Overview of the Second Shared Task on Indian Native Language Identification (INLI) *

Anand Kumar M[1][0000−1111−2222−3333], Barathi Ganesh H[2,3] Ajay S G[2], and Soman K P[3]

Department of Information Technology, NITK Surathkal m_anandkumar@nitk.edu.in
Amrita Vishwa Vidyapeetham, Coimbatore
{abc,lncs}@uni-heidelberg.de

Abstract. This overview paper describes the second shared task on Indian Native Language Identification (INLI) that was organized by FIRE 2018. Given a corpus with comments in English from various Facebook newspapers pages, the objective of the task is to identify the native language among the following six Indian languages: Bengali, Hindi, Kannada, Malayalam, Tamil, and Telugu. Altogether, 31 approaches of 14 different teams are evaluated. In this paper, we report the overview of the participant's systems and the results of second INLI shared task. We have also compared the results of the first INLI shared task conducted with FIRE-2017.

Keywords: Native Language Identification · Text Classification · Author Profiling.

1 Introduction

This paper explains the overview of the second INLI (Indian Native Language Identification) shared task conducted co-joined with FIRE2018. Native Language Identification (NLI) is the task of automatically classifying the L1 of a writer based only on his or her text written in another language[1]. The research in the native language identification is emerged in recent years because of its applications in Digital forensics and language learning. This is the first foremost task which is conducted particularly for Indian languages. It is a continuation of the previous shared task INLI-2017 conducted with the FIRE2017 conference. The objective of the task is defined as the set of user comments needs to be classified to an Indian native language. We have collected the user comments written in the English language from the regional news pages of Facebook. We assume that only the native persons will see the Regional news pages of Facebook. The motivation of the shared task is to create the first ever corpora for Indian native language identification in social media and to provide the environment to directly compare the different pre-processing methods, features, and the algorithms. Even though the researchers and industries showing an emerging interest

* Supported by organization x.

towards the native language identification, the development of such systems are slow down by the primary issue which is getting the right annotated corpora. Assessing the NLI system needs a corpus consists of texts in a language other than the native language of the user. The problem with the collecting essays and students assignments for native language identification is that even though the person belongs to a particular region or native language, we cannot assure that the person speaks or reads the native languages. Most of the Indians will speak their native language but not all will read and write their native language. Lack of such corpora in Indian languages induced us to collect the smaller size of INLI corpora and evaluating the participant's systems. Few of the prominent applications of native language identification is given below.

Error correction and language proficiency: The language proficiency of the region can be identified and analyzed with the help of native language identification system. It is known that people from different region and mother tongue will do a different kind of errors when they are learning the other language. The native language identification system will give the targeted feedback to language learners.

Marketing: Categorizing the geographical region and native language of authors who providing the opinions may help to improve the marketing strategies.

Politics: The comments of the user who likes the Govt. policies and whose dislike the policies and the region-specific people opinion can be identified automatically without looking to their profile. Getting the exact profile of the person is difficult in social media. Native language is a part of the user profile. So we need the mechanism to find the native language automatically by analyzing the usage of another language.

Person identification - Fake news identification: Analyzing the Fake news can be helped to find out the which region or native person created the Fake news or threatening messages.

In this overview of the shared task paper, we describe the task and the data sets used, the features and classifiers in which the participants used, the results and its comparisons with INLI-2017. The paper is organized as follows, section 2 explores the related works in the NLI and section 3 describes the task descriptions. Section 4 deals with the statistics of the INLI corpora used in the shared task. Section 5 shows the system descriptions of the participants and the various features used. Section 6 explains the Results and discussions and section 7 concludes the paper.

