International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

AN AUTOMATED CLASSIFICATION OF EMAIL CONTENTS BASED ON SEMANTIC MINING

S.Divya[1], Dr.G.Maria Priscilla[2]

[1]*Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science (Formerly SNR Sons College) (India)*

[2]*Professor & Head, Department of Computer Science, Sri Ramakrishna College of Arts and Science (Formerly SNR Sons College) (India)*

ABSTRACT

Email Classification is one of the vital problems in the email management due to its impact on the usage. Despite several applications like messengers such as Watsup, Kaizala and social media networks such as Facebook and Twitter, the importance of email was kept exploring. In order to increase the performance of management, it has become mandatory to automate the classification of the email against relevant and irrelevant emails. The email management poses several challenges to email servers and as an important and growing problem for individuals and organizations. Classification is one of the commonly used tasks in the data mining applications. It is theoretically and practically impossible to organize the content with historical data for training as large of feature extracted makes email undistinguishable. In this proposed system, analyses a novel automated email classification based on semantic mining towards its contents length and content evolution. Content evolution is common phenomena in the emails which occur as result of changes in the concepts. The feature is extracted using TF-IDF in order to convert textual data into vector space model for further processing. This proposed method equipped with Wordnet tool for semantic classification on the emerging emails. Semantic Classification uses Naive Bayes classifier model in order to generate the novel class to the email contents based on the feature similarity. To best of knowledge, the proposed model is the first work to develop a semantic classification to generate the novel class based on its contents evolution. The Experimental results demonstrate that proposed System outperforms existing state of art approaches in terms of probability, prediction and performance measures such as accuracy, precision, and recall.

Keywords: Data Preprocessing, Email Classification, Naive Bayes, Semantic Mining, WordNet.

I.INTRODUCTION

Due to the explosive growth of email contents in the email servers and it increases the attention of the researcher to classify the email contents among the huge amount of the email [1]. The dynamic and evolving nature of email content requires efficient and effective techniques that are significantly different from static data classification techniques. It can be named as Streaming Emerging New Classes (SENC) problem [2]. It is a fast and continuous phenomenon, it is assumed to have infinite length. In Existing, the common approach is to treat it as a classification problem and solve it using either a supervised learner or a semi-supervised learner among

feature selection methods without considering its semantic which have been shown to work well in unsupervised learning and supervised learning independently[3].

In this paper, an automated email classification technique has been proposed with utilization of Wordnet and Naive Bayes algorithm. Proposed algorithm allows methods to distinguish among two or more novel classes. The classification of the email is organized into several categories based on the feature extracted using TF–IDF. The extracted feature is represented as vector space model. With this approach, it is possible to distinguish different criteria for the email classification. Here, techniques are applied on Enron dataset [4]. The model empirically shows the effectiveness of this approach.

The remainder of the paper is organized as follows: Section 2 discusses the related works in email classification and its impacts against the performing classification under feature evolution, Section 3 briefly discusses the proposed technique in terms of automated email classification through semantic mining and Section 4 presents the experimental results on a data set. Section 5 discusses conclusions and future work.

II.RELATED WORK

There exist many techniques to classify the email which were designed and implemented efficiently. Each of these techniques follows some sort of class categorization, among few performs nearly equivalent to the proposed framework, which is described as follows,

2.1 Bag of Discriminant Words Based Email Classification

Many of the words in a given email either deliver facts depending on the topics they are involve. The different words have varying degrees of discriminative capacity in delivering the objective sense or the subjective sense with respect to their assigned topic. In this part, it utilizes the bag of Discriminant words mechanism as feature selection model towards email classification in order to produce the efficient discriminancy between the classes [5]. The essential idea underlying is that a pair of objective and subjective selection variables is explicitly employed to encode the interplay between topics and discriminative power for the words in email in a supervised classification.

2.2 Data Classification using Naive Bayes Classifier

Automated feature selection is important for text Classification as it reduces feature size and used to speed up learning process of classifiers. Naive Bayes algorithm is employed in to predict the class for the selected features. Naive Bayes algorithm predicts the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction. Naive Bayes classifier predicts the multi classes quickly. Also it has higher success rate as compared to other algorithms [6].

III PROPOSED SYSTEM

In this section, data preprocessing and automated classifier of email content using Wordnet in order to achieve Semantic Mining are described.

