Volume 6, Issue 4, April 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

News Classification Using Naïve Baye's Classifier

[1]Jasneet Kaur, [2]Seema Bhagla
[1] Research Scholar, [2]Assistant Professor,
[1, 2] Computer Engineering, Yadavindra College of Engineering, Talwandi Sabo, Punjab, India

Abstract –With the RAPID growing rate of techniques for manipulation in Real Time Data. News classification has increased the interest in the research of text mining. Correctly identifying the news into particular category is still presenting challenge because of large and vast amount of features in the dataset. In regards to the existing classifying approaches, Naïve Baye's is potentially good at serving as a document classification model due to its simplicity. This paper proposed the news classification using Naïve Baye's classifier in which several types ofdifferent news has been classified like politics, business, entertainment and health. The whole implementation has been taken place in Visual Basic 2010 by using language C#.

Keywords - News Classification, Naïve Baye's Classifier, Text Mining.

I. INTRODUCTION

News classification is a growing interest in the research of text mining. Correctly identifying the news into particular category is still presenting challenge because of large and vast amount of features in the dataset. In regards to the existing classifying approaches, Naïve Baye's is potentially good at serving as a document classification model because Naïve Baye's model is very simple and is also potentially good due to its simplicity.

With the rapid growth of online information, text categorization has become one of the key techniques for handling and organizing text data.Text categorization techniques are used to classify news stories, to find interesting information on the World Wide Weband to guide a user's search through hypertext. In these days, most of the available contents are in digital form. To manage such data is big challenge.The textual revolution has seen a tremendous change in the availability of online information. Finding information for just about any need has never been more automatic. Therefore, Text Classification is the task in which sortingis done automatically to classify the documents into predefined classes. Manual text classification is an expensive and time-consuming method, as it become difficult to classify millions of documents manually. Therefore, automatic text classifier is constructed using labeled documents and its accuracy is much better than manual text classification and it is less time consuming too.

The proposed work includes the use of Naïve Baye's for online news classification. In the proposed work four types of news has been classified like business, sports, entertainment, political and health. And the whole implementation has been taken place in Visual Basic 2010 in language C# by Microsoft.

Punjabi is an Indo-Aryan Language, spoken in both western Punjab (Pakistan) and eastern Punjab (India). It is 10th most widely spoken language in the world. Also it is the official language of Indian state of Punjab. In comparison to English language, Punjabi language has rich inflectional morphology but very little work has been done for text classification with respect to Indian languages, due to the problems faced by many Indian Languages such as: no capitalization, non-availability of large gazetteer lists, lack of standardization and spelling, scarcity of resources and tools. E.g. English verb "Play" has 4 inflectional forms: play, played, playing, plays; whereas same word in Punjabi. This depends upon gender, number, person, tense, phase, transitivity values in a sentence.

Rest of the paper is organized as: Section II presents the literature survey, Section III provides the system model, Section IV displays the proposed work, Section V shows the implementation results and finally section VI shows the conclusion and future scope for the proposed research work.

II. RELATED WORK

Table-1 Previous Techniques

Nidhi and Gupta, (2012) [9]	Studied that the Classification of text documents become a need in today's world due to increase in the availability of electronic data over internet and also investigated the Punjabi Text Classification is the process of assigning predefined classes to the un labelled text documents because of dramatic increase in the amount of content available in digital form.
Nidhi and Gupta, (2012) [10]	Studied that the Text Mining is a field that extracts hidden, not yet discovered, useful information from the text document according to user's query.
Nidhi and Gupta (2012) [11]	Investigated that the Punjabi Text Classification is the process of assigning predefined classes to the unlabeled text documents because of dramatic increase in the amount of content available in digital form.

