Text Mining from News Website using Machine Learning with Naïve Bayes Algorithm

Aye Aye Myint Aung, Su Wai Hlaing

Abstract— **Text Mining has gained a great deal of attention in recent years due the tremendous amount of text data, which are created in a variety of forms such as social networks, patient records, news outlets, etc. Unstructured text is easily processed and perceived by humans, but is significantly harder for machines to understand. Therefore, efficient and effective techniques and algorithms are required to define useful formats. Traditional algorithms struggle at processing these unstructured documents, but machine learning comes to the rescue. This research is based on machine learning techniques: a general inductive process automatically builds a classifier by learning the characteristics of the categories from a set of pre-classified documents. The system is built using Naïve Bayesian classifier and Python programming language. The main approach to this text categorization system that fall within the machine learning paradigm is described in this paper. The advantages of this knowledge-based engineering approach are very good effectiveness, considerable savings in terms of expert manpower, and straightforward portability to different domains.**

Index Terms— **Machine Learning, Naïve Bayesian, News Classification, Text Mining**

1) INTRODUCTION

Text classification or text categorization is the process of assigning text documents to one or more predefined categories. Text categorization process deals with large amount of electronic data such as blogs, newspaper, emails, online books etc. The tasks of managing electronic data are very challenging task as these data increase day by day. Currently, online news are also provided by many dedicated newswires such as Reuters. It will be useful to gather news from these sources and classify them accordingly for ease reference in the classification system.

Moreover, manually categorizing millions of documents is time consuming and expensive process. Thus, automatic text categorization is an important task due to large amount of electronic documents. Automatic text categorization has many applications including spam filtering, text filtering, e-mail routing, language identification, news classification, contextual search and genre classification. Document retrieval, categorization, routing and filtering can all be formulated as classification problems. The complexity of natural languages and the extremely high dimensionality of the feature space of documents cause the challenge in this classification problem.

This news classification system allows users to find desired information faster by searching only the relevant categories and not the entire information space. To automate the classification process, machine learning methods have been introduced. In a text classification method based on machine learning, classifiers are trained with a set of training documents. The trained classifiers can therefore assign documents to their suitable categories.

This paper is organized with six sections. Section 1 describes the research situation of text classification. Section 2 presents the literature review which is related to news classification in different language with different methods. Section 3 also presents background theory of this documents classification system. Section 4 is design and implementation of the system. Experimental result and performance evaluation are discussed in Section 5 and concludes this paper in Section 6.

2) RELATED WORKS

Many statistical learning methods with various languages have been applied in the field of text categorization in the recent years. This includes regression models, nearest neighbor classifiers, Bayes belief networks, decision trees, rule learning algorithms, neural networks, and inductive learning techniques[1] [2][3].

Young joong Ko and Jungyun Seo proposed [4] "Automatic Text Categorization by Unsupervised Learning". In this paper, unsupervised learning method was used for Korean Language text classification. The training documents were created automatically using similarity measurement and Naïve Bayes algorithm was implemented as the text classifier. The proposed method shows a similar degree of performance, compared with the traditional supervised learning methods. Therefore, This method can be used in areas where low-cost text categorization is needed. It also can be used for creating training documents.

Riyad Al-Shalabi, Ghassan Kanaan and Manaf H. Gharaibeh [5] provided a study of "Arabic Text Categorization Using KNN Algorithm". He has reached 0.95 micro-average precision and recall stores. Also he uses 621 Arabic text documents belong to six different categories. He has used a feature set consist of 305 keywords and another one of 202 keywords. Selection of keywords based on Document Frequency threshold (DF) method.

Moreover, different approaches on text classification were applied by many researchers to improve the efficiency. Jihong Guan and Shuigeng Zhou [6] proposed the effective algorithm for training training-corpus pruning. By using this approach, noisy and superfluous documents in training

Manuscript received October, 2018.
Aye Aye Myint Aung*, is with the Department of Information Technology, Pyay Technological University, Pyay, Myanmar. (corresponding author to provide phone:+959452333706,)*
Su Wai Hlaing *, is with the Department of Information Technology, Pyay Technological University, Pyay, Myanmar. (corresponding author to provide phone:+959962495919,)*

corpuses can be cut off drastically, which leads to substantial classification efficiency improvement.

