A Machine Learning Approach to Indian Native Language Identification

D. Thenmozhi, S. Kayalvizhi, and Chandrabose Aravindan

Department of CSE, SSN College of Engineering, Chennai
theni_d@ssn.edu.in, kayalvizhi1704@cse.ssn.edu.in, aravindanc@ssn.edu.in

Abstract. NLI (Native Language Identification) determines the native language of the non-native users using their writings in a foreign language. It has several applications namely forensic and security, author profiling and identification, and educational applications. English is a most common language used in social media by many non-English people in the world to share their thoughts and ideas. They blend English with their native language for their posts and comments. Identifying the native language from the short text in English is still a challenging task. In this paper, we present a language agnostic approach without any language specific processing and employed machine learning approach with and without feature selection to identify the native language of a Indian speaker using their comments and posts in social network. The bag of word features are extracted from the text posted by the user and the feature vectors are constructed using TF-IDF score for the training data. We have used a statistical feature selection methodology to select the features that are significantly contributing to NLI task. The classifier with highest cross validation accuracy was used for predicting the native language of the user. Our approaches are evaluated using INLI@FIRE2018 shared task data set.

Keywords: Indian Native Language Identification · Language Recognition · Author Profiling · Machine Learning · Feature Selection · Text Mining.

1 Introduction

NLI (Native Language Identification) is the process of automatically identifying the native language of speakers using their speech or writing in different language. It has several applications namely forensic and security [7], authorship profiling and identification [6], and educational applications [19]. Several researches have been reported on text–based NLI [20, 11, 5, 8, 15, 16]. Currently, people use social media like YouTube, Facebook, Blogs and Tweets to share their thoughts, ideas and comments. English is the prominent language used by many non-English people by blending their native languages in their social media postings. In this line, Indians also use English predominantly in their comments and postings. Indian Native Language Identification (INLI) focuses

on identifying native language of Indians based on their English writings. Many shared tasks have been conducted on NLI since 2013 to identify the native language from English text. Recently, shared tasks on INLI are also evolving since 2017 [2]. Their focus is to research and develop techniques to identify the native language namely Tamil, Hindi, Kannada, Malayalam, Bengali and Telugu from the set of Facebook comments. Several methodologies have been reported on INLI. N-gram approach [13], Machine learning approaches with Support vector machines [17, 3, 12], ensembling approaches [17, 9] and deep learning approaches [21, 4] have been used to identify the Indian native languages. In this research, our focus is on the shared task of INLI@FIRE2018 [1] which identifies the native language (Tamil, Hindi, Kannada, Malayalam, Bengali or Telugu) of Indians based on their comments posted in social media. INLI@FIRE2018 is a shared Task on Indian Native Language Identification (INLI) collocated with the Forum for Information Retrieval Evaluation (FIRE), 2018.

2 Related Work

Native language identification is an author profiling task. PAN 2017 [18] focuses on language variety identification tasks. The shared tasks on INLI are also evolving since 2017 [2]. This section describes the methodology used for INLI tasks. Nayel and Shashirekha [17] normalized the text by removing the emoji, special characters, digits, hash tags, mentions and links. They used the techniques namely removal of stop word using the NLTK stop words package, manual stop word collection and other resources (Python stop words) to preprocess the data. They used TF-IDF scores to construct feature vectors and employed SVM to classify the native language of the user. Bharathi et al. [3] and Lakshmi et al. [13] also used TF-IDF for feature construction and SVM for classification for this task. Also, Lakshmi et al. used character n-gram and word n-gram while computing TF-IDF score. However, they have not applied any preprocessing techniques. Kosmajac and Keselj [12] normalized the text similar to [17] used TF-IDF with character n-gram and word n-gram for feature construction and employed SVM for classification. Jain et al. [9] considered non-English word and nouns phrase while computing TF-IDF scores without applying any preprocessing techniques. They used Logistic Regression, SVM, Ridge Classifier and Multi-Layer Perceptron (MLP) as base classifiers and employed an ensembled approach for language identification. Bhargava et al. [4] used a deep learning approach using hierarchical attention with bi-directional GRU architecture for this task. Thenmozhi et al. [21] also employed a neural network approach with 2 hidden layers for this task. They normalized the text similar to [17] and handled shortened words as part of pre-processing. They have considered only the nouns and adjectives present in the text to extract the features. In this paper, we propose a language agnostic approach in which we have not used any language specific (or linguistic related) processing to extract the features. Thus, we simply took bag of words to consider all the words in the text and went for statistical based feature selection to extract the most significant features for language identification task.

