Overview of the INLI PAN at FIRE-2017 Track on Indian Native Language Identification

Anand Kumar M, Barathi Ganesh HB,
Shivkaran Singh and Soman KP
Center for Computational Engineering
and Networking (CEN)
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India

Paolo Rosso
PRHLT Research Center,
Universitat Politècnica de València,
Spain

ABSTRACT

This overview paper describes the first shared task on Indian Native Language Identification (INLI) that was organized at FIRE 2017. 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, 26 approaches of 13 different teams are evaluated. In this paper, we give an overview of the approaches and discuss the results that they have obtained.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Language resources**; **Feature selection**;

KEYWORDS

Author Profiling, Indian Languages, Native Language Identification, Social Media, Text Classification

1 INTRODUCTION

Native Language Identification (NLI) is a fascinating and rapidly growing sub-field in Natural Language Processing. In the framework of the author profiling shared tasks that have been organized at PAN[1], language variety identification was addressed in 2017 at CLEF [17]. NLI requires instead to automatically identify the native language (L1) of an author on the basis of the way she writes in another language (L2) that she learned. As her accent may help in identifying whether or not she is a native speaker in that language L1, in a similar way the way the language is used when she writes may unveil patterns that can help in identifying her native language [19]. From a cybersecurity viewpoint, NLI can help to determine the native language of an author of a suspicious or threatening text.

The native language influences the usage of words as well the errors that a person makes when writing in another language [19]. NLI systems can identify the writing patterns that are based on the author's linguistic background. NLI has many applications and studying the language transfer from a forensic linguistics viewpoint is certainly one of the most important. The first shared task on native language identification was organized in 2013 [21]. The organizers made available a large text corpus for this task. Other works approach the problem of native language identification using as well speech transcripts [30]. In the Indian languages context, this is the first NLI shared task. In India there are currently 22 official languages with English as an additional official language. In this shared task, we focus on identifying the native language of Indian authors writing comments in English. We considered six languages, namely, Bengali, Hindi, Kannada, Malayalam, Tamil and Telugu for the shared task.

Since comments over the internet are usually written in social media, the corpora used for the shared task was acquired from Facebook. English comments from Facebook pages of famous regional language newspapers were crawled. These comments were further preprocessed in order to remove code-mixed and mixed scripts comments from the corpus. In the following sections we present some related work (Section 2), we describe the corpus collection (Section 3), we give an overview of the submitted approaches (Section 4), finally we show the results that were obtained (Section 5). Finally, in Section 6 we draw some conclusions.

2 RELATED WORK

As said in [14], one of the earliest works on identifying native language was by Tomokiyo and Jones (2001) [23] where the author used Naive Bayes to discriminate non-native from native statements in English. Koppel et. al (2005) [25] approached the problem by using stylistic, syntactic and lexical features. They also noticed that the use of character n-grams, parts of speech bi-grams and function words allowed to obtain better results. Tsur and Rappoport (2007) [11] achieved an accuracy of about 66% by using only character bi-grams. They assumed that the native language phonology influences the choice of words while writing in a second language.

Estival et. al [8] used English emails of authors with different native languages. They achieved an accuracy of 84% using a Random Forest classifier with character, lexical, and structural features. Wong and Dras [27] pointed out that mistakes made by authors writing in a second language is influenced by their native language. They proposed the use of syntactic features such as subject-verb disagreement, noun-number disagreement, and improper use of determiners to help in determining the native language of a writer. In their later work [28], they also investigated the usefulness of parse structures for identifying the native language. Brooke and Hirst [4] used word-to-word translation of L1 to L2 to create a mappings which are the result of language transfer. They use this information in their unsupervised approach.

Torney et. al [24] used psycho-linguistic feature for NLI. Syntactic features showed also to play a significant role in determining the native language. Other interesting studies in the NLI field are [29]

[1] http://pan.webis.de

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 1: 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	185	1480	26653	5647	144.1	18.0	30.5	3.8
HI	251	2008	37232	6616	148.3	18.5	26.4	3.3
KA	74	592	12225	3477	165.2	20.7	46.9	5.9
MA	92	736	16805	4658	182.7	22.8	50.6	6.3
TA	100	800	14780	4192	147.8	18.5	41.9	5.2
TE	81	648	14692	3989	181.4	22.7	49.2	6.2

Table 2: Test data statistics

[20] [5]. In 2013 a shared task was organized on NLI [20]. The organizers provided a large corpus which allowed comparison among different approaches. In 2014 a related shared task was organized on Discriminating between Similar Languages (DSL[2]) [31]. The organizers provided six groups of 13 different languages, with each group having similar languages. In 2017 another shared task on NLI was organized. The corpus was composed by essays and transcripts of utterances. The ensemble methods and meta-classifiers with syntactic/lexical features were the most effective systems [15].

