Statistical testing based feature selection for Native Language Identification INLI@FIRE2018

B. Bharathi and Bhuvana J

[1] SSN College of Engineering, Chennai, Tamilnadu
[2] {bharathib, bhuvanaj}@ssn.edu.in

Abstract. Native Language Identification is the process of determining the language native to the author from the written text. We have proposed a system that uses machine learning algorithms to identify the native language from the written text. We extracted Term frequency-inverse document frequency (Tf-idf) as the feature from the given document and used statistical based measures such as Analysis of Variance -F value measure, Chi - square measure for selecting the best features. The selected features are fed to Multi Layer Perceptron and Stochastic Gradient Descent classifier to classify the native language into one of 6 listed Indian languages. This work was submitted to Indian Native Language Identification task INLI@FIRE2018. We have investigated the performance of the proposed system using three classifiers namely Multi Layer Perceptron (MLP) classifier with Analysis of Variance -F value measure, MLP classifier with Chi - square measure and Stochastic Gradient Descent classifier with Chi - square measure. From the results we have observed that SGD classifier with Chi - square measure has performed better than the other two classifiers.

Keywords: Native Language Identification · Tf-idf · MLP · SGD · Analysis of Variance-F value measure · Chi - square.

1 Introduction

Native Language Identification (NLI) is the process of identifying the native language of the author based on written texts in an another language. Influence of one language can affect the usage of other language by a same speaker referred to as cross-linguistic influence (CLI). CLI has played an important role in Second Language Acquisition (SLA) which studies and examines the effects of one language on other learned languages.

The influence of the native language will be reflected in the text through the usage of patterns. Identifying such patterns forms the basics of NLI. NLI has been identified as a multiclass classification task, for which various traditional machine learning approaches can be used to identify the native language. NLI find its applications in educational scenarios, where feedback can be provided to language learners, in authorship profiling, security, personalized grammar correction etc.

The proposed work for NLI of 6 languages has used three models such as Multi Layer Perceptron classifier with Analysis of Variance -F value measure, MLP classifier with Chi - square measure and Stochastic Gradient Descent classifier with Chi - square measure. This proposed work is submitted for the shared task on INLI @FIRE 2018 [3]. Against the existing trend of using deep learning, we have used tradional machine learning algorithms, since the size of the datasets used by INLI @FIRE 2018 shared task are too small. Related work in this field of research is given in section 2, the proposed system is elaborated in section 3. Section 4 provides the details of experimental setup with the analysis on performance in section 5 and section 6 concludes the paper.

2 Related Work

This section gives the overview of similar work done in the filed of NLI. Most of the widely used features for NLI [11] are part-of-speech, lexical features such as sentence length, document length, type/token ration, character and word n-grams, syntactic features namely parse tree and dependency-based features, lexical, and stylistic features, character n-grams, POS bi-grams, with some spelling mistakes character, word n-grams, writing quality markers, tree substitution Grammars and dependency features. The classifiers commonly used in literature are SVM classifier, Nave Bayes classifier, logistic regression learners etc.

In [8] the authors have used Tf-idf weighting schemes on features such as n-gram words/characters/POS tags with linear classifiers like support vector machines, logistic regressions and perceptrons. This work was submitted to 2013 NLI Shared Task in the closed-training track and has achieved 84.55% accuracy by 10 fold cross-validation testing on the TOEFL11 corpus. Authors of [4] have also used Tf-idf for feature extraction and SVM as classifier which reported a overall accuracy of 43.60%. Their work was one of the submissions of INLI@FIRE2017 tasks.

In [5], the authors have integrated several approaches into an ensemble for NLI. Two Resnets have been applied on linguistic features whose outputs are combined and given to a fully connected layer. A sentence level bidirectional LSTM was used to capture syntactic patterns over tokens and related POS tags from the Stanford CoreNLP toolkit were used. For NLI classification, features are extracted from misspelled words as well and are given to a logistic regression classifier. Continuous bag-of-words (CBOW), which is the mean of embedding of all words in the essay was used as a feature in simple neural network. It has been observed that the tradional methods have yield better results in the required task.

