Multinomial Naive Bayes Classification Model for Sentiment Analysis

Muhammad Abbas[†], Kamran Ali Memon[††,†††], Abdul Aleem Jamali[†††], Saleemullah Memon[††††], Anees Ahmed[†††††]

[†]School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
[††,†††]State Key Laboratory of Information Photonics and Optical Communications (IPOC), School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China
[†††]Department of Electronic Engineering, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan.
[††††]School of Information & Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China
[†††††]Department of Software Engineering, Iqra University Karachi, Pakistan.

Abstract
Automatic document sorting becomes increasingly important as handling and organizing documents manually is a time consuming and not a viable solution on given the number of documents is very huge. The Naive Bayes method is very well-known method for text classification due to its effective grating assumptions, quick and easy implantation. In this article, we propose the simple, heuristic solutions to some problems with multinomial Naive Bayes (MNB) that address both systemic problems and those problems that arise due to reason that text is not actually the case generated according to a multinomial model. An MNB classifier is a type of NB classifier and is often used as a baseline for text classification but here it is applied for Sentiment Analysis (SA). We have used a dataset of movie reviews from the site. In each review contains a notice in the form of text and a numerical score (0 to 100 scale). The Exhaustive experiments with a large number of widely used reference data sets for text classification confirm the effectiveness of our proposed algorithm. Thus, accuracy can be greatly improved with Multinomial Naive Bayes classifier.

Key words:
Naive Bayes, Text Categorization Techniques, Bag of Words, Tokenization, Multinomial Naive Bayes model.

1. Introduction

Text categorization is the task of determining a document it belongs to a series of pre-specified class documents. The automatic classification scheme can greatly promote the classification process. Along with the rapid growth of information on the Internet, the classification of texts is a general and important research field trend in the search for information. Most approaches dealing with text classification issues [1] – [4] have been proposed to improve the accuracy of the text classifiers. To handle text classification tasks, documents are characterized by words appearing in the text. One technique is to use machine learning to classify documents and treat the absence of each word as a logical attribute as in one of the initial statistical model of language: Multivariate Bernoulli Naive Bayes model (BNB) [5] is done. BNB is well thought out but cares only about the appearance of words which make it a baseline for text classification. In BNB, the word when appears in the document, the value of the attribute equivalent to that word is written either one otherwise zero. As an improved method of BNB, multinomial Naive Bayes (MNB) [6] was proposed. MNB assumes that the document is a bag of words and takes word frequency and information into account. In order to overcome the system problems MNB faces, Complement Naive Bayes (CNB) [1] was proposed.

Nowadays, the multinomial models are considered to be the dominant modeling approach and it is more efficient than multivariate Bernoulli model which introduces language modeling in information retrieval. It is found that the multivariate Bernoulli models are significantly better than multinomial model in sentences search tasks. Since, sentence would be a short sentence and it is likely to be used in multiple languages. The Bernoulli model takes non-query terms into account [2].

This paper [7] focuses on the adaptation of simple MNB text classification for the sentiment analysis. The main contribution is that with the help of MNB because of that the bit of have issues it can works abnormal, slow and too much over fit while the data sets are small so, and multinomial performance is better than Bernoulli when it compared to Bernoulli model. In more details, multinomial is always a preferred method for any sort of text classification (spam detection, topic categorization, sentiment analysis) as taking the frequency of the word into consideration, and get back better accuracy than just checking for word occurrence[6].

The classifiers of Naive Bayes (NB) are a family of classifiers based on bayes' popular probability theorem,

Manuscript received March 5, 2019
Manuscript revised March 20, 2019

known as to create simple powerful models, particularly in the areas of document classification and disease prediction. The textual classification of NB is most often used for categorizing text as it is quick and easy to implement. The less faulty algorithms tend to be slower and more complex. We examine the reasons for NB, not the poor performance of others. We look at the NB as a linear classifier and find opportunities for improvement classification. We have better adapted the assign text to distribution, the NB has taken. Further, we work with the MNB classification model and discuss various systemic issues with it.

