Bharathi_SSN@INLI-FIRE-2017:SVM based approach for Indian Native Language identification

B. Bharathi, M. Anirudh, J. Bhuvana
SSN College of Engineering
Chennai, Tamil Nadu
bharathib@ssn.edu.in,anirudh15058@cse.ssn.edu.in,bhuvanaj@ssn.edu.in

ABSTRACT

Native Language Identification (NLI) is the task of identifying the native language of a writer or a speaker by analyzing their text. NLI can be important for a number of applications. In forensic linguistics, native language is often used as an important feature for authorship profiling and identification. Nowadays due to the huge usage of social media sites and online interactions, receiving a violent threat is a common issue faced by commuters. If a comment or post poses any type of threat, then identifying the native language of the person will be one of the significant measures in finding the source. In this paper, we present our methodology for the task of identifying the native language of an Indian writer. We have extracted TF-IDF feature vectors from the given document and used SVM classifier to identify the native language of the document given by shared task on Indian Native Language Identification@FIRE2017. The performance is measured in terms of accuracy and we have obtained overall accuracy of 43.60%.

KEYWORDS

Indian Native Language Identification, Classification, Support Vector Machine, TF-IDF

1 INTRODUCTION

Native Language Identification (NLI), is the well-known task that focuses on identify the native language of the non-native speakers. In India, English is the most important language and has a status of the associated language. After Hindi, it is the most commonly spoken language in India and certainly the most read and written language. The number of second language speakers of English has constantly been on the increase and this has also contributed to its rich variation. English is blended with most of the Indian languages and is used as a second language or the third language frequently. Regional and educational differentiation, distinguish the language usage and shows the stylistic variations in English. Spoken English shows great variation across the states of India and it is relatively easy to identify the native speaker using their English accent. But finding the native language of the user based on the comments or posts written in English is a challenging task in the current scenario. NLI has been invariantly used in various applications and domains. In [2], experiments on language identification of web documents, focusing on which combination of tokenisation strategy and classification model achieves the best overall performance. Native Language identification for the NLI Shared Task 2013 using features based on n-grams of characters, words, Penn TreeBank and Universal Parts of Speech tagsets, and perplexity values of character of n-grams to build four different models are presented in[3]. In [3], the above mentioned four models are combined to create ensemble approach and achieved an accuracy of 75%. In [5], for NLI used a Maximum Entropy classifier, with the features such as character and chunk n-grams, spelling and grammatical mistakes, and lexical preferences. In [1], normalized lexical, syntactic and dependency features with SVM classifier has been used to identify the native language for NLI shared task 2013. For NLI task, the features used in [4] are n-grams of words, parts-of-speech as well as lemmas. In addition to normalizing each text to unit length, the authors also applied a log-entropy weighting schema to the normalized values, which gives the accuracy of 83.6%. An L2-regularized SVM classifier was used to create a single-model system in [4] .

Many of the research works on NLI system used lexical, syntatic features with different classifiers for the document specific to particular domain written by different native speakers. In this work, we haved experimented the shared task of INLI@FIRE2017 which aims to identify the native language of an Indian user based on their comments in social media [6]. The text used in the shared task is not specific to any particular domain. The training documents given by INLI@FIRE2017 is taken from social media. Our focus is to identify the native language using machine learning approach with Term Frequency-Inverse Document Frequency (TF-IDF) feature vector.

2 PROPOSED APPROACH

We have implemented supervised machine learning approach for this INLI task. The steps for the proposed approach are as follows:

- Data preparation
- Extract TF-IDF features from the given text document
- Train the SVM classifer using the features extracted from the training text corpus
- Predict class label for the instance as any of the six languages namely Tamil, Hindi, Kannada, Malayalam, Bengali or Telugu using the trained SVM model

The steps involved in the experimented approach is depicted in the Fig.1.

2.1 Data Preparation

The data used for our research are Facebook comments which is present in the form of embedded XML files. Hence, the data from these XML files are to be extracted and the special symbols, punctuation symbols are removed before they can be fed into mathematical model for training. The XML files contained various tags out of which only comment tag was of interest. Libraries such as *minidom* from XML.dom package was used to parse through the XML files for extracting the text within the comment tags. The comments which were encoded using UTF-8 encoding scheme were decoded

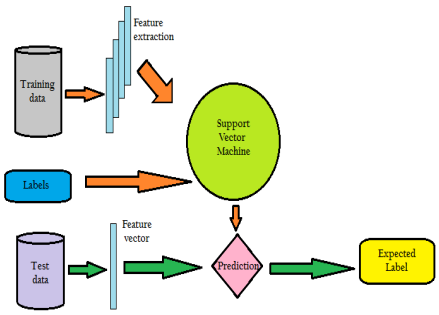

Figure 1: Experimented system architecture

and converted into a python lists with their native language. It was found that the number of Hindi, Bengali, Kannda, Telugu, Tamil and Malayalam comments used for training the model are 211, 202, 203, 210, 207 and 200 respectively.