3 INLI-2017 CORPUS

Many corpora have been created from social media (Facebook, Twitter and WhatsApp) for performing language modeling [9], information retrieval tasks [6], and code-mixed sentiment analysis [10]. A monolingual corpus based on the TOEFL[3] data is available for performing the NLI task for Indian languages such as Hindi and Telugu [16]. The INLI-2017 corpus includes English comments of Facebook users, whose native language is one among the following: Bengali (BE), Hindi (HI), Kannada (KA), Malayalam (MA), Tamil (TA) and Telugu (TE). The dataset collection is based on the assumption that, only native speakers will read native language newspapers. To the best of our knowledge, this is the first corpus for native language identification for Indian languages. The detailed corpus statistics are given in Table 1 and Table 2.

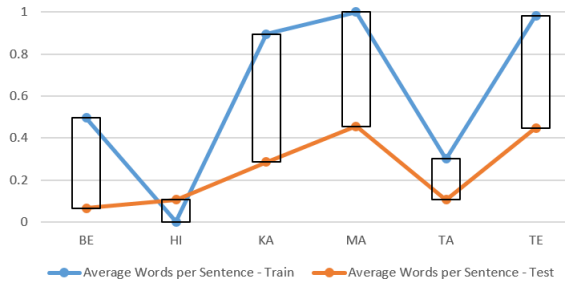

Figure 1: Variance b/w training and test corpus

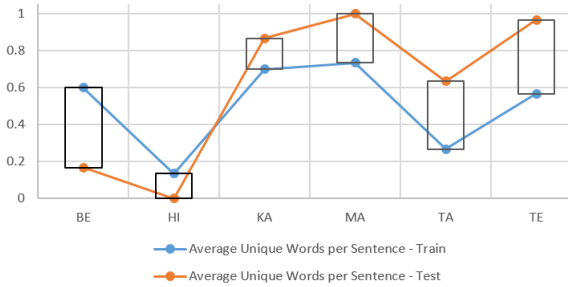

Figure 2: Variance b/w training and test corpus

[2] http://corporavm.uni-koeln.de/vardial/sharedtask.html
[3] https://www.ets.org/toefl

The texts for this corpus have been collected from the users comments in the regional newspapers and news channel Facebook pages. Around 50 Facebook pages were selected and comments written in English were extracted from these pages. The training data have been collected in the period of April-2017 to July 2017. The test data has been collected later on. It was expected that participants will focus on native language-based stylistic features. As a result, we removed code-mixed comments and comments related to the regional topics (regional leaders and comments mentioning the name of regional places). Comments with common keywords discussed across the regions were considered to avoid the topic bias. These common keywords observed were Modi, note-ban, different sports personalities, army, national issues, government policies, etc. Finally, the collected dataset was randomized and written to XML files randomly to avoid user bias.

From Table 1 and Table 2, it can be observed that except for BE and MA, the remaining languages have nearly the same ratio of average words per sentence. It is also visible that the test data was properly normalized in order to have the average words per sentence and average unique words per sentence. The variance between average of words per sentence and average of unique words per sentence for the training and the test data is shown in Figure 1 and Figure 2, respectively. This corpus will be made available after the FIRE 2017 conference in the web page of our NLP group website[4].

4 OVERVIEW OF THE SUBMITTED APPROACHES

Initially, 56 teams registered at the INLI shared task at FIRE, and finally 13 of them submitted a total of 26 runs. Moreover, 8 of them submitted their system description working notes[5]. We analysed their approaches from three perspectives: preprocessing, features to represent the author's texts and classification approaches.

4.1 Preprocessing

Most of the participants have not done any preprocessing [2, 7, 13, 18, 26]. Text is normalised by removing the emoji, special characters, digits, hash tags, mentions and links [1, 12, 22]. Stop words are removed using the nltk stop words package[6], other resources[7] and manual stop words collection [1]. White space based tokenization has been carried out by all other participants except [7]. The participant [22] handled the shortened words (terms such as n't, &, 'm, 'll are replaced as 'not', 'and', 'am', and 'will' respectively).

