DalTeam@INLI-FIRE-2017: Native Language Identification using SVM with SGD Training

Dijana Kosmajac
Dalhousie University, Faculty of Computer Science
Halifax, Nova Scotia, Canada
dijana.kosmajac@dal.ca

Vlado Keselj
Dalhousie University, Faculty of Computer Science
Halifax, Nova Scotia, Canada
vlado@cs.dal.ca

ABSTRACT

Native Language Identification (NLI), as a variant of Language Identification task, focuses on determining an author's native language, based on a writing sample in their non-native language. In recent years, the challenging nature of NLI has drawn much attention from the research community. Its application and importance are relevant in many fields, such as personalization of a new language learning environment, personalized grammar correction, and authorship attribution in forensic linguistics. We participated in the INLI Shared Task 2017 held in conjunction with FIRE 2017 conference. To implement a machine learning method for Native Language Identification, we used Character and Word N-grams with SVM (Support Vector Machines) classifier trained with SGD (Stochastic Gradient Descent) method. We achieved F1 measure of 89.60% (using 10-fold cross validation), using provided social media dataset and 48.80% was reported in the final testing done by INLI workshop organisers.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; **Classification and regression trees**; • **Social and professional topics** → *Cultural characteristics*;

KEYWORDS

Native Language Identification, Support Vector Machines, Stochastic Gradient Descent, N-Grams, Text Classification

1 INTRODUCTION

Since the 1950s there is a discussion in linguistic literature whether and how the native speakers of particular languages have characteristic patterns in sentence generation in their second language. This has been investigated in different domains and from different aspects, including qualitative research in Second Language Acquisition (SLA), more recently through predictive computational models in NLP [7] and in linguistic forensics [16].

In addition, the speaker's native language can have an effect on the types of errors they make. A study by Flanagan et al. [3] investigates the characteristics of errors by native language. They identified the differences and similarities of error co-occurrence characteristics of the following native languages: Chinese, Japanese, Korean, Spanish, and Taiwanese. They have shown that some languages have greater differences than another (Korean and Japanese tend to make similar mistakes).

This has motivated research in Native Language Identification (NLI), which was first defined as a Text Classification task by Koppel et al. [9], using a classifier with a set of lexical features such as function words, character n-grams, and Part-of-Speech (PoS) n-grams. The task, in general, focuses on the goal to identify speaker's native language from the samples of text written in a second language.

One of the main challenges for this task is the lack of corpora in appropriate size, class balance and topic homogeneity. So far, there are a couple of datasets which were used in the past research. International Corpus of Learner English (ICLE)[1] corpus is one of the first appearing in the early studies. Released in 2002 and updated in 2009, it became commonly used in research into native language prediction of learner writing. Brooke et al. [1] suggested that ICLI has problems that can lead to drop in performance when evaluated. They proposed additional corpora that might be useful in the task of native language prediction. They used data from a language learning SNS — *Lang-8.com* — and they show improved performance. Another corpus [17] was presented in a shared task on Native Language Identification of learners. The corpus was named TOEFL11, which contains essays in English by learners from 11 different native languages.

The approach we present is based on a linear Support Vector Machine classifier trained using Stochastic Gradient Descent method. As features, we used character and word n-grams. In addition, we used *tf-idf* weighting technique with χ^2 feature selection. We used a dataset provided by the Workshop organisers.

The rest of the paper is organised as follows: in Section 2 we present some of the most recent and relevant research to our experiments. Section 3 gives a short description of the dataset, using the information provided by the organisers. In the Section 4 we presented the experimental setup with details on data preprocessing, feature selection and weighting and classifier setup. Section 5 shows and discusses the results. In Section 6 we outline conclusions and further work.

2 RELATED WORK

The research in NLI domain is fairly recent. We present some of the most relevant to our experiments.

Kochmar et al. [8] study presented experiments on prediction of the native languages of Indo-European learners through binary classification tasks using with linear kernel SVM. They divided the native languages into two main groups: Germanic and Romance, with intergroup prediction performance accuracy 68.4%. The features used for prediction were words and n-grams,different error types that had been manually tagged within the corpus.