Three advanced ensemble models are evaluated in [10] on three data sets. SVMs with a Radial Basis Function (RBF), Logistic regression, Perceptron, ridge regression, Decision Trees, Linear Discriminant Analysis, Quadratic Discriminant Analysis, k-nearest Neighbors and Nearest Centroid are used as meta-classifiers for NLI. The outputs generated by the meta-classifiers are either discrete labels or continuous values. Authors have used McNemars test a non-parametric method as statistical testing for comparing NLI systems. The meta-classifier, mean probability combiner, the tree-based method namely Random forest and LDA-based method are observed to outperform in their performances.

The other work that supports the ensemble of classifiers for NLI is [9], which was the extension of work submitted to 2017 Native Language Identification shared task. To avoid excessive use of hand engineered features, [9] has used simple word and sub-word features and are used to train Naive Bayes classifier with 11 classes. Also used Gated Recurrent Units (GRU) as complementary model for invsetigation.

In [6], recurring word-based n-grams, part-of-speech, dependencies, lemma realization and syntactic parse tree are used as features for NLI with accuracy for open and close task namely 84.5% and 82.2% respectively. This work was a participant task submitted to NLI Shared Task 2013 and used an ensemble that integrates features and are evaluated on TOEFL11 and International Corpus of Learner English (ICLE), BUiD (British University in Dubai) Arab Learner Corpus (BALC), International Corpus Network of Asian Learners of English (ICNALE), Tbingen Telugu NLI Corpus (TTEL-NLI) and Non-TOEFL11 (NT11) corpora.

Another work that used ensemble of classifiers that also a participant of NLI Shared Task 2017 is reported in [7]. Along with lexical features, function words and Spelling errors, phonemes were also extracted for their task. These features are used to train both Support Vector Machines (SVM) and Fully Connected Neural Networks (FCNN) as classifiers. Outputs of these calssifiers are integrated using a voting scheme. Mean, Median and Plurality vote are the three different voting schemes used to combine the outputs of two classifiers. Also observed that the performance of CNN specifically for NLI is not good enough when compare with traditional machine laerning algorithms.

3 Proposed approach

The task of Indian Native Language Identification (INLI) has been carried out using machine learning algorithms. Three runs have been submitted for this INLI@FIRE2018 task. First submission has used Tf-idf features with Analysis of Variance (ANOVA) F-values for selecting best features. These features are trained using Multi Layer Perceptron (MLP) classifier. Second submission used Tf-idf features with Chi - square value feature selection method. These features are trained using same MLP classifier. Third submission used the Tf-idf features with Chi - square value feature selection method. These features are trained using Stochastic Gradient Decent (SGD) classifier. The details of the tasks have been explained in the following sections. The proposed approach uses the following steps to carry out the INLI task.
(i) Data preparation
(ii) Feature extraction

(iii)Feature selection and
(iv) Classification
The architecture of the proposed system is shown in figure 1.

