SPAM MAIL FILTERING

Ms. Nisha Advilkar **Ms. Pranali Mane** **Prof. Dhanraj Walunj**
SCOE, Kharghar **SCOE, Kharghar** **SCOE, Khargha**

***Abstract*— Email has been an efficient and popular communication mechanism as the number of internet user's increase .We us this in Security Informatics Application for detecting deceptive Communication in email.**

In this paper, we proposes to apply Association Rule Mining for Suspected theory suggests that deceptive writing is characterized by reduced frequency of first person pronouns and exclusive words and elevated frequency of negative emotion words and action verbs We apply this model of deception to the set of Email dataset, then applied Apriori algorithm to generate the rules .The rules generated are used to test the email as deceptive or not. In particular we are interested in detecting emails about criminal activities. After classification we must be able to differentiate the emails giving information about past criminal activities(Informative email) and those acting as activities(Informative email) and those acting as differentiation is done using the features considering the tense used in the emails. Experimental results show that simple Associative classifier provides promising detection rates.

***Index Terms*— Data Mining, Deceptive Theory, Association Rule Mining, Apriority algorithm**

I. INTRODUCTION

E-mail has become one of today's standard means of communication. The large percentage of the total traffic over the internet is the email. Email data is also growing rapidly, creating in this era. It is need for automated analysis, to detect crime spectrum techniques detect of techniques should be applied to discover and identify patterns and make predictions. Data mining has emerged to address problems of understanding ever-growing volumes of information for structured data, finding patterns within data that are used to develop useful knowledge. As individuals increase the usage of electronic communication there has been research into detecting deception these new forms of communication. Models of deception assume that deception leaves a footprint. Work done by various researches suggests that deceptive writing is characterized by reduced used apriori algorithms to generate a classifier categorize the email as deceptive or not. Or we can say Email spamming refers to sending email to thousands and thousands of users – similar to a chain letter. Spamming is often done deliberately to use network resources. Email spamming may be combined with email spoofing, so that it is very difficult to determine the actual originating email address of the sender. Some email systems, including our Microsoft Exchange, have the ability to block incoming mail from a specific address. However, because these individuals change their email addresses frequently, it is difficult to prevent some spam from reaching your email inbox.

A. Spam Mail

Spam is usually considered to be electronic junk mail or junk newsgroup postings. Some people define spam even more generally as any unsolicited email. However, if a long-lost brother finds your and sends you a message, this could hardly be called spam, even though it is unsolicited. Real spam is generally email advertising for some product sent to a mailing list or newsgroup.

Electronic mail, also known as **email** or **e-mail**, is a method of exchanging digital messages from an author to one or more recipients. Modern email operates across the Internet or other computer networks. Some early email systems required that the author and the recipient both be online at the same time, in common with instant messaging. Today's email systems are based on a store-and-forward model. Email servers accept, forward, deliver and store messages. Neither the users nor

ISSN: 2278 – 1323 All Rights Reserved © 2016 IJARCET

their computers are required to be online simultaneously; they need connect only briefly, typically to an email server, for as long as it takes to send or receive messages. An Internet email message consists of three components, the message envelope, the message header, and the message body. The message header contains control information, including, minimally, an originator's email address and one or more recipient addresses. Usually descriptive information is also added, such as a subject header field and a message submission date/time stamp.

B. Association Rule

In data mining, **association rule learning** is a popular and well researched method for discovering interesting relations between variables in large databases. For example, the rule {Onion,Potato}=>{Burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including web usage mining, intrusion detection and bioinformatics. As opposed to sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

C. *Useful Concepts*

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence.

- The supportsup(X) of an item set X is defined as the proportion of transactions in the data set which contain the item set. In the example database, the item set {milk, bread, butter} has a support of 1/5=0.2 since it occurs in 20% of all transactions (1 out of 5 transactions).

- The confidence of a rule is defined conf (X=>Y) = supp (X U Y) / supp (X). For example, the rule { milk, bread} => {butter} has a confidence of 0.2 / 0.4 = 0.5 in the database, which means that for 50% of the transactions containing milk and bread the rule is correct (50% of the times a customer buys milk and bread, butter is bought as well). Be careful when reading the expression: here supp (X∪Y) means "support for occurrences oftransactions where X and Y both appear", not "support for occurrences of transaction swhere either X or Y appears", the latter interpretation arising because set union is equivalent to logical disjunction. The argument of supp() is a set of preconditions, and thus becomes more restrictive as it grows (instead of more inclusive).

