Indian Journal of Science and Technology, Vol 10(29), DOI: 10.17485/ijst/2017/v10i29/116525, August 2017
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Urdu Documents Classification using Naïve Bayes

Sohail Abdul Sattar*, Saman Hina, Nimra Khursheed and Anaum Hamid

Department of Computer Science and Software Engineering, NED University of Engineering and Technology, Karachi 75270, Pakistan; sattar@neduet.edu.pk, saman.hina@gmail.com, nimra.khursheed@gmail.com, anaumhamid@gmail.com

Abstract

Objectives: The purpose of this conceptual paper is to highlight the process involved in handling of editorials based on Urdu morphology for better classification purpose. **Methods**: The first step is to collect editorials belongs to different categories, Corpus will be formed by the collected data, preprocessing activities makes corpus more reliable and relevant. Naïve Bayes will be used for classification purpose. Naïve Bayes is identified as best technique for serving as a document classification model, it produces fastest and accurate results as well as very robust to irrelevant features. **Findings:** Handling Urdu morphology is one of the biggest tasks of our research, to handle this problem we need to encode corpus by using utf-8 encoding and thereby changing system locale Urdu easily appear in readable form. **Application:** The main purpose of this approach is to work on classification of Urdu documents, as Urdu is a South Asian Language, which is among the widely spoken in the sub-continent. Urdu document classification involves all the pre-processing activities such as Language processing tasks, labeling and tagging; the tool explored will be R. It will be very helpful for all the firms who manage and manipulate data in Urdu languages e.g. this approach can be implemented on all Urdu news editorials so that all the editorials will be classified to different sub-categories so that user can extract the information he is looking for. This research is still in progress and may very time to time.

Keywords: Classification, Classification using Naïve Bayes, Documents Classification, Naïve Bayes Classification, Urdu Documents Classification

1. Introduction

In past years, a lot of research and work has been done in the field of natural language processing, but almost all focused around English language processing as it is easier to process English language data set than any other languages. Some of the Asian continent languages because of their more morphologically rich structure are becomes more challengeable for researchers and Urdu language is one of them. Urdu Language is very well-known language in Asian region and there are several millions of Urdu speakers around the world. Due to increasing demand and use of information technology massive amount of data generating from World Wide Web continuously, nowadays it becomes in need to extract information from huge data set efficiently and accurately. Scarcity of resources made not enough research has been done in this domain. The idea behind the Urdu Document classification is that it provides an ease to user while reading an Urdu text since document classification is vital and has various applications such as email categorization, spam detection, website classification, ontology mapping etc. The purpose of this document is to address the challenges and progress in Urdu document classification. Because of its complex morphological structure, it becomes difficult to process such a complicated language, to fulfill this challenge some Urdu specific pre-processing steps are required in order to split the text into standard lexicon and then processed through Naïve Bayes algorithm. The remaining part of this paper includes the research methodology. Steps (Section A) that are required before to process data set through Naïve Bayes classifier such as Corpus Collection, Tokenization, Stop words removal and Stemming. Section 1.3 explains how the data set is process through Naïve Bayes algorithm[1]. Section 1.4 describes the sequence of process diagrammatically. Section 3 is based

*Author for correspondence

on Evaluation measures; Section 4 contains challenges in the research domain and the risk factors, Section 6 will summarize the whole document.

1.1 Research Methodologies

To create a model for Document Classification, the beginning of methodology starts with data preprocessing. Section 1.2 provides the details about Urdu language corpora, then Section 1.5 will describe the work on useless data which do not affect the classification result (i.e. removal of stop words, word stemming etc.) and will gives the detailed about Stemming (i.e. to identify most common attributes of each category on which classification based on. For example, the term 'شراب'categorized in weather class has the highest score in term frequency Table, thus we can remove the less important attributes to improve computational time significantly) and Feature Extraction respectively. As we discuss, for classification purpose many classifiers are engaged to generate the model (such as DT, SVM, Naïve Bayes). Generic Classification approach is shown in Figure 1. However, our Independent Study Project is only targeted to Document Classification using Naïve Bayes classifier. After evaluation of model with the help of training set, model will predict the test documents and will identify from which category it belongs too. Finally, the model will have evaluated by a set of testing data. In order to test the classification ability of the model, several evaluation measures (such as precision, recall, and F-measure) will adopted. Furthermore, to interpret whether Naïve Bayes is best to use as the classifier, its testing result will be compared with other classifiers results as well.

1.2 Corpus

First step is to work on corpus. These phases include the collection of different Urdu news editorials and then separate those editorial in desired categories and then make 70% of dataset as training data while the remaining dataset is to be treated as test dataset. The important key point to be considered in this phase is that each category should be approximately the same amount of documents to evaluate the model.

1.3 Datasets

Dataset is the collection of data in any form giving the statistics exactly about related work. In our case, a dataset is a whole document, represented in our proposed system as a single object at the index of array and obviously array depicts our whole corpus. Each one holding a class and available to do macro level classification.