Wong[19] analyzed learner writing with an extension of Adaptor Grammars for detecting co-locations at the word level, as well as

[1] https://uclouvain.be/en/research-institutes/ilc/cecl/corpora.html

Table 1: INLI training dataset statistics

Language	Number	Percentage
Hindi (HI)	211	17.11%
Telugu (TE)	210	17.03%
Tamil (TA)	207	16.79%
Kannada (KA)	203	16.46%
Bengali (BE)	202	16.38%
Malayalam (MA)	200	16.22%
Total	1233	100%

for POS and functional words. Classification was performed at the document level by parsing individual sentences of the learner's writing to detect the native language with the final prediction based on a majority score of the sentences. Some notable characteristic features of languages extracted by this method were also discussed.

Bykh[2] discussed the use of recurring n-grams of variable lengths as features for training a native language classifier. They also incorporated POS features. They claim that their approach outperformed previous work under comparable data setup (ICLE corpus), reaching 89.71% accuracy for a task with seven native languages.

Jarvis et al. [6] was the best performing participant in earlier mentioned workshop by Tetreault [17]. They analyzed a set of features such as: word n-grams, POS n-grams, character n-grams, and lemma n-grams. On top of it, they used an SVM classifier. The prediction performance was evaluated on several different models with varying combinations of features.

Malmasi et al. [12–15] presented the first NLI experiments on Arabic[2] (Arabic Learner Corpus - ALC), Chinese (Chinese Learner Corpus [18]), Finnish and Norwegian languages data using a corpus of examination essays collected from learners of Norwegian. Given the differences between English and aforementioned languages, the main objective was to determine if NLI techniques previously applied to second language English can be effective for detecting native language transfer effects in second language.

3 DATASET

The dataset used in the experiment was provided by the organizers of the INLI Workshop [10]. Organizers identified the official Facebook pages of prominent regional language newspapers of the each region and extracted the comments. It consists of six classes: six languages of Indian subcontinent originating from different Indian states. As shown in Table 1, dataset is divided into classes named TA, MA, HI, BE, TE and KA. The dataset has following characteristics:

- It's balanced in the terms of the number of samples for each language;
- The native and mixed script text is removed from the comments;
- The comments are related to the general news in all over India in order to avoid topic bias.

[2]http://www.arabiclearnercorpus.com/

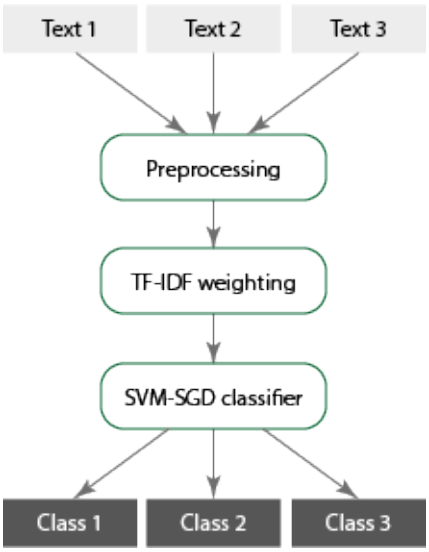

Figure 1: Architecture of the system.

4 EXPERIMENTAL METHODOLOGY

This paper presents a supervised multi-class classification approach. The training data texts are labeled with classes according to the author's native language. Figure 1 shows a diagram of the classifier components.

4.1 Data Preprocessing

4.1.1 Cleaning. Preparing and normalising the dataset are the first and necessary subtasks prior to the selection and classification. It includes filtering and adjusting the raw texts to make them suitable for the input of the next subtask. In general, social media user-generated texts are likely to be very noisy, containing textual elements irrelevant to the observed Classification Task. Hence, some parts of the comments were not considered as part of the feature set including hashtags, mentions and links.

4.1.2 Feature Extraction. Our model uses character n-grams of order 2–5. These n-grams capture small and localised syntactic patterns within a word of language production. Additionally, we used word n-grams of order 1–2. Our preliminary experiments showed that this n-gram lengths give best accuracy (possible reason is due to the data sparsity).