4.2 Features

Two of the participants directly used the Term Frequency Inverse Document Frequency (TFIDF) weighs as their features [1, 2], non-English words and noun-chunks are taken as the features while computing TFIDF [18], character n-grams of order 2-5 and word n-grams of order 1-2 have been used as features while computing the TFIDF vocabulary [7, 12, 13]. Only the non-English word counts have been taken as features in [26]. Nouns and adjective words have been taken as feature in [22]. Part of Speech n-grams, average word and sentence length have been used as the features in [7]. Distributional representation of words (pre-trained word vectors) have been used in [7].

Team	Run	P	R	F	Rank
IDRBT	1	**96.4**	**57.3**	**71.9**	1
MANGALORE	1	56.5	79.5	66.1	2
	2	54.0	84.9	66.0	
	3	**59.2**	**78.4**	**67.4**	
DalTeam	1	**56.2**	**83.2**	**67.1**	3
SEERNET	1	59.4	70.3	64.4	3
	2	57.6	74.1	64.8	
	3	**60.7**	**75.1**	**67.1**	
Baseline	-	**58.0**	**79.0**	**67.0**	-
Bharathi_SSN	1	**50.3**	**80.5**	**62.0**	4
SSN_NLP	1	**46.2**	**76.2**	**57.6**	5
Anuj	1	**56.6**	**50.8**	**53.6**	6
	2	56.5	47.0	51.3	
	3	45.5	18.9	26.7	
ClassyPy	1	**67.9**	**40.0**	**50.3**	7
	2	66.7	40.0	50.0	
	3	40.6	22.2	28.7	
DIG (IIT-Hyd)	1	55.2	45.9	50.1	7
	2	**55.6**	**45.9**	**50.3**	
	3	45.5	10.8	17.5	
Bits_Pilani	1	39.7	15.7	22.5	8
	2	**56.3**	**38.4**	**45.7**	
	3	39.4	23.2	29.3	
BMSCE_ISE	1	40.0	29.2	33.8	9
	2	**38.9**	**55.1**	**45.6**	
JUNLP	1	**8.3**	**7.0**	**7.6**	10
team_CEC	1	**0.0**	**0.0**	**0.0**	11

Table 3: BE-NLI results

4.3 Classification Approaches

Support Vector Machine (SVM) has been used as a classifier by most of the participants [1, 2, 7, 12, 13]. Two of the participants followed the ensemble based classification with Multinomial Bayes, SVM and Random Forest Tree as the base classifiers in [22] and Logistic Regression, SVM, Ridge Classifier and Multi-Layer Perceptron (MLP) as the base classifiers in [18]. Other than this the authors in [7] used the Logistics Regression, authors in [26] used Naive Bayes, authors in [3] used hierarchical attention architecture with bidirectional Gated Recurrent Unit (GRU) cell and authors in [22] employed the neural network classifier with 2 hidden layers, Rectified Linear Unit (ReLU) as the activation function and Stochastic Gradient Descent (SGD) as the optimizer.

[4] http://nlp.amrita.edu:8080/nlpcorpus.html
[5] ClassPy team did not submit any working notes, although a brief description of the approach was sent by email.
[6] http://www.nltk.org/book/ch02.html
[7] pypi.python.org/pypi/stop-words

Team	Run	P	R	F	Rank
team_CEC	1	**32.1**	**100.0**	**48.6**	1
Anuj	1	59.8	19.5	29.4	2
	2	52.4	17.5	26.3	
	3	**41.8**	**27.5**	**33.2**	
JUNLP	1	**26.1**	**37.8**	**30.9**	3
DIG (IIT-Hyd)	1	49.5	19.1	27.6	4
	2	**50.0**	**19.5**	**28.1**	
	3	34.1	11.6	17.3	
SSN_NLP	1	**49.4**	**16.3**	**24.6**	5
ClassyPy	1	**50.6**	**15.5**	**23.8**	6
	2	43.7	15.1	22.5	
	3	30.2	11.6	16.7	
DalTeam	1	**69.2**	**14.3**	**23.8**	6
Bits_Pilani	1	**24.0**	**19.5**	**21.5**	7
	2	23.9	6.8	10.6	
	3	19.0	8.8	12.0	
Baseline	-	**57.0**	**13.0**	**21.0**	-
SEERNET	1	**50.0**	**9.6**	**16.1**	8
	2	50.0	8.4	14.3	
	3	54.8	9.2	15.7	
MANGALORE	1	60.7	6.8	12.2	9
	2	**60.0**	**7.2**	**12.8**	
	3	66.7	4.8	8.9	
BMSCE_ISE	1	50.0	0.8	1.6	10
	2	**54.5**	**7.2**	**12.7**	
Bharathi_SSN	1	**51.9**	**5.6**	**10.1**	11
IDRBT	1	**25.0**	**0.4**	**0.8**	12