Huang Ke and Ma Shaoping [7] combine concept indexing and principal component analysis to aggressively reduce dimensionality of document vector space without sacrificing categorization accuracy.

Makoto Iwayama and Takenobu Tokunaga [8] proposed a cluster-based search with a probabilistic clustering algorithm and evaluated on two data sets. This cluster-based search was more effective than the full search owing to the generalization of training documents.

M. Ali Fauzi et al [11] used Bayesian approach for Indonesian news classification. Before the model is trained with Bayesian classifier, two stages of feature selection were performed; information gain and maximum marginal relevance for feature selection. They also used vector space model for document representation and TF*IDF for weighting scheme. This new method could lower the complexity of MMR-FS but still retain its accuracy.

Amritpal Singh and Sunil Kumar Chhillar [13] used distinctive bag of word model for document representation taking highest rank keyword as feature and used artificial neural network for the classification of news categories. The experimental results shows high classification rate in describing category of a news document.

Mengke Feng and Guoshi Wu [14] presented a distributed Naïve Bayesian classifier for text document classification. Proposed algorithm computed the probability that a word belonging to a class by using its related words base on word embedding. Author used Map Reduce and Hadoop to implement the model and improved the precision in text classification and also processes efficiently.

In this paper, bag-of-words model is used to convert sentences into vectors and TF-IDF (Term Frequency – Inverse Document Freuency) weighting scheme is presented. Supervised machine learning methods are used to build the classifier and evaluate the resulting model.

3) BACKGROUND THEORY

1) Text Classification

Text classification or categorization has been broadly studied in different communities such as data mining, database, machine learning and information retrieval, and used in vast number of applications in various domains such as image processing, medical diagnosis, document organization, etc. Text classification aims to assign predefined classes to text documents [10].

The problem of classification is defined as follows. In a training set D= $\{d_1, d_2, \ldots, d_n\}$of documents, such that each document d_i is labeled with a label ℓ_i from the set $L = \{\ell_1, \ell_2, \ldots, \ell_k\}$. The task is to find a classification model (classifier) f where $f(d) = \ell$ which can assign the correct class label to new document d (test instance).

Based on the number of classes in the problem, classification can be divided into binary classification and multiclass classification, where binary classification categorizes instances into exactly one of two classes and multiclass classification deals with more than two classes.

Based on the number of classes that can be assigned to an instance, classification can be divided into single-label classification and multi- label classification. In single-label classification, one and only one class label is to be assigned to each instance, while in multilabel classification; more than one class can be assigned to an instance.

Based on the type of class assignment in multiclass system, classification can be divided into hard classification and soft classification. In hard classification, an instance can either be or not be in a particular class, without an intermediate state, while in soft classification, an instance can be predicted to be in some class with some likelihood.

Many of the classification algorithms have been implemented in different software systems. To evaluate the performance of the classification model, a random fraction of the labeled documents is set. After training the classifier with training set, classifying the test set, comparing the estimated labels with the true labels and measuring the performance is done. The portion of correctly classified documents to the total number of documents is called accuracy.

2) Machine Learning Approach

In Machine learning approach, sample labeled documents are used to define equations automatically. Text categorization is based on machine learning methods and machine learning approach to the text categorization is based on bag of words. Collection of documents is divided into two sets: training set and testing set, where training set is a pre-classified set of documents which are tagged manually by the experts and used for training the classifier. The testing set is required to determine the accuracy of the classifier whether the set is having correct or incorrect. The performance of the system depends on good training set. The use of concepts for text categorization increases its overall performance specifically when considering categorization of domain specific corpus.

3) Representation of documents using vector space model

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval, also known as the vector space model. In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.

This model is commonly used in methods of document classification where the frequency of occurrence of each word is used as a feature for training a classifier.

In the classis vector space model [11] [12], the term-specific weights in the document vectors are products of local and global parameters. To calculate the weight value of words, the following functions are needed. The model is known as Term Frequency-Inverse Document Frequency (TF-IDF) model.