2 Related Works

Most commonly NLI is done as a supervised classification task, where features are extracted from the text produced by non-native speakers. NLI is a recent,

but rapidly growing, area of research. While some early research was conducted in the early 2000s, most work has only appeared in the last few years. The work of Koppel et al.[11] (2005) was the first in the field and they explored a multitude of features, many of which are employed in several of the systems in the shared tasks. These features included character and POS n-grams, content and function words, as well as spelling and grammatical errors (since language learners have tendencies to make certain errors based on their L1 (Swan and Smith, 2001)). An SVM model was trained on these features extracted from a subsection of the ICLE corpus consisting of 5 L1s. N-gram features (word, character and POS) have figured prominently in prior work. Not only are they easy to compute, but they can be quite predictive. Wong and Dras (2011)[12] utilized character and part-of-speech (POS) n-grams as well as cross-sections of parse trees and Context-Free Grammar (CFG) features, i.e., local trees. Their approach with a binary representation of non-lexicalized rules (except for those rules lexicalized with function words and punctuation) outperformed a setup using only lexical features, such as n-grams, on data from the International Corpus of Learner English (ICLE; Granger et al., 2002)[13]. Swanson and Charniak (2012)[14] used binary feature representations of CFG and Tree Substitution Grammar (TSG) rules replacing terminals (except for function words) by a special symbol. TSG outperformed CFG features in their settings. gs2 being widely noted (Brooke and Hirst, 2012a)[15]. More recently, TOEFL11, the first corpus designed for NLI was released (Blanchard et al., 2013)[16]. While it is the largest NLI dataset available, it only contains argumentative essays, limiting analyses to this genre. Research has also expanded to use non-English learner corpora (Malmasi and Dras, 2014a; Malmasi and Dras, 2014c)[17]. Recently, Malmasi and Dras (2014b)[17] introduced the Chinese Learner Corpus for NLI and their results indicate that feature performance may be similar across corpora and even L1- L2 pairs.

3 Task Descriptions

The shared task is the second version of the INLI-2017[1]. Given an XML file which contains the Facebook comments written in the English language, the task is to identify the native language of the author of comments. The native languages considered in the shared task are Hindi, Tamil, Malayalam, Kannada, Telugu, and Bengali. The highest accuracy obtained in the first shared task INLI-2017[1] is 48.80 which is comparably less. We felt that there are a lot of avenues to improve the performance of the INLI system. So we conducted the second version of the shared task with same Training data set and different test set.

Training Data : In this shared task, the training data set of the INLI-2017[1] shared task is used as it is an extended version of the earlier shared task. Totally, 1233 XML files for 6 Indian natives where each file contains 8 to 10 Facebook

comments written in English. Facebook comments are collected during the period of April 2017 to July 2017.

Test Set-1 : Test set-1 represents the test set used in the earlier INLI-task [1]. Totally 874 XML documents which are also collected in the same period of the Training data set. In order to compare the results of earlier results, we asked participants to test their systems with this test set.

Test Set-2 : Test set-2 the new set which is collected during the period of May 2018 to June 2018. The regional bias comments are removed in order to avoid the Topic bias. Here the author bias also removed so as expected the performance of the participants is comparably less.

The training data was released on 15th May 2018 and the unlabeled testing data set released almost one month later. The training set is categorized to the folders which are named as six Indian languages correspondingly. Each team allowed to submit up to 3 different runs of the test set-1 and test set2. task. This allowed participants to experiment with different variations of their developed system. The participants are only ranked based on the test set 2.

4 Corpora Statistics

Collecting corpora is an important challenge in the INLI. We have collected the comments posted in English on the top regional news pages of Facebook. In order to avoid the topic bias, we removed the comments with the regional flavor. We concentrate only on the comments on national importance like "Budget", "Modi", "BJP" and "Election" etc. The training dataset which is used in the INLI-2018 shared task is the same as the data set used in the INLI-2017. But, the test set is different in which it is collected recently in the time period of May 2018 to June 2018. In order to compare the previous shared task results, the participants are asked to test their systems with the INLI-2017 test set also. The detailed dataset statistics are given in table.1 and 2.

Figure 1 and 2 explain the word cloud of the training data set and testing set. Each language from the training data is represented separately in the word cloud.

