empty subset of U and $\emptyset \neq P \subseteq Q$. The P-upper approximation and the P-lower approximation of X in S are defined, respectively, by:

$$\overline{P}(x) = \bigcup_{x \in X} I_P(x)$$

$$\underline{P}(x) = \{x \in U : I_P \subseteq X\}$$

The elements of $\underline{P}(x)$ are only those objects ($x \in U$) which belong to the equivalence classes obtained by the indiscernibility relationI_P, contained in X, and the elements of $\overline{P}(x)$ are all and only those objects $x \in U$which belong to the equivalence classes obtained by the indiscernibility relationI_P, containing at least one object of $x \in X$. In another context we can say that,$\underline{P}(x)$is the smallest union of the P-elementary sets included in X, while $\overline{P}(x)$ is the largest union of the P-elementary sets containing X.The P-boundary region of Xin P, is denoted by$Bn_P(x) = \overline{P}(x) - \underline{P}(x)$. So, we can concludethat every object belonging to$\underline{P}(x)$must be an object of$\overline{P}(x)$. $Bn_P(x)$constitutes the "doubtful region" of$X$. If the P-boundary region of X is empty i.e. $Bn_P(x) = \emptyset$, then the set X is a crisp set with respect to P, i.e. it can be represented as the union of a certain number of P-elementary sets; otherwise, if $Bn_P(x) \neq \emptyset$, the set$X$is a rough set with respect to$P$and may be characterized by means of the approximations$\overline{P}(x)$and$\underline{P}(x)$. The accuracy of the approximation of X ($X \neq \emptyset$)is defined as$\alpha_P(x) = \frac{|\underline{P}(x)|}{|\overline{P}(x)|}$. Obviously, if$\alpha_P(x) = 1$, then$X$ is a crisp set with respect to P; if$\alpha_P(x) < 1$, then $X$is a rough set with respect to P.

C. Dependency and reduction of attributes

Dependence of the attributes is a very important concept for developing any concrete applications. Intuitively, a set of attributes $T \subseteq Q$totally depends on a set of attributes $P \subseteq Q$($P \rightarrow T$) if the set of values assigned to the attributes of$T$can be uniquely determined from the values assigned to the attributes of P, i.e. whether there exists any functional dependency between the evaluations of the attributes from P and T.Therefore,T totally depends on$P$iff$I_P \subseteq I_T$. So,T is totally (partially) dependent on Pif all (some) elements of the universe Umay be univocally assigned to classes of the partition$[U|I]_T$. The subset$P'$of the attributes set $P$is a reduct of P with respect to$P$, if and only if the following conditions are satisfied:

i) $Lower_{[\frac{P'}{D}]} = Lower_{[\frac{P}{D}]}$

ii) $\forall R \subset P', Lower_{[\frac{R}{D}]} \neq Lower_{[\frac{P}{D}]}$

There may exist more than one Y-reduct (or reducts) of P in the data table. The set consisting of all the indispensable attributes of P is known as the Y-core. Formally the relation between CORE and reduct can be shown as:
$CORE_Y(P) = \cap RED_Y(P)$.
Obviously, since the Y-core is the intersection of all the Y-reducts of P, so it should be included in every Y-reduct of P. It is the most significant set of attributes of Q, because removing or substituting any of its elements can deteriorate the quality of classification.

D. Decision table and decision rules

In decision table, the attributes of set $Q$are divided into two sets, condition attribute set($C \neq \emptyset$)and decision attributeset ($D \neq \emptyset$) where, $C \cup D = Q$ and $C \cap D = \emptyset$. The decision attributes induce a partition of Udeduced from the indiscernibility relation $I_D$such that it becomes independent from the conditional attributes. In the decision table,D-elementary sets are known as the decision classes. We try to reduce the set C while keeping all important relations between C and D, so that we can reach a decision based on the smallest amount of information available. If$f(x, q_1)$ is equal to$r_{q_1}$and $f(x, q_2)$ is equal to $r_{q_2}$and $f(x, q_P)$ is equal tor_{q_p}, then $x$belongs to r_{j_1}orr_{j_2}or r_{j_k}, where$\{q_1, q_2, \ldots\ldots, q_n\} \subseteq C$and
$\left(r_{q_1}, r_{q_2}, \ldots\ldots, r_{q_p}\right) \in V_{q_1} \times V_{q_2} \times \ldots\ldots\ldots \times V_{q_p}$ and $Y_{j_1}, Y_{j_2}, Y_{j_k}$are some of the decision classes of the considered classification (D-elementary sets). If the consequence is univocal, i.e. $k = 1$, then the rule is exact, otherwise it is approximate or ambiguous.

