International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2019): 7.583

Comparison of Various Models in the Context of Language Identification (Indo Aryan Languages)

Salman Alam

School of Language Science, EFL University, Hyderabad, India
salman.alm[at]rediffmail.com

Abstract: ***Automatic language detection is a text classification task in which language is identified in a given multilingual text by the machine. This paper compares the different models of machine learning algorithm in the context of language identification. The corpus includes five major Indo-Aryan Language which are closely related to each other like Hindi, Bhojpuri, Awadhi, Maghahi and Braj. In this paper I have compared models like Random forest classifier, SVC, SGD Classifier, Multi-nominal logistic Regression, Gaussian Naïve Bayes and Bernoulli Naïve Bayes. Out of these models Multi-nominal Naïve Bayes has attained the best accuracy of 74%.***

Keywords: Hindi, Magahi, Bhojpuri, Braj, Awadhi, SVC, Multinominal NB, RNN, Linear SVC, SGD Classifier.

1. Introduction

Indo-Aryan language family is a major language family in Indian sub-continent. It has more than 900 million speakers. It includes languages like Hindi, Bengali, Urdu, Bhojpuri, Awadhi, Magahi, Maithili, Braj etc. Hindi and Urdu together called 'Hindustani'. Hindi has the highest number of speaker, Bengali is on second position. Indo-Aryan is a branch of Indo-Iranian, which itself is a branch of Indo-European family.

Automatic detection of language(s) which is present in a document based on the text of the document is Language Identification task. Language identification techniques generally presuppose that every document is written in one of a closed set of known languages for which there is training data, and is thus formulated as the task of predicting the nearest likely language from the set of training languages. In this era most of us are multilingual so we can speak more than one language and we can also identify the language we know in a given multilingual text for example in the text below many of us can recognize at least one, two or all languages.

Table1: Example sentences

Language	Text
Hindi	पीएम की इस अपील पर पूरा देश आज पूरा देश रात 9 बजे नौ मिनट के लिए लाइट बंद करके दीया, मोमबत्ती या टॉर्च जलाएगा.
Bhojpuri	वर्तमान पीढ़ी कुछ ज्यादा व्यसत बिया बाकिर पहिले नीयन कुछ खास नइखे लउकत
English	Is corona virus good for our mother Earth?
Magahi	इ आदमी हर चीजवा के पैसा दे देथुन।

But this task is quite challenging for computers and it has been one of the favorite topics of research for more than fifty years for researchers. Researchers in this area basically try to make machines/models to imitate this ability of human and incorporate the same in computer. Many efficient models for this task has been developed in the past for various languages but when it comes to Indian language, the works are very limited In this task, the problem of language identification in documents that contain text from more than one language from the candidate set is addressed. For this work I modeled various models which detect the languages being used in a text document.

Most of the NLP tasks rely on the monolingual dataset whether it is information retrieval, Machine Translation, POS Tagger or Parsing. Due the immense rise of social media nowadays, we don't have proper structured data set. People use Multilanguage for better communication or other reason. If we feed this data set without knowing the language being used in the data the system will fail or will produce relatively poor prediction for the specific tasks. Automatic language detection of multilingual text can be used as preprocessing for NLP tasks which can improve the quality of tasks.

This task targets to identify 5 closely-related languages of the Indo-Aryan language family – Hindi, Braj Bhasha, Awadhi, Bhojpuri, and Magahi. Starting from Hindi and Braj Bhasha spoken in Western Uttar Pradesh to Awadhi and Bhojpuri soken in Eastern Uttar Pradesh and Bhojpuri and Magahi spoken in the neighboring Eastern state of Bihar, all these languages form a part of a continuum. For this task, I have used the corpus of 9000 sentences, mainly from the domain of literature, published over the web as well as in print.

2. Previous Work

Researchers have employed multiple types of algorithms like Naïve Bayes, SVM, N-gram, PPM etc. to detect the language in multilingual text document.