Brutlag and Meek [7]	Studied that the interactive classification of email into a user-defined hierarchy of folders is a natural domain for application of text classification methods.
McCallum and Nigam [2]	Examine that the text classification have used two different first-order probabilistic models for classification, both of which make the naive Bayes assumption.
Durga, Govardhan (2011) [1]	Introduce a new method of ontology based text classification for Telugu documents and retrieval system.
Frank1 and Bouckaert [8]	Develop that the Multinomial naive Baye's (MNB) is a popular method for document classification due to its computational efficiency and relatively good predictive performance.
Raghuveer and Murthy [4]	Presents their work on automatic text categorization in Indian languages. Here author use purely corpus based machine learning techniques
Ali and Ijaz [3]	Authors compare statistical techniques for text classification using Naïve Baye's and Support Vector Machines, in context of Urdu language.

III. SYTEM WORK MODEL

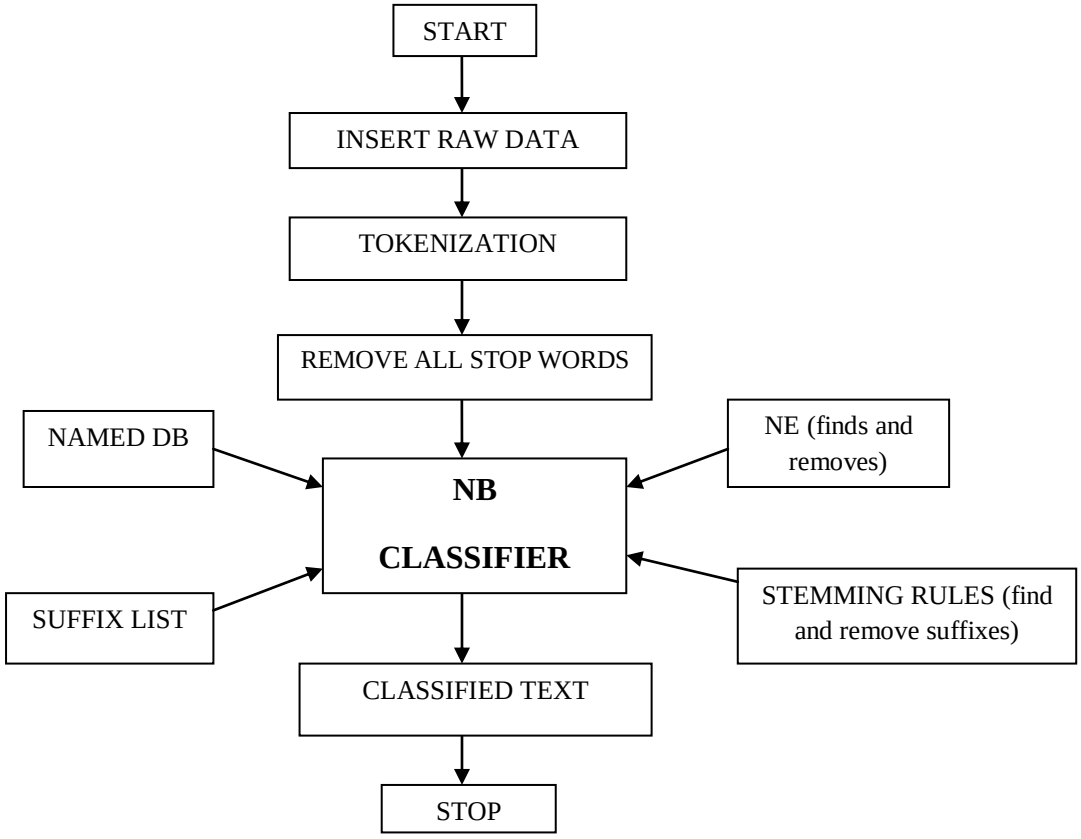

Figure 1: Flow Chart

IV. PROPOSED WORK

Hybrid models are basically combination of rules based and statistical models. In Hybrid NER system, approach uses the combination of both rule-based and ML technique and makes new methods using strongest points from each method. It is making use of essential feature from ML approaches and uses the rules to make it more efficient.

