# A Supervised Leaning Technique for Language Identification

Muhammad Rizwan Rashid Rana, Muhammad Aun Akbar, Tauqeer Ahmad and Sahira Gulfam

***Abstract*—** With the rapid expansion in internet technology and research, people are more vocal about their belongings and accomplishments. It could be due to social media and easy access to those websites where they can conveniently/freely share their views and opinions. People find it easy to converse in their native languages. These websites present flood of data presenting the latest interests of communities around the world. To identify the native language from this bunch of data is very important to transform and use this information. In this paper, we proposed a conceptual model for native language identification (NLID) based on supervised learning and Naïve Bayes classifier. Experiments are conducted on comments taken from different public websites and any document related to any native language. Results show an average accuracy of 97.50 % which is high from other methods. A web application is developed for public use for identifying a language.

***Index Terms*—** Native language identification, Naïve Bayes classifier, Normalization, User reviews

## I. INTRODUCTION

Due to innovation in internet, people especially youth are now more interested in sharing their opinion about anything or more vocal about events and occurrences around. It could be due to social media networking or easy excess to those websites where they can conveniently and spontaneously share their views in their native language. People find it easy to communicate in their native language producing bunch of data, all in different language. There are hundreds and thousands of different languages across the world even across the nation with little or large differences in scripts or literals. Native Language Identification (NLID) is more important in online business and in ecommerce. Where the product vendors are very much interested knowing about their products. Even a politician will be very much interested in knowing about his likeness across the nation. Different algorithm and technologies can be used to analyze this data which is in different languages.

There are millions of comments are presents in social networking sites, blogs, forum, ecommerce sites etc. [1]. Identification of language in these comments is not an easy task in the presence of thousands of languages. Machine learning techniques such as Naıve Bayes and Support Vector Machine are widely used as a classifier because of their ability to "learn" from the training dataset to make decisions with on-line data and to provide real-time analysis with relatively high accuracy [2]. In this research, the aim is to evaluate the scalability of Naïve Bayes classifier (NBC) in language identification system for identifying English, Urdu Arabic, Chinese and Polish languages. Naïve Bayes classifier is implemented here to achieve the better results in language identification system.

The rest of paper is organized as follows. Section II introduces the background study. Section III illustrates the proposed system design for language identification. System includes three major steps Dataset, Preprocessing, and Naïve Bayes classifier. Section IV shows the experiment setup and results. Conclusions are addressed in section V and then Section VI is for references.

Manuscript received August 9, 2016

Muhammad Rizwan Rashid Rana is with the Department of Computer Science, Pir Mehr Ali Shah Arid Agriculture University, Rawalpindi, Pakistan (e-mail: rizwanrana315@ gmail.com).

## II. BACKGROUND STUDY

For native language identification (NLID) variety of algorithms have been tried like Naive Bayes, Support Vector Machine (SVM), Neural Networks, prediction partial matching (PPM) and many with multiple classifier but the best accuracy achieved are still in the lower ninety percent.

Shervin Malmasi and Mark Dras et.al presents discriminative models for differentiate Dari and Persian language at sentence level [3]. Linear Support Vector Machine classifier was used for text classification which results 96% accuracy. 14k per-language sentences are used as training data set. Testing was conducted on cross corpus 79k sentences resulting in 87% accuracy but this data set is out of domain. Şengül BAYRAK HAYTA et.al presented a paper in which they used characters, words and n-gram sequences with different machine learning techniques [4]. They used a sequence of n-gram frequencies. They used five different classification algorithms SVM, Centroid Classifier, Multilayer Perceptron, Fuzzy C-Means and k-Means methods and analyze the frequency of these algorithms on documents those belong to five different languages. They used n-gram feature based method that is used to extract feature vector that is belonging to languages. For experimenting, they used a dataset that is selected from ECI multilingual corpus. After the experiment the accuracy of Centroid Classifier and SVM classifier has provided best accuracy and k-Means has the lowest performance among these five algorithms. Dattesh B Naik, Jeevan R Patil and Pravin P Maske et.al proposed a system for identifying different types of Indian language scripts [5]. By using supervised machine learning proposed system will identify the language of input text. The classifiers ANN or SVM and Random Forest approach were used by the author. The limitation of this system are time consuming while handling large data sets and it is difficult to process neologisms and non-standard words.

