# Language Identification Using Combination of Machine Learning Algorithms and Vectorization Techniques

Anushri Bhansali[1]
Department of Computer Science and Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Anand, India.
[1]anushribhansali@gmail.com

Amit Chandravadiya[2]
Department of Computer Science and Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Anand, India.
[2]amittckn7@gmail.com

Brijeshkumar Y. Panchal[3]
Department of Computer Science and Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Anand, India.
[3]panchalbrijesh02@gmail.com

Mohammed Husain Bohara[4]
Department of Computer Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Anand, India.
[4]mohammedbohara.ce@charusat.ac.in

Amit Ganatra[5]
Department of Computer Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Anand, India.
[5]amitganatra.ce@charusat.ac.in

*Abstract*— **Language Identification refers to the process of ascertaining and discerning the language found in a particular text or document. In this work, approaches for language identification, using Machine Learning Algorithms and Vectorization methods have been compared and contrasted. Three machine learning algorithms, along with two vectorization techniques have been used. The ML Algorithms used are Naïve bayes, Logistic Regression, and SVM (Support Vector Machine), and the vectorization techniques used are Term Frequency-Inverse Document Frequency (TF-IDF), and Count Vectorizer (Bag of Words (BoW)). This research put forwards the contrast and comparison of the above-mentioned classification algorithms and vectorization methods. It is also a web development-based work**.

*Index Terms*-- **Bag of Words, Classification, Language Identification, TF IDF.**

## I. Introduction

There are numerous cases and circumstances, where in the source language of a particular content is not known. Identifying the language becomes crucial prior to performing machine translation or even other information retrieval tasks. Identifying the languages automatically using computational methods can turn out to be of great use in such instances. This automatic way of recognizing the language present in the

content is known as Language Identification. The unceasing escalation of data has led the Natural Language Processing

(NLP) domain to gain a universal desirability. The objective of this paper is to use these NLP vectorization techniques such as Bag of Words and TFIDF along with different classification models: Logistic Regression, Naïve Bayes, and SVM, and compare and contrast each combination of approach**.**

## II. Related Work

In June 2013, Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes' study focused on the use of TF-IDF weighting schemes. The use of TF- IDF weighing schemes on features was a unique facet of the proposed system. The approach of recognizing the native language was taken as a kind of text classification.

In In this kind of approach, choices and decisions were made at three levels, which in- volved using of the training and development data at first level, feature selection at second level, selecting the appropriate machine learning algorithms and parameters at third level. On first level, the TOEFL11 dataset was used. On the second level, the unigrams, along with the bigrams of words were looked into separately and in combi- nations. Character n grams were also explored and the TFIDF method was used. For the last

stage, 3 (linear) classifiers were used: linear support vector machines, logistic regression and perceptron. As a result, the best accuracy was given by the TFIDF feature combination of unigrams and bigrams of words, followed

2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) | 978-1-6654-3789-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACITE53722.2022.9823628

by TF- IDF of 6-gram characters. It resulted in 84.09% of accuracy. Both, the linear SVM and logistic regression classifiers performed almost tantamount. SVM gave an accuracy of 84.55% and logistic regression gave an accuracy of 84.45%. Finally, the authors also urged to look into various other features which can help to differentiate typologically or geographically related languages [1].

In January, 2015, the study by Marcos Zampieri studied and experimented about language identification. ML classifiers as well as BoW focusing on different language varieties were used. Bag of Words is a simple technique of representing data, such that the words are represented in form of a word vector. This word vector comprises of the frequency of occurrences of words present in the content. Multinomial Naive Bayes (MNB), Support Vector Machines (SVM) and J48 were the three ML algorithms used for the study. The domain of Natural Language Processing and text classification sees an extensive use of these algorithms. There is a significant difference in the way classification is performed by these methods. 1,000 documents were used in form of da- taset. As a result, the method outperformed the word unigram method in 2 out of 3 cases. Also, the outcomes extended from 0.988(for Portuguese), using MNB and 0.865(for Spanish), using the J48 classifier. The MNB classifier resulted in the best average per- formance. It yielded 0.968 of accuracy [2].