3.1 Data preprocessing

Data preprocessing is the key initiative in the data Classification. Data preprocessing is carried out using the following sub-process such as Tokenization, Stop word removal, Stemming and Lemmatization.

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

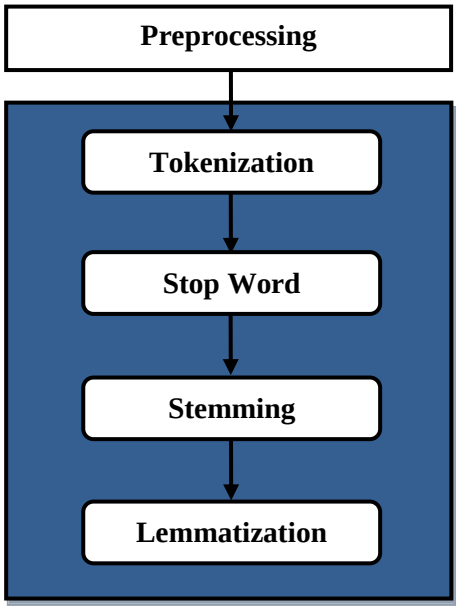

Figure.1 Preprocessing Operations

3.1.1 Tokenization

Tokenization is the process of breaking a stream of textual content into words, terms, symbols, or other meaningful elements known as tokens. The lists of tokens taking it as input for in additional processing including parsing or text mining. Tokenization is beneficial both in linguistics and Computer science. In general, the process of tokenization occurs at the word level. But, it's sometimes tough to explain what's meant by a "word". Frequently a tokenizer commits on simple heuristics, for example:

- Punctuation and whitespace may or may not be integrated into the resulting list of tokens.
- All nearby strings of alphabetic characters are part of one token; in a similar way with numbers.
- Tokens are divided by the way of whitespace characters, such as space or line break, or by punctuation characters.

3.1.2 Stop Word Removal

Stop Word removal is a process of eliminating the words which are considered as less important in the classification. Removing the Stop words reduces the dimensionality of the term space. The most common words to be removed are Pro noun and Prepositions of the each sentence [7]. Example for Stop words: the, in, a, an, with, etc. Stop words are removed from contents of the email because those words are not considered as keywords. The Stop Removal Methods are categorized into four types among that classic method is applied which is based on removing stop words obtained from pre-compiled lists.

3.1.3 Stemming

Stemming is a process of identifying the root/stem of a word. The function of this method is to remove different suffixes, to decrease the number of words, to have accurately matching stems and also to save time and memory space [22]. For example, the words Implement Implemented, Implementing and Implementation all can be

stemmed to the word "Implement". This is illustrated in FIG.2. In stemming, translation of morphological types of a word to its stem is done assuming each one is semantically related.

There are two points are considered while using a stemmer:

- Words that do not have the same meaning should be kept separate
- Morphological forms of a word are assumed to have the same base meaning and

hence it should be mapped to the same stem.

These two rules are good and sufficient in text mining or language processing applications. Stemming is usually considered as a recall-enhancing device. For languages with relatively simple morphology, the power of stemming is less than for those with a more complex morphology. Most of the stemming experiments done so far are in English and other west European languages.

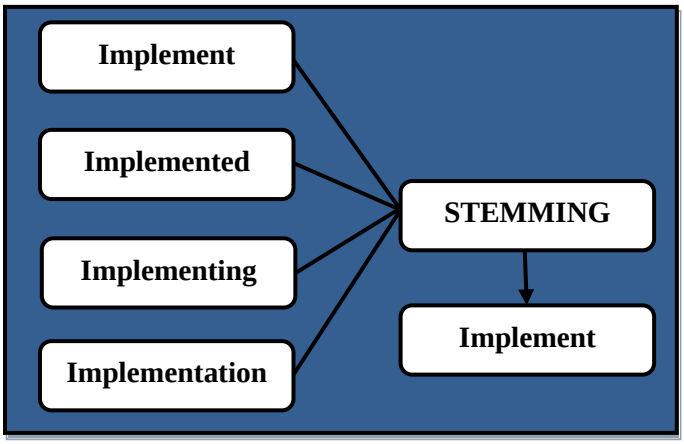

Figure.2 Stemming Process

3.1.4 Lemmatization

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of words that is called as the Lemma. If confronted with the token saw, lemmatization would attempt to return either see or saw depending on whether the use of the token was a verb or a noun. To extract the proper lemma, it is necessary to look at the morphological analysis of each word. This requires also having dictionaries of every language to provide that analysis.