Karen Shiells and Peter Pham et.al proposed an approach of unsupervised technique for language identification [6]. Author's uses Twitter as a data source. The technique is based on Chinese Whispers algorithm with some improvement. Microsoft Translator API was used to construct data set for unsupervised learning and evaluation process. Main contribution of the research was to develop or improve a system of identification for short text cluster using unsupervised learning. Shervin Malmasi and Aoife Cahill et.al proposed a novel approach for native identification language [7]. The proposed solution is a function that measure and analyze the features independence of native language. Authors shows that 1-skip bigrams as a useful variant and also be a new native language identification feature.

Priyank Mathur et.al presented a paper in which they use a technique called Stanford Language Identification Engine (SLIDE) [8]. They used different methods for this purpose. The first method they use is Multinomial Naive Bayes model. They use this because it is quick to prototype and provide fast and decent results. For implementation, they choose languages that are very little in common e.g. Portuguese spoken in Portugal and Brazil. They experimented with both word and character n-grams. The experiment shows that the performance of character-level n-gram better. Secondly, they use Logistic Regression and results show that the character-level n-gram again performs better results. Both MNB and LR are not good for the languages that are close to each other and share a lot of words between them, therefore, they used another method for this purpose that's called Recurrent Neural Network. They use 5 different RNN each built using different feature set, namely, from char 2-gram to char 5-gram and last one that is uni-gram. They combine these 5 RNN model called SLIDE. Results show that MNB, LR and SLIDE have the accuracy 0.9452, 0.9449 and 0.9512 respectively.

P. Barlas, D. Hebert, C. Chatelain, S. Adam, and T. Paquet et.al presents an automatic system for identification of language in complex and heterogeneous documents [9]. Proposed system was divided into script identification, writing type identification and language identification. The methods for script discrimination and writing type recognition are based on analyzing the connected components while the language identification requires a recognition engine. Author's implements its proposal on public data set and evaluated it through Google plug-in. Shubham Saini, Bhavesh Kasliwal and Shraey Bhatia et.al proposed a method of G-LDA [10]. This process works on the concepts of Genetic Evolution techniques and Latent Dirichlet Allocation (LDA). G-LDA indicates the words that present in any given document more than one time. Leipzig Corpora was the sub data set used by the author for testing of the technique. JGibbLDA package which is java implementation of LDA using Gibbs sampling techniques was used to generate the document of sentences from five languages. When applied to five training languages (English, Arabic, Italian, Hindi and Gujarati) of the 15 web-pages, gave accurate results. The recognition rate decreases when applied on small data set i-e words less than 1000.

## III. PROPOSED SOLUTION

Different algorithm and technologies can be used to analyze data which is in different languages because people find it easy to converse in their native language producing bunch of data, all in different language. One can identify language of those comments by using different techniques. For language identification we will use a conceptual model which is shown in Fig 1. This model includes dataset, preprocessing, supervised learning algorithm, results. Supervised leaning is the type of machine leaning technique which uses the dataset for training to make predictions [11]. For better predictions we have to classify our data set from raw and unstructured form to a classified form. This classification will do predefine data in to classes based on training data set [12]. There are many algorithms that can be used for NLID native language identification process such as Naive Bayes, Support vector machine etc. we will prefer Naive Bayes over because Naive Bayes algorithm is most simple and efficient algorithm.

Naive Bayes algorithm is a classification technique that is based on Bayesian theorem with an assumption of independence among predictors [13]. In short, a Naive Bayes classifier assumes that the occurrence of a specific feature in a class is distinct to the occurrence of any other feature. To understand naïve Bayes we take an example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'. It is easy and very fast to predict a class and it also performs well during multiclass prediction [14]. It performs well as compared to logistic regression and when you need less training data. It's not sensitive to irrelevant features.