Table 4: HI-NLI results

Team	Run	P	R	F	Rank
DalTeam	1	**40.5**	**66.2**	**50.3**	1
SEERNET	1	**38.1**	**71.6**	**49.8**	2
	2	37.1	62.2	46.5	
	3	37.0	68.9	48.1	
Baseline	-	**39.0**	**64.0**	**48.0**	-
IDRBT	1	**40.0**	**59.5**	**47.8**	3
MANGALORE	1	**38.4**	**58.1**	**46.2**	4
	2	**40.4**	**54.1**	**46.2**	
	3	34.8	64.9	45.3	
Bharathi_SSN	1	**33.3**	**64.9**	**44.0**	5
SSN_NLP	1	**39.6**	**48.6**	**43.6**	6
Bits_Pilani	1	**30.4**	**45.9**	**36.6**	7
	2	26.0	45.9	33.2	
	3	20.8	59.5	30.8	
ClassyPy	1	22.2	77.0	34.4	8
	2	**23.7**	**77.0**	**36.2**	
	3	19.7	60.8	29.7	
DIG (IIT-Hyd)	1	**21.8**	**59.5**	**31.9**	9
	2	21.7	59.5	31.8	
	3	21.1	40.5	27.8	
Anuj	1	19.4	40.5	26.2	10
	2	**20.3**	**41.9**	**27.3**	
	3	27.5	14.9	19.3	
BMSCE_ISE	1	11.7	27.0	16.3	11
	2	**19.0**	**44.6**	**26.6**	
JUNLP	1	**17.9**	**13.5**	**15.4**	12
team_CEC	1	**0.0**	**0.0**	**0.0**	13

Table 5: KA-NLI results

5 EXPERIMENTS AND RESULTS

Accuracy was used as measure to evaluate the performance of the systems[8]. In baseline system, the Term Frequency-Inverse Document Frequency (TF-IDF) features with SVM linear kernel and default parameters were used.

Each team was allowed to submit up to three systems. For the final ranking the best performing system is considered. We have evaluated 26 submissions from 13 participants. The submissions are ranked per language and the final ranking is based on the overall accuracy of the system across all the languages.

The ranking of the submitted systems for Bengali (BE) is given in Table 3. The maximum F-measure scored for this language is 71.9%, which is 4.9% greater than the base line system. The lowest F-measure scored for this language is 7.6% and this is 59.4% less than the baseline. Among the all the other languages, this is the highest variation with respect to the baseline.

The ranking of the systems submitted for Hindi (HI) is given in Table 4. The maximum F-measure scored for this language is 48.6%, which is 27.6% higher than the baseline. The lowest F-measure scored for this language is 0.8% and this is 20.8% less than the baseline. This is the lowest F-measure across all the languages.

The ranking of the submitted systems for Kannada (KA) is given in Table 5. The maximum F-measure scored for this language is 50.3%, which is 2.3% greater than the baseline. The lowest F-measure scored for this language is 15.4% and this is 32.6% less than the baseline.

The ranking of the systems submitted for Malayalam (MA) is given in Table 6. The maximum F-measure scored for this language is 51.9%, which is 0.9% greater than the baseline. Among the all the other languages, this is the lowest variation with respect to the baseline. The lowest F-measure scored for this language is 1.8% and this is 49.2% less than the baseline.

The ranking of the submitted systems for Tamil (TA) is given in Table 7. The maximum F-measure scored for this language is 58.0%, which is 12.0% greater than the baseline. The lowest F-measure scored for this language is 13.2% and this is 32.8% less than the baseline.

The ranking of the systems submitted for Telugu (TE) is given in Table 8. The maximum F-measure scored for this language is 50.5%, which is 8.5% greater than the baseline system. The lowest F-measure scored for this language is 2.4% and this is 39.6% less than baseline.