4.1.3 χ^2 feature selection. The formula for χ^2 feature selection can be expressed as follows:

$$\chi^2(M,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \quad (1)$$

where M is a message (a Facebook comment), t is a feature and c is a class. N is the observed frequency in M and E the expected frequency. Subscript e_t and e_c can take values 0 or 1. For example, $N_{e_t=1, e_c=0}$ means feature t is in N messages and is not in class c. We selected 50,000 features.

4.2 TF-IDF Weighting

Tf-idf (term frequency - inverse term frequency) is one of the best-known weighting algorithms. Several newer methods adapt *tf-idf* for use as part of their process, and many others rely on the same fundamental concept. *Idf*, being the measure's key part, was introduced in a 1972 paper by Karen Spärck Jones. As suggested in study by [5], we opted for using *tf-idf* measure in our experiment.

Tf-idf is the product of two measures, term frequency and inverse document frequency. In literature, different variations can be found. In this work we have used normalized term frequency to reduce bias towards different lengths between text samples.

$$ntf(t,d) = \frac{f_{t,d}}{\max\{f_{t',d} : t \in d\}} \tag{2}$$

$$idf(t,d) = \log_{N_{comments}} \frac{N_{comments}}{1 + \sum ntf(t, d_{comments})} \tag{3}$$

The final weight is expressed as follows:

$$weight(t,d) = ntf(t,d) \cdot idf(t,d) \tag{4}$$

4.3 Classifier

In the experiments we used a linear SVM (Support Vector Machine) to perform multi-class classification. SVM was chosen primarily because it shows effectiveness for this particular task [17] and we confirmed that in our preliminary experiments. The implementation is based on *Python* library *scikit-learn*, where we used linear SVM with SGD (Stochastic Gradient Descent) training.

The textual training samples $\mathbf{x}$ are represented as a d-dimensional vector. The vector $\mathbf{x}$ is classified by looking at the sign of a linear scoring function $\langle \mathbf{w}, \mathbf{x} \rangle$. The goal of learning is to estimate the d-dimensional parameter $\mathbf{w}$ so that the score is positive if the vector $\mathbf{x}$ belongs to the positive class and negative otherwise.

$$\ell_i(\langle \mathbf{w}, \mathbf{x} \rangle) = \max\{0, 1 - y_i \langle \mathbf{w}, \mathbf{x} \rangle\} \tag{5}$$

$$E(\mathbf{w}) = \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i \langle \mathbf{w}, \mathbf{x} \rangle\}. \tag{6}$$

$$E(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} E_i(\mathbf{w}), \quad E_i(\mathbf{w}) = \frac{\lambda}{2} \|\mathbf{w}\|^2 + \ell_i(\langle \mathbf{w}, \mathbf{x} \rangle). \tag{7}$$

SGD can be used to learn an SVM by minimizing $E(\mathbf{w})$. SGD performs gradient steps by considering at each iteration one term $E_i(\mathbf{w})$ selected at random from this average. Conceptually, the algorithm is:

(1) Start with $\mathbf{w}_0 = 0$;
(2) For $t = 1, 2, \ldots, T$;
  (a) Sample one index i in $1, \ldots, n$ uniformly at random;
  (b) Compute a sub-gradient $\mathbf{g}_t$ of $E_i(\mathbf{w})$ at $\mathbf{w}_t$;
  (c) Compute the learning rate η_t;
  (d) Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t$.

We used variable learning rate (in *scikit-learn* 'optimal'), which is computed as follows:

$$\eta_t = \frac{1}{\alpha \cdot (t + t_0)} \tag{8}$$

Table 2: Stratified 10-fold cross-validation

Folds	F1
#1	0.896
#2	0.904
#3	0.896
#4	0.901
#5	0.869
#6	0.907
#7	0.913
#8	0.861
#9	0.918
#10	0.892
Mean	0.896
St.D.	0.018

where α represents a constant that multiplies regularization term, and is used in learning rate calculation.

