Word Level Language Identification on Code-Mixed English-Bodo Text

Nayan Jyoti Kalita, Ankita Goyal Agarwala, Jayprakash Das
Department of Computer Science and Engineering,
The Assam Royal Global University
E-mail: nayan.jk.123@gmail.com, ankita89.goyal@gmail.com, dasjay084@gmail.com

Abstract. Since social media has become an active part of one’s life, people express their views freely in mixed informal languages on such platforms. So, in a multi-lingual country like India, it becomes really difficult for conventional language detectors to identify such languages. This paper mainly aims to detect the language at word level where the code mixed text can be in English-Bodo-Assamese. The data for the same is collected from some related Facebook pages and various classification algorithms are used to predict and compare the accuracy with which the detection is done.

Keywords: Code Mixing, Classification Algorithm, English, Assamese, Bodo

1. Introduction
With the increasing use of social media like Facebook, WhatsApp, etc., people often tend to express their ideas, feelings, statements, etc. in mixed languages like English-Hindi or Hindi-English-any Regional language together. Identifying the languages of the texts extracted from such platforms becomes very difficult. Social media facilities the sharing of ideas, thoughts and information and the building of the virtual networks and communities. However, it has adversely affected the language identification task as social media users writes texts in short or with spelling mistakes or with different phonetic typing, acronyms, or mixed with different languages, etc. Also, instead of using only Unicode, social media users makes use of phonetic typing, Roman script or transliteration for writing texts and have developed the habit of mixing different languages more frequently to express their views and thoughts.

The main purpose of Language Identification is to examine the extracted text of each document to determine the primary language and any secondary languages present in that document.

2. Motivation
The text or data collected from any social media platform is highly unstructured with a lot of non-standard abbreviations. Moreover, these data can also include texts written in some regional languages and the traditional language detectors cannot determine the language behind these texts because of their multilinguality. And Assam being rich for its diverse set of languages, the problem of language identification becomes all the more severe. So, the main objective of this piece of work is to identify the language of the code mixed text which will be written in English-Bodo-Assamese language.
3. Organisation of the Paper
The paper is organized as follows: Section 4 depicts a summary of the various work done on code mixing using social media text. Section 5 then describes the corpus that has been built from posts collected from Facebook groups and pages and which is code mixed among English, Bodo and Hindi language. Section 6 describes the different classification algorithms used along with the methods used for word level language detection, based on character n-grams, dictionaries. The results and discussion are reported in Section 7, while Section 8 concludes this piece of work and proposes some future line of work.

4. Background Study
A lot of work has been done so far on the language identification of text collected from social media platforms out of which some of them are summarized as follows:

In [1], Ted Dunning in 1994 had built a language identification model based on Markov model and it got 92% accuracy using 20 bytes size of test text data and 50 thousand training data set.

Arjen Poutsama in [2] introduced a language identification method which was based on Monte Carlo sampling in 2002. The proposed technique performed somewhat less than best-performing classification techniques, but it worked much faster.

Heba Elfardy et al. [3] in 2013 presented a model which could perform sentence level identification between modern standard Arabic and Egyptian Dialectal Arabic. They used token level to derive sentence level features. Their system achieved an accuracy of 85.5% on online - community dataset.

William B. Cavnar et al. in [4] proposed an N-Gram- Based Text Categorisation system. This system was based on calculating and comparing profiles of N-gram frequencies. They achieved as high as an 80% correct classification rate and concluded that N-gram frequency profiles provides a simple and reliable way to categorise documents in a wide range of classification tasks.

In [5], Utsab Barman et al. proposed a model at word-level classification in 2014 using a simple dictionary-based method, linear kernel support vector machines (SVMs) and a k-nearest neighbour approach. Based on these experiments, finally they selected SVM-based system and presented results for the Nepali-English and Spanish-English datasets.

In 2014, Amitava Das et al. in [6] proposed a model using n-gram with weights and support vector machine(SVM) along with additional features such as dictionary based for the English Bengali Hindi code mixing corpus and they achieved 94% accuracy on their system.

Nayan Jyoti Kalita et al. in [7] in 2018 proposed a model on social media code mixed corpus on English Assamese language and they found 89% accuracy using Support Vector Machine with additional features such as n-gram with weights, dictionary based and word position checking.