IV. SPAM DISCOVERY SCHEME DEVELOPMENT

We have shown below the algorithm, which has been developed for detecting SPAM messages in a chat application. This application has been developed in java language.

A. SPAM detection algorithm

In this paper, we have implemented the following algorithm, which detects whether a message is spam or not. We have deployed this algorithm in the server itself so that the clients do not need to install it in their system. So, it becomes very easy to maintain and update the code and also provides better security. Any number of users can connect to the server and then he/she can start sending messages. The pseudo code of the spam-detection algorithm is as follows:

```
Begin
Read the input from the client
Split the input wherever blank space is available
Insert all the split words in database
Initialize the string array Check with all the split words
FOR each token in the Check array
Set j=0
        Select all the words present in the database
                WHILE words present in the database
                    IF words in check array matches with
                     database THEN
                 Set genre equals to the genre of that word in
                          the database and Set j= 1
                     IF genre = "sexual" || "offensive" ||
                     "drug" THEN
                              FOR n=0 to word length
                                  SET star += "*"
                            END FOR
ELSE
                          Display token
```

```
 END IF
                END IF
        END WHILE
                        IF j = 0 THEN
                      SET genre="normal"
                        END IF
END FOR

Initialize a1, a2, a3, a4, a5, a6, a7 equals to zero
FOR i = 0 to token length
            IF genre equals "expression" THEN
                        SET a1 = 1
                         ELSE
                  IF genre equals "symbol" THEN
                           SET a2 = 1
                            ELSE
                 IF genre equals "normal" THEN
                           SET a3 = 1
                              ELSE
                IF genre equals "offensive" THEN
                           SET a4 = 1
                                ELSE
                            IF genre equals "sexual"
                 THEN
                              SET a5 = 1
                                  ELSE
                                IF genre equals
                  "drug" THEN
                                    SET a6 = 1
                                     END IF
                           END IF
                END IF
                 END IF
 END IF
    END IF
END FOR
          IF a4 = 0 && a5 = 0 && a6 = 0 THEN
                              SET a7 = 0
                        ELSE
                      IF a4  = 1 THEN
                         SET a7 = 1
                           ELSE
                          IF a5  = 1 THEN
                            SET a7 = 1
                               ELSE
                        IF a4 = 0 && a5 = 0 && a6 = 1
                        THEN
                              SET a7 = 1
                              END IF
                              END IF
                           END IF
                      END IF
   Insert the values of a1, a2, a3, a4, a5, a6, a7 in the database.
Display the dataset table
Stop
```

Algorithm 1: Pseudo code of the spam-detection algorithm

V. EXPERIMENTAL SETUP AND RESULTS

The data set used for this classification purpose is generated from the program that we developed for spam filtering i.e. whenever a user sends a message; our program classifies it into different categories. The different classification parameters for this paper are influenced by corpus representation taken from the paper written by *Perez Diaz et al*[21]. For first set of data we have named each messages as {O1, O2, O3, O4, O5, O6, O7, O8, O9, O10, O11, O12, O13, O14,O15} and each message has seven attributes, out of which six attributes are conditional and the last one is decisional attribute.

TABLE1: PATTERN OBTAINED FROM 15 CHAT MESSAGES (O1,…., O15) AND SEVEN ATTRIBUTES (A_1,….., A_7)

	Expression (a_1)	**Symbol (a_2)**	**Regular word (a_3)**	**Abusive (a_4)**	**xual (a_5)**	**Drug (a_6)**	**Spam (a_7)**
O1	1	1	0	0	0	0	0
O2	1	0	1	0	0	0	0
O3	1	1	1	0	0	0	0
O4	0	0	1	0	0	0	0
O5	0	0	1	1	0	0	1
O6	0	0	1	0	1	0	1
O7	0	0	1	0	0	1	1
O8	0	1	1	1	1	0	1
O9	1	0	1	0	1	1	1
O10	1	1	0	1	0	1	1
O11	1	0	1	1	1	1	1
O12	0	0	1	1	1	1	1
O13	1	0	1	1	0	0	1
O14	1	1	1	1	1	1	1
O15	0	0	1	0	0	1	0

In Table 1, different chat messages are represented as a feature vector in different rows where the value assigned to each attribute a_i belonging to $\{a_1, \ldots, a_{n-1}\}$ is 1 when the chat message contains the term a_i, and 0 otherwise. Likewise, the decision attribute (spam) value is 1 iff there is any spam content present in the message and 0 for legitimate ones. So, we can say that a decision table is a pair $S = (U, A)$, where Udenotes a finite, non-empty set called the universe (e.g. all the chat messages included in the Table 1), and A represents finite, non-empty set of features which are already defined. In Table1, $A = (C \cup D)$= {Expression, Symbol, Regular word, abusive, sexual, Drug}∪{Spam}.For any table,$S = \{U, P\}$ and any set $X \subseteq U$ can be defined by the use of two sets called lower and upper approximations. The lower approximation$\underline{P}(x)$, denotes the set of elements in Uwhich can be classified with full certainty as elements of X using the set of attributes P, and the upper approximation, denoted by$\overline{P}(x)$, represents all the elements which may or may not be classified with full certainty as elements of X.

