Hinglish Text Classification

Arnav Panchal
Student, Dept. of Computer Engineering, PHCET, Rasayani, Maharashtra, India

***Abstract*- Text Classification can be done in easily in English, but it's difficult to perform in various other languages. Not much of the work is done for Indian languages like Hindi, Marathi, Bengali. Due to the incredible growth in the internet users most of the people are comfortable with Hinglish which is the combination of Hindi and English. This paper identifies of language, sentiment analysis and the new classification based on Hinglish language is proposed.**

Index Terms- Hinglish, Text Classification, Naive Bayes, Tokenization, Diacritics Removal, Words Stemming.

I. INTRODUCTION

Text classification is always been an important application and research topic. Today, text classification is a necessity due to large amount of text document that we have to deal with daily. With the list of references at the end of the paper. Texts can be written in various genres, for instance: scientific articles, news articles, movie reviews.

In this paper, we look at classifying text for, identifying whether the given text is Hinglish or English, sentimental analysis determining whether the text is positive or negative and news classification for determining what type of news it is, is it Political, Sports, Entertainment, Auto.

II. THE SYSTEM MODEL

This paper introduces the principle of the text classification and design system based on machine learning algorithm. This model trains the classifier according to the existing data, and then classifies the unlabeled data by it. Fig.1 is an example of this model.

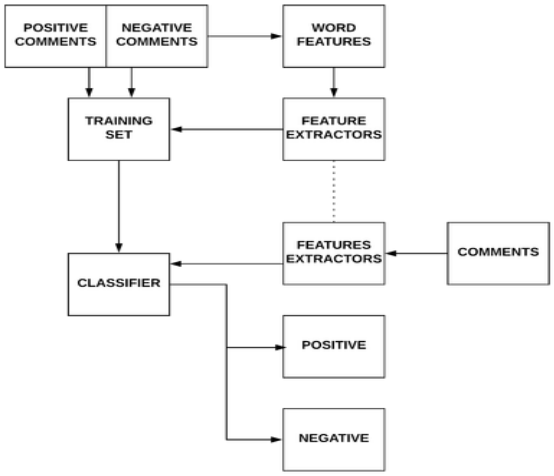

Fig -1: Text Classification model (Sentimental analysis)

IV. TEXT CLASSIFICATION WORKFLOW

Classification is a challenging task in the field of text mining as it requires pre-processing steps to prepare the textual data into structured form which is initially available in un-structured form.

V. GATHERING DATA

This is the first step of text classification, which involves gathering data in hinglish from various sources. These sources can be anything from newspapers, electronic media, print media and much more.

VI. PRE-PROCESSING

Text pre-processing is a primary step in news headlines classification process. Text is pre-processed effectively where the unstructured text data is initially obtained, which mostly is the combination of both garbage and useful data. All of the data comes from variety of data gathering sources and is to be cleaned. Firstly, the text data is made free from

all noisy and useless information, which include punctuation marks, semicolons, irrelevant texts, quotes, exclamation marks, dates etc.

Tokenization: Breaking huge text into small tokens or segments is said to be text tokenization. Each word in the headline is treated as a string and headlines are broken in small tokens. Final output obtain is then served as input for further processing of text mining. All documents are combined and a set of words is obtained. This process is called dictionary [1,2].

Diacritics Removal: Diacritics are defined according to the language. All irrelevant words that include commas, semi-colons, quotes, double quotes, full stops, underscores, special characters, dashes, and brackets etc. are removed. Simple way to implement this is to replace all those words, which are diacritics with simple space [1, 3]. SAS text miner was used by M.W Pope where he removed diacritics for classifying news headlines in an automated way [4].

Words Stemming: One of the most important step in pre-processing is stemming. Stemmer reduces the word to its root by removing affixes. Stemming is adopted to cluster words as per document topics. Purpose of performing stemming is to save time and space to make classification efficient and fast[5]. Using this approach, a stemming algorithm is applied, which stems normal words from to their root forms. The most common stemmers being discussed in literature are S Stemmer, Porter, and Paice/Husk Stemmers.

VII. INDEXING

This is considered as the most important step for text classification because each word in data is considered to be a vector and bag of words approach is used to reduce complexity as well as difficulty of text classification. Usually high dimensionality and loss of correlation are minor issues faced in Bag of words approach, but still this approach is useful for indexing news headlines. Each word in the form of vector is shown and a complete matrix is made for that purpose.

