IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 6, June 2017

Sender Recognition of E-mail & Email Categorization

George Philip C.[1], Bansi R. Savaliya[2]

Associate Professor, Information Science Engineering, M.S. Ramaiah Institute of Technology, Autonomous Institute Affiliated to VTU, Bangalore, Karnataka, India[1]

M. Tech, Information Science Engineering, M.S. Ramaiah Institute of Technology, Autonomous Institute Affiliated to VTU, Bangalore, Karnataka, India[2]

Abstract: Tackling irrelevant emails have become part of every email user's activity. Emails that seem valid are received in the inbox and, sometimes relevant emails are directed to spam. Another aspect of the problem is that due to very high number of incoming emails, it is very difficult to identify the required ones easily. In this process, users waste so much of their time, energy and efforts by sifting through irrelevant mails also in which they have no interest. Sometimes users also get frustrated getting such junk mails frequently. To support ease of access, emails are to be categorized based on the type of information they contain, which will help a person to identify required mails even before opening it. This paper involves development of a feasible solution to this problem by identifying the real sender using past email patterns and features. The proposed project uses this solution to solve the problem of email categorization also. This paper uses machine learning algorithm for detecting the actual Email composer. Different Semantic, Syntactic, and Lexical features of the incoming will be considered to implement the project. Features like Ngram, Lemmatization, creating personalized vocabulary, and observation of patterns are utilized. Algorithms like Lesk will be used to find the meaning according to the context of the text. A database like WordNet helps to find relevant words in the text. Machine learning will be used to learn the different features and create the training data. After the framework is trained, testing data will be used to assess the system. As the system is tested, it will continue to learn from the input data to make it better. Once the machine is trained, the system can start working as a fraud detection system which identifies the real sender, and categorize emails.

Keywords: Text Mining, Ngram (Unigram, Bigram, Trigram), Lemmatization, Wordnet, Enron E-mail Dataset.

I. INTRODUCTION

Email spam also known as junk, or unsolicited bulk email have created havoc in user's life. It has grown up from an avoidable issue to an unavoidable one. Email spam is one of the social issues confront every day. Spam email turns into a major issue in email communication. It makes consistent issues the system manager and security master to prevent it effectively. There are many sorts of strategies utilized by the spammer to send inconvenience email messages.

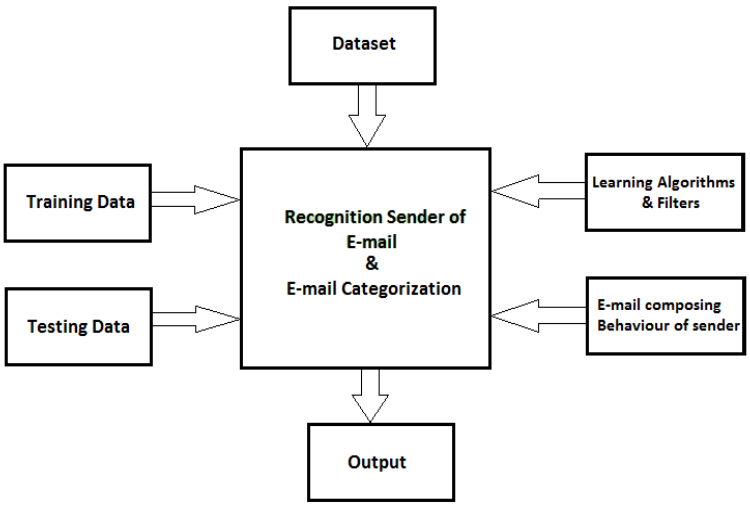

Figure 1- 1 System Architecture

DOI10.17148/IJARCCE.2017.6621

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 6, June 2017

Fig. 1 describes the brief behaviour of the framework. The system will be prepared in the wake of giving it input dataset of an E-mail. In the wake of preparing the framework, testing input data will be given as an input to distinguish the genuine sender of an E-mail.