3 Proposed Methodology

We have used a supervised approach with three variations namely a) term-frequency (TF) without feature selection, b) TF-IDF (term-frequency inverse-document-frequency) without feature selection and c) TF-IDF with statistical feature selection for the INLI task. The steps used in our approach are given below.

- Preprocess the data
- Extract bag of words (BOW) features from training data
- Construct feature vectors using TF or TF-IDF with and without χ^2 feature selection
- Build the models using a classifier for the three variations
- Predict any of the six languages namely Tamil (TA), Hindi (HI), Kannada (KN), Malayalam (ML), Bengali (BE) or Telugu (TE) as class label for the instance using the model

The steps are explained below in detail.

3.1 Feature Extraction

The data for INLI task is given as XML file. The given text is preprocessed by extracting only the textual part of the content present in XML file. All the punctuations are removed before extracting the features. Since, the texts are collected from social network sites, many terms are in transliterated form and many terms are in short-hand notations like pls, sry, tc, tks, etc. Hence, we did not apply stop word removal and stemming as preprocessing steps. The unique terms present in the text are considered as features in our first two variations. The feature vectors for the training data is constructed using term-frequency in the first variation. TF-IDF is used to construct feature vectors in the second variation. However, the number of extracted features may be more. We have employed a χ^2 feature selection in our third variation to extract the useful features that are contributing to native language identification. The details of feature selection are explained below.

3.2 Feature Selection

In our third variation, we have used χ^2 feature selection. INLI task involves six categories namely "BE", "HI", "KN", "ML", "TA" and "TE". Hence, 2×6 CHI table (Table 1) or contingency table [14, 10, 22, 23] is constructed for all the feature f_x. Table 1 contains the observed frequency (O) of feature f_x for every category "BE", "HI", "KN", "ML", "TA" and "TE".

The observed frequencies (O) are used to compute the expected frequencies (E) for the feature f_x using Equations 1.

$$E(x,\ y) = \frac{\Sigma_{a\in\{f_x,\neg f_x\}}O(a,y)\Sigma_{b\in\{BE,HI,KN,ML,TA,TE\}}O(b,y)}{n} \tag{1}$$

Table 1. Feature-Category CHI Table for Language Identification

	BE	HI	KN	ML	TA	TE
f_x	$O(f_x, BE)$	$O(f_x, HI)$	$O(f_x, KN)$	$O(f_x, ML)$	$O(f_x, TA)$	$O(f_x, TE)$
$\neg f_x$	$O(\neg f_x, BE)$	$O(\neg f_x, HI)$	$O(\neg f_x, KN)$	$O(\neg f_x, ML)$	$O(\neg f_x, TA)$	$O(\neg f_x, TE)$

where n is the total no. of training instances, x indicates whether the feature f_x is present or not, y represents whether the training instance belongs to any of the six languages namely "BE", "HI", "KN", "ML", "TA" or "TE".

The expected frequencies namely $E(f_x, BE)$, $E(f_x, HI)$, $E(f_x, KN)$, $E(f_x, ML)$, $E(f_x, TA)$, $E(f_x, TE)$, $E(\neg f_x, BE)$, $E(\neg f_x, HI)$, $E(\neg f_x, KN)$, $E(\neg f_x, ML)$, $E(\neg f_x, TA)$ and $E(\neg f_x, TE)$ are calculated using Equation 1 for language identification. Then, we have calculated the χ^2 value for each feature f_x using Equation 2.

$$\chi^2_{stat} f_x = \Sigma_{x \in \{f_x, \neg f_x\}} \Sigma_{y \in \{BE, HI, KN, ML, TA, TE\}} \frac{(O(x,y) - E(x,y))^2}{E(x,y)} \tag{2}$$

The set of features whose χ^2_{stat} value is greater than $\chi^2_{crit(\alpha=0.01, df=5)} : 9.24$ are considered to be significant features for language identification. These selected features are used to build a model with a classifier in our third variation.