Term Frequency (TF): Each term or feature *t* is assumed to have important proportional to the number of times it occurs in the document.

$$TF(t) = \frac{\text{Frequency of word or feature}}{\text{Maximum number of frequency of word}} \quad (1)$$

Inverse Document Frequency (IDF): This looks at feature occurrence across a collection of documents. The importance

of each feature is assumed to be inversely proportional to the number of documents that contain the feature.

The IDF factor of a feature *t* is given by:

$$IDF(t) = \log \frac{N}{n} \quad (2)$$

where, *N* is the number of document in each class and *n* is the number of document that contains the feature *t*.

Weight of a word: It is the combination of TF and IDF to weight terms. The combination of weight of a word t in a document is given by:

$$Weight(t) = TF(t) * IDF(t) \quad (3)$$

3) Naïve Bayes Classification

Naive Bayesian Classifier, or simply naive Bayes, is one of the most effective and efficient classification algorithms. Bayes theorem provide a method to calculate the probability of a hypothesis based on its prior probability, the probability of observing various data given the hypothesis, and the observed data itself.

Then, considering a set of documents *D* belonging a set of known classes *C*, the most probable classification of a document instance is obtained combining the predictions of all hypotheses (the prior probabilities of each one of them) weighted by their posterior probabilities.

The naive part in this approach is the assumption of word independence: the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category. This assumption makes the computation in Eqn. (4), which is more efficient than considering word combinations as predictors.

$$P(c_j \mid d) = \frac{P(c_j) \times P(d \mid c_j)}{P(d)} \quad (4)$$

where d is the unknown document to classify, c_j is the hypothesis such as t_k is the term value contained in the document, $P(c_j \mid d)$ is the probability of a document $d \in D$ belongs to $c_j \in C$, $P(d \mid c_j)$ is the probability of *d* condition of c_j, $P(c_j)$ is the probability of c_j, $P(d)$ is the probability of d.

$$P(d, c_j) = \sum (TF(t_k, d) * P(t_k, c_j)) \times (S_{ik} / S_i) \quad (5)$$

The reason for this is that according to the bag-of–word model, the ordering of words is ignored. $P(t_k, c_j)$ is the probability of a term $t_k \in W$ belong to class c_j.

W is the word of all terms that appear in *D*. S_{ik} is the number of words in d belongs to class and S_i is the number of words contained d. In order to classify an unknown sample *d*, $P(d, c_j)P(c_j)$ is evaluated for each class c_j, d is assigned to the class c_j for which $P(d, c_j)P(c_j)$ is the maximum in Eqn. (5).

4) OVERVIEW OF PROPOSED SYSTEM

This section is intended to present the design and implementation of news classification system using machine learning approach. The BBC news dataset has been used for this experiment [9].This original dataset consists of 2225 documents from the BBC news website corresponding to stories in five topical areas.

But, balancing the data with analyzing features of news documents is done in this dataset. The refined dataset has totally 1967 documents.

1) Overall System Deign

Fig.1 is the workflow diagram for the news classification using Naïve Bayesian algorithm.

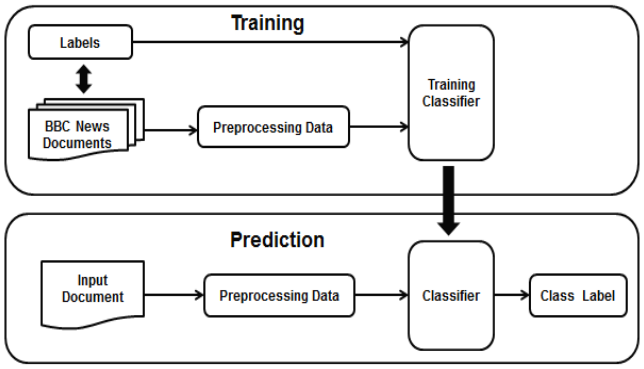

Figure: 1 Proposed System Design

Now, the detail of tasks related to this system in processing of news documents with machine learning approach is described.

- **Loading Data**

Data is the essential resource for machine learning tasks. BBC dataset is directly downloaded from its original source and exported to Comma Separated Value (CSV) file.

- **Preprocessing news documents**

In the preprocessing stage, the following processes are involved.

