RESEARCH ARTICLE

NAIVE BAYES CLASSIFIER WITH MODIFIED SMOOTHING TECHNIQUES FOR BETTER SPAM CLASSIFICATION

Gurneet Kaur*, Er. Neelam Oberai

*Research Scholar Masters of Technology, Computer Science, Department of Computer Science, Maharishi Markandeshwar University, Sadopur (AMBALA)
Assistant Professor of Masters of Technology, Computer Science, Department of Computer Science, Maharishi Markandeshwar University, Sadopur (AMBALA)
* gur_neet_kaur66@yahoo.com; Neelamoberoi1030@gmail.com

Abstract: Text Mining has become an important research area due to the glorification of electronic documents available on web. Spam (junk-email) identification is one of the important application areas of Text Mining. Naive Bayes is very popular in commercial and open-source anti-spam e-mail filters. There are, however, several forms of Naive Bayes, something the anti-spam literature does not always acknowledge. A good spam filter is not just judged by its accuracy in identifying spam, but by its overall performance. It has been found that it largely depends on the smoothing method, which aims to adjust the probability of an unseen event from the seen event that arises due to data sparseness. The aim is at enhancing the performance of Naive Bayes Classifier in classifying spam mails by proposing a modification to Jelinek-Mercer Smoothing and Dirichlet Smoothing method against the Laplace method of traditional Naive Bayes Classifier. To overcome these issues, Naive Bayes Classifier is implemented with the modification in Smoothing techniques for calculating the collection probability for the model. The modified smoothing method calculates the collection probability by using the uniform distribution probability. The improved method shows the high performance in case of large data set, with precise number of keywords, with variations in smoothing factor. The improved method shows the high performance in case of varying data set, varying number of keywords and variations in smoothing factor based on the data set used.

Keywords: Naïve Bayes Classifier, Text Classification, Smoothing Methods, Spam Classification
1. INTRODUCTION-Text Mining[1]

The discovery of new and previously unknown information from a large amount of different unstructured textual resources is known as text mining. In text mining the patterns are extracted from natural language text rather than other databases. As in data mining, new unknown information patterns are generated by various techniques like classification, clustering, summarization, prediction, etc., when applied on data stored in data warehouse, similarly in Text Mining, new information is extracted by applying these methods on text documents. Various techniques used in text mining are Classification, prediction, clustering, aggregation, regression, etc[1].

1.2 NAÏVE BAYES CLASSIFIER

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (naive) independence assumptions.[3] An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters necessary for classification.

Assumption: A Naïve Bayes Classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature.

1.3 SMOOTHING METHODS

It refers to the adjustment of maximum likelihood estimator for the language model so that it will be more accurate. At the very first, it is not required to assign the zero value to the unseen word. It plays two important roles: 1) Improves the accuracy of the language model. 2) Accommodate the generation of common and non informative words.

General Model:
The maximum likelihood generator generally under estimate the probability of unseen words. So the main purpose of the smoothing is to provide a non-zero probability to unseen words and improve the accuracy of probability estimator. The general form of smoothed model is of the form:

\[ P(w|d) = \begin{cases} \frac{Ps(w|d)}{\alpha} & \text{if } w \text{ is seen} \\ \frac{P(w|c)}{\alpha} & \text{otherwise} \end{cases} \]

Where \( Ps(w|d) \) is the smoothed probability word seen in the document and \( P(w|d) \) is the collection language model and \( \alpha \) is the coefficient controlling the probability assigned to unseen words so that probabilities sum to one.

Generally, Smoothing methods differ in choice of \( Ps(w|d) \). A Smoothing method can be as simple as adding extra count or more complex where words of different count are treated differently.
1.4 CLASSIFICATION OF SPAM
In this era of rapid information exchange, electrical mail has proved to be an effective means to communicate by virtue of its high speed, reliability and low cost to send and receive [4]. Also, in recent years, the increasing popularity and low cost of e-mail have attracted the attention of direct marketers as they use to send blindly unsolicited messages to thousands of recipients at essentially no cost [7]. While more and more people are enjoying the convenience brought by e-mail, an increasing volume of unwanted junk mails have found their way to users' mail boxes [4]. This explosive growth of unsolicited e-mail, commonly known as spam, over the last years has been deteriorating constantly the usability of e-mail [8]. Unsolicited bulk e-mail, electronic and Spam messages posted blindly to thousands of recipients, is becoming alarmingly common. For example, a 1997 study by Cranor & LaMacchia, 1998 found that 10% of the incoming e-mail to a corporate network was spam [5]. Junk mail, also called unsolicited bulk e-mail, is Internet mail that is sent to a group of recipients who have not requested it [4]. The task of junk mail filtering is to rule out unsolicited bulk e-mail (junk) automatically from a user's mail stream.