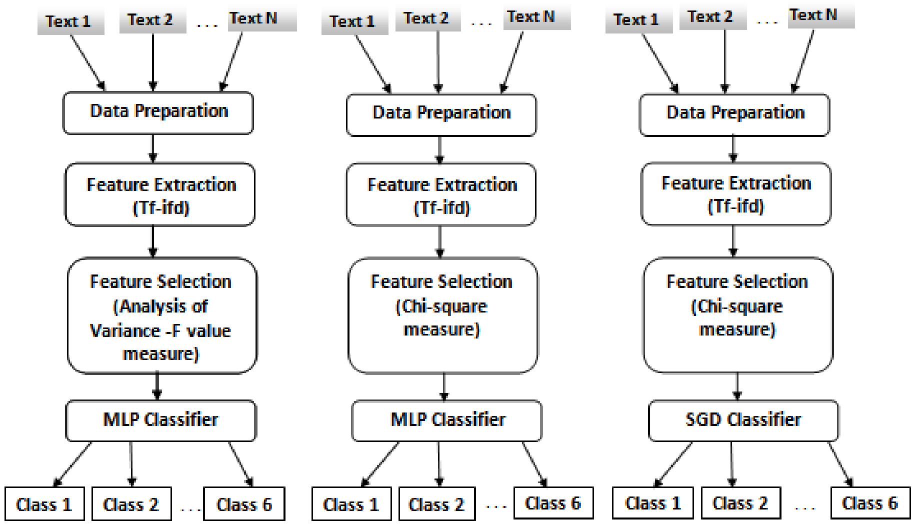

Fig. 1. Architecture of the proposed system

3.1 Data preparation

The dataset used for this research are Facebook comments which are presented in the form of XML files. From this XML file, the relevant text portions, that is, text given between the comment tags are extracted. From the extracted text, special symbols, punctuation symbols are removed before processing the input text. Minidom library from XML.dom package [1] was used to extract the text presented between comment tag. The text in each file is then converted into a Python lists. In this work six lists were created for each of the language given in the dataset.

3.2 Feature extraction

Tf-idf (Term frequency-inverse document frequency) is used in the proposed approach to represent the feature of the given text. The Tf-idf weight is the one that is used widely in natural language processing applications. It is a statistical measure which is used to assess how significant is a word is to a document in a corpus. The significance increases as the number of times a word occurs increases in the document. The feature extraction is done using the tool TF-IDF vectorizer method from the *scikit* learn library.

3.3 Feature selection

From the extracted Tf-idf features, the best features are selected using two statistical feature selection algorithms namely Analysis of Variance -F value measure (ANOVA) and Chi-square methods in the proposed approach. ANOVA and Chi-square methods are filter based feature selection methods. These methods select the best discriminative features from the training data. These statistical feature selection methods are applied on the categorically independent features. The analysis of variance is a statistical inference test that compares multiple groups at the same time. The chi-square test is a statistical test of independence to determine the dependency of two variables. Chi-square statistics between every feature variable and the target variable observes the existence of a relationship between the variables and the target. If the target variable is independent of the feature variable, those features are discarded. If they are dependent, the feature variable is considered to be very important. The implementation of feature selection is carried out using *f_classif* and *chi2* packages from *scikit-learn* Python library.

3.4 Classification

The proposed approach uses Multi Layer Perceptron (MLP) classifier for submission 1 & 2. Stochastic Gradient Decent algorithm is used for submission 3. MLP is a kind of feedforward neural network with layers namely, input layer, hidden layers and output layer. MLP forms a directed graph where the input vector flows only in one direction through the intermediate layers. Each node has a neuron with a nonlinear activation function excepting the input nodes. With multiple layers, a MLP performs back propagation which is a supervised learning technique that classifies non linearly separable samples. For training with multi layer perceptron network, the following configurations were used. Number of hidden layers for the network is 2, number of hidden nodes in each hidden layer is 25 as shown in figure 2, the activation function used in the hidden layer is RELU (Rectified Linear Unit) and Adam optimizer is used for weight optimization.

Stochastic Gradient Decent classifier is a linear classifier used for supervised learning. SGD is the variant of Gradient Decent algorithm which is used to minimize the least square error. The problem with Gradient Descent is when the data set is huge, parameter calculation becomes expensive. The weight is optimized using the following equation.

$$w = w - \lambda \triangledown Q_i(w) \tag{1}$$

In equation 1, $Q_i(w)$ refers to the gradient of the prediction error for the model on the training data, w is the weight being optimized and λ is the learning rate. In SGD, a sample of training set or one training value is used to calculate the parameters instead of the entire sample space on each iteration. When compared to other classifiers, this works much faster. By combining several binary classifiers