Basically, the NB algorithm is a machine learning algorithm. It is mainly used to categorize text, including multidimensional training data sets. Some examples are famously document classification, span filtration, sentimental analysis, and using the NB algorithm, one can quickly create models and quickly predict models. To estimate the required parameters, a small amount of training data is required. The NB algorithm is called "naive" because it assumes that the appearance of a feature is irrelevant to the appearance of other features.

To select the Naive Bayes classifier (NBC) would be more desirable because of its high speed. Even it is suggested that use NB, rather than other algorithms, for this size of problem, as they have a parallel map-reduce implementation for it. The NBC has excellent results for text data analysis. i.e. natural language processing, etc.

2. Literature Review

There have been many researches works available on the use of Naive Bayes (NB) for sentiment analysis. In [8] paper offers the possibility to use the KNN algorithm with TF-IDF method and text classification framework. This framework allows for classification by various parameters, measurement and analysis of results. The assessment of the framework focused on the speed and quality of the classification. The test results have been shown positive and negative points in the algorithm and given some pointers to further develop a similar framework. The main contribution of this study aims to develop a framework concept concentrating on KNN and TF-IDF modules[9], [10]. The structure offers the possibility to update and extend the existing integrated evaluation algorithms.

In [3], author proposed a weight classification algorithm based on k-NN classification paradigm for Weight Adjusted k-nearest neighbor (WAKNN) correction. In WAKNN, an iterative algorithm is used to learn the weight of the feature. In the weight adjustment step, to identify the update of maximum improvement of the objective function and the corresponding weight. It is shown in the results that it is superior to the most advanced classification algorithms. Rather using KNN algorithm because of a significant disadvantage behind the KNN. one of the main disadvantages of the similarity used in k-NN is to use all the features in distance calculation.

Additionally, the researcher has worked on problem as how to avoid local minima, in searching for the optimal weight carrier with experimental results in [3]. In [11], the author proposed method to solve the text classification problem statically. This solution includes an equivalence function to find the reliability of documents in categories known in the past. As it is based on a simple weighting approach, which is easy to modify so it is an efficient and simple algorithm. In addition, this article shows an k-NN algorithms to resolve the problem of text rating to a stable solution. Uses the predefined set of documents of algorithms known. Your number is set on runtime. With their help, the type of input document is found [11].

This article [12] presents a method called tree fast-K-nearest-neighbor (TFKNN), which allows one to quickly search for the nearest k neighbors. The traditional KNN has a fatal error because of the similar competitive computation time is applied to the KNN algorithm for classification which is too long for the large size of text and big samples. Author proposed the new algorithm to overcome the computational cost, that is TFKNN algorithm which allows one to immediately close the neighbor. Behind that, the SSR tree is created to search for nearby neighbors, all child nodes of each non-leaf node are ranked according to the distance between their principal points and the focus of their parents. Since that, the space has been shrunk in the form of trees. After that competitive computing time went very short [12]. Then, it retains the advantages of the KNN algorithm rules. In this research paper, improving the KNN approach is to accelerating all about. If apply the KNN approach for text classification, it needs to adjust the accuracy of the classification [12].

There exist many algorithms in data mining field and their usability on the basis of requirement for the text classifications. In this paper [13], researcher has proposed a new strategy which is based on the weakness and strength of each method after identifying. Author has also analyzed the algorithms and identified some of the disadvantages of both. Based on the analysis, author developed a new approach called kNN (kNNModel) model based algorithm, which combines the strengths of the k-nn and the Rocchio classifier[14]. Finally, all experimental results showed that it is superior to classifiers k-NN and Rocchio and comparable to support vector machine (SVM).