A. Upper Approximation, Lower Approximation and Boundary message detection using Rough Sets for first set of SPAM data set

Lower boundary for (spam=0): {1, 2,3, 4}
Lower boundary for (spam=1): {5,6,8,9,10,11,12,13,14 }
Total lower boundary: {1,2,3,4,5,6,8,9,10,11,12,13,14}
Let, Total lower boundary= P1
Upper boundary for (spam=0): {1,2,3,4,7,15}
Upper boundary for (spam=1) : {5,6,7,8,9,10,11,12,13,14,15}
Total upper boundary={1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
Boundary region: {7,15}
Lower boundary for spam=0 for a_4,a_5,a_6: {1,2,3,4}
Lower boundary for spam=1 for a_4,a_5,a_6: {5,6,8,9,10,11,12,13,14 }
Lower boundary for a_4,a_5,a_6: {1,2,3,4,5,6,8,9,10,11,12,13,14}
Let, Lower boundary for a4, a5, a6 is denoted by P2 and now, we can clearly see that P1=P2
Therefore,$Lower_{[\frac{C'}{D}]} = Lower_{[\frac{C}{D}]}$ i.e. the first condition is satisfied and now let us consider that, the lower boundary of a_4,a_5,$a_6$are P3,P4,P5 respectively and it can be calculated that P3,P4,P5 are not equal to P1. Therefore,$\forall R \subset C', Lower_{[\frac{R}{D}]} \neq Lower_{[\frac{C}{D}]}$ i.e. the second condition is also satisfied. So,{a_4,a_5,a_6}is the reduct subset and also the core for this dataset. So, from the experimental data, we can generate a few rules, which will be helpful for filtering the spam content in chat messages. These rules can also be used to update the existing algorithm dynamically, which will be helpful to detect different types of intrusion techniques and provide better security.

TABLE 2 : RULES

Condition1	Condition2	Condition3	Decision	No of times
(a_4=0)	(a_5=0)	(a_6=0)	a_7=0	4
(a_4=1)	(a_5=0)	(a_6=0)	a_7=1	1
(a_4=0)	(a_5=1)	(a_6=0)	a_7=1	1
(a_4=0)	(a_5=0)	(a_6=1)	a_7=1	1
(a_4=0)	(a_5=0)	(a_6=1)	a_7=0	1
(a_4=1)	(a_5=1)	(a_6=0)	a_7=1	2
(a_4=0)	(a_5=1)	(a_6=1)	a_7=1	1
(a_4=1)	(a_5=0)	(a_6=1)	a_7=1	1
(a_4=1)	(a_5=1)	(a_6=1)	a_7=1	3

B. Comparison of our method with naïve Bayesian classifier and Logistic Regression Classifier for first set of SPAM messages

In this subsection we have compared our proposed rough set model with the other existing techniques for spam detection like Naïve Bayes Classifier and Logistic Regression Classifier. From the above dataset, we can clearly observe that *7* and *15* are the only two feature vector which belong to the boundary region and this condition of uncertainty occurs only when a_1=0, a_2=0, a_3=1,a_4=0, a_5=0 and a_6=1. So, we have applied the other two spam classifying methods on the above dataset and calculated the probability to detect spam in boundary region cases i.e. where a_1=0, a_2=0, a_3=1,a_4=0, a_5=0 and a_6=1 and compared the result with our proposed method.

TABLE3: PROBABILISTIC PREDICTION AND CLASSIFICATION OF SPAM AND HAM MESSAGES

	Our approach	Naïve Bayes Classifier	Logistic Regression Classifier
SPAM detection probability for boundary cases	0.5	0.47	0.5
HAM detection probability for boundary cases	0.5	0.53	0.5

So, from Table 3, we can see that logistic regression classifier and our hybrid rough set theory model has the same probability to detect a SPAM or HAM message in boundary cases but in Naïve Bayes method, the probability to detect SPAM in boundary cases is relatively low as compared to our approach. Also our proposed method can easily detect SPAM and HAM and along with that, understanding of the computability process and implementation style is relatively lucid.