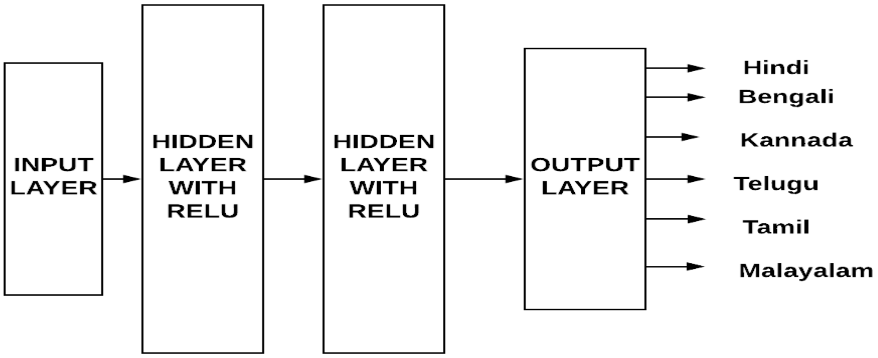

Fig. 2. Multi Layer Perceptron

together in one versus all manner, SGD performs a multi-class classification. A binary classifier is trained to learn to distinguish between one and that of the remaining N-1 classes of the total N classes. During the testing phase, for each of the classifier, signed distance to the hyperplane which is referred to as confidence score is calculated and the class with the maximum confidence is chosen.

4 Experimental setup

The proposed approach used the data set provided by the INLI@FIRE 2018 shared task [2]. The data set consists of Text/XML files which in turn have a set of Facebook comments in English language. Six Indian languages are proposed to consider for this shared task on native language identification. They are Tamil, Hindi, Kannada, Malayalam, Bengali and Telugu. The details of the number of training data for each of the language is given in Table 1.

Table 1. Number of training documents for each of the six language

Language	Training dataset
Hindi (HI)	211
Bengali (BE)	202
Kannada (KA)	203
Telugu (TE)	210
Tamil (TA)	207
Malayalam (MA)	200

5 Performance analysis

Two test sets have been provided by INLI@FIRE 2018 shared task for testing. The number of documents in test set 1 is 783. The number of documents in test set 2 is 1185. The cross validation accuracy (with 5 folds) of the proposed work for the three submissions are given in Table 2.

Table 2. Cross validation accuracy of the proposed system with different approaches

Method	Cross validation accuracy in %
MLP classifier with ANOVA -F value measure	94.3
MLP classifier with Chi - square measure	95.1
SGD classifier with Chi - square measure	90.4

The lack of accuracy or the errors are mainly due to the lack of reasonable number of training samples in the given data set. The performance of the proposed system for Indian language identification task have been measured using the metrics namely precision, recall and F1-measure for each language for the two test sets. The overall accuracy of proposed system has also been estimated. The performance of the proposed system using MLP classifier with Analysis of Variance -F value measure is given in Table 3.

Table 3. Proposed system performance using MLP with ANOVA-F value measure

Class	Test set 1			Test set 2		
	Precision in %	Recall in %	F1-measure in %	Precision in %	Recall in %	F1-measure in %
BE	63.90	71.90	67.70	43.40	49.30	46.20
HI	67.90	15.10	24.80	10.00	5.10	6.70
KA	31.40	66.20	42.60	38.10	40.40	39.20
MA	31.30	55.40	40.00	35.90	39.50	37.60
TA	36.40	43.00	39.40	26.80	42.10	32.80
TE	37.80	38.30	38.00	41.10	28.80	33.90
Overall accuracy		44.1%			35.4%	

The model using MLP classifier with Analysis of Variance -F value measure, Bengali language has achieved highest F1 score than the other languages in both the test sets.

The performance of the proposed system using MLP classifier with Chi - square measure is given in Table 4.

Table 4. Performance of the proposed system using MLP with Chi - square measure

Class	Test set 1			Test set 2		
	Precision in %	Recall in %	F1-measure in %	Precision in %	Recall in %	F1-measure in %
BE	65.80	68.60	67.20	45.50	50.70	47.90
HI	53.70	11.60	19.00	11.10	4.30	6.20
KA	30.90	62.20	41.30	38.00	42.00	39.90
MA	26.00	54.30	35.20	33.20	39.50	36.10
TA	39.30	48.00	43.20	27.40	41.40	33.00
TE	49.30	44.40	46.80	47.70	33.20	39.20
Overall accuracy		42.9%		36.8%		

Similar to the previous proposed model, the model using MLP classifier with Chi - square measure has also identified Bengali language with highest F1 score than the other languages in both the test sets.