2. PROBLEM STATEMENT

The main issues of Spam Classification using Naive Bayes Classifier are data Sparsity and cost of classifying spam without much reduction in recall are handled by using modified Jelinek-Mercer Smoothing and Modified Dirichlet Smoothing methods. The Main Parameter we have in mind to explore them out are:

I. Growing Need
II. To Handle Issues
III. High Performance and Accuracy

3. METHODOLOGY OF WORK
Methodology used in Improved Naïve Bayes Classifier with enhanced Smoothing Methods for Spam Classification contains the following steps:
Step 1, the documents selected as training data are imported at the back end. Instead of the complete document, we store the path of the document in the data set table.
Step 2, apply the various preprocessing steps such as removing the stop words, stemming rules and lemmatization on the data-set documents and then store the preprocessed documents as training set.
Step 3, based on the training set documents, generates the basic Naïve Bayes Classifier using Laplace method and using old Jelinek-Mercer and old Dirichlet Smoothing Techniques and modified new versions of Jelinek-Mercer and Dirichlet Smoothing.

Step 4, test the generated model on the documents selected as Test Data set. The results for the old and modified methods are stored in the table.

Step 5, compare the results stored in the table to check which method performs well.

4. PROPOSED ALGORITHM

In the already existing JM and Dirichlet Smoothing methods, the probability of word \( w_k \) in collection language model is calculated as,

\[
P(w_k/C) = \frac{\sum_{j=1}^{m} count(w_k,c_j)}{\sum_{k=1}^{m} \sum_{j=1}^{m} count(w_k,c_j)}
\]

Where \( m \) is the total number of classes and \( n \) is the total number of vocabulary words. Thus, above equation estimates total occurrences of word with respect to each class to the total number of occurrences of each vocabulary word with respect to each class. In the modified version, probability of word in collection model is not considered, rather it is considered as a function of word, which is a uniform distribution probability multiplied by the occurrence of word in collection model and is given by:

\[
P(w_k/C) = P_{\text{unif}}(W) \sum_{j=1}^{m} count(w_k,c_j)
\]
Where \( P_{\text{unif}}(W_k) = \frac{1}{|V|} \) \(|V|\) is the total number of vocabulary words and \( \sum_{j=1}^{m} \text{count}(w_k, c_j) \) is the total number of occurrences of word \( w_k \) in all classes. So, above equation becomes:

\[
P(w_k/C) = \frac{\sum_{j=1}^{m} \text{count}(w_k, c_j)}{|V|} \tag{38} \]

With the replacement of total word count of each vocabulary word with respect to each class, overhead for calculating the probability of with respect to whole collection has been reduced.

The above modifications in the smoothing techniques used for spam classification with Naïve Bayes Classifier is checked whether it reduces the cost factor without much reduction in recall and can it be able to handle zero probability problem of unseen words over the seen words shown by the results in next chapter. The experiment results obtained from the modified method is shown in next chapter. Thus, the application of the modified smoothing methods with Naïve Bayes Classifier for spam classification is checked that whether it enhances the overall performance of classification of spam or not.

### 5. RESULTS

Naïve Bayes Classifier with modified Smoothing methods and existing Smoothing methods is implemented for better classification of spam from the legitimate mails based on the text area of the mail. The results based on the varying data set size, varying number of keywords and varying the smoothing factor with respect to the accuracy, recall and cost of classification.

#### 5.1 PARAMETERS TO BE DISCUSSED:

I. ACCURACY

II. RECALL

III. COST OF CLASSIFICATION

The goal here is to measure the performance of Naïve Bayes Classifier with Enhanced Smoothing Method and whether the incorporated method helps to improve classification accuracy. Also, the performance of modified model is evaluated and compared with a Naïve Bayes classifier. At the same time the effect of number of keywords, training-corpus size, on the model’s performance, smoothing factors has been explored.