In 2018, Ermelinda Oro, Massimo Ruffolo and Mostafa Sheikhalishahi presented a paper, whose objective was to present a LID (Language Identification) model. The main aim was to differentiate and discriminate between similar languages accurately when applied to the written text. The proposed system used the monolingual training datasets: Wikipedia Dataset,DSL 2015 Dataset, TweetID Dataset, and VOA Dataset. The newly proposed method made use of the word vector representation (Word2vec) as well as the classification-based Long Short-Term Memory recurrent neural network (LSTM RNN). First, the text corpus of each language was collected. After the pre-processing, the text was fed to Word2vec, that gave a list of vectors which were related to each word in input text as output. Further, a lookup table which matched each word of the dataset with its related vector was obtained. While in the training phase, the classifier corresponding to LSTM RNN took the input vectors of the dataset. After the training, the test phase was performed, where the input was taken from test phase. Finally, the accuracy and precision of the built model were calculated. This model, when compared with prior works, outperformed and obtained better outcomes. The obtained accuracy was 97%. However, several modifications could still be tested to improve the proposed method. Other features extraction techniques, or other recent neural network- based classifier, or more datasets could be used. This work has just shown how the combination of recent deep learning and vector representation techniques allows to getting better results on the problem of language identification of (short) texts[3].

Tomasz Walkowiak, Szymon Datko, and Henryk Maciejewski, in 2019, through their study, put forwarded a comparison on 4 approaches to depict text documents with the help of feature vectors. There were two methods which were mainly used in the study, each of those sharing a common domain of Natural Language Processing, in order to preprocess raw text. Both of these methods made use of "frequency-of-words" or "frequency-of-topics representation" of the given content. These methods were Bag of Words (BoW) and LDA (nouns). Two more methods not requiring any language knowledge were also studied. These methods represent the words present in the document as feature vectors by either using their vector representation (Word2Vec method) or by making use of the frequencies of terms extracted from the original text. These methods were fast Text and LDA (raw). Three completely different datasets of text content were taken and utilized from Press, Wiki and Web. It was clearly indicated that language knowledge is crucial. Nevertheless, if the training set comprises adequate examples, the language knowledge no longer remains important. Henceforth, such datasets proved to be comparable with Natural Language Processing based methods. Further, the Natural Language Processing techniques in two situations can prove to be more promising in terms of getting better outputs. One of the situations is where there are not many training examples, and the second situation being when there is presence of a large number of categories which are not easily distinguished. Yet, the results from suggested system convey that from mentioned methods, no particular approach surpasses other methods in all tests, even so, tests having more amount of exercise remarks seem to favor the NLP-free Word2Vec methods [4].

Rosemol Thomas, Anu George and Leena Mary, in their study focused on identifying five Indian languages by developing an LID system. Deep Neural Networks were used for the same. Indian languages share similar origins and also intersecting phoneme sets. However, each language has eccentricity with respect to the phonotactic constraints. Considering their idiosyncratic effect on the phonotactic constraints, this uniqueness was put to use, to develop LID systems. The proposed LID system included speaker variability as well as multiple utterances per speaker. The result conveyed that this helped in rescinding the impact of speakers in language, and thus proved fit for LID. Further, context DNN, context free DNN, as well as autoencoder structures were gauged. Testing was done by using speech utterances duration in range of 10,15 and 20 seconds. Comparing and contrasting of results of the structures used informed that the autoencoder structure gave the best result. Also, it was further concluded that as the duration of speech utterance increased; error of classification reduced. This LID system can also be broadened and used for including a greater and diverse number of Indian Languages [5].

Deepu S , Pethuru Raj and S.Rajaraajeswari, through their work presented a convenient framework for the extensive use of BagofWords model. A straightforward example was also demonstrated by using the framework to depict its usability, which can be further used across in various other related cases.The BOW model (and the BOW technique) usually involves three important steps: Reading the Unstructured Text Data, Data Pre- processing, and Knowledge Discovery. This is a generic framework, where in advanced algorithms can be easily sneaked in. Two probabilistic models: the multinomial model, and the multivariate Bernoulli model were used. Each of these represent the given con- tent using Naïve Bayes (NB) assumption and together are used in the "Bernoulli Document Model", which has been leveraged in the proposed system. Entire methodology of the developed system involved the Loading of the R-package and text (Dataset ingestion), followed by cleansing and text processing. Leveraging the Bernoulli Docu- ment was the next step, followed by obtaining the outputs. The prominent steps also involved creation of document as well as the transpose, exploration of the words, crea- tion of the word cloud, and also the recognition of most widely used critical. This is how an easy-to-use framework was developed [6].