Marcos Zampiri (2013), used bag of word model for the task and compares three different classifiers with it, namely Multinominal Naïve Bayes, Support Vector Machine and J48. In the work, Unigram method performed better because in two out of three cases it attained better accuracy.

In another work Tommi Jauhiainen et.al provide an extensive literature survey of Language Identification over a period of time and discusses various problems, methods, tools and techniques for the task. The work also enlightens the importance of Language Identification for natural language processing.

Marco Lui et.al 2014, presented a model of Language Identification which uses Generative Mixture Model that is influenced by Supervise Topic Modeling Algorithm. Authors experimented the model with synthetic as well as real life data scraped for Web. The model was able to predict the quantity of language text available in the document, accurately.

Priya Rani Et.al 2018 developed an Automatic Language Identification tool for two closely related languages Hindi and Magahi. For this task authors have used traditional method, Rule based method. The model comprises extensive linguistics rules for both languages. It has scored the accuracy of 86.34%.

Indhuja Et.al proposed a language Identification model for five languages which follows Devanagari script i.e. Hindi, Bhojpuri, Marathi, Nepali and Sanskrit. The proposed system is based on supervised machine learning which uses n-gram including word n-gram and character n-gram as its feature. The paper shows that the system works better with the unigram feature because most of the words are unique to the language.

Most of the work in the literature is based on N-gram. In this work I am also using n-gram as feature and experimenting with different algorithm of machine learning. The aim of this paper is to see the classification efficiency of various algorithms in the context of closely related Indian languages using Bag of word model.

3. Dataset

The data set was provided by "Shared task on language identification", VarDial 2018. The original dataset contains more than 67000 sentences. I have taken a part of the dataset for the task of comparison. The truncated dataset is of around 9000 sentences as we can see in the diagram given below. Dataset contains sentences and their respective Language tag (MAG for Magahi, AWA for Awadhi, BHO for Bhojpuri, BRA for Braj and HIN for Hindi.). Sentence and language tag were separated by tabs provided in a text file. It was well structured data. The dataset also consisted of a subset (dev.txt) of the overall training data which we utilized for hyper-parameter tuning. Each entry given in the dataset is a full sentence. The sentences are extracted from journalistic corpora and written in one of the mentioned languages. Then the sentences are tagged with the language group accordingly. A similar set of mixed language instance was provided as well for adding noise to the data. A separate gold test data was provided for the final evaluation (test.txt). The numbers of sentences per language provided in the dataset are not same for every language.

In the Fig 1, in form of chart we can see the number of sentences as per languages in new dataset. It has one thousand sentences for Awadhi language, around sixteen hundred sentences for Bhojpuri Language, fifteen hundred sentences, for Hindi language it has sixteen hundred sentences and for Magahi language it has around sixteen hundred sentences.

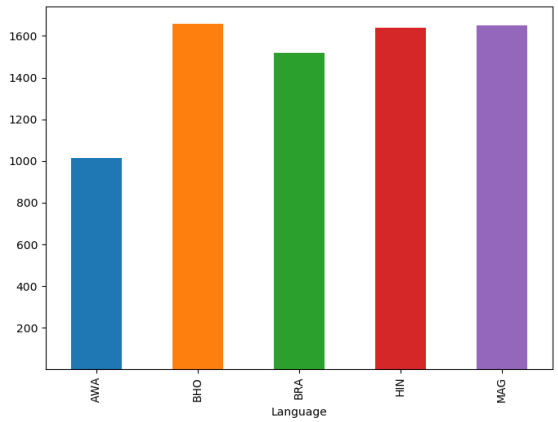

Figure 1: Density of sentences per language

4. Data Preprocessing

The corpus was well structured but still there were some sentences which were not appropriate for the task. Those sentences were removed manually. With this dataset I have created another dataset of unigram, bigram and trigram of the common terms for every single language. Since the original dataset was huge it would have taken a lot of time and algorithms would have to undergo multiple unnecessary mathematic calculations with the data to predict the language. The data which I had created was smaller compared to the original one but it had the appropriate items required for the algorithms in the prediction task.