© 2014, IJCSMC All Rights Reserved
In figure 2.1, the results are shown on the basis no. of 100 documents corpus. The accuracy and recall of the NB with modified JM and Dirichlet Smoothing methods gets improved by 10% and 20% respectively as compare to Naïve Bayes Classifier with existing Smoothing methods. The cost of NB with modified smoothing methods is lower than Naïve Bayes Classifier with existing Smoothing methods.

It has been shown in figure 2.2 the results are shown on the basis no. of 150 documents corpus. Again, the accuracy and recall of the NB with modified JM and Dirichlet Smoothing methods gets better respectively as compare to Naïve Bayes Classifier with existing Smoothing methods. The cost of NB with modified smoothing methods is lower than Naïve Bayes Classifier with existing Smoothing methods.
Figure 2.2: Spam Classification using Naïve Bayes with Enhanced Smoothing Techniques. From the above graph, it is clear that enhanced JM method is proved to be best to improve the overall performance.

Figure 2.3: Spam Classification using Naïve Bayes with modified Smoothing Techniques in terms of Accuracy, Recall and Cost for varying Smoothing Factor $\lambda=0.5$ and $\mu=0.5$.

As depicted in figure 2.3, describes the performance of already existing methods with the modified methods in terms of Accuracy, recall, cost and Smoothing factors. In case of NB with Jelinek-Mercer Smoothing old, JM method has the highest value of accuracy and recall at $\lambda=0.5$. NB with Dirichlet Smoothing old, Dirichlet method new has highest accuracy and recall.
at $\mu=0.5$. In the case of NB with existing smoothing methods, modified smoothing methods has low cost of classification at $\lambda=0.5$ and $\mu=0.5$.

![Figure 2.4: Spam Classification using Naïve Bayes with modified Smoothing Techniques in terms of Accuracy, Recall and Cost for 1200 keywords.](image)

The graph shown in figure 2.4 describes the performance of already existing methods with the modified methods in terms of Accuracy, recall, and cost and 1200 keywords. the Naïve Bayes with enhanced Dirichlet Smoothing method is proved to be best in case of number of keywords; also the cost of Classifying Spam is less in this method with respect to other techniques, without much reduction in recall rate. With the precise number of documents, the overall performance has been increased by almost 10% in each classifier.

6. CONCLUSION

- With varying data set size, the performance of classifying spam increases by 5-10%. The Naïve Bayes Classifier with modified smoothing method achieves the highest performance as compare to Naive Bayes Classifier with already existing smoothing methods.

- For precise number of keywords, Naive Bayes Classifier with the enhanced Dirichlet smoothing method achieves the highest performance. Also, the overall performance of the system increases with precise number of keywords as compare to large dictionary size.

- In case of varying Smoothing factors and based on the studied data-set, the Naive Bayes with enhanced JM Smoothing shows the highest performance for spam classification at smoothing factor $\lambda = 0.5$. The results obtained by using enhanced Jelinek-Mercer
Smoothing method at $\lambda = 0.5$ is same at $\lambda = 0.9$. The Naive Bayes Classifier with enhanced Dirichlet Smoothing method shows the better result at $\mu = 0.7$, $\mu = 0.5$.

- To compare both the Naive Bayes Classifier with enhanced Jelinek-Mercer and Naive Bayes Classifier with enhanced Dirichlet Smoothing methods, it can be said that enhanced Jelinek-Mercer Smoothing method is more accurate than enhanced Dirichlet Smoothing Method based on Personal data-set used in this project.

7. **FUTURE WORK**

As in this solution, we have used the modified smoothing techniques, there are number of techniques that can be used as future work such as:

- There are various good Classification Algorithms other than Naive Bayes such as Support Vector Machine, Centroid Based, Nearest Neighbor, etc. Such techniques can be applied for Spam Classification task to see the improvements.

- The modified smoothing with Naive Bayes Classifier can be used to classify the mails into not just spam but also in number of folders.

- There are other various smoothing techniques Good Turing, Katz-Backoff and Witten-Bell that can be applied to Spam Classification to check the performance issues. These smoothing methods can be implemented as n-gram models, which represent the relation between the different features of the vector.

- The uniform probability distribution method, in case of probability of a word with respect to whole collection, can be embedded with the above smoothing techniques.

- Naive Bayes Classifier with Modified Smoothing Techniques can be used in other application areas such as documents Classification, News filtering and Organization, and document Organization and Retrieval.

- The Naive Bayes Classifier with modified Smoothing methods can be implemented in hierarchical manner to check the further improvements.

**REFERENCES**