Andre Lynum, presented a study which showed that lexicalized features can form the cornerstone of a sturdy Native Language Identification system. This method requires abandoning linguistic considerations, so that an efficient and accurate system is build. Features used in the submitted Native Language Identification Systems were Bareword- feature, character n-gram feature, Bareword directed collocation features, Suffix directed collocation features. The proposed system used a linear SVM multiclass classifier. SVM was used, because it can train models with many features accurately and expeditiously. Also, it has been successfully used to build many high- dimensional models for various Natural Language Processing related works. Four systems were submitted to the shared task. Out of the different high dimensional models created using SVM, three possessed same feature types. The DF cutoff used to filter individual features were different for each of them. Also, the fourth system made use of character n-grams along with the features present in the rest of the systems. All the four systems showed equally satisfying performance. On comparing the performance of these systems, it was found that there was only some difference. It was also concluded that the lexical features used in the systems could be used for competitive systems by themselves, as they predicted well enough in isolation also. However, the features based on POS tags were very less predictive when compared with the lexical features[7].

Vadim Andreevich Kozhevnikov, Evgeniya Sergeevna Pankratova, in 2020 in their study focused on different classification algorithms and vectorization methods. Classification by using neural networks was also done (specifically with convolution neural networks). The different vectorization algorithms included vectorization method based on occurrence of words in corpus, direct coding (or binary), TF-IDF method and disturbed vector representation method using Word2vec. These methods were further also reviewed with its implementations in scikit-learn frameworks. Classification algorithms used in their study were SVM, KNN and neural networks. It was concluded that out of all the methods implemented, neural networks proved to be the best tool for text classification. The reason was that CNN uses the idea of weight distribution, due to which the number of parameters requiring training was significantly reduced which led to improved generalization. This further led the model to study the data better without undergoing overfitting [8].

Marco Lui, Jey Han Lau, Timothy Baldwin in their study focused on introducing a method to recognize the languages present in a multilingual document. They also focused on estimating their proportions. Synthetic data and real-time multilingual content collected from web was used to demonstrate the effectiveness of their method. The model used in their study posited that such content which comprises of many languages together is generated as an unknown language fusion of languages from the training set. In this study, a Gibbs sampler was introduced in the system for any set of languages. It was used to select the language which maximized the posterior probability of the document. In all, the presented system made used a generative mixture model. As a result, the presented system outperformed other approaches on synthetic data, as well as on real- time data. Their system finally could also efficiently predict the portion of the document written in each of the identified languages [9].

## III. DATASET

The dataset used for the study was acquitted using Kaggle.com. Datasets of total of 17 languages were collected, out of which 4 are Indian languages. The table below shows the languages considered for this study. Each of these languages have a separate .csv file, which is further combined into a single .csv file. It comprises of 10367 datasets in all, out of which 600 datasets belong to Indian languages like Tamil, Hindi, Kannada, Malayalam etc.

## IV. VECTORIZATION TECHNIQUES

### *A. Bag of Words*

The Bag of words is an extensively used model for the resolution of text categorization. The model learns from vocabulary from the given content, and further models every document by tallying the amount of times each word appears present in the content or given document. It has proved to be a very straightforward method for representing data, such that no independence among the words present in the document is assumed. Hence it is one of the most simplifying representation techniques used in NLP as well as information retrieval tasks.

As the name of the model suggests, the text from the dataset or content is arrayed as a “bag” of words. Although the word order and grammar is indifferent for the model, the multiplicity is crucially considered.

Example: Sentence A: This book is written in English. Sentence B: This book is expensive and is interesting. From the above sentences, following is the vocabulary formed: {this, book, is, written, in, English, expensive, and, interesting}

In order to get the “bag” of words, frequency of each word is counted in each of the sentences.

Sentence A: {1, 1, 1, 1, 1, 1, 0, 0, 0}

Sentence B: {1, 1, 2, 0, 0, 0, 1, 1, 1}

In sentence A, “this”, “book”, “is”, “written”, “in”, and” English” occur once, so the frequency of the respective words is depicted by number 1 in the feature vector. Since “expensive”, “and”, “expensive” do not appear in the sentence, their absence is shown by 0. Similarly, the features of sentence B can be represented as Sentence B: {1, 1, 2, 0, 0, 0, 1, 1, 1}. Since “is “occurs twice in sentence B, its presence is marked as 2 in the feature vector.

### *B. Term Frequency-Inverse Document Frequency*

TF-IDF (Term Frequency-Inverse Document Frequency) is a technique that quantifies words from a given document or content. A score for each word is computed to highlight its significance in the corpus. TF-IDF is widely used for text mining as well as Information Retrieval.