Algorithm

It is probabilistic classifier that considers each term independent of each other.this algorithm consider each Punjabi Text Document d as Bag of words i.e. d= (w1, w2,……wn) where wn is the nth word in the document and then for classification calculate the posterior probability of the word of the document being annotated to a particular class.

Training Set

Prepare training set for the classifier in which folders represent class and each folder contains set of documents called labeled documents. Punctuations, special symbols are removed from the document. Then, documents are segmented into meaningful units called words. Stop words, Name entities such as names, locations, date/time, counting etc. are removed from the document as they are irrelevant to the classification task.

- Step1: Calculate total words in each class in the Training set.
- Step2: Calculate total words in Training set.
- Step3: Calculate P(c) the prior probability of a document occurring in each class c.

P(Ci) = (Total docs in Ci) / (total docs in training set).

Test set:

- Step4: After preprocessing and feature extraction steps, each unlabeled document are represented as list of words i.e. w1, w2 ….wn, where wn is the nth word of the document.

Calculate probability of the document to belong to the particular class using equation.

P(Ci|document) = (P(Ci|w1, w2……wn)/ n

Where n is the total word in the input document.

Assign class Ci to the document if it has maximum posterior probability with that class.

P(Ci|document) = max (P(Ci)*P(wj|Ci))/ n

Where

P(wj|Ci) = (1+freq. of wj in class Ci)/(total words in Ci + total words in training set).

Evaluation Parameters in Research Work

$$\text{Recall} = \frac{\#\ of correct output return by the system}{\#\ of total files}$$

$$\text{Precision} = \frac{\#\ of correct output returns of system}{\#\ of Actual(True) predictions}$$

Total Files Tested: 100

$$\mathbf{R} = \frac{18+20+16+18}{100} = \frac{72}{100} = \mathbf{0.72}$$

$$\mathbf{P} = \frac{72}{92} = \mathbf{0.78}$$

$$\mathbf{F_1}\ \textbf{–Score} = 2 * \frac{(R*P)}{(R+P)}\ \ 2 * \frac{0.72*0.78}{0.72+0.78} = \mathbf{0.74}$$

Rule Based Approach

It uses linguistic grammar-based techniques to find named entity (NE) tags. It needs rich and expressive rules and gives good results. It requires great knowledge of grammar and other language related rules. Good experience is needed to come up with good rules and heuristics. It is not easily portable and has high acquisition cost. It is very specific to the target data.

Research Flow

The whole implementation has been taken place in Visual Basic 2010 by using language C#. Below figure shows the implementation of the classification of news of various types like business, politics, entertainment and health.