1.4 Preprocessing Steps

To process the test data, we have to clean it and make them able to digestible for our proposed system. For that purpose, we applied some techniques before going towards exact classification methodologies. there is multiple type of preprocessing techniques, like tokenization, lemmatization and stemming, POS tagging, chunking, Stop words removal etc., depending upon the nature of system that

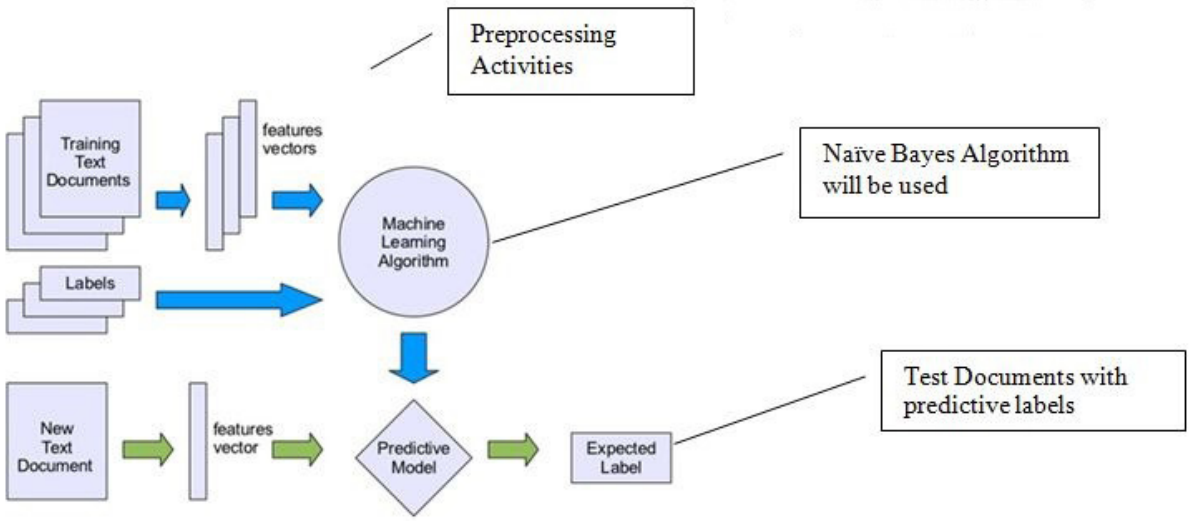

Figure 1. Generic classification approach.

is going to take input. Below are the preprocessing steps required to a document for classification from our classifier model:

1.5 Tokenization and Stop Words Removal

First phase of this step is to tokenize the dataset. Tokenization is the task of chopping up the data into pieces, called tokens. The second phase of this step is to remove stop words. Stop words are easy to find out and the removal of these kinds of words will not affect the classification result, thus this step includes the removal of useless words (such as a, an, able, about, above, according, accordingly, behind and across).

1.6 Stemming

By adopting this step, model will work more efficiently as words will be reduced after applying this activity we will apply stemming process. In this step all relevant words will bind to a root word. For example, 'روشراب' will bind to its root word which is 'شراب'. 'Again by adopting this process, editorials can show a better representation and even the dataset will be reduced which will help us for achieving faster processing time.

1.7 Feature Extraction or Selection of Top Terms

It is one of the most important preprocessing steps in classification task. This technique is use to reduce noise in features. Redundant words and useless words may remove or deleted from the dataset, only key words or top terms will only remain in data which increase the efficiency and speed of Model.

2. Adoption of Naïve Bayes Classifier

Data will be more significantly reduced and more precise for evaluating model after performing the above two steps. For classification purpose "Naïve Bayes multinomial classifier" will be used, because its performance is good in classification of documents as well as it is a simplified technique[2–11]. This technique is probabilistic model which based on the probabilities and probabilities can be obtained by frequency Table[4,5]. Frequency Table can be generating by occurrence of words from single data set. Each category has its own dataset and this data set contains some words which generally lie on the specific categories. When the new data arrives, system classifies this data by matching words from dataset. Formula of Bayes Algorithm used is shown in Figure 2.

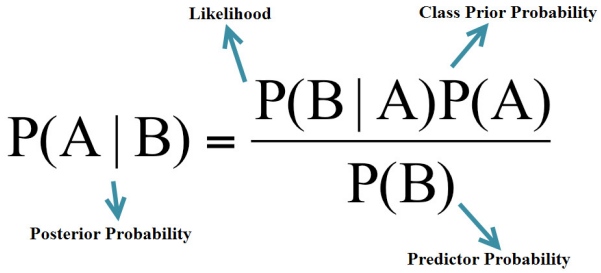

Figure 2. Naive Bayes calculation.

As shown in Figure 3, project flow for classification of Urdu text documents will include the following steps; Documents collection and Corpus creation, apply all necessary pre-processing activities and use Naïve Bayes multinomial classifier to generate a classifier model . Figure 2 is the high level overview of system for better understanding.