*1) Term Frequency (TF):* The number of times a particular term is mainly stated as frequency mode. The term which occurs the most helps specify the content in a better way, rather than the term which is present lesser number of times.

*2) Inverse Document Frequency (IDF):* Terms occurring frequently in the given document are regarded as important. However, this is not the sufficient condition, as these words are not equally important because the term's eventual relevance is also determined by how uncommon it is in other portions and areas of the corpus and document. This is handled by Inverse Document Frequency. Giving more weight to frequently occurring terms (by TF) and de-escalating this weight (by IDF) in case the term appears in numerous essays, is combined together to form TF-IDF.

## V. MACHINE LEARNING ALGORITHMS USED

### *A. Support Vector Machine*

Support Vector Machine is a machine learning algorithm that is supervised. Although it may be used for both regression and classification issues, it is more suited to the latter. This algorithm's main goal is to locate a hyperplane in an N-dimensional space. The data points are neatly grouped in this hyperplane. The dimension of the hyperplane is determined by the number of features. In the case of two input characteristics, the hyperplane, for example, is linear. The hyperplane is a two-dimensional plane when three characteristics are present. However, as the number of characteristics grows, determining the hyperplane gets more difficult.

### *B. Naïve Bayes Classifier*

The Naïve Bayes algorithm is a basic supervised as well as probabilistic machine learning method that is also one of the most effective. As a result, it is a probabilistic classifier because it predicts based on the likelihood of items. Every occurrence of a feature is presumed to be independent of the occurrences of other features. It all comes down to the Bayes theorem, which says the below formula:

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)} \tag{1}$$

Where, A and B are two different types of events. P (A | B) is the probability of event A occurring given the occurrence of event B. The previous autonomous probability is P(A) (probability of event before evidence is seen). P (B | A) is the probability of B given occurrence A, i.e., the probability of B after seeing evidence A.

### *C. Logistic Regression*

One of the most well-known Machine Learning algorithms is logistic regression, which belongs to the category of Supervised Learning methods. It is a technique to identify from a collection of independent variables, a categorical dependent variable. Linear Regression is used to solve regression tasks, while Logistic Regression is used to handle classification tasks. Rather than developing a regression line in logistic regression, a "S" formed logistic function is generated, and 0 and 1 are the maximum values obtained. The chance of anything happening is represented by the logistic function's curve. Using continuous and discrete data, it may generate probabilities and categories fresh data. Logistic regression can easily determine the most appropriate classification scheme. It can also classify observations using a range of data sources.

## VI. SYSTEM FLOW

Data cleaning and text processing were next steps taken after data acquisition. This involved exempting the data from symbols and numbers, and converting lower case characters into upper case characters. The processed data was then used for feature selection for label encoding. Further, vectorization techniques: Bag of Words, and TF-IDF were implemented onto the selected features from processed data. Finally, six combination model were built by putting together three classification models: SVM, Logistic Regression, and Naïve Bayes, with the two vectorization techniques discussed above. The models were compared and contrasted based on the performance accuracy and outputs.

Working of the system:

- The web application is loaded on the browser.
- Homepage is loaded or opened.
- Text is entered in the given text input space. Entered text is sent to backend of the web application.
- Text sent to backend is processed before loading and feeding to Machine Learning model. Extra characters such as #$%^&*()_ and numbers are removed from the text.
- A machine learning model is loaded. This model is pre-trained and ready to do predictions.
- The processed text is given to the machine learning model in the backend of the web application.
- The machine learning model makes predictions and gives output which is displayed on homepage of the web application.

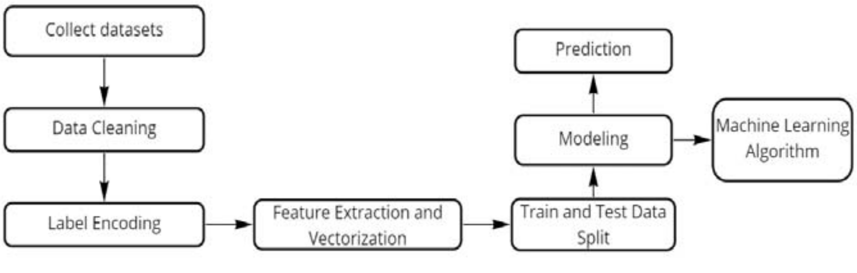

Collect datasets
Data Cleaning
Label Encoding
Feature Extraction and Vectorization
Train and Test Data Split
Modeling
Prediction
Machine Learning Algorithm