Emails are widely used as a means of communication on web. A major issue that is faced by users is irrelevant mails. Emails are mainly originated in newsgroups, web pages, chat rooms and infection that reside in users address books. Spam usually contains irrelevant information, advertisements, malware, viruses and phishing links. So, it is usually preferred to avoid those mails. Sometimes, relevant mails are labelled as spam which are originated from various sources. Another relating critical issue is the enormous spam sent from the legitimate home PCs. Many spam messages contain URLs to sites and many of them are commercial in nature however may contain duplicate links that seems familiar sites but in fact lead to phishing sites or containing malware. Spam email may incorporate malware as scripts or other executable document connections. Another aspect of the problem is that due to very high number of incoming emails, it is very difficult to find one. To support ease of access, emails are needed to be categorized based on the type of information they contain which will help a person to know the content before even opening it.

The goal of this paper is detecting the real Email composer. This can be applied for Email Fraud Detection. In Daily life, humans can recognize an anonymous note by any sender. This can be applied to computers using the available text mining and machine learning techniques. The values of user grammar, mail structure, content and frequently used words can be inferred as attributes/features for learning algorithm. System creates user vocabulary database and recognize the user.

II. THEORETICAL BACKGROUND OF APPLIED METHODOLOGY

This part covers the blend of Text Mining [5] and Machine Learning [1] approach. That suggests how framework can use Text Mining and Machine Learning approach both together and hint at change yield. Text Font of Entire Document

A. Text Pre-processing

Pre-processing aims to shape corpus and dictionary from all slithered records. Corpus essentially implies an assortment of content utilized for aggregating measurements on characteristic dialect content. Dictionary is a vocabulary, a lexicon, a rundown of words or potentially expressions and sense definitions. Content ought to be changed over into a vector space to discover numeric word weights, word frequencies and word similitudes. Prior to the content is translated as vectors, pre-handling is authorized to clean and configuration the information. Reprocessing changes over crude content information into an all-around characterized semantic word set that is sanitized from babble words and this stage covers the disposal of accentuation and numbers, bring down packaging of capital letters, tokenization and stemming and stop words expulsion. At the point when the pre-handling stage is closed, the information winds up plainly accessible for finding comparative words and displaying themes that will be communicated in different areas.

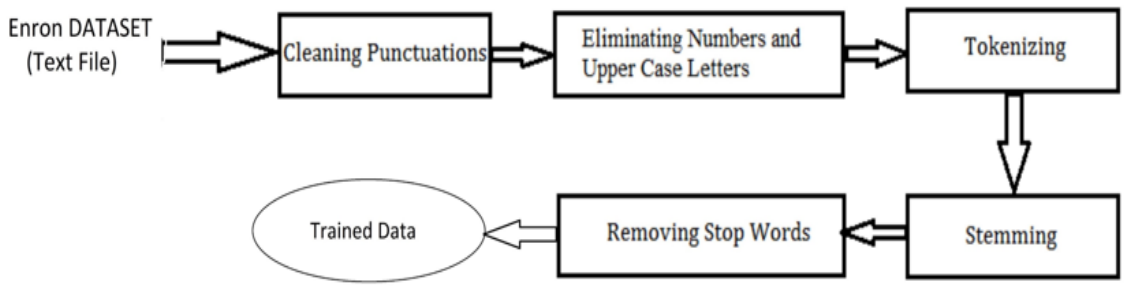

Figure 2-1 Data Pre-Processing

Cleaning Punctuations

Cleaning punctuation is the initial step of the pre-processing stage. Punctuations don't change word implications and they cause a commotion on info. Punctuations can likewise change the word check since a word with accentuation and without accentuation are numbered in an unexpected way. Thusly, commas, dabs, shout marks, triple spots, semicolons, punctuations and different Punctuations ought to be expelled from the informational index. In addition, letters that are by the punctuations are expelled from the informational index in light of the fact that these letters don't present a sense word in Turkish. For instance, a word with a punctuation "okuldan" (from school) is assumed as "okul" (school) as indicated by end of babble word in this review.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 6, June 2017

Eliminating Numbers and Upper Case Letters
It is observed that numbers are utilized to speak to year, budgetary portrayal, measures and so on. Since the numbers are pointless to break down for the situation, numbers ought to be expelled from the information. Consequently, numbers ought to be expelled from the information. Also, a word with a lower case and capitalized is including diversely a case delicate numbering, even though they have a similar importance. Thus, capital letters are changed over to their lower case.