V. LIVE WINDOWS

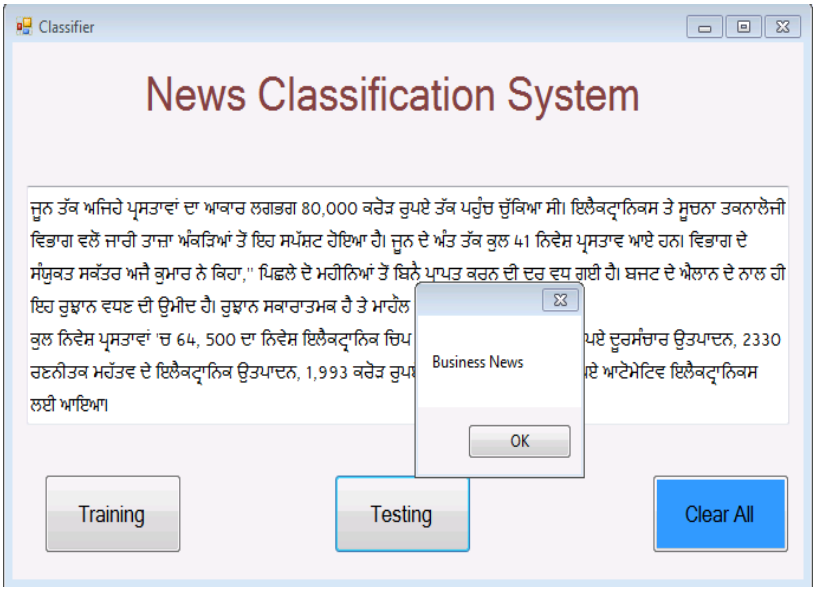

Fig: 1 Text Classification Script Related to Business

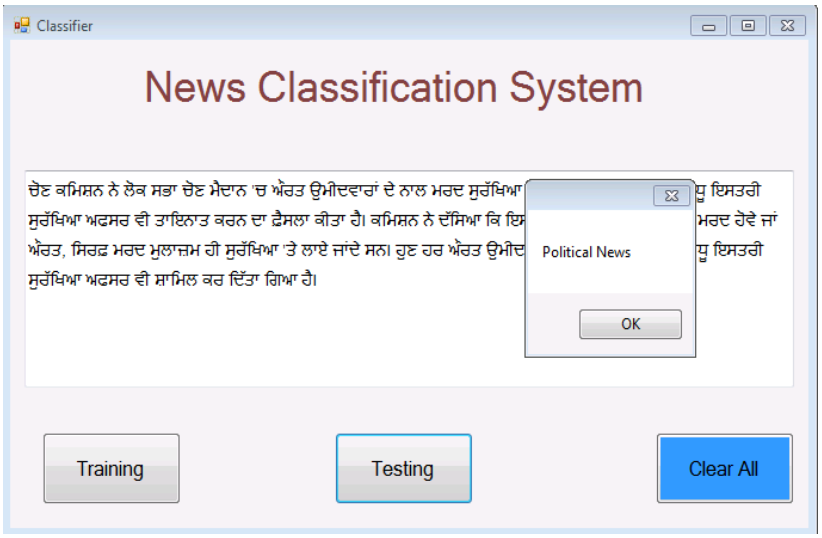

Fig: 2 Text Classification Script Related to Political News I

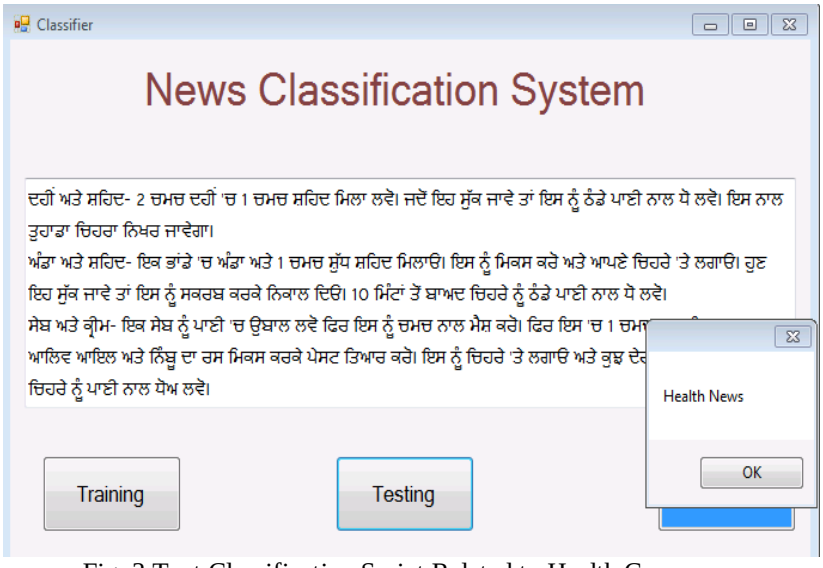

Fig: 3 Text Classification Script Related to Health Concerns

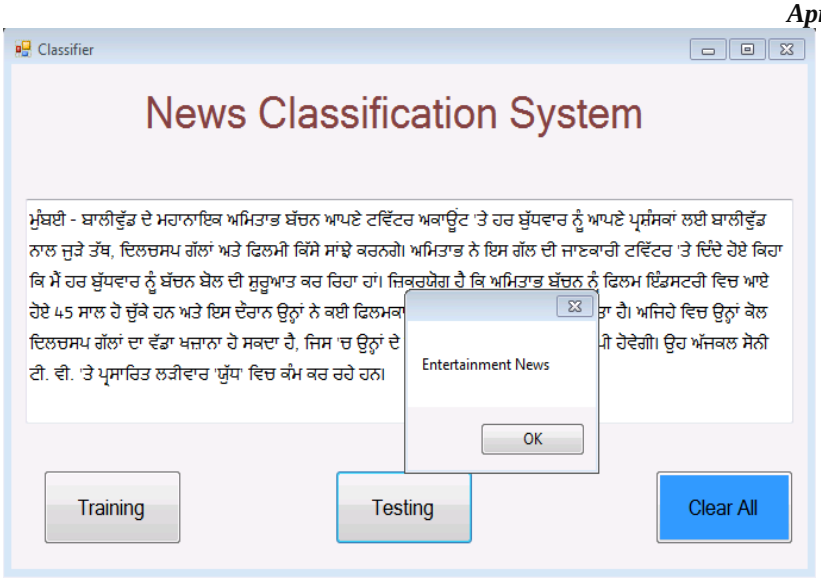

Fig: 4 Text Classification Script Related to Entertainment News

VI. CONCLUSION& FUTURE SCOPE

This proposed work presented an approach to classify online news classification system. Results showed that there are four types of categories has been proposed like politics, entertainment, health,business, in categories where enough good training examples were present the user did not change the automatically preselected category that often.The future scope lies in the use of the hybridization of naïve bay's with another classifier or other technique so that high accuracy can be gained by classifying news according to their subset.

REFERENCES

[1] A. K. Durga, and A. Govardhan, September-2011 "Ontology based text categorization - telugu documents", International Journal of Scientific & Engineering Research, Volume 2 Issue 9, ISSN 2229-5518.
[2] A. McCallum and K. Nigam "A comparison of event models for naive bayes text classification".
[3] A. R. Ali and M. Ijaz, 2009 " Urdu text classification", ACM.