The results rank per language is given in Table 9. The team_CEC has not identified any language apart from Hindi. The overall ranking for the submitted systems are given in Table 10. The maximum accuracy scored is 48.8%, which is 5.3% greater than the baseline.

[8] http://www.nltk.org/_modules/nltk/metrics/scores.html

Team	Run	P	R	F	Rank
MANGALORE	1	40.4	70.7	51.4	1
	2	**42.7**	**66.3**	**51.9**	
	3	32.6	78.3	46.0	
Baseline	-	**42.0**	**65.0**	**51.0**	-
DalTeam	1	**46.7**	**54.3**	**50.3**	2
SEERNET	1	38.5	59.8	46.8	3
	2	**41.0**	**64.1**	**50.0**	
	3	39.5	53.3	45.4	
Bharathi_SSN	1	**36.4**	**60.9**	**45.5**	4
ClassyPy	1	**34.3**	**53.3**	**41.7**	5
	2	34.5	52.2	41.6	
	3	33.7	31.5	32.6	
Anuj	1	48.4	33.7	39.7	6
	2	**51.7**	**32.6**	**40.0**	
	3	26.7	21.7	24.0	
DIG (IIT-Hyd)	1	**37.9**	**39.1**	**38.5**	7
	2	37.5	39.1	38.3	
	3	21.4	19.6	20.5	
Bits_Pilani	1	20.0	28.3	23.4	8
	2	15.5	31.5	20.8	
	3	**39.4**	**34.6**	**36.8**	
IDRBT	1	**18.1**	**84.8**	**29.9**	9
BMSCE_ISE	1	**17.3**	**64.1**	**27.2**	10
	2	22.3	31.5	26.1	
SSN_NLP	1	**31.7**	**21.7**	**25.8**	11
team_CEC	1	**100.0**	**1.1**	**2.2**	12
JUNLP	1	**5.3**	**1.1**	**1.8**	13

Table 6: MA-NLI results

Team	Run	P	R	F	Rank
MANGALORE	1	**58.0**	**58.0**	**58.0**	1
	2	**58.0**	**58.0**	**58.0**	
	3	54.4	49.0	51.6	
SEERNET	1	**50.4**	**59.0**	**54.4**	2
	2	47.9	57.0	52.1	
	3	46.8	59.0	52.2	
Bharathi_SSN	1	**48.6**	**51.0**	**49.8**	3
DalTeam	1	**51.1**	**48.0**	**49.5**	4
Baseline	-	**42.0**	**50.0**	**46.0**	-
ClassyPy	1	**41.0**	**41.0**	**41.0**	5
	2	38.7	43.0	40.8	
	3	30.4	41.0	34.9	
Anuj	1	**28.3**	**63.0**	**39.0**	6
	2	27.3	66.0	38.6	
	3	14.3	57.0	22.9	
DIG (IIT-Hyd)	1	**33.3**	**45.0**	**38.3**	7
	2	32.8	44.0	37.6	
	3	17.6	74.0	28.4	
SSN_NLP	1	**27.5**	**49.0**	**35.3**	8
Bits_Pilani	1	**26.6**	**37.0**	**31.0**	9
	2	22.5	39.0	28.6	
	3	21.5	40.0	28.0	
BMSCE_ISE	1	53.3	8.0	13.9	10
	2	**21.8**	**26.0**	**23.7**	
IDRBT	1	**81.2**	**13.0**	**22.4**	11
JUNLP	1	**10.2**	**19.0**	**13.2**	12
team_CEC	1	**0.0**	**0.0**	**0.0**	13

Table 7: TA-NLI results

The lowest accuracy scored is 17.8% and this is 25.2% less than the baseline.

6 CONCLUSION

In this paper we presented the INLI2017 corpus, we briefly described the approaches of the 13 teams that participated at the Indian Native Language Identification task at FIRE 2017, and the results that they obtained. The participants had to identify the native language of the authors of English comments collected from various newspaper pages and television pages in Facebook. Six have been the native languages that have been addressed: Bengali, Hindi, Kannada, Malayalam, Tamil and Telugu. 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.

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). A two layer based neural networks with document vectors built from TFIDF and Recurrent Neural Networks (RNN) with word embedding have been used from the field of deep learning. However, deep learning approaches obtained lower accuracy than the baseline.