Tokenizing
The objective of tokenization is to investigate words in a sentence. In this manner, tokenization parts archives into words with spaces. A grouping of tokens is uncovered and the words can be checked effectively. Tokenization intends to create scanty vectors for records that parsed into vector space.

Stemming
Stemming process works on word roots. On the off chance that a base type of a word and word with an informative supplement have a similar significance however in the diverse frame, they can be expected as a one of a kind word. Stemming procedure is required to build precision and diminish word variation. For example, "okuma" (perusing) and "okumak" (to peruse) are gotten from a similar word "oku" (read). These words convey a similar importance. On the off chance that words are gotten from a similar root yet don't have a similar significance, these words ought to be isolated. Zemberek venture is utilized to stem Turkish composed content to gather every one of a kind word into a vocabulary.

Removing Stop Words
Stop words are normal words that don't convey any implications and no conclusive rundown. They ought to be expelled amid the pre-preparing. Stop words are pointless words, for example, assistant words, some conjunction words, verb modifiers and pronouns. A little rundown containing of stop words are exhibited in here: {he, she, is, are, his, her, and, from, before}. Stop word disposal builds the exactness and adds to diminish the extent of an element space.

For identification of email sender and categorization of mails, some algorithms are required to process the system such as unigram, bigram, trigram explained in next section.

B. Equation
unigrams (single-word variables) Unigram algorithm which working on the following equation.

ngram text, degree (1) threshold (1)
The alternative degree (1) alludes to unigrams. That is, factors contain checks of individual words (instead of from multi-word successions). At the point when words seem occasionally, relating unigrams are regularly disposed of. Option threshold (1) implies that all subsequent factors are held regardless of the possibility that an individual word just shows up in a solitary content.

Bigrams
Bigram calculation is done with this command:

ngram text, degree (2) threshold (1)
The option degree (2) alludes to Bigrams. That is, factors contain numbers of couple of words (instead of person). Alternative edge (1) implies that all subsequent factors are held regardless of the possibility that an individual word just shows up in a solitary content.

III. IMPLEMENTATION & DATASET

The framework will use the database of messages from Enron open mail corpus [6] which contains 500000 E-sends of 150 clients. The basic part is Each mail is sorted by customer. This dataset contains 2.23 GB of E-mail Text orchestrated data. It fuses archive (txt) records of each customer.

To increase the prospect of database, framework used WordNet dictionary from which it took the definitions and usage examples of each keyword.

The system has been made which recognizes the genuine sender of the email. Diverse lexical, semantic and syntactic elements had been utilized as a part of request to make the framework. Utilizing this, we have prepared the framework to discover the words and its number in every database. To enhance the execution of framework, all the superfluous

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 6, June 2017

parts from the email like html labels and stop words are evacuated. Once more, to further expand the execution, WordNet database had been utilized to simply keep the most vital words in the database. The idea of unigram likelihood, bigram likelihood and trigram likelihood utilized as a part of request to recognize the genuine sender of email. Ngram (uni, bi, tri) likelihood is figured on a weighted premise which will be clarified in further segment.

WordNet database has been used at multiple places in this paper and filter out our database, to extract semantics of words/sentences, to extract the sense of the maximum probable word from the test e-mail, etc.

While training database for features, a word was only allowed to be stored in the database if it was found in WordNet dictionary. By doing this, all un-important words are prevented to be stored in the database. WordNet database was also used to self-train our database for creating the database for different categories like travel, finance, sports, geography and job/occupation. Finally, it was used to extract the head word in the test e-mail to provide the concept of the e-mail.

IV. RESULT

The result output is the accuracy which contains the ratio of the correct users predicted to the total number of test mails. For testing the algorithm, ~1000 mails of 100 users are selected and tested with the algorithm. It gives an accuracy of ~60%.