Figure.1 shows the top 50 words of the training data set using the word cloud visualization. Tamil, Malayalam, Kannada, and Telugu are spoken in the southern part of India. Bengali is spoken in the eastern part and Hindi is most common in the northern region. Interestingly all the keywords in the Hindi language are present in the all other five languages. So identifying the Hindi native language is difficult compared with other languages.

Each language comments are visualized separately to understand the most frequently used words by the native speakers. The figure shows the common words like "India", Country, People, Modi, money and politics and government. Even though we removed the region-specific words. Some of the posts still reflect the region information. This also depends upon the news item where the

Fig. 1. Top 50 content words of the training data set of INLI corpora.

Fig. 2. Top 50 content words of the testing data set of INLI corpora.

Table 1. INLI Training data statistics

Language	# XML docs	# Sentences	# Words	# Unique Words	Avg. # Words/ XML docs	Avg. # Words/ Sentence	Avg. # Unique Words/ XML docs	Avg. # Unique Words/ Sentence
BE	202	1616	37623	8180	186.3	23.3	40.5	5.1
HI	211	1688	28983	6285	137.4	17.2	29.9	3.7
KA	203	1624	45738	8740	225.3	28.2	43.1	5.4
MA	200	1600	47167	8854	235.8	29.5	44.3	5.5
TA	207	1656	34606	6716	167.2	20.9	32.4	4.1
TE	210	1680	49176	8483	234.1	29.3	40.4	5.0

Table 2. INLI-2018 Test data statistics

Language	#XML docs	# Sentences	# Words	#Unique Words	Avg.# Words/ XML docs	Avg.# Words/ Sentence	Avg.# Unique Words/ XML docs	Avg.# Unique Words/ Sentence
BE	207	1656	23548	5889	113.8	14.2	28.4	3.6
HI	138	1104	22150	6248	160.5	20.1	45.3	5.7
KA	250	2000	39095	9513	156.4	19.5	38.0	4.8
MA	200	1600	27065	8093	135.3	17.0	40.4	5.1
TA	140	1120	17935	5327	128.1	16.0	38.0	4.8
TE	250	2000	44009	11178	176.0	22.0	44.7	5.6

comments have been collected. Compare to other languages the word "farmers" are more in the Tamil region and similarly, most of the border region consists of the word "army".

5 System Descriptions of Participants

In total, 14 teams were submitted their runs. Each team is restricted to 3 runs. Totally, we received 31 submissions from participants for test set-1 and test set-2 data.

Ajees et.al[3] from CUSAT team applied the Convolutional neural network for native language identification. Four convolution layers, three max-pooling layer, and two dense layers were used on the CNN network. Instead of treating the problem as a document classification, they converted to sentence/comment classification where each comment are tagged with the corresponding native language. Each post in the XML file of the test set is tagged in the model. Since we created the training and testing data where each document contains an equal number of comments, the developed model will not be affected based on the number of comments in each document. The maximum number of prediction