[1] introduced the simple heuristic solutions to certain problems of Naive Bayes (NB) classifiers, because text is not created according to a multinational model, both systematic and systematic problems occur. It turned out to be easy, fixed method results in fast algorithm to compete with the latest classification algorithm such as SVM. In addition, various techniques have been described modifying the application of NBC to text data. There are various methods of error correction in the application of the NBC to

text data. Here author proposed a series of transformations for finite frequencies. Each of them is attempting to solve another problem by assuming a Naive Bayesian model[1].
At present, many algorithms are presented by authors which focus on ways to get the information from web by the use of base command. As the researcher worked in [15], where Query-based k-NN method is used for accessing document related information via retrieval of the most relevant information among the documents presented on the Internet. Followed by, the feedback from the server side via the query base rather than regular content-based classification was obtained in [16].
Primarily, the text categorization method based on a query on text documents is proposed with the KNN approach in [12], thus giving a better margin of maneuver for optimality and functionality-based categorization. This method significantly reduced the response time to a query, increasing the accuracy and degree of relevance. It is useful to find the constraints that are most appropriate for interrelated documents presented on the Internet rather than the usual classification of content, and to classify documents based on this query[15].

3. Proposed Sentiment analysis model

One of the most common uses of machine learning is the analysis of categorical data, specifically text data.

3.1 Class Distribution

In Eq. 1, we have calculated the fraction of documents in each class π_c where the class c, word w at a word frequency f, $N = \{n_1, \dots, n_n\}$ where N is the total number of words and n represents the each word.

$$\pi_c = \frac{class_c}{\sum_{n=1}^{N} class_n} \quad (1)$$

3.2 Probability of each word per class

For calculating our probability, we have found the average of each word for a given class. For class c and word w, the average is given in Eq. 2

$$P(w|c) = \frac{word_{wc}}{word_c} \quad (2)$$

However, in Eq. 3 since some words had 0 counts, then we have performed a Laplace Smoothing with low α value, where α alpha presents the smoothing value for the unseen words that don't appear in the training data.

$$P(w|c) = \frac{word_{wc}+\alpha}{word_c+|V|+1}, \alpha = 0.001 \quad (3)$$

Where the V is an array of all the words in the vocabulary

3.3 Multinomial Naive Bayes Classifier

Combining probability distribution of P with fraction of documents belonging to each class, given in Eq. 4.

$$\Pr(c) \, \alpha \, \pi_c \prod_{w=1}^{|V|} \Pr(w|c)^{f_w} \quad (4)$$

In order to avoid underflow, we have used the sum of logs as in Eq. 5 and Eq. 6

$$\Pr(c) \, \alpha \, \log(\pi_c \prod_{w=1}^{|V|} \Pr(w|c)^{f_w}) \quad (5)$$

$$\Pr(c) = \log \pi_c + \prod_{w=1}^{|V|} f_w \log(\Pr(w|c)) \quad (6)$$

One issue is that, if a word appears again, the probability of it appearing again goes up. In order to smooth this, we take the log of the frequency (Eq. 7)

$$\Pr(c) = \log \pi_c + \prod_{w=1}^{|V|} \log(1 + f_w) \log(\Pr(w|c)) \quad (7)$$

Also, in order to take stop words into account, we have added here an Inverse Document Frequency (IDF), t is the term frequency in the document doc, the word t_w number of times term t appears in a document doc in (Eq. 8) and for further process put in (Eq. 9).

$$t_w = \log(\frac{\sum_{n=1}^{N} doc_n}{doc_w}) \quad (8)$$

$$\Pr(c) = \log \pi_c + \prod_{w=1}^{|V|} f_w \log(t_w \Pr(w|c)) \quad (9)$$

Even though the stop words have already been set to 0 for this specific use case, the IDF implementation is being added to generalize the function. As we can see, IDF has little effect as we removed the stop words. However, for the smoothing it makes the model more accurate. Hence, our optimal model is mathematically expressed in Eq. 10

$$\Pr(c) = \log \pi_c + \prod_{w=1}^{|V|} \log(1 + f_w) \log(\Pr(w|c)) \quad (10)$$

4. Experiments and Result Analysis

In this paper, the sentiment analysis is performed using statistics, natural language processing and machine learning which extracts and categorizes the content of feelings from a text extract. In this article, we used the multinomial Naive Bayes (MNB) classification algorithm for classifying movie reviews based on overall sentiment (positive/negative). each with a positive or negative sentiment label. We use the bag wherein disregard word order and focused only on the

number of occurrences of each word. Each document is denoted as a "bag" of words composed of several sets of word expressions to train the MNB classifier on the data and test the model performance with the holdout set. We have applied following steps.