Part of Output:
File name: 1, original sender: zufferli-j, predicted sender: zufferli-j
File name: 107, original sender: zufferli-j, predicted sender: lokey-t
File name: 128, original sender: zufferli-j, predicted sender: zufferli-j
File name: 2, original sender: zufferli-j, predicted sender: zufferli-j
File name: 3, original sender: zufferli-j, predicted sender: zufferli-j

Accuracy = 60.28225806451613 true prediction= 598.0 false prediction= 394.0

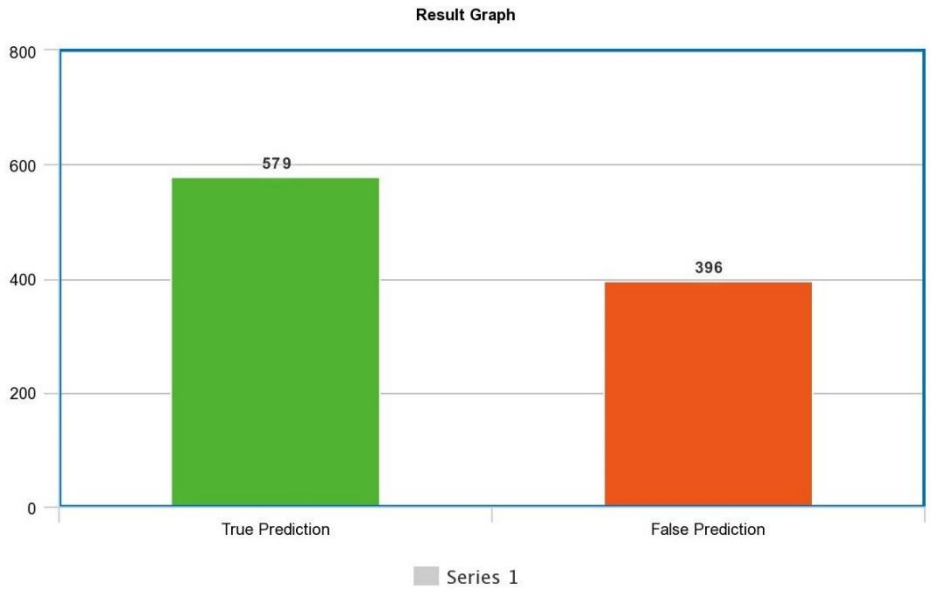

Figure- 3-1 Graphical representation of output

$$\text{Accuracy} = \frac{\text{Count}_{\text{CorrectOutput}}}{\text{Count}_{\text{TestSize}}} * 100$$

The result output is the accuracy which contains the ratio of the correct users predicted to the total number of test mails.

V. CONCLUSION

This paper therefore details about the well and best-known Machine learning methods or concepts and a captious review of the key ideas, various approaches and techniques were explained to build an appropriate paper. Machine

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 6, June 2017

learning will be used to learn the different features from the training data. Different Semantic, Syntactic, and Lexical features will be used to implement this system. Features like Ngram, Lemmatization, creating personalized vocabulary and observation of pattern are utilized. Algorithms like Lesk will be used to find the meaning per the context of the text. After the system is trained, testing data will be used to assess its results. As the system is tested, it will continue to learn from the input data to make it better. At the end system, will become an appropriate and give the correct result as mentioned before.

REFERENCES

[1] Identification of Spam Email Based on Information from Email Header - Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference
[2] A Comprehensive Study on Machine Learning Concepts for Text Mining - 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]
[3] Spam Mails Filtering Using Different Classifiers with Feature Selection and Reduction Techniques - 2015 Fifth International Conference on Communication Systems and Network Technologies
[4] E-mail Sender Identification through Trusted Local Deposit-Agents - 2011 International Conference on Network-Based Information Systems
[5] A Proposed Investment Decision Support System for Stock Exchange Using Text Mining Method - 2016 Al-Sadeq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA) – IRAQ (9-10) May
[6] Recommender System using Category Correlations based on WordNet Similarity - 2015 International Conference on Platform Technology and Service
[7] A Review on Social Audience Identification on Twitter using Text mining methods - IEEE WiSPNET 2016 conference
[8] A Survey and Evaluation of Supervised Machine Learning Techniques for Spam E-Mail Filtering - Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference
[9] An Approach to Automatic Text Summarization using WordNet - Advance Computing Conference (IACC), 2014 IEEE International
[10] Text mining using n-grams- Department of Statistics and Actuarial Sciences University of Waterloo Waterloo, ON, Canada
[11] Enron public mail corpus - https://www.cs.cmu.edu/~./enron/