5. Feature Selection

Each document is converted into a vector where each entry counts the number of times a particular n-gram is present in the document. This is analogous to a bag-of-words model, where the vocabulary of "words" is a set of n-gram sequences that has been selected to distinguish between languages. The exact set of features is selected from the training data using Information Gain (IG), an information-

theoretic metric developed as a splitting criterion for decision trees (Quinlan, 1993). IG- based feature selection combined with a naive Bayes classifier has been approved to be predominantly effective for task of language identification.

An example of feature selection n-gram is given below in form of chart.

Table 2: Feature selection N-gram 1

Language	Feature in unigram	Feature in bigram
AWADHI	उलट्रहण्,ययन	मनरह,नहकह,अध्ययन
MAGAHI	पस्,बहल्,ठल	जबतक्,बहपर्, एक्अपन
BHOJPURI	बदलत्,अह्,हड	हमआप्,क्अपन्न्न्, पतअपन
BRAJ	आ,क्रग्, तकन	वतकह्,गलकर्, औरलन
HINDI	भरप्,लहर्, सरल	हमआप्,क्करकर्, करअन

6. Algorithms

Here in this task I have chosen seven different classifiers which are Random forest classifier, Linear SVC, Multinomial Naïve Bayes, Logistic Regression, Gaussian Naïve Bayes, Bernoulli Naïve Bayes and SGD Classifier. Brief introductions of classifiers are mentioned below:

a) Random forest classifier: Random forest classifier is based on decision trees. The classifier generates multiple decision trees which work as an ensemble. Every tree in the classifier gives a class prediction and the most common class wins out. RFC is one of the most rigorous learning algorithm. The advantage of this algorithm is that it can run on huge data smoothly, it has an effective method of estimating the missed out data which helps in predicting better results. Disadvantage of this algorithm is that it works badly with the dataset which has categorical variables because it is biased to the attributes which has more levels.

b) Linear SVC: Linear SVC creates a hyper-lane based on the dataset which divides the different classes for classification. The advantage of this algorithm is, it is faster than non-linear classifier. Once the model is trained we don't need the training data for predictions.

c) SGD Classifier: SGD classifier is an implementation of regularized linear models with Stochastic Gradient Descent. It deals very well with the large corpus and uses less memory. It requires multiple hyper-parameters e.g. regularization parameter. It is also biased to feature scaling which is one of the drawbacks of the algorithm.

d) Multinomial Naïve Bayes: Multinomial Naïve Bayes comes under the family of Naïve Bayes classifier. It applies Bayes theorem with the strong assumption that each factor of the data is autonomous to predict the class. This algorithm is good for discrete dataset and multinomial (multi-label) classification task.

e) Logistic Regression: Logistic regression is a kind of linear regression which uses complex cost function or sigmoid function. It is a supervised machine learning algorithm, which is commonly used for classification task based on probability. It is less inclined to over fitting, easy to interpret, implement and efficient. This algorithm is good for predicting discrete function.

f) Gaussian Naïve Bayes: This Naïve Bayes is based on Gaussian/normal distribution which reduces the sum of squared error. This algorithm is good for continuous data.

g) Bernoulli Naïve Bayes: This is another variant of Naïve Bayes, this algorithm is based on Bernoulli distribution and good for data which has feature of binary.

7. Result

In the figure 2 below We can see the accuracies of the model. For random forest classifier I have got an accuracy of 56%, for linear SVC accuracy is around 73%, for Multinomial Naïve Bayes accuracy is around 74%, for logistics regression accuracy is around 72.5%, for Gaussian Naïve Bayes accuracy is 58%, for Bernoulli Naïve Bayes accuracy is around 72%, and finally for SGD Classifier accuracy is 73%.