Lemmatization is more formal or "Proper" because instead of just chopping the end of words to try and get to a base, there is more analysis done with lemmatization to get the true lemma of the word. For instance, "go" is the lemma of the words "goes", "gone", "going", and "went". Lemmatization is also much more complicated than stemming because it also uses vocabulary and morphological analysis to understand context before deriving root word or lemma. For example, lemmatization would differentiate word like "player" based on context to mean either a person that plays (lemma=player) or playing sports (lemma=play).

3.1.5 Term Frequency-Inverse Document Frequency

Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistic which reveals that a word is how important to a document in a collection. The TF-IDF uses the whole token within the dataset as vocabulary. The frequency of incidence of a token from vocabulary in every document consists of the term frequency and a number of documents within token occur to determine the Inverse document frequency.

If a token occurs frequently during a document that token can have high TF but if that token occurs frequently in majority of documents then it reduces the IDF, so stop word like: an, the, so, of, that etc., that occur frequently are penalized and important words that contain the essence of document get boost. TF–IDF is often effectively used for stop-words filtering in numerous subject fields together with summarization and classification.

TF–IDF is the product of two statistics that are Termed Frequency (TF) and Inverse Document Frequency (IDF),

- **TF:** Term Frequency that measures how frequently a term occurs in a document. Since every document is completely different in length, it is feasible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is generally divided by the document length (the total number of terms within the document) as a way of normalization:

$$\mathbf{TF(t)} = \frac{\mathbf{(Number\ of\ times\ term\ t\ appears\ in\ a\ document)}}{\mathbf{(Total\ number\ of\ terms\ in\ the\ document)}}$$

- **IDF:** Inverse Document Frequency that measures how important a term is. While computing TF, all terms are considered equally important. It is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have very little importance. Thus it needs to weight down the frequent terms while scale up the rare ones, by computing the following:

$$\mathbf{IDF(t)} = \log \frac{\mathbf{(Total\ number\ of\ documents)}}{\mathbf{(Number\ of\ documents\ with\ term\ t\ in\ it).}}$$

3.2 Feature Evolution Prediction & Feature Set Formation Based Wordnet- Semantic Mapping

The Feature evolution of the features derived from the TF-IDF model is represented in terms of Vector Space Model (VSM). Vector Space is to create a dictionary of the terms expert the stop words present in the email content. It acts as dimensions from which the index vocabulary is created to the words of the feature set with the help of the Wordnet tool. Here, Vector Space model uses the TF-IDF process.