Tokenizing: This is the task of breaking text documents into pieces called tokens.

Filtering: Stop words (words that don't have particular meanings), punctuation and unnecessary character will be removed to from the training document to improve the overall accuracy of the system.

Feature Extraction: Every textual document is represented in the form of a vector to start training on those documents. This is done through Vector Space Modelling (VSM).

Weighting: This is a task of scoring the frequency of a token. To calculate the weight value of the words, Term Frequency-Inverse Document Frequency (TF-IDF) is used.

- **Training**

Learning process using machine learning algorithm is done and then the classifier model with Naïve Bayesian algorithm is built. Now system is ready to classify news documents.

- **Classification**

For the classification of incoming text documents, the system needs to process some steps as in training phase. Each document need to be preprocessed. Finally, the system can classify the incoming news document into corresponding category.

2) Implementation of the Proposed System Design

Fig. 2 describes the form of the news classification system in which an uncategorized news document (.txt) file is inputed by copying text from anywhere or browsing local file system to classify.

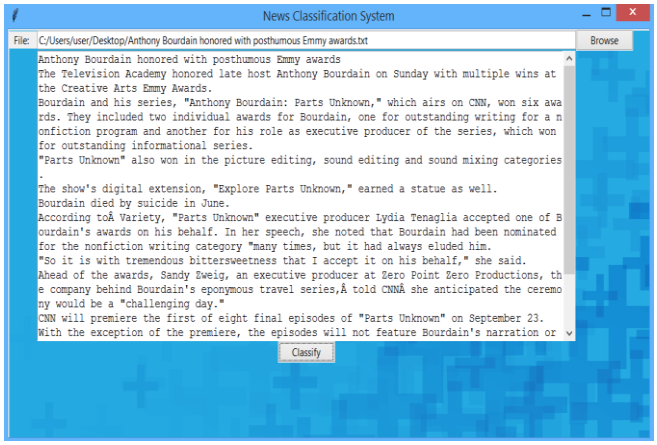

Figure: 2 Form of news classification system showing input documents

Fig. 3 describes the form of showing the category result based on the content of the input documents.

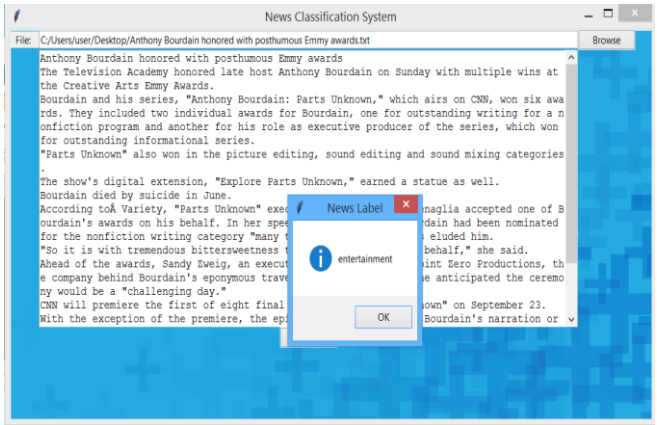

Figure: 3 Form of news classification system showing classification result

5) EXPERIMENTAL RESULTS

The following dataset is used by the system. Total 5 categories are used for training data and total number of documents is 1967.

TABLE 1. Dataset

Categories	Numbers of Documents
Business	380
Entertainment	389
Politics	417
Sport	380
Technology	401
Total	**1967**

The evaluation of the system has been done by using hold out method. In this method, given data is randomly partitioned into two independent sets.

- Training multi-set (e.g., 2/3 of data) for the statistical model construction, i.e. learning the classifier.
- Test set (e.g., 1/3 of data) is hold out for the accuracy estimation of the classifier.

Random sampling is a variation of the hold out method: Repeat the hold out *k* times, the accuracy is estimated as the average of the accuracies obtained.

Accuracy is taken as the average accuracy overall categories for the training and testing dataset where accuracy is computed as the percentage of correct category predication.