4.1 Tables and Figures

The data set can be easily added as a pandas Data Frame with the help of 'read_csv' function. The encoding to 'latin-1' is set as the text had many special characters.

4.2 Data Cleanup

This step removes blank rows in data, if any, and remove stop words, non-Numeric and perform Word Stemming/Lamenting.

4.3 Word Tokenization

The first step is to split the text into proper units (letters, words, expressions, etc.). These units called tokens, and use tokenization at the word level. "she's funny person seems earth holly like able person out their personality".
output like as:
['she', 'is', 'funny', 'person', 'seems', 'earth', 'holly', 'like', 'able', 'person', 'out', 'their', 'personality']

4.4 Bag of Words (Bow)

The succeeding step is to create a numeric vector element for each document. Bow calculates the number of tokens collected in each document and returns a matrix with sequential properties.

4.5 Partitioning the Data

There are two stages in the classification: the learning phase and the evaluation phase. At the learning stage, classifier trains its model on a given dataset and in the evaluation phase, it tests the classifier performance. Performance is evaluated on the based on various parameters such as accuracy, error, precision, and recall rate. After completing the review of the text data, we have used scikit-learn's train_test_split to split into a training and testing set. As I have shown below the figure 1. ("MNB Model") uses 25% of the record for testing.

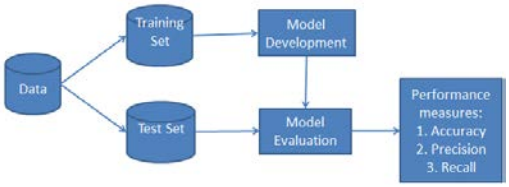

Fig. 1 MNB Model

4.6 Building a Classifier

The Naive Bayesian method is a probabilistic learning method based on Bayes' theorem. There are several variants of this algorithm, but here we have used MNB algorithm.

4.7 Multinomial Naive Bayes (MNB)

First attempt is to use a simple (MNB) and We get a 91% accuracy shown below the figure 2. ("Multinomial Naive Bayes Model Result").

	precision	recall	f1-score	support
0	0.84	0.87	0.85	3111
1	0.86	0.84	0.85	3139
micro avg	0.85	0.85	0.85	6250
macro avg	0.85	0.85	0.85	6250
weighted avg	0.85	0.85	0.85	6250

Fig. 2 Multinomial Naive Bayes Model Result

4.8 Confusion Matrix with using Logistic Regression

After that we have applied to increase the efficiency through NB classifier and then learn them from data using Logistic Regression and to accuracy goes to 93.1% and it is the good performance than before it shown below the Figure 3. ("Confusion Matrix by Logistic Regression accuracy").

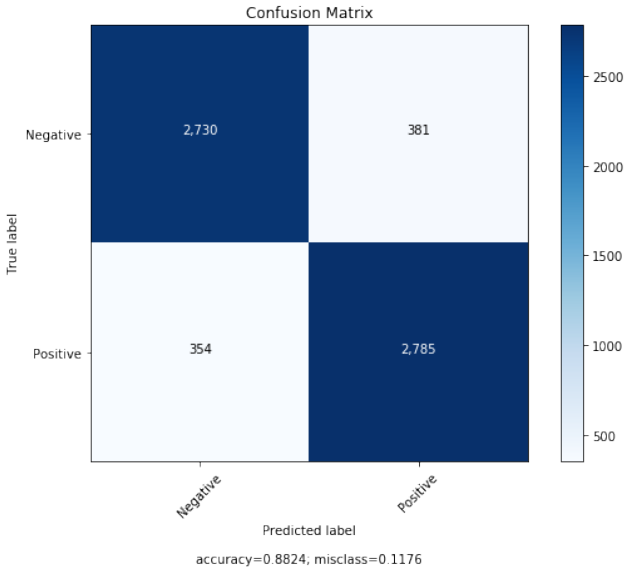

Fig. 3 Confusion Matrix by Logistic Regression accuracy

Let's plot most relevant words that the algorithm used to classify a text in positive or negative it shown below in the figure 4. ("Show relevant positive and negative words")