2.2 Feature Extraction

The data used for training the model are essentially Facebook comments written by non-native speakers of English language. By virtue of which the grammar and diction are not considered to be above par, which makes it unfit for applying commonly on native language identification algorithms such as Prediction by Partial Matching (PPM) algorithm, word-length algorithm, syntactic structure, error analysis algorithm and phonetic algorithm.

This model exploits the fact that an author's native language will dispose them towards particular language production patterns in their second language. This theory can also be extended to the errors made by the authors native to a particular language which clearly confides that if bag of words feature is used to extract the proper English words will further lessen probability of qualifying the desired features to predict the native language. Hence, the data is taken as such for training, keeping the writing error and diction patterns of the different author groups intact. The feature extraction is done using the tool TF-IDF vectorizer method from the *scikit* learn library which yields the highest accuracy. This extraction tool first analyses the common words in a document and also counts the words as well. Then the data is transformed using "$TF - IDF_{vectorizer}$" method before training the model.

2.3 Support Vector Machine

The Support Vector Machine (SVM) algorithm is used here for classification as it is well suited for text classification with colossal data and features. SVM performs multi-class classification through one-against-one on the six classes. The Radial Basis Function (RBF) kernel is used in training which fits the patterns produced by the authors in different groups better than poly and linear kernels. To classify the training examples correctly, we set the "C" parameter in the SVM to 10,00,000 and gamma value as 0.1, which gives the freedom to the model built for selecting more samples as support vector. We achieved cross validation accuracy of 84.61%.

Table 1: Performance analysis of INLI task

Class	Precision(in %)	Recall(in %)	F1-measure(in %)
BE	50.30	80.50	62.00
HI	51.90	5.60	10.10
KA	33.30	64.90	44.00
MA	36.30	60.90	45.50
TA	48.60	51.00	49.80
TE	40.40	28.40	33.30

2.4 Language Identification

The feature vectors for the test documents are derived similar to training data using TF-IDF features. The trained multiclass SVM was used to predict the language for the test documents. Each test document was predicted as one of the six languages namely Tamil, Hindi, Kannada, Malayalam, Bengali or Telugu.

3 PERFORMANCE ANALYSIS

Our approach for Indian native language identification has been evaluated based on the metrics namely precision, recall and F1 measure for each language with an overall accuracy. The results reported for our approach are given in Table 1.

We have obtained an overall accuracy of 43.60% using multiclass SVM based approach for Indian native language identification task.

4 CONCLUSIONS

We have presented an approach to identify the native language of the Indian speaker from the text posted in the social media. In the experimented methodology, TD-IDF features were extracted from the text documents. Then a multiclass Support Vector Machine is trained using the extracted feature vectors. The experimented system is evaluated using the test instances given by INLI@FIRE2017 shared task organizers for the six languages. We have obtained an overall accuracy of 43.60% using our experimented multiclass SVM based approach. The system could further be improved by removing or replacing the lexically incorrect terms such as plz, buzz, Y(why), r(are) into lexically correct terms in order to enhance the accuracy.

REFERENCES

[1] Amjad Abu-Jbara, Rahul Jha, Eric Morley, and Dragomir Radev. 2013. Experimental Results on the Native Language Identification Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Atlanta, Georgia, 82–88. http://www.aclweb.org/anthology/W13-1710

[2] Timothy Baldwin and Marco Lui. 2010. Language Identification: The Long and the Short of the Matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 229–237. http://dl.acm.org/citation.cfm?id=1857999.1858026

[3] Binod Gyawali, Gabriela Ramirez, and Thamar Solorio. 2013. Native Language Identification: a Simple n-gram Based Approach. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*.

[4] Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing Classification Accuracy in Native Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for

Computational Linguistics, Atlanta, Georgia, 111–118. http://www.aclweb.org/anthology/W13-1714

[5] Thomas Lavergne, Gabriel Illouz, Aurélien Max, and Ryo Nagata. 2013. LIMSI's participation to the 2013 shared task on Native Language Identification.. In *BEA@ NAACL-HLT*. 260–265.

[6] Anand Kumar M, Barathi Ganesh HB, Shivkaran S, Soman K P, and Paolo Rosso. 2017. Overview of the INLI PAN at FIRE-2017 Track on Indian Native Language Identification. In *Notebook Papers of FIRE 2017,FIRE-2017, Bangalore, India, December 8-10*. CEUR Workshop Proceedings.