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

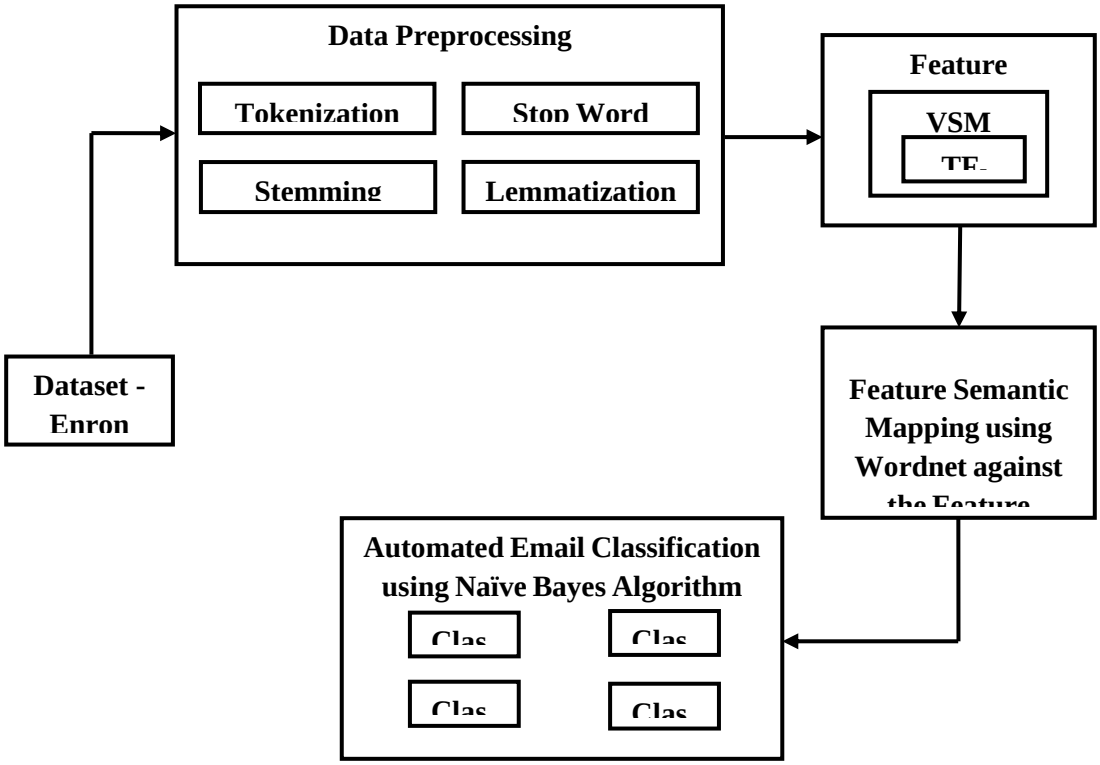

Fig.3: Architecture Diagram of the Automated Email Classification System

The Feature evolution is predicted using the Semantic Mining mechanism which each streaming word from the email is computed with existing feature set for its semantic similarity, if semantic similarity is existed then grouping the particular feature into existing feature space else grouping that as new feature [9].

3.3 WordNet

WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, abd records a number of relations among these synonym sets or their members. WordNet can thus be been as a combination of dictionary and thesaurus.

3.4 Automated Email Classification using Naive Bayes

Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Feature space is converted into frequency table. From the frequency table, likelihood table is generated by finding the probabilities. Naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability which is given as follows

$$\mathbf{P(c \mid x)} = \frac{P(X|C)\,P(C)}{P(X)} \text{ ---------- (1)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \ldots\ldots\ldots \times P(x_n|c) \times P(c) \text{ ---------- (2)}$$

Where,

P(c | x) is the posterior Probability

P(x) is the Predictor prior probability

P(c) is the Class prior Probability

P (x | c) is the likelihood

Naive Bayesian equation is used to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction [10]. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

IV.EXPERIMENTAL ANALYSIS

In this section, the experimental results of the proposed model against the existing approach are described.

4.1 Dataset Description – Enron Dataset

Dataset was collected and prepared by the CALO project. It contains data from about 150 users, mostly senior management of Enron, organized into folders. The corpus contains a total of about 0.5M messages.

4.2 Performance Evaluation

The proposed Framework is evaluated against the following measure termed as Prediction probability. The prediction probability of the class generated is computed against the proposed and existing method.

4.2.1 Classification Using Naive Bayes without Semantic Mining

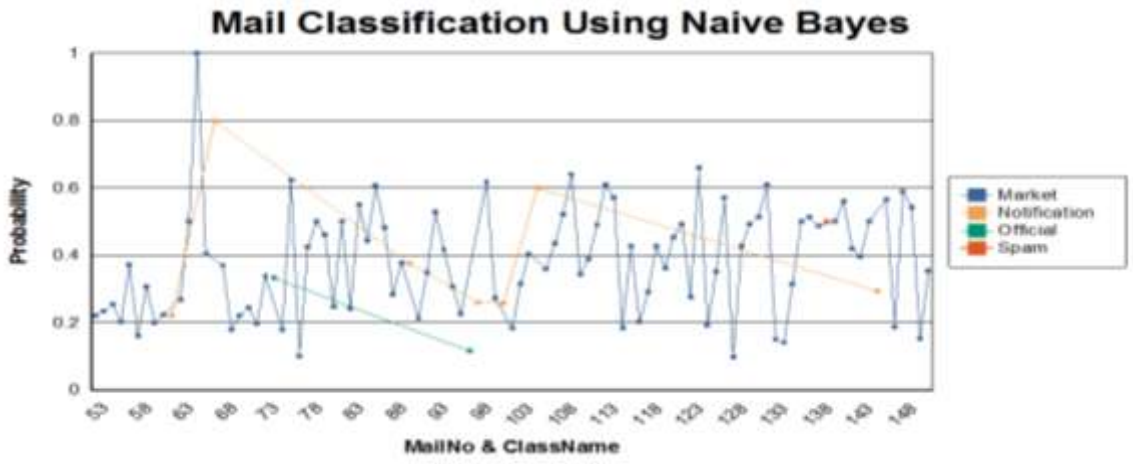

Figure.4 Existing Model - Classification Using Naive Bayes without Semantic Mining