Overall the best performance system obtained an accuracy of 48.8%, which is 5.8% greater than the baseline. Overall four of the systems performed better than the baseline. These systems have used the following features: character and word n-grams, non-English words, and noun chunks. It is notable that all these systems have used TFIDF for representing the features. The smallest overall accuracy was 17.8%, which is 25.2% less than the baseline. Among the top performing systems, two of them used an ensemble method and all the systems employed SVM. As future work, we believe that native language identification should be addressed taking into account also socio-linguistics features to improve further.

ACKNOWLEDGEMENT

Our special thanks goes to F. Rangel, all of INLI's participants, students in Computational Engineering and Networking Department for their efforts and time in developing INLI-2017 corpus. The work of the last author was in the framework of the SomEMBED TIN2015-71147-C2-1-P MINECO research project.

REFERENCES

[1] Hamada A. Nayel and H. L. Shashirekha. 2017. Indian Native Language Identification using Support Vector Machines and Ensemble Approach.. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, 8th - 10th December.*

[2] B. Bharathi, M. Anirudh, and J. Bhuvana. 2017. SVM based approach for Indian native language identification.. In *Working notes of FIRE 2017 - Forum for*

Team	Run	P	R	F	Rank
IDRBT	1	**43.4**	**60.5**	**50.5**	1
SEERNET	1	37.6	54.3	44.4	2
	2	37.1	53.1	43.7	
	3	**37.1**	**56.8**	**44.9**	
ClassyPy	1	**40.2**	**48.1**	**43.8**	3
	2	39.4	45.7	42.3	
	3	30.1	50.6	37.8	
Baseline	-	**40.0**	**44.0**	**42.0**	-
DalTeam	1	**33.3**	**55.6**	**41.7**	4
MANGALORE	1	**32.8**	**49.4**	**39.4**	5
	2	32.5	48.1	38.8	
	3	39.4	34.6	36.8	
Anuj	1	**34.4**	**39.5**	**36.8**	6
	2	32.6	37.0	34.7	
	3	28.6	9.9	14.7	
Bits_Pilani	1	28.8	44.4	35.0	7
	2	30.5	35.8	33.0	
	3	**43.6**	**29.6**	**35.3**	
Bharathi_SSN	1	**40.4**	**28.4**	**33.3**	8
DIG (IIT-Hyd)	1	29.0	35.8	32.0	9
	2	**29.3**	**35.8**	**32.2**	
	3	57.1	4.9	9.1	
BMSCE_ISE	1	**26.7**	**38.3**	**31.5**	10
	2	15.4	12.3	13.7	
SSN_NLP	1	**27.0**	**21.0**	**23.6**	11
JUNLP	1	**100.0**	**1.2**	**2.4**	12
team_CEC	1	**0.0**	**0.0**	**0.0**	13

Table 8: TE-NLI results

Team	Rank					
	BE	HI	KA	MA	TA	TE
DalTeam	3	6	1	2	4	4
MANGALORE	2	9	4	1	1	5
SEERNET	3	8	2	3	2	2
BharathiSSN	4	11	5	4	3	8
SSN_NLP	5	5	6	11	8	11
ClassyPy	7	6	8	5	5	3
Anuj	6	2	11	6	6	6
IDRBT	1	12	3	9	11	1
DIG (IIT-Hyd)	7	4	9	7	7	9
team_CEC	11	1	13	12	13	13
Bits_Pilani	8	7	7	8	9	7
BMSCE_ISE	9	10	10	10	10	10
JUNLP	10	3	12	13	12	12

Table 9: NLI results rank per language

Team	Run	Accuracy	Rank
DalTeam	1	**48.8**	1
MANGALORE	1	**47.3**	2
	2	47.6	
	3	45.2	
SEERNET	1	46.6	3
	2	46.4	
	3	**46.9**	
Bharathi_SSN	1	**43.6**	4
Baseline	-	**43.0**	-
SSN_NLP	1	**38.8**	5
ClassyPy	1	**38.2**	6
	2	37.9	
	3	28.9	
Anuj	1	**38.2**	6
	2	36.8	
	3	25.5	
IDRBT	1	**37.2**	7
DIG (IIT-Hyd)	1	**36.7**	8
	2	**36.7**	
	3	22.3	
team_CEC	1	**32.2**	9
Bits_Pilani	1	26.9	10
	2	**28.0**	
	3	26.7	
BMSCE_ISE	1	22.2	11
	2	**27.8**	
JUNLP	1	**17.8**	12

Table 10: Overall results

Information Retrieval Evaluation, Bangalore, India, 8th - 10th December.