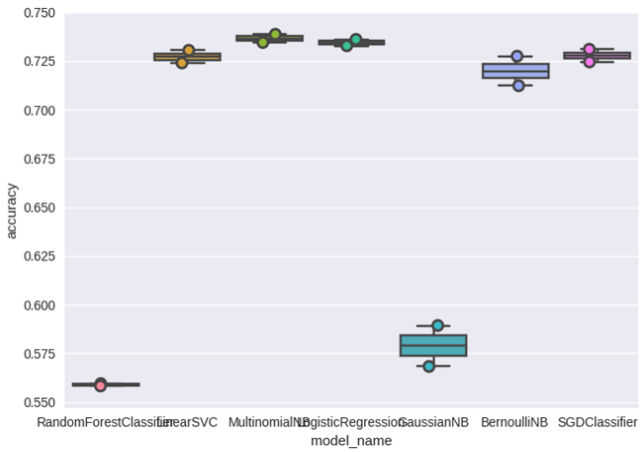

Figure 2: Model-wise result

For this task I have used multiple machine learning algorithms as I have mentioned above. I also used RNN (Recurrent Neural Network) which gave me accuracy of 73%. From the results above we can easily conclude that Multinomial Naïve Bayes gives better result compare to other algorithm, this is because the algorithm is designed

for classification of multi label problems and my dataset contains five different labels of Indo-Aryan Language. Random Forest classifier performs badly with this data set, this is because we don't have same no. of label for every language and it gets biased to the label which is more in number.

8. Conclusion

Language identification for multilingual text in a text document is a multi-label classification problem, in which a document can be mapped onto any number of labels from a closed set. For this task I have used various models and techniques of NLP and the highest accuracy that I have achieved is 74%.

Acknowledgement

I am very thankful for VarDial 2018 and the organizers of VarDial 2018 to allow me to use the dataset for this work.

Reference

[1] Bi Y., Bell D., Wang H., Guo G., Greer K(2004) Combining Multiple Classifiers Using Dempster's Rule of Combination for Text Categorization, MDAI, 127-138.
[2] Brank J., Grobelnik M., Milic-Frayling N., Mladenic D. (2002) Interaction of Feature Selection Methods and Linear Classification Models, Proc. of the 19th International Conference on Machine Learning, Australia, 2002.
[3] Jurafsky, D. and Martin, J. H. (2002) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Preason Education Series
[4] Kumar, Ritesh & Lahiri, Bornini & Alok, Deepak & Ojha, Atul & Jain, Mayank & Basit, Abdul & Dawer, Yogesh. (2018). Automatic Identification of Closely-related Indian Languages: Resources and Experiments. WILDRE – 4.
[5] Marco Lui, Jey Han Lau and Timothy Baldwin (2014). Automatic Detection and Language Identification of Multilingual Documents, Transactions of the Association for Computational Linguistics Volume 2, 2014
[6] Marcos Zampieri (2013) Using Bag-of-words to Distinguish Similar Languages: How Efficient are They?, IEEE 14th International Symposium on Computational Intelligence and Informatics (CINTI)
[7] Rani, Priya & Ojha, Atul & Jha, Girish. (2018). Automatic Language Identification System for Hindi and Magahi.
[8] Sousa P., Pimentao J. P., Santos B. R. and Moura-Pires F (2003) Feature Selection Algorithms to Improve Documents Classification Performance LNAI 2663, 2003, pp. 288-296
[9] Tommi Jauhianen, Marcos Zampieri, Timothy Baldwin and Krister Linden, (2018), Automatic Language Identification in texts: A Survey, Journal of Artificial Intelligence Research · April 2018.
[10] William B. Cavnar and John M. Trenkle. (1994) N-gram-based text categorization. In Proceedings of the Third Symposium on Document Analysis and Information Retrieval, pages 161–175, Las Vegas, USA.