The goal of the SGD algorithm is to bring the primal suboptimality below a threshold ϵ_P:

$$E(\mathbf{w}_t) - E(\mathbf{w}^*) \leq \epsilon_P \tag{9}$$

4.4 Evaluation Measure

As suggested by INLI 2017 organisers, we used macro-averaged F1 score for evaluation measure (Eq. 10).

$$\begin{aligned} P &= \frac{TP}{TP+FP}, \\ R &= \frac{TP}{TP+FN}, \\ F1 &= 2 \cdot \frac{P \cdot R}{P+R} \end{aligned} \tag{10}$$

where *TP* are true positive predicted values, *FP* are false positive predicted values, *FN* false negative predicted values, *P* represents *precision* and *R* represents *recall*.

5 RESULTS

The results of our final experiment for distinguishing non-native Indian authors of the Facebook comments are shown in the accumulated confusion matrix on Fig. 2. The results show that features we used are useful for discriminating among non-native comments, achieving 89.60% F1 measure. The result is based on the mean performance of 10-fold validation.

The testing set from the organisers was a separate dataset from the one which was provided to the Workshop participants. The test results from the organisers shown in Table 3 report macro-averaged F1 measure 48.80%. The best performing class is BE (Bengali) giving the accuracy of 67.10%. The *recall* for this class is significantly higher compared to the other classes. The worst performing class is HI (Hindi) with the accuracy 23.80%. This is due to the very low *recall* value of 14.30%. Compared to the results of 10-fold cross-validation, we can see that HI class was performing worst. However, arguably due to the original dataset size and topic bias, overall

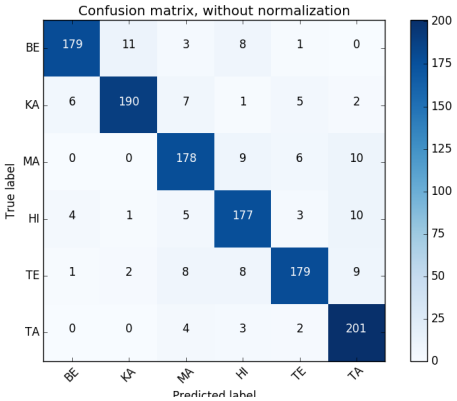

Figure 2: Accumulated confusion matrix from 10 fold cross validation on INLI dataset.

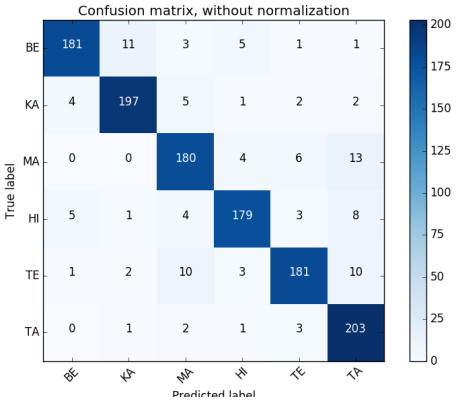

Figure 3: Accumulated confusion matrix from LLO validation on INLI dataset.

Table 3: Class-wise accuracy provided by the organisers

Class	Precision	Recall	F1
BE	56.20%	83.20%	67.10%
HI	69.20%	14.30%	23.80%
KA	40.50%	66.20%	50.30%
MA	46.70%	54.30%	50.30%
TA	51.10%	48.00%	49.50%
TE	33.30%	55.60%	41.70%
Overall			**48.80%**

system performance dropped significantly with the new test set. Additional datasets should be considered in the future.

6 CONCLUSION AND FURTHER WORK

In this paper, we experimented on the task of Native Language Identification (NLI). We used two different types of features: character and word n-grams. We use these features in a machine learning setup using a Support Vector Machine (SVM) classifier with Stochastic Gradient Descent (SGD) training on data from the INLI corpus which consists of six different native languages of Indian subcontinent.

There are a couple of directions for future work. In the related literature there are some relevant NLI approaches that could be tested on the data explored this paper. Some of them are analyses of feature diversity and interaction [11], and common error analysis by language [4]. Due to the lack of corpora for the languages investigated in this study, the application of more linguistically sophisticated features is limited, but to be explored in the future. For example, the use of a English parser could be used to study the overall structure of grammatical constructions as captured by context-free grammar production rules using parser designed for social media texts[3]. Another possible improvement is the use of classifier ensembles to improve classification accuracy. This has previously been applied to English NLI [17] with good results.