VI. CONCLUSIONS

In this paper we aimed to have a new spam detection scheme for chat system based on a rough set algorithm. The main objective is to show a comprehensive study on efficient utilization of rough set theory as main classifier for spam filtering. We have proposed our own technique to filter the data shared via chat systems. Also, analyzed the message content to check whether the message is SPAM or HAM. By examining some of the previous research works, we have figured out that majority of those works are only modest analyses using corpora with an insufficient preprocessing. From all the research works carried out, we have found out very interesting conclusions. Rough set based schemes can be a very suitable substitute for AdaBoost, SVM and Naive Bayes classifier. Although Rough Set based approaches perform well in spam filtering, but still new adaptive reduction techniques and rule execution methods should be developed to achieve more accuracy in results.

ACKNOWLEDGMENTS

This work was supported in the framework of the IT4 Innovations Centre of Excellence project, reg. no. CZ.1.05/1.1.00/02.0070 by operational programme Research and Development for Innovations funded by the Structural Funds of the European Union and state budget of the Czech Republic, EU.

REFERENCES

[1] L. Roberts, The evolution of packet switching, Proceedings of the IEEE 66 (11)(1978) 1307–1313, http://www.packet.cc/files/ev-packet-sw.html.

[2] SpamHaus Project Organization, TheSpamHaus Project, 1998. http://www.spamhaus.org.

[3] P. Bueno, T. Dirro, P. Greve, R. Kashyap, D. Marcus, S. Masiello, F. Paget, C. Schmugar, A. Wosotowsky, (McAfee Inc.), McAfee Threats Report, Third Quarter 2010.

[4] Message Labs Ltd., MessageLabs Intelligence: 2010 Annual Security Report.

[5] J. Klensin, Simple Mail Transform Protocol, RFC5321, 2008. [6] T.S. Guzella, W.M. Caminhas, A review of machine learning approaches to spam filtering, Expert Systems with Applications 36 (7) (2009) 10206–10222.

[6] B. Biggio, G. Fumera, I. Pillai, F. Roli, A survey an experimental evaluation of image spam filtering techniques, Pattern Recognition Letters 32 (10) (2011) 1436–1446.

[7] S.J. Delany, P. Cunningham, A. Tsymbal, L. Coyle, A case-based technique for tracking concept drift in spam filtering, Knowledge-Based Systems 18 (4–5)(2005) 187–195.

[8] Rough Set Approach to Spam Filter Learning

[9] MawuenaGlymin and WojciechZiarkoChouchoulas, A. (2004). A rough set-based approach to text classification. In Lecture notes in computer science

[10] Zhao, W., & Zhang, Z. (2005). An email classification model based on rough set theory. In Proceedings of the international conference on active media technology (pp. 403–408).

[11] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C.D. Spyropoulos, P. Stamatopoulos, Learning to filter spam e-mail: a comparison of a Naı¨ve Bayesian and a memory-based approach, in: Proc. of the workshop: Machine Learning and Textual InformationAccess, 2000, pp. 1–13.

[12] I. Androutsopoulos, J. Koutsias, K. Chandrinos, G. Paliouras, C.D. Spyropoulos, An evaluation of naive bayesian anti-spam filtering, in: Proc. of the Workshop on Machine Learning in the New Information Age: 11th European Conference on Machine Learning, 2000, pp. 9–17.

[13] X. Carreras, L. Marquez, Boosting trees for anti-spam email filtering, in: Proc. of fourth Int'l Conf. on Recent Advances in Natural Language Processing, 2001, pp. 58–64.

[14] H. Drucker, D. Wu, V.N. Vapnik, Support vector machines for spam categorization, IEEE Trans. Neural Netw. Vol. 10 (No. 5) (1999) 1048–1054.

[15] A. Kołcz, J. Alspector, SVM-based filtering of e-mail spam with content-specific misclassification costs, in: Proc. of TextDM'01 Workshop on Text Mining, 2001

[16] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz, A Bayesian approach to filtering junk e-mail, in: Learning for Text Categorization – Papers from the AAAI Workshop, 1998, pp. 55–62.

[17] X. Carreras, L. Márquez, Boosting trees for anti-spam e-mail filtering, in: Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing, 2001, pp. 58–64.

[18] S.J. Delany, P. Cunningham, A. Tsymbal, L. Coyle, A case-based technique for tracking concept drift in spam filtering, Knowledge-Based Systems 18 (4–5)(2005) 187–195.

[19] Z Pawlak- Rough Sets, International Journal of Computer and Information Sciences, 11, pp. 341-356, 1982.

[20] Roy, Sanjiban Sekhar, et al. "Applicability of Rough Set Technique for Data Investigation and Optimization of Intrusion Detection System." Quality, Reliability, Security and Robustness in Heterogeneous Networks. Springer Berlin Heidelberg, 2013. 479-484.

[21] Pérez-Díaz, Noemí, et al. "Rough Sets for Spam Filtering: Selecting Appropriate Decision Rules for Boundary Classification." *Applied Soft Computing* (2012).