The above FIG.4 illustrates classification result of an Existing model with different email records. Here, after data preprocessing Naive Bayes algorithm has been applied for email classification process. Each mail classified under one or more novel classes. A probability value of each mail under respective class has been calculated here. If the particular mail contains same probability value under more than one class, the mail will classify to the class based on priority.

Table 1 Probability value of Existing Model

Mail No	Class Name	Probability
58	Market	0.16
59	Market	0.31
60	Market	0.20
61	Market	0.22
62	Notification	0.22
63	Market	0.27
64	Market	0.50
65	Market	1.00
66	Market	0.40
67	Notification	0.80

TABLE 1 used to represent the probability value of each mail under the respective class name. The probability value may differ based on a word occurs frequently in one more records. Data has been extracted from the huge dataset by using Vector Space Model – TF IDF before applying Naive Bayes Classification for better classification.

4.2.2 Classification Using Naive Bayes with Semantic Mining

The below FIG.5 illustrates classification result of the proposed model with different email records. Here, after data preprocessing Semantic analysis has been applied on the extracted feature set, Where semantic analysis along with Word net tool that improves the probability of each record by giving the more suggested terms which give, same meaning but different words from non-class. Non-class represents the number of domain terms which does not belong to any class.

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

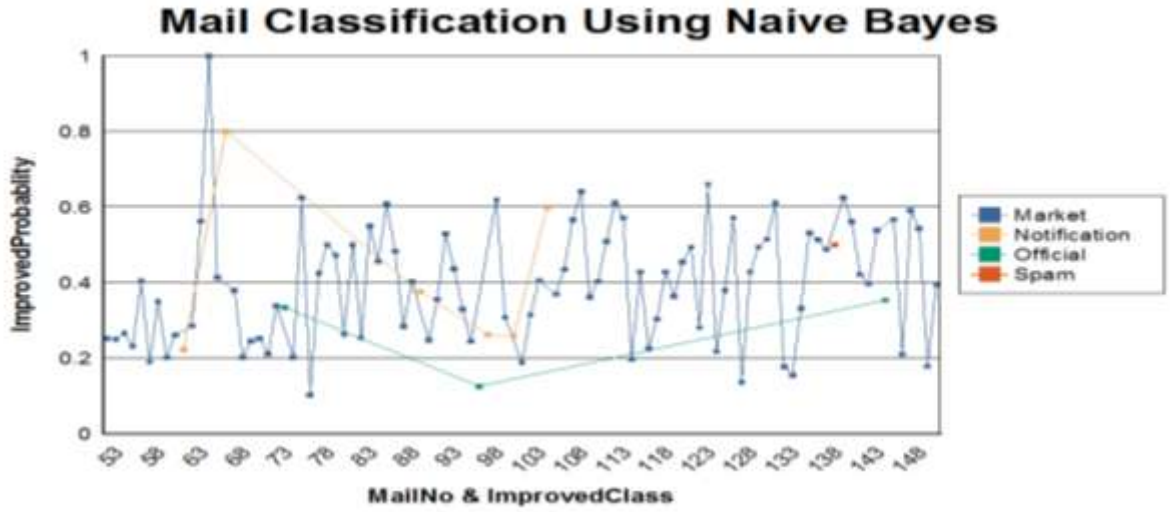

Figure.5 Proposed Model – Classification Using Naive Bayes with Semantic Mining

Table 2 Probability value of Proposed Model

Mail No	Improved Class Name	Probability
58	Market	0.19
59	Official	0.35
60	Market	0.25
61	Spam	0.26
62	Notification	0.28
63	Official	0.38
64	Notification	0.56
65	Market	1.00
66	Market	0.41
67	Notification	0.80

The TABLE 2 used to represent the probability value of each mail under the respective class name. List of suggested terms has given by Wordnet tool will be added in respective class from non-class that boost ups the probability of each record. A probability value of the proposed model is better than the probability value of the existing model is obtained as a result. Here, mail may swap their class name from one to another class based on a suggested terms that occur frequently in records and that improves probability too.