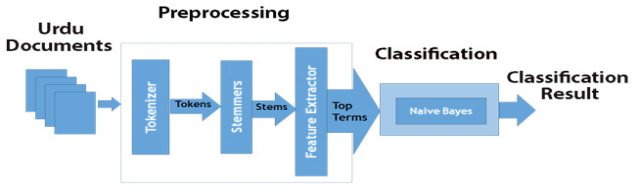

Figure 3. Project flow of classification.

3. Evaluation Measures

To evaluate model, 70% of data is based on test dataset and rest of the 30% will be served as test dataset. After evaluating results, some of the performance measures will be used to calculate system accuracy. Performance measure includes recall, precision, and F-measure.

4. Challenges and Risk Factors

Working on corpora is one of the biggest challenges and manually labeled process made it harder. In this step we separating documents from 500 total editorials into two categories initially picked two categories for baseline results, after that we must be increase the count of cat-

egory or editorials later for improvement of model. In this step following factors are more challengeable:

(i) Creating training set data: (manually) separated documents as per their category, making corpora of 100 documents and Collection of at least 100 documents for baseline results.

(ii) To process Urdu morphology to encode the array vector of Urdu editorials first.

(iii) To change system locale manually to handle encoded documents for better visualization of encoded data of Urdu. By changing System Locale we can easily have visualized Urdu processing.

(iv) To apply all steps of preprocessing activities on Urdu documents.

(v) Feature selection will be done and then Naïve Bayes model will be trained on encoded Urdu documents to evaluate classes of Urdu documents.

5. Discussion

The Ultimate outcome from a classifier is to categorize some piece of text or document to one or more predefined classes or categories. Text or Document could be anything like New Editorial, Emails, tweets, Search queries, Support tickets, Customer feedback, Product reviews etc. Applications of Urdu Text Classification are Categorizing Urdu New articles and new wire content into Topics, Organizing Urdu web Pages into hierarchical Categories, filtering Urdu Spam (either emails or broadcast), Sentiment Analysis, Predicting User intent from search queries, routing support tickets and Capture market rating by analyzing Customer feedbacks.

The main cognition behind the aforementioned proposal is the notion to implement the classification of documentation in Urdu language as research for this specific linguistic branch is still in progress and underdeveloped. Being the national language of Pakistan, a need will arise in the near future to classify the documents which are presented in this language as there will be a vast amount of data and no way to identify whether the data is useful or not. Therefore, we decided to undertake this idea in our Independent Study Project. Our classifier when implemented will make it easier for any editorial written in Urdu to be classified in to their respective branches like economy, sports, health, entertainment, technology etc. This classifier will be of great help to different print media or electronic media for identifying the data in to valuable piece of information in a blink of eye and hence will make many people's jobs far easier.

6. Conclusion

The conducted research reported the classification of Urdu documents by Naive Bayes multinomial algorithm and statistically it is observed that Naive Bayes is generating efficient results by different features selection on R-Studio. No work has been recorded on behalf of Classification of Urdu documents using this approach though it can generate desirable outcomes. In future, we will work to classify Urdu document through different statistical approach like SVM in Supervised Learning and Centroid based as Unsupervised Learning Technique.

7. References

1. Ting SL, Ip WH, Albert HC, Tsang. Is Naïve Bayes a good classifier for document classification? International Journal of Software Engineering and Its Applications. 2011 Jul; 5(3):1–10.
2. Ali AR, Ijaz M. Urdu text classification. ACM, New York, NY, USA; 2009. Crossref
3. Gogoi M, Sarma SK. Document classification of Assamese text using Naïve Bayes approach. International Journal of Computer Trends and Technology. 2015; 30(4):1–5.
4. Vidhya KA, Aghila G. A survey of Naive Bayes machine learning approach in text document classification. International Journal of Computer Science and Information Security. 2010; 7(2):1–6.
5. Ramdass D, Seshasai S. Document classification for newspaper articles, Spring; 2009 May 18. p. 1–12.
6. Korde V, Mahender CN. Text classification and classifiers a survey. International Journal of Artificial Intelligence & Applications. 2012 Mar; 3(2):1–15. Crossref
7. Hussien MI, Olayah F, Al-Dwan M, Shamsan A. Arabic text classification using SMO, Naïve Bayesian, J48 Algorithms International Journal of Recent Research and Applied Studies. 2011 Nov; 9(2):1–11.
8. Sathyadevan S, Athira U, Sarath PR, Anjana V. Improved document classification through enhanced Naive Bayes algorithm. International Conference on Data Science & Engineering; 2014. p. 100–4. Crossref
9. Rakholia RM, Saini JR. Classification of Gujarati documents using Naïve Bayes classifier. Indian Journal of Science and Technology. 2017 Feb; 10(5):1–9. Crossref
10. Zhang Z. Naïve Bayes classification in R. Annals of Translational Medicine. 2016 Jun; 4(12):241. Crossref. PMid:27429967 PMCid:PMC4930525
11. Rajeswari RP, Juliet K, Aradhana. Text classification for student data set using Naive Bayes classifier and KNN classifier. International Journal of Computer Trends and Technology. 2017 Jan; 43(1):1–5.