4.9 Enhancements

As there are words in several documents of both classes, they do not provide any relevant information. To overcome this problem, there is a convenient way called term frequency inverse document frequency (TF-IDF)[17]. It does not only take into account the frequency, but also the uniqueness of the word. In addition, each token in the Bow model we created, it represents a single word. This is called the unigram model. If the token is a pair of consecutive words, you can also add bigrams. And finally, the accuracy of predicting a classification model is given by the proportion of the total number of correct predictions.

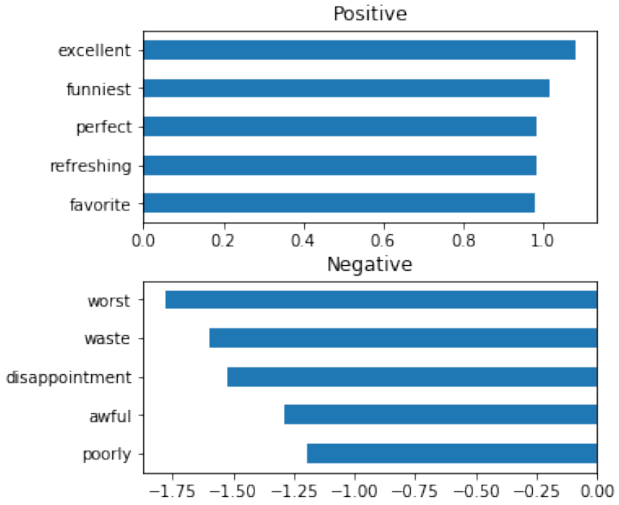

Fig. 4 Show relevant Positive and Negative words

The accuracy for this model turns out to be ~90% and showing the Confusion Matrix in the figure 5 with title final model result.

```
accuracy=0.8992; misclass=0.1008


confusion matrix:
 [[2767  344]
 [ 286 2853]]
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      3111
           1       0.89      0.91      0.90      3139

   micro avg       0.90      0.90      0.90      6250
   macro avg       0.90      0.90      0.90      6250
weighted avg       0.90      0.90      0.90      6250
```

Fig. 5 Final model result

We find that, despite many simplifying assumptions, the NB algorithm reasonably predicts the correct emotional category.

5. Conclusion and Future Work

In this paper, we present a text classification framework. for Sentiment Analysis based on MNB classification Algorithm and Method TF - IDF. The main motivation of this study is to develop a framework concept oriented towards the MNB algorithm and the TF - IDF module. Review and comparison of some modern Naive Bayesian classifiers is performed based on their ability to classify a large number of text documents efficiently. We achieved significant results in text categorization performance with the help of MNB Model. Basically, we improved the classification principles to improve the NB performance with standardize categorization and management the dependence of word occurrence. This change improved the performance of data sets, as experimental results reveal according as per numerical background of the datasets. MNB algorithm is a fast, easy-to-implement, almost modern text categorization algorithm. Proposed method and algorithm offers many possibilities for text categorizations. There are also some changes that can be made to our classifier for greater accuracy. It would be carried out in future and would involve use of artificial intelligence to improve the accuracy up to the best extent.

Acknowledgments

This work is supported by Natural National Science Foundation of China (NSFC) (61727817/ 61425022/ 61522501/61605013/61875248/61307086/61475024/61672290/61475094/61675030); the National High Technology 863 Program of China (2015AA015501, 2015AA015502), the Fund of State Key Laboratory of IPOC (BUPT). acknowledgment, if any.

References

[1] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, “Tackling the Poor Assumptions of Naive Bayes Text Classifiers,” no. 1973, 2003.

[2] D. E. Losada, “Language modeling for sentence retrieval: A comparison between Multiple-Bernoulli models and Multinomial models,” Bernoulli, pp. 1–9, 2005.

[3] E.-H. Han, G. Karypis, and V. Kumar, “Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification,” 2001.

[4] F. E. T. Al, “Locally Weighted Naive Bayes,” 2003.