Fig. 1. Accuracy comparison among models

TABLE 1. Accuracy comparison between Bag of Words and Term Frequency-Inverse Document Frequency

| **BoW** | **TFI-DF** |
|---|---|
| 97.34%<br>(Naïve Bayes) | 81.33%<br>(NaïveBayes) |
| 95.69%<br>(Logistic Regression) | 98.45%<br>(Logistic Regression) |
| 93.56%<br>(SVM) | 93.08%<br>(SVM) |

The highest accuracy achieved was of 98.45 % (by Logistic Regression and TF-IDF combination model), followed by 97.34% (Bag of Word and Naïve Bayes combination model). All the three models performed different under different vectorization techniques. It was observed that Naïve Bayes performed well when the vectorization technique used was Bag of Words.

Naïve Bayes assumes all the features to be conditionally independent. Therefore, if some of the features are dependent then it might result into poor accuracy. In TF-IDF, when there is a class imbalance or more occurrences in one class, the recurring class's strong word characteristics risk getting a smaller IDF, and therefore the main qualities end up with far less weight. Hence Naïve Bayes gives less accuracy with TF-IDF as compared to Naïve Bayes with BoW. Therefore, in regard to the dataset used, second highest accuracy of 97.34% was obtained from the Bag of Word and Naïve Bayes combination model. However, there was not much difference in the performance in SVM model when BOW and TF-IDF techniques were used. But SVM and Bag of Words combination model gave a marginally better accuracy. The Bag of Words model is the most basic type of numerical representation of text. If newer phrases comprise unfamiliar words, the vocabulary grows, and hence the size of vectors grows as well. Since the vectors will increase the number of features will also increase. SVM algorithm works well in high dimensional spaces and therefore good accuracy was obtained by the Bag of Words and SVM combination model.

Logistic Regression performed well when the vectorization technique used was TF-IDF. Logistic Regression does not assume that features are conditionally independent, hence Logistic Regression performs better with TFIDF than BoW. It divides feature space in a linear manner and works effectively when some characteristics are connected. Another benefit of Logistic Regression is that it is super quick at sorting unknown records and can easily be extended to numerous classes. The Logistic Regression and TF-IDF combination model therefore gave the highest accuracy of 98.45%.

The TF-IDF and Logistic Regression model was then deployed and hosted on web.

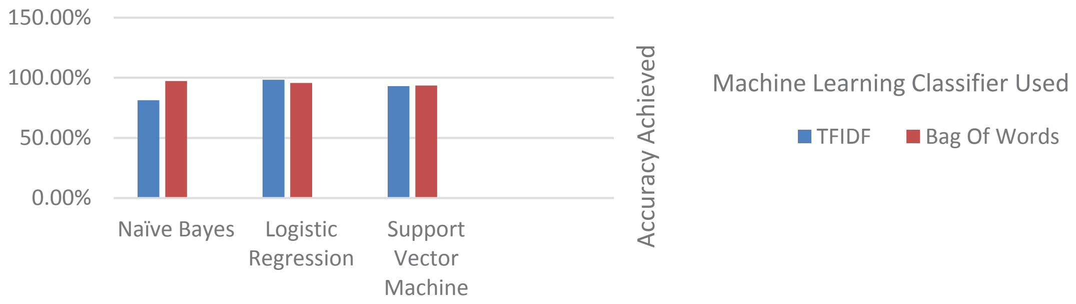
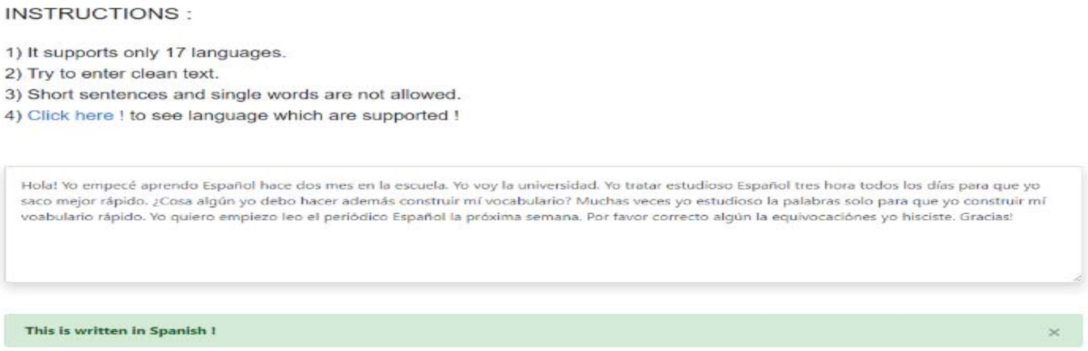


Fig. 2. Accuracy comparison among models



INSTRUCTIONS :

1) It supports only 17 languages.
2) Try to enter clean text.
3) Short sentences and single words are not allowed.
4) Click here ! to see language which are supported !