- Confidence can be interpreted as an estimate of the probability P (Y|X), the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

II. LITERATURE SURVEY

A. Email Spam and the CAN-SPAM Act

In December 2003, the CAN-SPAM Act was enacted and took effect in January 2004 (Lee, 2005). The Act was enacted in an attempt to regulate interstate commerce by imposing penalties and limitations on sending unsolicited commercial email via the Internet (Year gain et al., 2004). The Federal Trade Commission (FTC) was authorized to enforce provisions provided in the CAN-SPAM Act, and unsolicited commercial emails that fail to comply with its regulations were declared criminal. The punishment could be a fine up to $16,000 for each separate email in violation of the CAN-SPAM Act (FTC, 2009), or it could be imprisonment (Year gain et al., 2004). In 2008, the so-called “Spam King” Robert Soloway was convicted under the CAN-SPAM Act for sending fraudulent emails along with two other charges and was sentenced to 47 months in federal prison (Rabinovitch, 2007). The Act also allows states and Internet service providers to file civil lawsuits against spammers (Ford, 2005; Year gain et al., 2004).

Email spam is a rampant activity in cyberspace. This study performed a qualitative forensic analysis on 3,983 spam emails with respect to their content,

ISSN: 2278 – 1323 All Rights Reserved © 2016 IJARCET

format, techniques, and their compliance with the CAN-SPAM Act. The findings suggested spammers show little interest in complying with the CAN-SPAM Act, and different purposes of spam determine the format and the techniques spammers use. The rationales behind these spam choices were discussed. Legal and research implications were suggested.

B. Existing System

Apriori algorithm for Association -rules generation

- The association rules generated is in numerical values hence the visualized output with respect to the output column of the preprocessing. Association Rule mining searches for interesting association or correlation relationships among items in a given large data set. We model email messages as transaction where items are words or phrases from the email. After preprocessing a email message, by eliminating stop words and stemming.
- It's hard to remember what our lives were like without email. Ranking up there with the web as one of the most useful features of the Internet, billions of messages are sent each year. Though email was originally developed for sending simple text messages, it has become more robust in the last few years. So, it is one possible source of data from which potential problem can be detected. Thus the problem is to find a system that identifies the .deception in communication through emails.

C. Name of Different Methods

- Content-Based Filters

Content-based filtering, also referred to as cognitive filtering, recommends items based on a comparison between the content of the items and a user profile. The content of each item is represented as a set of descriptors or terms, typically the words that occur in a document. The user profile is represented with the same terms and built up by analyzing the content of items which have been seen by the user.

Several issues have to be considered when implementing a content-based filtering system. First, terms can either be assigned automatically or manually. When terms are assigned automatically a method has to be chosen that can extract these terms from items. Second, the terms have to be represented such that both the user profile and the items can be compared in a meaningful way. Third, a learning algorithm has to be chosen that is able to learn the user profile based on seen items and can make recommendations based on this user profile.

The information source that content-based filtering systems are mostly used with are text documents. A standard approach for term parsing selects single words from documents. The vector space model and latent semantic indexing are two methods that use these terms to represent documents as vectors in a multidimensional space.

Rather than enforcing across-the-board policies for all messages from a particular email or IP address, content-based filters evaluate words or phrases found in each individual message to determine whether an email is spam or legitimate.

- Word-Based Filters

A word-based spam filter is the simplest type of content-based filter. Generally speaking, word-based filters simply block any email that contains certain terms. Since many spam messages contain terms not often found in personal or business communications, word filters can be a simple yet capable technique for fighting junk email. However, if configured to block messages containing more common words, these types of filters may generate false positives. For instance, if the filter has been set to stop all messages containing the word "discount," emails from legitimate senders offering your non-profits hardware or software at a reduced price may not reach their destination. Also note that since spammers often purposefully misspell keywords in order to evade word-based filters, your IT staff will need to make time to routinely update the filter's list of blocked words.