With this detailed study, it was noted that apart from the classification algorithms, the accuracy of identification also greatly depends on the language itself, the source from where the data is taken and the size of the corpus. Therefore, same classifier may have different accuracies with respect to these factors. In this work, the identification will be done using three different classifiers and the results will then be compared with each other.

5. Corpus Design
A lot of work has been done so far under language identification of texts collected from Social Media. Most of these have central languages like English, Hindi, some Arabic languages, etc. However, not much has been done with regional languages. Here, the code mixing using English-Bodo-Assamese language. The Bodo language is the official language of the Bodoland Autonomous region and co-official language of the state of Assam. It is also one of the 22 official languages of India.

The English Bodo code mixing language was collected from two Bodo pages of Facebook which contains 15,022 tokens and 6,602 unique words. Since Hindi is the official language of
India so it acts as a bridge for communication between different tribes. In the corpus collected, a small percentage of Hindi of about 1% was also found.

5.1. Annotation
The collected corpus was annotated at word level with 8 different tags. The annotation is shown below in the table 1.

5.2. Level of Code Mixing in the Corpus
The code mixing occurs at different level. The distribution of different code switching present in the corpus is as follows: 31.59% from intra sentential, 65.15% from inter sentential and 3.36% from intra word. The examples of sentences present in the corpus in the stated three forms are as follows:

- Inter-sentential: Mwjang/BD jadwang/BD Sir/EN.
- Intra-sentential: Banlgadeshi/NE Sir/EN Jwb/BD nama/BD.
- Intra-word: Publickho/WM aakhai/BD phwnaghwnai/BD nonga/DB.

In the intra-word case, publickho is the mixing of two words from two different languages, public(English) and kho(Bodo).

6. Classification Algorithms Used
The language identification here is done using three classification algorithms namely the Decision Tree, Naive Bayes and the Multilayer Perceptron. These three classifiers are described as follows:

6.1. Decision Tree Classifier
Decision trees [8] tries to evaluate the instances of data by creating a tree in a greedy manner and traverses as deep as possible until a prediction can be made. Once the tree is made, it is pruned in order to improve the model’s ability to generalize to new data.

6.2. Naive Bayes Classifier
Naive Bayes [9] is a simple classifier that used the Bayes Theorem as its base. For the implementation, this method calculates the the posterior probability for each of the seven mentioned classes and makes a prediction for the class with the highest probability.
### 6.3. Multi Layer Perceptron Classifier

A multilayer perceptron (MLP) [10] is a classifier that generally works with three layers namely: the input layer, hidden layer and output layer. A technique called the backpropagation is used for training the network and some activation functions are used for the hidden layers.

For the said study, English Bodo code mixing language were selected through various post and comment from some Facebook pages. The collected corpus was annotated at the word level with different tags. Weka V3.9[8] tool was used to implement three different classifiers with the default parameters and it was trained with the features like:- n-gram with weight, dictionary based, Acronym check, etc. The reason behind choosing three different classifiers was mainly to test their accuracy. Also the chosen classifiers show better results in comparison to others across multiple domains including language identification. Following are some of the features used to train the model:

- **N-gram with weights:** This feature is carried out using the bag-of-words principle. If there are $n$ unique n-grams for a language pair, then consider them as $n$ unique features. For example, that $in$ is the $i^{th}$ bigram in the list. In a given word $w$ (e.g., training), a particular n-gram occurs $k$ times (twice for $in$ in raining). If the pre-calculated weight of $in$ is $t^w_i$, the feature vector is $1, 2, ..., (t^w_i * k), ..., (n-2), (n-1), n$

- **Dictionary Based:** There are two English dictionaries, so there are two binary features. 1 is used to mark the presence of a word in a specific dictionary and 0 is used to mark its absence. Lexical Normalisation Dictionary and ABC (Australian Broadcasting Commission 2006)3 are used.

- **Acronym Check:** For an acronym and named entity, we use there Boolean features. In general, named entity begins with a capital letter. Again, acronyms are written in all capital letters. Therefore, we have checked whether all the letters are capitalized, only the first letter is capitalized or any letter of the word is capitalized.