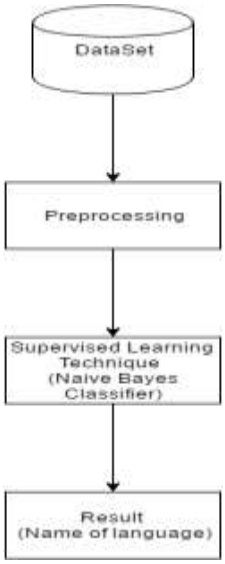


Fig. 1. Proposed Solution

### A. Dataset

We take two types of datasets to calculate the accuracy of language identification system. First one is Abainia dataset[1] and second one is the real-time data from user's comments and views.

Main reason behind real time dataset is to check accuracy on real-time comments uploaded from public users. We develop a responsive web interface for public users to write their text in any of five languages (i.e in English, Urdu, Polish, Arabic and Chinese) and our system judge accuracy on run time using naive Bayes classier. We collect this data from website interface in the form of user's comments and document's. Figure 2 elaborates the interface.


Fig. 2. Interface

### B. Preprocessing

Preprocessing is the most essential step of any technique. The preprocessing of user comments and reviews is a very important part of this research. Using preprocessing we get the data in our required format [15]. The data from user comments and text documents contain noise such as URLs, scripts, HTML tags, and symbols such as asterisks, hashes, etc., which do not have an impact and are not useful for machine learning. These comments also contain one word in many forms, for example in upper and lower case, in a misspelled form, with character repetition. These have to be removed in order to keep only the text so as to improve the performance of the classifier. Steps involves preprocessing are Tokenization, Normalization, Stemming and Filtering

Tokenization is the process in which we spilt the user's comments or text from document into a sequence of tokens. The goal of the tokenization is the exploration of the words in a sentence [16]. Normalization of data includes removing special characters (i.e. @, #, & etc), removing of hash tags from the user comments and text document data [17]. In order to get rid of the multiple forms of a single word we use Stemming. There are many algorithms available for stemming like Porter Stemming Algorithm [18] etc. We also use streaming for removing of repeating characters like happppppiiiieeeeee etc. Filtering is the function that's filters English stop word from a user comments document by removing every token which matches a word from a built-in stop words list. These stops words decrease the efficacy on any machine leaning techniques. Stop words are words that are not critically necessary to the sentence or opinion

### C. Classifier

Naive Bayes classifier is the most accurate classifier which works efficiently in many cases. Naive Bayes has proven to be an effective and much simple supervised leaning method. It is even optimal in some cases. Suppose there are n possible classes $C = \{c1, c2\ldots c^n\}$ for a domain of documents $D = \{d1, d2, \ldots, d^n\}$. Let $W = \{w1, w2,\ldots,w^n\}$ be the set of unique words, each of which appears at least once in one of the documents in D. The probability of a document d being in class c can be computed using Bayes' rule:

$$P(c|x) = P(x/c)P(c)/P(x) \quad (1)$$

In the given formula (1) P(c|x) is the posterior probability of class, P(c) is the prior probability of class, (x|c) is the likelihood which is the probability of predictor given class and P(x) is the prior probability of predictor. Naive Bayes classifier is the better classifier in many real time situations [15]. Before applying Naïve Bayes classifier we first train the classifier on training dataset. Training is the most essential step without this training Naïve Bayes cannot gave us accurate results.

## IV. Results

We tested our proposed solution on Abainia dataset which is also standard dataset used by many researchers in their research for identifying language. It includes more than 500 reviews on scripts written in English, Urdu, Polish, Arabic and Chinese. Experimental results show average accuracy of 98.5 % which is better than any other classifier identifying language.

We also test results on real time environment on our web-application. This is running on internet for public use for identifying languages (English, Urdu, Polish, Arabic and Chinese). We provide a user friendly interface compatible with every device for better experience by users. Peoples can write or just copy paste their comments in any language or upload document of any language in our application and our application successfully identify the language using Naive Bayes Classifier.