[4] E. Frank1 and R. R. Bouckaert "naive bayes for text classification with unbalanced classes".
[5] V. Gupta and G. S. Lehal (2011), "Punjabi Language Stemmer for nouns and proper name", South and Southeast Asian Natural Language Processing(WSSANLP), IJCNLP, Chiang Mai, Thailand, pp. 35–39.
[6] http://www.scholarpedia.org/article/Text_categorization
[7] http://en.wikipedia.org/wiki/Document_classification
[8] J. D. Brutlag and C. Meek, "Challenges of the email domain for text classification", Microsoft Research, Redmond, WA, 98052 USA.
[9] K Raghuveer and K. N. Murthy "Text categorization in indian languages using machine learning approaches" Department of Computer and Information Sciences,University of Hyderabad,Hyderabad.
[10] Nidhi and V. Gupta, 2012 "Punjabi text classification using Naïve Bayes, Centroid and Hybrid Approach", Sundarapandian et al. (Eds): CoNeCo,WiMo, NLP, , pp. 245–252.
[11] Nidhi and V. Gupta, December-2012 "Domain based classification of punjabi text documents using ontology and hybrid based approach", Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), COLING, pp. 109–122.
[12] Nidhi and V. Gupta,January 2012 "Algorithm for punjabi text classification", International Journal of Computer Applications (0975 – 8887), No.11, Volume 37, pp. 30-35.
[13] Irina Rish. An empirical study of the naive bayes classifier. In IJCAI2001 workshop on empirical methods in artificial intelligence, pages 41–46, 2001.