- Heuristic Filters

Heuristic filtering works by subjecting email messages through thousands of pre-defined rules against the message envelope, header and content. Each rule assigns a numerical score to the probability of the message being spam. The result of the final

ISSN: 2278 – 1323 All Rights Reserved © 2016 IJARCET

equation is known as the Spam Score .The spam score is then measured against the user's desired level of spam sensitivity - whether low, medium or high sensitivity. Setting a higher level of sensitivity leads to more spam being captured, but has the adverse effect of filtering legitimate emails as spam. This is known as a false-positive.

Heuristic (or rule-based) filters take things a step beyond simple word-based filters. Rather than blocking messages that contain a suspicious word, heuristic filters take multiple terms found in an email into consideration. Heuristic filters scan the contents of incoming emails and assigning points to words or phrases. Suspicious words that are commonly found in spam messages, such as "Rolex" or "Viagra," receive higher points, while terms frequently found in normal emails receive lower scores. The filter then adds up all the points and calculates a total score. If the message receives a certain score or higher (determined by the anti-spam application's administrator), the filter identifies it as spam and blocks it. Messages that score lower than the target number are delivered to the user.

Heuristic filters work fast — minimizing email delay — and are quite effective as soon as they have been installed and configured. However, heuristic filters configured to be aggressive may generate false positives if a legitimate contact happens to send an email containing a certain combination of words. Similarly, some savvy spammers might learn which words to avoid including, thereby fooling the heuristic filter into believing they are benign senders.

- Bayesian Filters

Bayesian filters, considered the most advanced form of content-based filtering, employ the laws of mathematical probability to determine which messages are legitimate and which are spam. In order for a Bayesian filter to effectively block spam, the end user must initially "train" it by manually flagging each message as either junk or legitimate. Over time, the filter takes words and phrases found in legitimate emails and adds them to a list; it does the same with terms found in spam.

D. A Review on Different Spam Detection Approach

Rafiqul Islam and Yang Xiang[3] performed classification of user emails form penetration of spam. In their paper," Email Classification Using Data Reduction Method" an effective and efficient email classification technique based on data filtering method is presented. They have introduced an innovative filtering technique using instance selection method (ISM) to reduce the pointless data instances from training model and then classify the test data. The objective of ISM is to identify which instances (examples, patterns) in email corpora should be selected as representatives of the entire dataset, without significant loss of information. They have used WEKA interface in our integrated classification model and tested diverse classification algorithms. Their empirical studies show significant performance in terms of classification accuracy with reduction of false positive instances.

AsmeetaMali[4] performed a work," Spam Detection using Bayesian with Pattern Discovery". In her paper she presents an effective technique to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Using Bayesian filtering algorithm and effective pattern Discovery technique we can detect the spam mails from the email dataset with good correctness of term.

Vandana Jaswal[5] proposes an image spam detection system that uses detect spam words. In her work," Spam Detection System Using Hidden Markov Model" filtering method are used to detect stemming words of spam images and then use Hidden Markov Model of spam filters to detect all the spam images

III. PROPOSED SYSTEM

Based on the theory of deception a deceptive email will have highly emotional words and action verbs. So, such words are set as keywords and extracted from the input dataset. Example for highly emotional words and action verbs are "lifeless", "anger", "kill", "attack", etc. The future tense denoting keywords such as will, shall, may, might, should, can, could, would are used to indicate that the suspicious email is of the type alert. The past tense denoting keywords such as was, were, etc. are used to indicate that the suspicious email is of the informative type. After

ISSN: 2278 – 1323 All Rights Reserved © 2016 IJARCET

these email is given to the preprocessing program. This system gives the pure classification of the email message on the users screen. So that user can easily categorize the received emails at his/her side.