[1] https://github.com/xprogramer/DLI32-corpus


Fig. 3. Interface with results

From past few days by providing training data set we get the average accuracy of 97.36%. A bar code graph of everyday use is presented by plotting accuracy on x-axis and day on y-axis. This accuracy graph is shown in Fig 4.

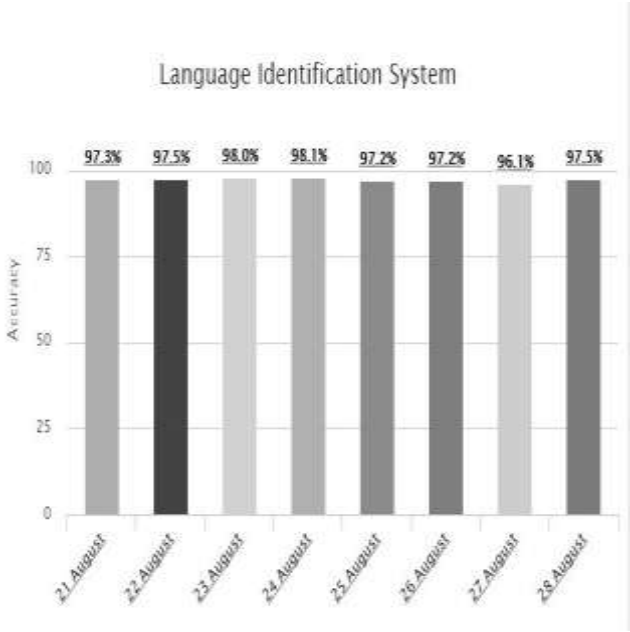


Fig. 4. Accuracy Graph

## CONCLUSION

In this paper, we presented a simple, efficient and complete system for identifying language from user's views and comments datasets using a Naive Bayes classifier. Moreover we run Naïve Bayes Classifier on real time user comments and it also judge language from any document. Our results show that Naïve Bayes Classifier provides efficient results on both datasets. Because of our simplified setup, the average accuracy on Abainia dataset stays 97.50% in all cases and for real time environment users posted comments and files get an accuracy of 97.36%. An intelligent filter in preprocessing might be helpful to increase the accuracy.

We believe that our work is just a beginning of employing machine learning techniques in real time environment of user comments and uploaded documents. Future work will include using our framework by applying on user spoken words or language identification from user's voice.

## REFERENCES

[1] Ashish A. Bhalerao, Sachin N. Deshmukh and Sandip D. Mali, "Predicting Sentiment of User Reviews," International Research Journal of Engineering and Technology (IRJET), vol: 03 issue: 05, May 2016, pp. 985-989.

[2] Bingwei Liu, Erik Blasch, Yu Chen, Dan Shen and Genshe Chen, "Scalable Sentiment Classification for Big DataAnalysis Using Na¨ıve Bayes Classifier," IEEE International Conference on Big Data, 2013.

[3] Shervin Malmasi and Mark Dras et.al, "Automatic Language Identification for Persian and Dari texts", Held at the Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, co-located with COLING 2014.

[4] Şengùl BAYRAK HAYTA, Hidayet TAKÇI, Mùbariz EMİNLİ, Ş.B. HAYTA "Language Identification based on n-gram Feature extraction method by using classifiers", IU-JEEE Vol. 13, 2013.

[5] Dattesh B Naik, Jeevan R Patil and Pravin P Maske et.al, "Language Identification System for Indian Languages" in International Journal for Scientific Research & Development, vol. 4, issue 01, 2016.

[6] Karen Shiells and Peter Pham et.al, "Unsupervised Clustering for Language Ide+ntication" , December , 2010

[7] Shervin Malmasi and Aoife Cahill et.al, "Measuring Feature Diversity in Native Language Identification" , Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 49–55,2015.
https://doi.org/10.3115/v1/W15-0606

[8] Priyank Mathur ,Arkajyoti Misra ,Emrah Budur, "Language Identification from Text Documents", 2015

[9] P. Barlas, D. Hebert, C. Chatelain, S. Adam, and T. Paquet et.al, "Language Identification in Document Images",Journal of Imaging Science and Technology, 2016.
https://doi.org/10.2352/J.ImagingSci.Technol.2016.60.1.010407

[10] Shubham Saini, Bhavesh Kasliwal and Shraey Bhatia et.al, "LANGUAGE IDENTIFICATION USING G-LDA", International Journal of Research in Engineering and Technology, 2013.