for that particular document is considered as a label for the XML document. Bharathi et.al[4] from SSNCSE team used the statistical test based feature selection method for identifying the native language of the document. There have been submitted three runs for the INLI 2018 task. They have used TFIDF as the common initial feature for all the submissions. Each submission is differentiated with the feature selection method and classifier. In the first run, Analysis of Variance (ANOVA) F-values for selecting best features and trained using Multi-Layer Perceptron (MLP) classifier. The second submission is Chi - square value based feature selection method and the MLP classifier. The third submission is with Chi-square based feature selection and trained using Stochastic Gradient Descent (SGD) classifier. For MLP classifier, RELU (Rectified Linear Unit) is used as activation function and Adam optimizer is used for weight optimization. The SGD supports multi-class classification by combining multiple binary classifiers in a one versus rest fashion. Their third submission SGD classifier with Chi-square feature selection methods outperforms the other submission submitted on the shared task. Thenmozhi et.al[5] from SSNNLP team also used the feature selection method with the traditional classifiers for native identification. As a preprocessing, they removed the punctuation and they have not applied the stemming and stop-word removal. To extract the useful features that are contributing to native language identification, they have used Chi-Square feature selection method. They tried with different combinations of features and machine learning classifiers and recorded the cross-validation results. Finally, the MLP classifier with TF-IDF features (without feature selection) and Multinomial Naive Bayes classifier with Chi-Square feature selection methods are submitted for evaluation. The results clearly show that the performance of Chi-Square feature selection method is comparably lesser than the TFIDF features with no feature selection. Soumik Mondal et.al[6] from Corplab team designed an INLI system with TFIDf features and linear SVM classier with three different strategies. In preprocessing they removed the non-ASCII characters and replaced multiple occurrences of some characters like "....." or "sorryyyyyyy" with "." or "sorry". In the first submission, they have used one-vs-rest classifier and in the second and third submission, they have used the Pairwise Coupling strategies proposed in [6]. Ian markov et.al[7] from CIC-IPN team proposed a system with the SVM classifier on rich feature set including the emotion-based features. They used the word and character n-grams, part-of-speech (POS) tag n-grams, character n-grams from misspelled words, punctuation mark n-grams, and emotion-based features. The features are weighted using the log-entropy weighting scheme . They have used the Emotion polarity features similar to the features proposed in [16]. The well-known NRC emotion lexicon [17] is used in the features. Aman Gupta[8] from Team WebArch proposed a system using n-gram based TFIDF features extracted the given data set and trained with logistic regression. He divided the training data set into train, test and validation data set. He has calculated the validation accuracy, for with stop words and without stop words, different n-grams and TFIDF and Count Vectorizer. Rajesh Kumar et.al[9] from NLPRL team developed an INLI system using Hybrid

gated LSTM-CNN. The Glove pre-trained word embeddings are used to find the initial level word representation of tokens in the sentences. The word level input is converted into sentence level input by using a bidirectional LSTM. This is achieved by linearly combining the last hidden state of forwarding and backward LSTM. The entire network is trained by Adam optimizer with epoch and mini-batch size of 15 and 10 respectively. The proposed model retrieved more relevant documents for the Tamil language as compared to other languages during the testing phase. For Hindi and Tamil language, the proposed model achieves highest F1-score for Test set1 data. Ashish Patel et.al[10] from IIITV team proposed a Hyper-dimensional Computing (HDC) as a supervised learning model for identifying Indian Native Language from the user's social media comments written in English. HDC represents language features as high dimensional vectors called hyper vectors. Initially, comments are broken in character bi-grams and tri-grams which are used for generating comment hyper vectors. These hyper vectors are further combined to create different language profile vectors. Profile hyper vectors are then used for classification of test comments. They have removed of non-English characters, special characters and converting the text in lowercase (alphabets). Hamada et.al[2] from Mangalore University team used Artificial Neural Network (ANN) model and Ensemble approach. The traditional TFIDF features have been used to represent comments. The ANN-based classifier is designed for the first and second submissions. The hidden layer of the first submission contains 70 neurons and the second submission contains 80 neurons and the activation function is the logistic function. Ensemble approach using majority voting technique has been used in the third submission.

Table 3 explains the various features used by the participating teams. Table 4 shows the important preprocesing techniques, feature selection methods and classifiers used by the participants.