Hola! Yo empecé aprendo Español hace dos mes en la escuela. Yo voy la universidad. Yo tratar estudioso Español tres hora todos los días para que yo saco mejor rápido. ¿Cosa algún yo debo hacer además construir mí vocabulario? Muchas veces yo estudioso la palabras solo para que yo construir mí voabulario rápido. Yo quiero empiezo leo el periódico Español la próxima semana. Por favor correcto algún la equivocaciónes yo hisciste. Gracias!

This is written in Spanish !

Fig. 3. Spanish language prediction using Logistic Regression (with TF-IDF) model.

## CONCLUSION

Out of the 6 combination models of classification and vectorization, the Model created using Logistic Regression and TF-IDF proved to yield highest accuracy of 98.45%, followed by the Naïve Bayes and Bag of Word model, which gave 97.34% of accuracy. The least accuracy was 81.33%, given by TF-IDF and Naïve Bayes model. The accuracy in both models of SVM classifier, resulted to be around 93.08% to 93.56%. The proposed system of approaches can be also used for a greater number of languages.

Along with these promising accuracies and identification of 17 languages, there were 2 shortcomings observed. One of those is the discrepancy in identification of language in case of very short sentences (those comprising only one or two words). Other short- coming is the case of multilingual document, regarding which future work and enhancement is urged to be done.

## FUTUREWORK

As concluded above, the systems and models compared and contrasted possess loopholes regarding identification of languages from content comprising of multilingual texts and a very short sentence alone. Work regarding the same can be done in future using a larger and diverse corpus, along with the use of LSTM with vectorization methods.

## REFERENCES

[1] Binyam Gebrekidan Gebre , Marcos Zampieri, Peter Wittenburg, and Tom Heskes, "Improving Native Language Identification with TF-IDF Weighting", Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 216–223

[2] Marcos Zampieri, "Using Bag-of-words to Distinguish Similar Languages: How Ef- ficient are They?", 2013 IEEE 14th International Symposium on Computational Intel- ligence and Informatics (CINTI).

[3] Ermelinda Oro , Massimo Ruffolo and Mostafa Sheikhalishahi, "Language Identifi- cation of Similar Languages using Recurrent Neural Networks" , ICAART 2018 - 10th International Conference on Agents and Artificial Intelligence.

[4] Tomasz Walkowiak, Szymon Datko, and Henryk Maciejewski," Bag-of-Words, bag- of-topics and word-to-vec based subject classification of text documents in polish - a comparative study", Springer International Publishing AG, part of Springer Nature 2019 W. Zamojski et al. (Eds.): DepCoS-RELCOMEX 2018, AISC 761, pp. 526–535, 2019

[5] Rosemol Thomas, Anu George and Leena Mary," Language identification using deep neural network for Indian languages", Proceedings of the International Conference on Microelectronics, Signals and Systems 2019 AIP Conf. Proc. 2222, 030018-1–030018-6; https://doi.org/10.1063/5.0004096

[6] Deepu, Pethuru Raj and S.Rajaraajeswari," A Framework for Text Analytics using the Bag of Words (BoW) Model for Prediction", International Journal of Advanced Networking & Applications (IJANA)

[7] Andre Lynum," Native Language Identification using large scale lexical features", Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educa- tional Applications, pages 266–269, Atlanta, Georgia, June 13 2013. c 2013 Associa- tion for Computational Linguistics

[8] Vadim Andreevich Kozhevnikov and Evgeniya Sergeevna Pankratova, "Research Of The Text Data Vectorization and Classification Algorithms Of Machine Learning", ISJ Theoretical & Applied Science, 05 (85), 574-585.

[9] Marco Lui, Jey Han Lau and Timothy Baldwin, "Automatic Detection and Language Identification of Multilingual Documents", Transactions of the Association for Com- putational Linguistics, 2 (2014) 27–40. Action Editor: Kristina Toutanova.