[3]http://www.cs.cmu.edu/ãrk/TweetNLP/

A LEAVE-ONE-OUT CLASSIFIER VALIDATION

In addition, we performed Leave-One-Out (LLO) cross-validation technique. This validation technique is appropriate, first, because training dataset is relatively small (consisting of approximately 200 samples per class). Second, the training set used for the final classifier is approximately equal to the training sets in LLO validation (all samples, but one). On Fig. 3 is shown accumulated confusion matrix from 1233 validation runs.

Final F1 measure is 90.90%.

REFERENCES

[1] Julian Brooke and Graeme Hirst. 2013. Native language detection with 'cheap' learner corpora. In *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, 37.

[2] Serhiy Bykh and Detmar Meurers. 2012. Native Language Identification using Recurring n-grams - Investigating Abstraction and Domain Dependence.. In *COLING*. 425–440.

[3] Brendan Flanagan, Chengjiu Yin, Takahiko Suzuki, and Sachio Hirokawa. 2014. Classification and clustering english writing errors based on native language. In *Advanced Applied Informatics (IIAIAAI), 2014 IIAI 3rd International Conference on*. IEEE, 318–323.

[4] Brendan Flanagan, Chengjiu Yin, Takahiko Suzuki, and Sachio Hirokawa. 2015. *Prediction of Learner Native Language by Writing Error Pattern*. Springer International Publishing, Cham, 87–96.

[5] Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with tf-idf weighting. In *the 8th NAACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA8)*. 216–223.

[6] Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing Classification Accuracy in Native Language Identification.. In *BEA@ NAACL-HLT*. 111–118.

[7] Scott Jarvis and Scott A Crossley. 2012. *Approaching Language Transfer Through Text Classification: Explorations in the Detection based Approach*. Vol. 64. Multilingual Matters.

[8] Ekaterina Kochmar. 2011. *Identification of a writer's native language by error analysis*. Ph.D. Dissertation. Master's thesis, University of Cambridge.

[9] Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 624–628.

[10] Anand Kumar M, Barathi Ganesh HB, Shivkaran S, Soman K P, and Paolo Rosso. 2017. Overview of the INLI PAN at FIRE-2017 Track on Indian Native Language Identification. In *Notebook Papers of FIRE 2017*. CEUR Workshop Proceedings, Bangalore, India.

[11] Shervin Malmasi and Aoife Cahill. 2015. Measuring Feature Diversity in Native Language Identification. In *Proceedings of the Tenth Workshop on Innovative Use*

of NLP for Building Educational Applications. Association for Computational Linguistics, Denver, Colorado, 49–55. http://aclweb.org/anthology/W15-0606

[12] Shervin Malmasi and Mark Dras. 2014. Arabic Native Language Identification. In *Proceedings of the Arabic Natural Language Processing Workshop (EMNLP 2014)*. Association for Computational Linguistics, Doha, Qatar, 180–186. http://aclweb.org/anthology/W14-3625

[13] Shervin Malmasi and Mark Dras. 2014. Chinese Native Language Identification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL-14)*. Association for Computational Linguistics, Gothenburg, Sweden, 95–99. http://aclweb.org/anthology/E14-4019

[14] Shervin Malmasi and Mark Dras. 2014. Finnish Native Language Identification. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*. Melbourne, Australia, 139–144. http://www.aclweb.org/anthology/U14-1020

[15] Shervin Malmasi, Mark Dras, and Irina Temnikova. 2015. Norwegian Native Language Identification. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2015)*. Association for Computational Linguistics, Hissar, Bulgaria, 404–412.

[16] Gerald R McMenamin. 2002. *Forensic linguistics: Advances in forensic stylistics*. CRC press.

[17] 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.

[18] Maolin Wang, Qi Gong, Jie Kuang, and Ziyu Xiong. 2012. The development of a chinese learner corpus. In *Speech Database and Assessments (Oriental COCOSDA), 2012 International Conference on*. IEEE, 1–6.

[19] 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.