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

4.2.3 Comparison of Class Probability for Existing and Proposed System

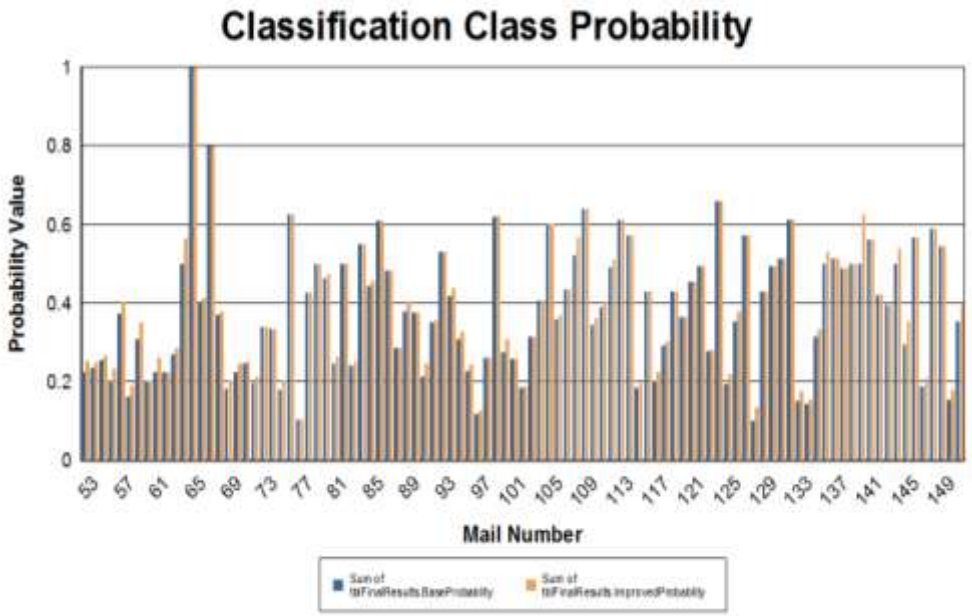

Figure.6: Performance Evaluation of proposed model against the existing approach

In FIG.6, prediction probability for existing and proposed model to the different record (mails) has been explained against the class generation. The divergence has been occurred due to the incorporation of Wordnet tool and semantic mining approach to the Naive Bayes algorithm. The value computed for each method is described in the TABLE 3.

Table 3: Performance Computation of the Email Classification Model

Mail No	Existing Class Name	Existing Probability	Proposed Class Name	Proposed Probability	Accuracy Improved
58	Market	0.16	Market	0.19	**0.03**
59	Market	0.31	Official	0.35	**0.04**
60	Market	0.20	Market	0.25	**0.05**
61	Market	0.22	Spam	0.26	**0.04**
62	Notification	0.22	Notification	0.28	**0.06**

The Accuracy has been improved in the proposed system due to inclusion of Wordnet and semantic mining approaches to exact class generation and distance between the each data points in the class is been obtained with least distance.