3.3 Model Building and Prediction

The models for the first two variations for language identification are built from training data using Multi Layer Perceptron (MLP) and the model for the third variation is built using Multinomial Naive Bayes (MNB) classifier with the selected features. The classifiers were chosen based on the cross validation accuracies. The class label as one among the six languages namely "BE", "HI", "KN", "ML", "TA" or "TE" is predicted for the test data instances by using the models.

4 Implementation

Our methodology was implemented in Python for this Shared Task on Indian Native Language Identification (INLI) task. The number of training instances are 202, 211, 203, 200, 207 and 210 for the languages namely Bengali, Hindi, Kannada, Malayalam, Tamil and Telugu respectively. Two sets of test data was given for the evaluations that consist of 783 and 1185 instances for test-set-1 and test-set-2 respectively. The textual part of data is extracted from XML file using xml.etree library. The punctuations are removed and the BOW (bag of words) features are extracted using the training instances. We have obtained a total of 21813 features from training data. Scikit–learn machine learning library was used to vectorize the training instances using CountVectorizer for the first variation and TfidfVectorizer for the second variation. We have implemented χ^2

feature selection algorithm to extract the significant features for native language identification. We have obtained a total of 1555 features by the feature selection with alpha=0.10 and degree of freedom 5 for the six classes.

We have employed several classifiers namely, Multinomial Naive Bayes, Gaussian Naive Bayes (GNB), Random Forest (RF), Decision Tree (DT), Extra Trees (ET), Ada Boost (AB), Stochastic Gradient Descent (SGD), Support Vector Machines (SVM), and Multi Layer Perceptron, and measured 10-fold cross validation to select the best classifier for all the three variations of our approach. Table 2 shows the cross validation output of various classifiers for all the three variations. This table shows that MLP performs better for the first two variations that are without feature selection and MNB performs better for the third variation which used feature selection. MNB performs better with less number of features that are selected using our chi-square feature selection. However, MNB was not able to perform well with all the features. This is because the likelihood would be distributed and may not follow the Gaussian or other distributions when huge feature set is used. When more features are there, they may affect each other's likelihood which reduces the performance. Hence, we have chosen MLP to build models for the first two variations and MNB to build model for the third variation. These models are utilized to predict the native language for the two sets of test instances.

Table 2. 10-fold cross validation accuracies.

Classifier	Term-Freq	Tf-IDF	ChiSquare
MNB	70.40	64.49	80.69
GNB	59.92	58.07	43.08
DT	44.92	42.5	37.48
RF	42.01	43.56	46.38
ET	51.01	48.58	49.23
AB	49.70	49.88	49.75
SUM	25.63	20.41	57.66
SGD	68.36	73.81	70.66
MLP	82.64	86.47	78.27

5 Results and Discussions

We have submitted our second variation (best among first two without feature selection) using MLP classifier and third variation (with feature selection) using MNB classifier as two runs for the shared task. The performance is measured in terms of precision (P), recall (R) and F1-measure. The results obtained by our approach for Run 1 on two test sets are shown in Table 3. The results show that our methodology which uses TF-IDF with MLP classifier does not perform well for Hindi language. We have obtained overall accuracies as 46.1% and 34.3% for test-set-1 and test-set-2 respectively.

Table 3. Test data Performance for Run 1.

Language	Test-Set-1			Test-Set-2		
	P	R	F1	P	R	F1
BE	65.4	65.4	65.4	43.2	35.3	38.8
HI	47.0	12.5	19.6	10.8	6.5	8.1
KA	29.9	58.1	39.4	37.3	40.0	38.6
MA	40.6	75.0	52.7	36.6	51.0	42.6
TA	46.4	58.0	51.6	24.4	41.4	30.7
TE	41.9	48.1	44.8	43.9	26.0	32.7
Overall Accuracy	46.1			34.3		

The results obtained by our approach for Run 2 on two test sets are shown in Table 4. The results show that our methodology which uses TF-IDF, χ^2 feature selection and MNB classifier improved the performance for Hindi and Tamil languages on test set 2. However, this method does not improve the performance for the other languages of test set 2 and for test set 1. We have obtained overall accuracies as 32.4% and 19.7% for test set 1 and test set 2 respectively.