### 7. Observations and Results

The model was trained by three different classifiers with n-gram and dictionary-based features. In word level evaluation, different accuracies were obtained from the code-mixed corpus. Table 2 gives a detailed accuracy of the model with respect to the different classifiers and also demonstrates their respective precision, recall, F-score and Root Mean Square (RMS) error values.

The confusion matrix of the experimental setup for all the tags is reported in the Table 3. Form the experimental set-up, it was observed that decision tree classifier correctly classified 60% of the original instances, Naive Bayes Classifier identified 78% of the instances correctly and MLP classified 65% instances correctly. The overall accuracy score was found a bit low in all the three classifiers as the size of the corpus was comparatively small because of the limited availability of structured English-Bodo code mixed data. Moreover, some of the words of the Bodo language happens to be same in either English, Assamese or Hindi language. This property again contributes to the wrong classification of the instances thereby reducing the overall accuracy.

---

**Table 2.** Detailed accuracy with different features

| Classifiers             | Precision | Recall | F-Score | RMS   |
|-------------------------|-----------|--------|---------|-------|
| Decision Tree           | 0.708     | 0.544  | 0.471   | 0.269 |
| Naive Bayes             | 0.762     | 0.693  | 0.659   | 0.299 |
| Multilayer Perceptron   | 0.335     | 0.425  | 0.372   | 0.288 |
5

|     | ACRO | UNIV | NE  | EN  | HI  | UN  | BD  | WM  |
|-----|------|------|-----|-----|-----|-----|-----|-----|
| ACRO  | 0    | 0    | 0   | 12  | 0   | 0   | 0   | 0   |
| UNIV  | 0    | 20   | 0   | 87  | 0   | 0   | 0   | 0   |
| NE    | 0    | 0    | 18  | 12  | 0   | 0   | 7   | 0   |
| EN    | 0    | 0    | 0   | 344 | 0   | 0   | 8   | 0   |
| HI    | 0    | 0    | 0   | 6   | 0   | 0   | 1   | 0   |
| UN    | 0    | 0    | 0   | 1   | 0   | 0   | 0   | 0   |
| BD    | 0    | 0    | 3   | 113 | 0   | 0   | 188 | 0   |
| WM    | 0    | 0    | 1   | 2   | 0   | 0   | 0   | 0   |

Table 3. Confusion Matrix for 8 class in Naive Bayes Classifier

8. Conclusion and Future Scope
The scope and dimension of language processing has progressed a lot with the more prevalent use of social media where culture and language are getting mixed more often. This paper presented a study on the detection of code-mixing in the context of social media texts. Since social media texts contain words in different languages, so identification of language becomes a challenging task. So this identification needs to be done at word level.

In this work, the corpus has been built with only the Facebook posts written in mixed English and Bodo language together. In the future, this model can be applied and experiment with other languages and other types of social media text, such as tweets. And there may be other features which can increase the accuracy. Facebook posts are generally restricted to 140 characters and tweets are even shorter. It would be interesting to investigate how this total character restriction of Facebook affects the level of code-mixing in tweets. It would be challenging to perform language identification with respect to tweets owing to its small size.

Also, this work has emphasized on code-mixing in romanized Indian social media texts, but there are other possible code-mixing cases such as Unicode and romanized Indian language text plus English.

References
[1] Dunning T 1994 Statistical identification of language (Computing Research Laboratory, New Mexico State University Las Cruces, NM, USA)
[2] Poutsma A 2002 Computational Linguistics in the Netherlands 2001 (Brill Rodopi) pp 179–189
[3] Elfardy H and Diab M 2013 Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) pp 456–461
[4] Cavnar W B, Trenkle J M et al. 1994 Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval vol 161175 (Citeseer)
[5] Barman U, Wagner J, Chrupala G and Foster J 2014 Proceedings of The First Workshop on Computational Approaches to Code Switching pp 127–132
[6] Das A and Gamb"ack B 2014 Proceedings of The 11th International Conference of Natural Language Processing
[7] Kalita N J and Saharia N 2018 Proceedings of the International Conference on Computing and Communication Systems (Springer) pp 433–440
[8] Quinlan J R 1986 Machine learning 1 81–106
[9] Rish I et al. 2001 IJCAI 2001 workshop on empirical methods in artificial intelligence vol 3 pp 41–46
[10] Rumelhart D E, Hinton G E and Williams R J 1985 Learning internal representations by error propagation Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science