VII. TRAINING MODEL AND CLASSIFICATION

Text classification mainly uses classifier to label the text of unknown category, so most important part of classification is the selection of classification algorithm. At present, the most commonly used algorithm is machine learning, which lets the computer to train the classifier according to the training set. And for a new classified document, the computer will make a judgment based on the previous experience, learning the rule of classification, and then give a category.

Naive Bayes: Naive Bayes [6] is a probabilistic classifier, which is entirely based on text features. Individual calculations are done for each feature using this classifier. It prepares class labels and calculates text probability within those classes. Use of Naive Bayes for news classification exists in literature at higher extents. Multivalued headlines classification [7] as well as full text news classification was performed by Andreas using Naive Bayes and results were compared. Due to incorrect parameter assessment of Naive Bayes, Poisson model [8] was proposed and in result improved accuracy measure was reported. Ignoring minor classification issues of naive bayes, it was declared better than other classifiers by S. Ting et. al.

The Naive Bayesian algorithm is a very simple classification algorithm based on Bayes theory. Bayes theory is calculating the frequency of occurrence of something in the past to estimate the future probability of its occurrence. Its results indicate that the probability distribution for the random variable and can also be interpreted as the possibility of different level of trust. The Bayesian formula is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

That is the probability of event *A* occurring on the premise that event *B* has already occurred. In general, we usually use another form:

$$\mathrm{P}(B \mid A) = \frac{P(AB)}{P(A)} = \frac{P(A \mid B) * P(B)}{P(A)}$$

In the Naive Bayesian classification algorithm, if the unclassified feature set is represented by *A*, the classification result set is represented by *B*, then:

$$P(category|feature) = \frac{P(feature|category) * P(category)}{P(feature)}$$

Thus, we can find out the category that makes P (category | feature) the biggest, this category is the classification result of the text.

VIII. RESULTS

The Naive Bayesian method algorithm takes less time, and it is easy to implement. However, the Naive Bayesian classification model needs to meet the requirement that the conditional independence assumption between features
Overall, the Hinglish text classification system got satisfactory results.

IX. CONCLUSION

Area of text mining is so vast that anything can be made possible for improving system accuracy results. This paper is constructing Hinglish text classification system model based on Machine learning algorithm. It achieves a high accuracy of classifications. We used the Naive Bayes algorithm which is based on probabilistic framework to handle our classification problem. However, there are many problems with the Hinglish text classifier like that the small class produces better results than big class, and that semantic information of text features is insufficient. Thus how to make the classification more effective is the next major content of the study.

REFERENCES

[1] K Anita. TEXT CATEGORIZATION: Building a k-NN classifier for the Reuters-21578 collection. December 4, 2006

[2] Andreas Hotho ―A Brief Survey of Text Mining_ 2005. LDV Forum- GLDV Journal for Computational Linguistics and Language Technology, 20, 19-62

[3] Joachims, T. Text categorization with support vector machines. Technical report, LS VIII Number 23, University of Dortmund, 1997.

[4] M. W. Pope, "Automatic Classification of Online News Headlines,"2007. School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfilment of the requirements for the degree of Master of Science in Information Science (November2007)

[5] Moral, C., de Antonio, A., Imbert, R. & Ramírez, J. (2014). A survey of stemming algorithms in information retrieval/Information Research, 19(1) paper 605. http://InformationR.net/ir/19-1/paper605.html

[6] Kim, S.B., Han, K.S., Rim, H.C., Myaeng, S.H.: Some EffectiveTechniques for Na¨ıve Bayes Text Classification. IEEE Trans. Knowl. Data Eng. 18(11), 1457–1466 (2006)

[7] Andreas Heb, Philipp Dopichaj and Christian Maab, "Multi-Value Classification of Very Short Texts.", In Proceedings of the 31st annual German conference on Advances in Artificial Intelligence, Springer-Verlag Berlin, 2008, pp.70-77

[8] S. Ting, W. Ip &A. H. Tsang, "Is Naive Bayes a good classifier for document classification?," International Journal of Software Engineering and Its Applications, vol 5 no 3, 2011