TABLE 2. Performance Evaluation

Dataset	Accuracy in Percentage (%)
Training	100
Testing	91

6) CONCLUSION

The final conclusion will be drawn from this research work is we have built very efficient and time saving system to classify the unseen documents with inductive machine learning approach. We can also prove sufficient conditions for the optimality of Naïve Bayes, one of the most efficient and effective inductive learning algorithms for machine learning and data mining. We intended to use this soft classification system in any task related to processing text documents, and other types of machine learning tasks. This work could further have extended with news documents in any other file format for classification process.

ACKNOWLEDGMENT

First of all, the author really thanks to Chairmen and Co-chairmen of Organization Committee for International Journal of Science, Engineering and Technology Research (IJSETR). I would like to express my deep gratitude to my research supervisor, Daw Su Wai Hlaing, for her patient guidance, enthusiastic encouragement and useful critiques of this research work. I would also like to thank Dr.Naychi Htun, Head of Information Technology Department, for her advice and assistance in keeping my progress on schedule. My grateful thanks are also extended to my teachers for their help in offering me the resources in running the program. Finally, I wish to thank my family for their strong moral and physical support, care and kindness throughout my study and anybody for overall supporting during my thesis.

REFERENCES

1) Y.Yang and X.Liu, "A re-examination of text categorization methods", In proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), pp.42-49, 1999.
2) Y.Yang and J.O.Pedersen, "A comparative study on feature selection in text categorization", In Proceedings of ICMC-97, 14th International Conference on Machine Learning, pp.412-420, 2002.
3) F.Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys (CUSR), vol.34 (1), pp.1-47, 2002.
4) Youngjoong Ko and Jungyun Seo, "Automatic Text Categorization by Unsupervised Learning", Proceedings of the 18th conference on Computational linguistics, vol. 1, pp.453-459, July 2007.
5) Riyad Al-Shalabi, Ghassan Kanaan, Manaf H. Gharaibeh, "Arabic Text Categorization Using KNN Algorithm", The International Arab Journal of Information Technology, vol.4, pp.5-7, 2015.
6) Jihong Guan and Shuigeng Zhou,"Pruning Training Corpus to speedup Text Classification", International Conference on Database and Expert Systems Applications, pp.831-840, 2015.
7) Huang Ke and Ma Shaoping, "Text Categorization based on Concept Indexing and Principal Component Analysis", Proc. TENCON Conference on Computers, Communications, Control and Power Engineering , pp.51-56, 2012.
8) Makoto Iwayama and Takenobu Tokunaga, "Cluster-Based Text Categorization",SIGIR'95 Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.273-280,2015.
9) **http://mlg.ucd.ie/datasets/bbc.html**
10) Tom M Mitchell, Machine Learning, 1997.
11) M. Ali Fauzi, Agus Zainal Arifin2, Sonny Christiano Gosaria3, Isnan Suryo Prabowo, "Indonesian News Classification Using Naïve Bayes and Two-Phase Feature Selection Model", Indonesia Journal of Electrical Engineering and Computer Science, Vol 8, No 3, December 2017.
12) Muthe Sandhya, Shitole Sarika, Sinha Anukriti, Aghav Sushila, "Automatic Text Categorization on News Articles", International Journal of Engineering and Techniques - Volume 2 Issue 3, May – June 2016.
13) Amritpal Singh, Sunil Kumar Chhillar, "News Category Classification Using Distinctive Bag of Words and ANN Classifier", International Journal of Emerging Research in Management and Technology, ISSN-2278-9359, Volume 6, Issue 6.
14) Mengke Feng, Guoshi Wu, "A Distributed Chinese Naive Bayes Classifier Based on Word Embedding", 4th International Conference on Machinery, Materials and Computing Technology (ICMMCT 2016).
15)https://ils.unc.edu/courses/2013_spring/inls509_001/lectures/06-VectoSpaceModel.pdf
16) https://en.wikipedia.org/wiki/Vector_space_model

Aye Aye Myint Aung received her BE (Information Technology) degree from Pyay Technological University, Pyay, Myanmar. She is doing postgraduate research for master degree at Information Technology Department, Pyay Technological University. Her research is concerned data mining technology. She is also lecturer at University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar.

Su Wai Hlaing (Lecturer), Department of Information Technology in Pyay Technological University, Pyay, Myanmar.