[11] Junwei Han,Dingwen Zhang,Gong Cheng,Lei Guo and Jinchang Ren, "IEEE Transactions on Geoscience and Remote Sensing", vol 53, issue 6, 2015.

[12] Neha Khan Mohd Shahid Husain and Mohd Rizwan Beg, "Big Data Classification using Evolutionary Techniques: A Survey", IEEE International Conference on Engineering and Technology, 2015

[13] Sharvil Shah, K Kumar and Ra. K. Saravanaguru,"Sentimental Analysis of Twitter Data Using Classifier Algorithms", International Journal of Electrical and Computer Engineering (IJECE) Vol. 6, 2016, pp. 357-366

[14] Sharvil Shah, K Kumar, Ra. K. Saravanaguru, " Sentimental Analysis of Twitter Data Using Classifier Algorithms", International Journal of Electrical and Computer Engineering (IJECE), February 2016,

[15] Banage T. G. S. Kumara, Incheon Paik, Jia Zhang, T. H. A. S. Siriweera and Koswatte R. C. Koswatte, "Ontology-Based Workflow Generation for Intelligent Big Data Analytics", IEEE International Conference on Web Services, 2015

[16] Tanu Verma, Renu, Deepti Gaur, "Tokenization and Filtering Process in RapidMiner ,International Journal of Applied Information Systems (IJAIS), April 2014

[17] Viktor Hangya, Richard Farkas, "Target-Oriented Opinion Mining from Tweets", 4th IEEE International Conference on Cognitive Infocommunications , December, 2013

https://doi.org/10.1109/coginfocom.2013.6719251

BIOGRAPHIES

**Muhammad Rizwan Rashid Rana** was born in Bahawalnagar, Punjab, Pakistan. He completed his early education from Army Public School and College Rawalpindi. He received his master degree in computer science (MCS) from Pir Mehr Ali Shah Arid Agriculture University, Rawalpindi. Currently he is studying in MS degree in Computer Science (MSCS) from PMAS-Arid Agriculture University, Rawalpindi. His specialization is in Artificial Intelligence (AI).

His research interests are expert systems, sentiment classification, opinion mining, big data classification, image retrieval systems, natural language processing and neural networks.

**Muhammad Aun Akbar** was born in Chakwal, Punjab, Pakistan. He completed his early education from Punjab Group of College. He completed his master degree in Information Technology (MIT) from Pir Mehr Ali Shah Arid Agriculture University, Rawalpindi. Currently he is studying in MS degree in Computer science from PMAS-Arid Agriculture University, Rawalpindi. His specialization in Artificial Intelligence (AI).

His research areas are social network data classification, statistical learning methods, expert systems, speech recognition and image retrieval systems.

**Tauqeer Ahmad** was born in Kot Adu, Punjab, Pakistan. He completed his early education from Punjab Higher Secondary School Kot Adu. He completed his graduation degree in Computer Science (BSCS) from The Islamia University of Bahawalpur. Currently he is studying in MS degree in Computer science from PMAS-Arid Agriculture University, Rawalpindi. He is doing specialization in Artificial Intelligence (AI).

His research areas are feature extraction, memory-based learning, speech recognition, expert systems and image retrieval systems.

**Sahira Gulfam** was born in Attock, Punjab, Pakistan. She received her Matric degree in 2007. In 2009 she received her Intermediate degree in pre-medical domain. Later she joined Arid Agriculture University Rawalpindi Pakistan for Bachelor Honors in Computer Science from 2009 to 2013. Now she is doing her Master Science of Computer Science in Artificial intelligence from Arid Agriculture University Rawalpindi Pakistan.

Her current research interests are social media networks, data mining, big data, native language identification, taxonomy generation and advanced topics in artificial intelligence