Table 4. Test data Performance for Run 2.

Language	Test-Set-1			Test-Set-2		
	P	R	F1	P	R	F1
BE	58.0	43.2	49.5	34.6	22.2	27.1
HI	32.3	8.4	13.3	10.2	7.2	8.5
KA	19.6	41.9	26.7	30.9	37.2	33.8
MA	29.2	62.0	39.7	25.7	33.5	29.1
TA	37.5	36.0	36.7	28.2	34.3	31.0
TE	22.1	35.8	27.4	32.9	29.2	30.9
Overall Accuracy	32.4			19.7		

The results obtained by various teams for this shared task are shown in Table 5.

It is observed from Table 5 that the maximum accuracy obtained for Test set 2 is 37%. This may be due to data set size for training the model. Thus, we have obtained a very low accuracy. The data set size may be improved further using Generative Adversarial Networks (GAN) to improve the performance.

6 Conclusions

We have presented a machine learning approach for identifying the Indian native language namely Bengali, Hindi, Kannada, Malayalam, Tamil or Telugu from the English comments posted in social media. We have presented the three variations of our approach namely term-frequency without feature selection, TF-IDF without feature selection, and TF-IDF with χ^2 feature selection for the language

Table 5. Results comparison.

Team	Accuracy (%)	
	Test-Set-1	Test-Set-2
SSN-NLP submission-1	46.1	34.3
SSN-NLP submission-2	32.4	28.4
Ajees submission-1	14.0	24.1
Ajees submission-2	15.2	20.9
Ajees submission-3	10.7	21.8
CIC-IPN submission-1	41.8	34.1
CIC-IPN submission-2	41.3	34.4
CIC-IPN submission-3	41.4	34.5
CorpLab submission-1	42.1	31.8
CorpLab submission-2	39.8	30.8
CorpLab submission-3	40.4	31.5
DNLP submission-1	29.6	22.9
HIMANIKHURANA submission-1	19.3	–
IDRBT-TEAM-A submission-1	14.8	19.7
IDRBT-TEAM-A submission-2	19.7	18.0
Leorius submission-1	31.5	29.0
MANGALORE submission-1	46.6	35.3
MANGALORE submission-2	45.5	35.3
MANGALORE submission-3	46.6	35.3
NLPRL submission-1	15.3	17.1
SSNCSE submission-1	44.1	35.4
SSNCSE submission-2	42.9	36.8
SSNCSE submission-3	46.2	37.0
TeamJosan submission-1	22.2	24.5
TeamJosan submission-2	31.7	30.5
WebArch submission-1	41.4	31.9
WebArch submission-2	28.2	21.7
WebArch submission-3	29.8	21.9

identification task. The data set of INLI@FIRE2018 shared task is used to evaluate our approach. We have submitted our second and third variations to the task and we have obtained overall accuracies of 46.1% and 34.3% for our first run on test-set-1 and test-set-2 respectively. We have obtained overall accuracies of 32.4% and 19.7% for our second run on test-set-1 and test-set-2 respectively. Our feature selection improved the F-measure for Hindi and Tamil for test-set-2. However, it does not improve for the other languages. Since our approach is language agnostic, we have not included any character level features at present. These features may be considered in future to improve the performance of NLI task. The performance may also be improved by incorporating word embedding techniques in future. Due to data set size for training, we have obtained very low accuracy. The data set size may be improved by using Generative Adversarial Networks (GAN) in future to improve the performance.