4.2.4 Classification Accuracy Improvement of Proposed Model

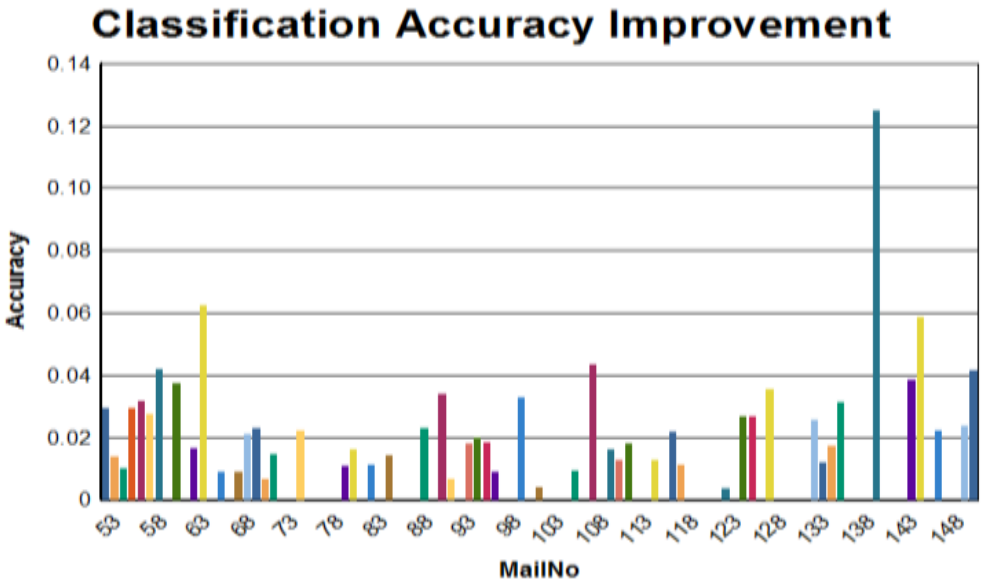

Figure.7 Classification Results of Proposed System

The above FIG.7 illustrates the classification results of the proposed system. Semantic mining along with Wordnet tool which smoothly increases the probability values of each record by suggesting more related terms. When probability value increases the accuracy level of email classification also increases simultaneously.

Table 4 Accuracy Computation of Proposed Model

Mail No	Improved Accuracy
58	**0.03**
59	**0.04**
60	**0.05**
61	**0.04**
62	**0.06**
63	**0.11**
64	**0.06**
65	**0.00**
66	**0.01**
67	**0.00**

TABLE 4 is used to represent the accuracy of email classification in the proposed system. Differentiate between probability value of existing and proposed system consider as an improved probability. When the probability increases that leads to improved accuracy of classification.

V.CONCLUSION AND FUTURE WORK

In this paper, an automated email classification based on semantic mining were designed and implemented. The Proposed work has given a great insight towards handling the streaming of data with infinite contents length and content evolution. Semantic mining and Wordnet tool has vital part in this classification. The use of different probability parameter within classifier allowed for tradeoffs between accuracy, Precision and recall. Future work will investigate the influence of feature reduction model on feature quality and classification accuracy.

REFERENCES

[1].Guangxia Li ; Steven C. H. Hoi ; Kuiyu Chang, Wenting Liu, Ramesh Jain *"Collaborative Online Multitask Learning"* IEEE Transactions on Knowledge and Data Engineering in *Volume: 26, Issue: 8, Aug. 2014*

[2]. S. Park and D. U. *An, "Automatic E-mail classification using dynamic category hierarchy and semantic features,"* IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India), *vol. 27, pp. 478-492, 2010.*

[3] J. C. Gomez, E. Boiy, and M. F. Moens, *"Highly discriminative statistical features for email classification,"* Knowledge and Information Systems, *vol. 31, pp. 23-53, 2012.*

[4]. https://www.cs.cmu.edu/~./enron/

[5]. Bo Tang, Steven Kay, Haibo He *"Toward Optimal Feature Selection in Naive Bayes for Text Categorization"* in IEEE Transactions on Knowledge and Data Engineering in *Volume: 28, Issue: 9, Sept. 2016*

[6].A. Harisinghaney, A. Dixit, S. Gupta, A. Arora, and Ieee, "*Text and Image Based Spam Email Classification using KNN, Naive Bayes and Reverse DBSCAN Algorithm,"* Proceedings of the 2014 International Conference on Reliabilty, Optimization, & Information Technology (Icroit 2014), *pp. 153-155, 2014.*

[7] Ms. Anjali Ganesh Jivani, *A Comparative Study of Stemming Algorithms,* Anjali Ganesh Jivani et al, Int. J. Comp. Tech. Appl., *Vol 2 (6), 1930-1938, ISSN:2229-6093.*

[8] Deepika Sharma, *Stemming Algorithms,* A Comparative Study and their Analysis, International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868, Foundation of Computer Science FCS, New York, USA, *Volume 4– No.3, September 2012 – www.ijais.org.*

[9]. Wen Hua, Zhongyuan Wang, Haixun Wang ,Kai Zheng,Xiaofang Zhou *"Understand Short Texts by Harvesting and Analyzing Semantic Knowledge"* in IEEE Transactions on Knowledge and Data Engineering in *Volume: 29, Issue: 3, March 1 2017*

[10]. Sang-Bum Kim,Kyoung-SooHan,Hae-Chang Rim, Sung HyonMyaeng *"Some Effective Techniques for Naive Bayes Text Classification"* In IEEE Transactions on Knowledge and Data Engineering in *Volume: 18, Issue: 11, Nov. 2010*