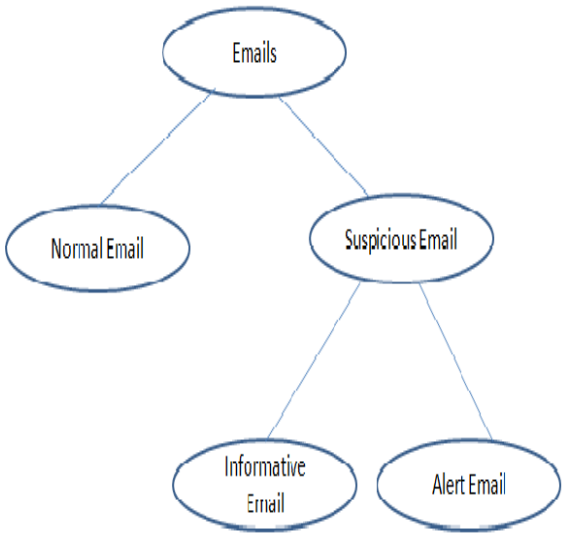

Figure Classification of Email

Example of suspicious and normal email.

Suspicious Email	Normal Email
Sender: X Sub: Bomb Blast Body: Today there will be bomb blast in parliament house and the US consulates in India at 11.46 am. Stop it if you could. Cut relations with the U.S.A. long live Osama Finladen Asadullah Alkalfi.	Sender: y Sub: Hi Body: Hope ur fine! How are u & family members?

Figure Types of email

Example of classifying Suspicious into Alert and informative email:

Alert Email	Informative Email
Sender: X Sub: Bomb Blast Body: Today there will be bomb blast in parliament house and the US consulates in India at 11.46 am. Stop it if you could. Cut relations with the U.S.A. long live Osama Finladen Asadullah Alkalfi.	Sender: y Sub: WTC Attacked Body: The World Trade Center was attacked on 9/11/01 by Osama Bin Laden and his followers.

Figure Difference between types of email

IV. CONCLUSION

In this project, we have deployed Association rule mining based classification approach to detect deceptive communication in email text as informative or alert emails. We can find it that the simple Apriori algorithm can provides better classification result for suspicious email detection The Proposed work (using keyword extraction and keyword attribute called tense) will be helpful for identifying the deceptive email and to get information in time to take effective actions to reduce criminal activities.

V. FUTURE SCOPE

In the near future, we plan to incorporate other techniques like different ways of feature selection, and classification using other methods. One major advantage of the association rule based classifier is that it does not assume that terms are independent and its training is relatively fast. Furthermore the rules are human understandable and easy to maintained or pruned by human being.

REFERENCES

[1] S.Appavu alias Balamurgan, Aravind, Athiappan, Bharathiraja, Muthu Pandian and Dr.R.Rajaram "Association Rule Mining for Suspicious Email Detection: A Data Mining Approach" 2007

[2] http://en.wikipedia.org/wiki/Apriori_algorithm

ISSN: 2278 – 1323 All Rights Reserved © 2016 IJARCET

[3] Rafiqul Islam and Yang Xiang ," Email Classification Using Data Reduction Method"

[4] Asmeeta Mali, " Spam Detection using Bayesian with Pattern Discovery"

[5] VandanaJaswal, " Spam Detection System Using Hidden Markov Model"

[6] R.Agrawal and R.Srikant, "Fast algorithms for mining association rules," In Proc. 20th Int. Conf. Very Large Data Bases (VLDB'94), pages 487-499, Santiago, Chile, 1994.

[7] G. Boone. "Concept features in re:agent, an intelligent email agent," In Proc. 2nd Int. Conf. Autonomous Agents (Agents'98), pages 141-148, New York, 1998.

[8] S. Chakrabarti, B. E. Dom, R. Agrawal, and P.Raghavan, "Using taxonomy, discriminants, and signatures for navigating in text databases," In Proc. 23rd Int. Conf. Very Large Data Bases, pages 446-455, Athens, GR, 1997.

[9] Zan Huang and Daniel D.Zeng, "A link Prediction approach to anomalous email Detection"

[10] J.Rennie ,"An Application of Machine Learning to email Filtering," In proc. KDD 2000 workshop on Text mining, Boston, MA 2000

Author 1:
Name: Nisha Advilkar.
Education: BE
College Name: Saraswati College Of Engineering, Kharghar.

Author 2:
Name: Pranali Mane
Education: BE
College Name: Saraswati College Of Engineering, Kharghar

Author 3:
Name: Dhanraj Walunj
Qualification: Assistant Professor at Saraswati College Of Engineering**,** Kharghar

ISSN: 2278 – 1323 All Rights Reserved © 2016 IJARCET