[3] Rupal Bhargava, Jaspreet Singh, Shivangi Arora, and Yashvardhan Sharma. 2017. Indian Native Language Identification using Deep Learning.. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, 8th - 10th December.*

[4] Julian Brooke and Graeme Hirst. 2012. Measuring Interlanguage: Native Language Identification with L1-influence Metrics.. In *LREC*. 779–784.

[5] Serhiy Bykh and Detmar Meurers. 2014. Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization.. In *COLING*. 1962–1973.

[6] Kunal Chakma and Amitava Das. 2016. Cmir: A corpus for evaluation of code mixed information retrieval of hindi-english tweets. *Computación y Sistemas* 20, 3 (2016), 425–434.

[7] Christel and Mike. 2016. Participation at the Indian Native language Identification task.

[8] Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*. 263–272.

[9] Anupam Jamatia, Björn Gambäck, and Amitava Das. 2016. Collecting and Annotating Indian Social Media Code-Mixed Corpora. In *the 17th International Conference on Intelligent Text Processing and Computational Linguistics*. 3–9.

[10] Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. 2016. Towards Sub-Word Level Compositions for Sentiment Analysis of Hindi-English Code Mixed Text.. In *COLING*. 2482–2491.

[11] Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. *Intelligence and Security Informatics* (2005), 41–76.

[12] Dijana Kosmajac and Vlado Keselj. 2017. Native Language Identification using SVM with SGD Training.. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, 8th - 10th December.*

[13] Sowmya Lakshmi B S and hambhavi B R. 2017. A simple n-gram based approach for Native Language Identification: FIRE NLI shared task 2017.. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, 8th - 10th December.*

[14] Shervin Malmasi. 2016. Native language identification: explorations and applications. *Sydney, Australia: Macquarie University* (2016). http://hdl.handle.net/1959.14/1110919

[15] Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A Report on the 2017 Native Language Identification Shared Task. In *Proceedings of the 12th Workshop*

on Innovative Use of NLP for Building Educational Applications. 62–75.

[16] Sergiu Nisioi, Ella Rabinovich, Liviu P Dinu, and Shuly Wintner. 2016. A Corpus of Native, Non-native and Translated Texts.. In *LREC*.

[17] Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working Notes Papers of the CLEF* (2017).

[18] Venkatesh Duppada Royal Jain and Sushant Hiray. 2017. Hierarchical Ensemble for Indian Native Language Identification.. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, 8th - 10th December.*

[19] Bernard Smith. 2001. *Learner English: A teacher's guide to interference and other problems*. Ernst Klett Sprachen.

[20] Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native tongues, lost and found: Resources and empirical evaluations in native language identification. *Proceedings of COLING 2012* (2012), 2585–2602.

[21] Joel R Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A Report on the First Native Language Identification Shared Task.. In *BEA@ NAACL-HLT*. 48–57.

[22] D. Thenmozhi, Kawshik Kannan, and Chandrabose Aravindan. 2017. A Neural Network Approach to Indian Native Language Identification.. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, 8th - 10th December.*

[23] Laura Mayfield Tomokiyo and Rosie Jones. 2001. You're not from'round here, are you?: naive Bayes detection of non-native utterance text. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics, 1–8.

[24] Rosemary Torney, Peter Vamplew, and John Yearwood. 2012. Using psycholinguistic features for profiling first language of authors. *Journal of the Association for Information Science and Technology* 63, 6 (2012), 1256–1269.

[25] Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*. Association for Computational Linguistics, 9–16.

[26] Ajay P Victor and K Manju. 2017. Indian Native Language Identification.. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, 8th - 10th December.*

[27] Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop*. 53–61.

[28] Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1600–1610.

[29] Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2012. Exploring adaptor grammars for native language identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 699–709.

[30] Marcos Zampieri, Alina Maria Ciobanu, and Liviu P Dinu. 2017. Native Language Identification on Text and Speech. *arXiv preprint arXiv:1707.07182* (2017).

[31] Marcos Zampieri, Liling Tan, Nikola Ljubešic, Jörg Tiedemann, and Nikola Ljube. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*. 58–67.