Table 3: Features

Features	TFIDF	S-Words	Word-ng	Char-ng	POS-ng	Emb	others
CUSAT						CNN	CNN
SSNc	✓						
SSNn	✓						
CorpLab	✓						
CIC			✓	✓	✓		Emotion
WebArch	✓	✓	✓				
NLPRL						Glove	LSTM
IIITV				✓			HDC
MU	✓						ANN

Table 4. Feature Selection and Classifier

Features	Preprocessing	Feature Selection	Classifier
CUSAT	-	CNN	Softmax
SSNc	-	S-test, Chi-Squ	MLP, SGD
SSNn	Punct	Chi-Squ	MNP, MLP
CorpLab	non-ASCII	-	SVM
CIC	-	-	SVM
WebArch	-	-	LR
NLPRL	-	-	LSTM
IIITV	non-Eng, Lowercase	HDC	-
MU	-	-	ANN

Table 5. Result Comparision

INLI 2017 Test set1 Results				***INLI 2018 Test set2 Results***		
Team	Run	Accuracy	Rank	Run	Accuracy	Rank
SSNCSE	1	44.1	1	1	35.4	1
	2	42.9		2	36.8	
	3	**46.2**		3	**37.0**	
MANGALORE	1	**46.6**	2	1	**35.3**	2
	2	45.5		2	35.3	
	3	46.6		3	35.3	
CIC-IPN	1	**41.8**	3	1	34.1	3
	2	41.3		2	34.4	
	3	41.4		3	**34.5**	
Baseline	-	**43.0**	-	-	**34.0**	-
SSN_NLP	1	**46.1**	4	1	**34.3**	4
	2	32.4		2	28.4	
WebArch	1	**41.4**	5	1	**31.9**	5
	2	28.2		2	21.7	
	3	29.8		3	21.9	
CorpLab	1	**42.1**	6	1	**31.8**	6
	2	**39.8**		2	30.8	
	3	40.4		3	31.5	
IITV	1	32.4	7	1	29.2	7
	2	**31.1**		2	**31.5**	
teamJason	1	22.2	8	1	24.5	8
	2	**31.7**		2	**30.5**	
OscarGaribo	1	35.1	9	1	29.5	9
	2	**36.0**		2	**29.6**	
Leorius	1	**31.5**	10	1	**29.0**	10
CUSAT	1	14.0	11	1	**24.1**	11
	2	**15.2**		2	20.9	
	3	10.7		3	21.8	
DNLP	1	**29.6**	12	1	**22.9**	12
IDRBT	1	14.8	13	1	**19.7**	13
	2	**19.7**		2	18.0	
NLPRL	1	**15.3**	14	1	**17.1**	14

6 Results and Discussions

The participants are asked to test their systems with two test sets. The accuracy of the first and second INLI shared task is given in the Table.5. The highest accuracy of INLI-2017 is 46.6 %. The highest accuracy of the same data set in the second shared task is 37.0 % which is less compare to the previous shared task. Table.6 describes the test set-1 results in the INLI shared task 2018. The highest accuracy is achieved by the TFIDF features and ANN classifier.

For the test set-2, the highest accuracy is achieved by SSN_CSE team. They have tried the TFIDF features and feature selection methods with MLP and SGD classifier.

Most of the teams tried the conventional TFIDF features. Teams are not considered the socio-linguistic features and preprocessing methods. As expected deep learning methods are not dominating the traditional methods due to the size of the training data set. Feature selection method on top of TFIDF shows the improvement over other methods.

The reasons for less accuracy of the shared task is as follows, The data set size is very small, which is one of the reasons that the accuracy of the participant's system not performed at the expected level. Facebook comments are also small in size compared to the essays which are used in the NLI shared task. The topic bias comments are removed, in order to give attention to only on the writing style of the user, which are the main evidence for identifying the native of the user.

7 Conclusions

For any language processing task collecting annotated corpora is the challenging part. The training data set of the INLI 2018 is same as the 2017 shared task data set. The dataset collection is based on the assumption that, only native speakers will read native language newspapers. Code-mixed comments and comments related to the regional topics were removed from the corpus, and comments with common keywords discussed across the regions were considered in order to avoid possible topic biases. To the best of our knowledge, this is the first corpus for native language identification for Indian languages. The participants used different feature sets to address the problem: content-based (among others: bag of words, character n-grams, word n-grams, term vectors, word embedding, non-English words) and stylistic-based (among others: words frequency, POS n-grams, noun and adjective POS tag counts). Participants have used hybrid gated LSTM-CNN, ANN etc and some have used Glove pre trained word embeddings. Overall the best performance system obtained an accuracy of 46.6%, which is 3.6% greater than the baseline. Overall three of the systems performed better than the baseline. These systems have used the bag of word features which are extracted from the text posted by the user and the feature vectors are constructed using TF-IDF score for the training data and Artificial Neural Network (ANN) model and Ensemble approaches. The smallest overall accuracy