[5] J. M. Ponte and W. B. Croft, “A Language Modeling Approach to Information Retrieval,” pp. 275–281.

[6] D. E. Losada and L. Azzopardi, “Assessing Multi-variate Bernoulli models for Information Retrieval,” no. February, 2014.

[7] L. Jiang, Z. Cai, D. Wang, and H. Zhang, “Knowledge-Based Systems Improving Tree augmented Naive Bayes for class probability estimation,” Knowledge-Based Syst., vol. 26, pp. 239–245, 2012.

[8] B. Trstenjak, S. Mikac, and D. Donko, “KNN with TF-IDF based framework for text categorization,” Procedia Eng., vol. 69, pp. 1356–1364, 2014.
[9] C. Friedman, T. C. Rindflesch, and M. Corn, “Natural language processing : State of the art and prospects for significant progress , a workshop sponsored by the National Library of Medicine,” J. Biomed. Inform., vol. 46, no. 5, pp. 765–773, 2013.
[10] K. Masuda and T. Matsuzaki, “Semantic Search based on the Online Integration of NLP Techniques,” vol. 27, no. Pacling, pp. 281–290, 2011.
[11] G. Toker and Ö. Kırmemiş, “Text Categorization Using k-Nearest Neighbor Classification,” Middle East Tech. Univ., 2013.
[12] Y. Wang and Z. O. Wang, “A fast KNN algorithm for text categorization,” Proc. Sixth Int. Conf. Mach. Learn. Cybern. ICMLC 2007, vol. 6, no. August, pp. 3436–3441, 2007.
[13] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “Using kNN Model-based Approach for Automatic Text Categorization,” Soft Comput., vol. 10, no. 5, pp. 423–430, 2006.
[14] T. Joachims, “A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.”
[15] S. Manne, “A Query based Text Categorization using K-Nearest Neighbor Approach,” vol. 32, no. 7, pp. 16–21, 2011.
[16] A. Arnold, “Query Dependent Ranking Using K-Nearest Neighbor ∗,” 2008.
[17] H. Ug, “Knowledge-Based Systems A two-stage feature selection method for text categorization by using information gain , principal component analysis and genetic algorithm,” vol. 24, pp. 1024–1032, 2011.

Authors

Muhammad Abbas is currently pursuing his MS degree at school of Computer science, Beijing University of Posts and Telecommunications (BUPT), China

Kamran Ali Memon, received his Bachelor Degree in Electronics Engineering (2009) from Mehran University of Engineering Technology, Jamshoro Pakistan and Master’s Degree in Communication (2015) from Quaid e Awam UEST Nawabshah Pakistan. He worked as a Lecturer/Assistant Professor in QUEST Pakistan for 08 years. Currently he is working toward his PhD at State Key Laboratory of Information Photonics and Optical Communications (IPOC), Beijing University of Posts and telecommunications, China. His research interests include optical and Wireless communications, PONs, Radio over fiber and WSNs.

Abdul Aleem Jamali received B.E. degree in Electronic Engineering in 2006 from Mehran University of Engineering and Technology, Jamshoro. After that he pursued for higher qualification to Germany where he received M.Sc. in Electrical and Communication Engineering and Ph.D. degrees from University of Kassel, Germany in 2011 and 2015, respectively. He is currently working as Assistant Professor, Department of Electronic Engineering, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah. His research interests include Engineering Electromagnetics, Plasmonics, Optoelectronics, and UWB antennas.

Saleemullah Memon received his B.E degree in electronic engineering from Quaid-e-Awam University of Engineering, Science and Technology (QUEST), Pakistan, in 2017. He is currently pursuing his MS degree at Key Laboratory of Universal Wireless Communication (Ministry of Education), Beijing University of Posts and Telecommunications (BUPT), China. His current research interests include simultaneous wireless information and power transfer (SWIPT) in cooperative relaying networks, MIMO systems and cognitive radio networks.

Anees Ahmed is presently affiliated with faculty of engineering science and technology, department of software engineering, Iqra university Karachi, Pakistan.