The performance of the proposed system using SGD classifier with Chi - square measure is given in Table 5. SGD classifier with Chi - square measure has classified Bengali language with highest F1 score in both test datasets with 65% and 44.4% respectively. When compared with the other two models of the proposed work, SGD classifier with Chi - square measure has achieved overall accuracy of 37% for test set 2 and 46.2% for test set 1 in NLI task. This might be because the linear classifier, SGD which is actually a Bayesian that relies on prior distribution that improves mean squared error and thereby the improving prediction accuracy. This performance of the proposed system using SGD with Chi-square measure has been evaluated to be the best performing model among the INLI@FIRE2018 submitted tasks.

Table 5. Performance of the proposed system using SGD with Chi - square measure

Class	Test set 1			Test set 2		
	Precision in %	Recall in %	F1-measure in %	Precision in %	Recall in %	F1-measure in %
BE	57.40	75.70	65.30	36.20	57.50	44.40
HI	69.20	17.90	28.50	9.10	3.60	5.20
KA	39.00	64.90	48.70	43.20	42.00	42.60
MA	29.50	55.40	38.50	37.20	45.00	40.70
TA	45.50	45.00	45.20	28.20	35.70	31.50
TE	41.80	40.7	41.20	49.60	27.60	35.50
Overall accuracy		46.2%		37%		

6 Conclusion

Our proposed approach has used a machine learning algorithm to identify the native language of the given document using Tf-idf features. From the extracted features we used statistical measures such as Analysis of Variance -F value measure, Chi - square measure for selecting the best features. These are given to MLP and SGD classifiers to classify the Indian Native Languages. We observed that SGD with chi-square measure has performed better than the MLP classifier with Analysis of Variance -F measure and MLP with chi-square measure. Performance of the proposed work can be improved further by using socio linguistic features from the text.

References

1. Lightweight DOM implementation,
https://docs.python.org/3.0/library/xml.dom.minidom.html
2. Anand Kumar, M., HB, B.G., Singh, S., Soman, K., Rosso, P.: 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.
3. Anand Kumar M, B.G.B., P., S.K.: Overview of the INLI@FIRE-2018 Track on Indian Native Language Identification. In: In workshop proceedings of FIRE 2018, FIRE-2018, Gandhinagar, India, December 6-9, CEUR Workshop Proceedings.
4. Bharathi, B., Anirudh, M., Bhuvana, J.: Bharathi SSN@ INLI-FIRE-2017: SVM based approach for Indian Native Language Identification. In: International conference on Forum of Information Retrieval Evaluation (FIRE 2017) CEUR proceedings. vol. 2036, pp. 110–112 (2017)
5. Bjerva, J., Grigonyte, G., Östling, R., Plank, B.: Neural networks and spelling features for native language identification. In: EMNLP. pp. 235–239. Association for Computational Linguistics (2017)
6. Bykh, S., Vajjala, S., Krivanek, J., Meurers, D.: Combining shallow and linguistically motivated features in native language identification. In: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 197–206 (2013)
7. Chan, S., Jahromi, M.H., Benetti, B., Lakhani, A., Fyshe, A.: Ensemble methods for native language identification. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 217–223 (2017)
8. Gebre, B.G., Zampieri, M., Wittenburg, P., Heskes, T.: Improving native language identification with TF-IDF weighting. In: the 8th NAACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA8). pp. 216–223 (2013)
9. Kepler, F., Astudillo, R., Abad, A.: Fusion of Simple Models for Native Language Identification. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 423–429 (2017)
10. Malmasi, S., Dras, M.: Native language identification using stacked generalization. arXiv preprint arXiv:1703.06541 (2017)
11. Malmasi, S., Dras, M.: Native language identification with classifier stacking and ensembles. Computational Linguistics (In Press), 1–70 (2018)