was 15.2%, which is 27.8% less than the baseline. As future work, we believe that native language identification should be addressed taking into account also socio-linguistics features to improve further.

References

1. Anand Kumar, M., et al. "Paolo Rosso. 2017. Overview of the INLI PAN at FIRE-2017 Track on Indian Native Language Identification." Notebook Papers of FIRE (2017): 8-10.
2. Hamada et.al."Artificial Neural Network and Ensemble Based Models for INLI".In Working notes of FIRE 2018 - Forum for Information Retrieval Evaluation. Gandhinagar, 6th - 9th December
3. Ajees et.al."A Native Language Identification System using Convolutional Neural Networks",In Working notes of FIRE 2018 - Forum for Information Retrieval Evaluation. Gandhinagar, 6th - 9th December
4. Bharathi et.al ."Statistical testing based feature selection for Native Language Identification",In Working notes of FIRE 2018 - Forum for Information Retrieval Evaluation.Gandhinagar ,6th - 9th December
5. Thenmozhi et.al ."A Machine Learning Approach to Indian Native Language Identi cation",In Working notes of FIRE 2018 - Forum for Information Retrieval Evaluation. Gandhinagar ,6th - 9th December
6. Soumik Mondal et.al ."Identification of Indian Native Language using Pairwise Coupling", In Working notes of FIRE 2018 - Forum for Information Retrieval Evaluation.Gandhinagar, 6th - 9th December
7. Ilia Markov et.al ."CIC-IPN@INLI2018: Indian Native Language Identi cation", In Working notes of FIRE 2018 - Forum for Information Retrieval Evaluation. Gandhinagar ,6th - 9th December
8. Aman Gupta ."Team WebArch at FIRE-2018 Track on Indian Native Language Identification", In Working notes of FIRE 2018 - Forum for Information Retrieval Evaluation. Gandhinagar ,6th - 9th December
9. Mundotiya et.al ."NLPRL@INLI-2018: Hybrid gated LSTM-CNN model for Indian native language identification", In Working notes of FIRE 2018 - Forum for Information Retrieval Evaluation. Gandhinagar ,6th - 9th December
10. Ashish Patel et.al ."IIITV@INLI-2018 : Hyperdimensional Computing for Indian Native Language Identification", In Working notes of FIRE 2018 - Forum for Information Retrieval Evaluation. Gandhinagar, 6th - 9th December
11. Koppel, Moshe, Jonathan Schler, and Kfir Zigdon. "Determining an author's native language by mining a text for errors." Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, 2005.
12. Wong, Sze-Meng Jojo, and Mark Dras. "Exploiting parse structures for native language identification." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.
13. Granger, Sylviane, Joseph Hung, and Stephanie Petch-Tyson, eds." Computer learner corpora, second language acquisition, and foreign language teaching." Vol. 6. John Benjamins Publishing, 2002.
14. Swanson, Ben, and Eugene Charniak. "Native language detection with tree substitution grammars." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics, 2012.

15. Brooke, Julian, and Graeme Hirst. "Native language detection with 'cheap'learner corpora." Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011). Vol. 1. Presses universitaires de Louvain, 2013.
16. Tetreault, Joel, Daniel Blanchard, and Aoife Cahill. "A report on the first native language identification shared task." Proceedings of the eighth workshop on innovative use of NLP for building educational applications. 2013.
17. Malmasi, Shervin, and Mark Dras. "Language identification using classifier ensembles." Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects. 2015.