References

1. Anand Kumar, M., Barathi Ganesh, B., Soman, K.P.: Overview of the INLI@FIRE-2018 track on Indian native language identification. In: In workshop proceedings of FIRE 2018. CEUR Workshop Proceedings, Gandhinagar, India, December 6-9 (2018)
2. Anand Kumar, M., Barathi Ganesh, H., Shivkaran, S., Soman, K., Rosso, P.: Overview of the INLI PAN at FIRE-2017 track on Indian native language identification. CEUR Workshop Proceedings **2036**, 99–105 (2017)
3. Bharathi, B., Anirudh, M., Bhuvana, J.: Bharathi SSN@ INLI-FIRE-2017: SVM based approach for Indian native language identification. In: FIRE-Working Notes. pp. 110–112 (2017)
4. Bhargava, R., Singh, J., Arora, S., Sharma, Y.: Bits_pilani@ INLI-FIRE-2017: Indian native language identification using deep learning. In: FIRE-Working Notes. pp. 123–126 (2017)
5. Bykh, S., Meurers, D.: Exploring syntactic features for native language identification: A variationist perspective on feature encoding and ensemble optimization. In: Proc. of COLING 2014, the 25th Int. Conf. on Computational Linguistics: Technical Papers. pp. 1962–1973. Dublin, Ireland (2014)
6. Estival, D., Gaustad, T., Hutchinson, B., Pham, S.B., Radford, W.: Author profiling for english emails. In: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics. pp. 263–272. ACL, Australia (2007)
7. Gibbons, J.: Forensic linguistics: An introduction to language in the justice system. Wiley-Blackwell (2003)
8. Ionescu, R.T., Popescu, M., Cahill, A.: Can characters reveal your native language? a language-independent approach to native language identification. In: Proc. of the 2014 Conf. on Empirical Methods in NLP (EMNLP). pp. 1363–1373. ACL (2014)
9. Jain, R., Duppada, V., Hiray, S.: Seernet@ INLI-FIRE-2017: Hierarchical ensemble for Indian native language identification. In: FIRE-Working Notes. pp. 127–129 (2017)
10. Janaki Meena, M., Chandran, K.: Naive bayes text classification with positive features selected by statistical method. In: Int. Conf. on Autonomic Computing and Communications, ICAC 2009. pp. 28–33. IEEE (2009)

11. Jarvis, S., Bestgen, Y., Pepper, S.: Maximizing classification accuracy in native language identification. In: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 111–118. ACL, Atlanta, Georgia (2013)
12. Kosmajac, D., Keselj, V.: Dalteam@ INLI-FIRE-2017: Native language identification using SVM with SGD training. In: FIRE-Working Notes. pp. 118–122 (2017)
13. Lakshmi, S., Shambhavi, B.: BMSCE_ISE@ INLI-FIRE-2017: A simple n-gram based approach for native language identification. In: FIRE-Working Notes. pp. 115–117 (2017)
14. Li Yanjun, C.L., Chung, S.M.: Text clustering with feature selection by using statistical data. IEEE Transactions on Knowledge and Data Engineering **20**(5), 641–652 (2008)
15. Malmasi, S., Dras, M.: Native language identification using stacked generalization. arXiv preprint arXiv:1703.06541 (2017)
16. Mohammadi, E., Veisi, H., Amini, H.: Native language identification using a mixture of character and word n-grams. In: Proc. of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 210–216. ACL, Copenhagen, Denmark (2017)
17. Nayel, H.A., Shashirekha, H.: Mangalore-university@ INLI-FIRE-2017: Indian native language identification using support vector machines and ensemble approach. In: FIRE -Working Notes. pp. 106–109 (2017)
18. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Working Notes Papers of the CLEF (2017)
19. Rozovskaya, A., Roth, D.: Algorithm selection and model adaptation for esl correction tasks. In: Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies-Volume 1. pp. 924–933. ACL, Portland, Oregon, USA (2011)
20. Tetreault, J., Blanchard, D., Cahill, A., Chodorow, M.: Native tongues, lost and found: Resources and empirical evaluations in native language identification. Proceedings of COLING 2012 pp. 2585–2602 (2012)
21. Thenmozhi, D., Kannan, K., Aravindan, C.: SSN_NLP@ INLI-FIRE-2017: A neural network approach to Indian native language identification. In: FIRE-Working Notes. pp. 113–114 (2017)
22. Thenmozhi, D., Mirunalini, P., Aravindan, C.: Decision tree approach for consumer health information search. In: FIRE-Working Notes. pp. 221–225 (2016)
23. Thenmozhi, D., Mirunalini, P., Aravindan, C.: Feature engineering and characterization of classifiers for consumer health information search. In: Forum for Information Retrieval Evaluation. pp. 182–196. Springer (2016)