Section A-Research paper
ISSN 2063-5346

Language Identification of English and Punjabi Code-Mixing and Code-Switching Sentences

Enjula Uchoi[1]
Department of School of Computer Science and Engineering
Lovely Professional University, Punjab[1]
enjulapaintoma@gmail.com[1]

Mandeep kaur[2]
Department of School of Computer Science and Engineering
Lovely Professional University, Punjab[2]
gagankaur005@gmail.com[2]

doi: 10.48047/ecb/2023.12.si6.367

Abstract **- People express their opinions freely on these platforms in a variety of informal languages because social media has become such an integral part of everyday life. As a result, it becomes quite challenging for traditional language detectors to recognise such languages in a multilingual nation like India. In this research, the primary goal is in order to determine the language at the word level of the code mixed sentences of English and Punjabi language. As per our knowledge very few researches has been done so far in English and Punjabi code-mixing and code-switching sentences. The suggested model combines a language dependent morphological dictionary-based model with a character n-gram language model based on frequency lexicons to accurately classify each word. With few dataset we could achieve the accuracy level of 88%.**

Index Terms – Character n-gram, Code-mixing, Code-switching, English, Punjabi

I. INTRODUCTION

With the widespread social media usage platforms like Facebook and WhatsApp, individuals increasingly choose to communicate their thoughts, feelings, and other sentiments in a combination of languages, such as English and Punjabi or Punjabi English, English and Hindi or Hindi English and any Regional language. It becomes exceedingly difficult to identify the languages of the texts that are taken from such sites. Social media makes it possible to create online networks and groups as well as to share ideas, thoughts, and information. The language recognition task has, however, suffered as a result of social media users' tendency to compose texts quickly, incorrectly spell, or in a variety of phonetic or acronymic ways. Additionally, social media users now regularly mix several languages to communicate their ideas and opinions rather than writing messages solely in Unicode; they do this by employing phonetic typing, Roman script, or transliteration. The main idea was that the Language Identification looks examines the each document's extracted text to identify the primary language and any more languages that are utilised in that document. Speech and written language identification are the two categories that make up language identification.
Signal processing methods and the use of phoneme-based structures are typically used for spoken language identification. The letter sequences, words, and n-gram frequencies used in document-based language identification play this role. Language identification is the technique of determining a language that is unknown by using characteristics and formulas present in the texts.

II. RELATED WORK

The use of generic phrases in recognising documental language is one of the easiest methods [1]. In order to accurately categorise each word, the proposed model combines a character-n-gram language model based on frequency lexicons with a language dependent morphological dictionary-based model [2]. Employed Bayesian classifier in documents as small as 20 bytes, even as high as 92% outcomes, and machine learning techniques for Spanish and English [3]. have suggested an HMM (Hidden Markov Model) based on character string language identification. On web documents, this method utilised automatic language detection. In the tests, 140 bytes of test data in English, German, French, Spanish, and Italian were employed, and 99% accuracy was attained [4]. A case study of textual sequences in the data was modelled using Markov. Claude Shannon's modelling study of letter and word sequences is another well-known piece of his work [5]. Extracting the entity of the code-mixed Hindi, Tamil, and English languages from a social media text was done by Rao and Devi in 2016 [6]. Ali Selamat and Nicholas Akosa, 2016 An algorithm based on the lexicon is proposed to conduct language identification at both the document and sentence levels [7]. Nguyen and Dogruoz (2013) are a couple of the

other methods. In their most recent efforts, they created character n-gram models for each language with a maximum length of 5 gramme to help with the analysis and processing of code-switched text. They also tested utilising linear-chain conditional random fields and a logistic regression model [8]. In R.V. Kumar et al. (2015) The Support Vector Machine was used to Label the terms as Language1, Language2, and sequence level and use Name entities, mixed script, and punctuation based on the Indian language of the mixed script[9]. By utilising character-sequencing approaches, Harshi Jhamtani et al. (2014) suggested a method to identify phrases that use a code switch [10]. Tommi Vatanen et al. (2010) used character n-gram models to identify brief texts and the Cavnar and Trenkle (1994) ranking method [11].

III. MOTIVATION

Any social media platform's language or data collection is largely unstructured and uses several non-standard abbreviations. These data may also contain writings various regional languages are used to write, and due to their multilingualism, conventional language detectors are unable to identify the language of these texts. Because punjabi is home to such a wide variety of languages, the issue of language identification is made even worse. The primary goal of this piece of work is to determine the language of the code mixed text, which will be written in the languages of English, Punjabi, roman Script. Even do the Punjabi has its own scripts but in today's generation the social media platform has been used as one of the most popular platform in conversion. While texting most of the speakers text with roman script instead of using their own scripts. This is the major challenges and issues were the texts are unstructured and in-formal sentences which occurs major errors in identifying the Word level language.

IV. METHODOLOGY

Few methodologies have been applied to identify the code-mixing and code-switching of English-Punjabi languages.

A. corpus design

The 45,628 tokens and 8,464 unique words that make up the English punjabi code mixing language were taken from Facebook, whatapps, group sites.

B. Tag Set

The languages of English, Punjabi and Hindi have been Tag as EN, PU, HI. Few others universal symbols are also been identified while training the set.

TABLE I
CODE MIX TAG SET

Sl. No	Tag Appearance		
	Tag and Meaning	Example	Total
1	English, EN	Can do	8714
2	Punjabi , PU	Mai hun	1623
3	Hindi , HI	Mai Ab	457
4	Universal/ UNIV	“,@,&,!,?	1230
5	Acronym/ACRO	RIP,GBU,CU,OMG	345

C. Algorithm Used

1. Dictionary based:- The word used in the dictionary-based method was taken directly from the corpora. When a word is not found in dictionaries, the system then searches dictionaries for it before selecting the language with the highest likelihood. In the event that English is the only option, English is selected as the language.
2. Character N-gram:- Regardless of their meanings or the sequence in which the words appear in the text, frequency data for word uni- and bi-grams were gathered for each language. To create word uni- and bi-grams, the countvectorizer module in Python was created. A Document Term Matrix X [i, j] was created, where i is the document's id, j is each word's dictionary index, and Wij is how frequently each word appears in the document. Each unigram and bigram from the test data was compared to how frequently those words appeared in documents written in all other languages. In this experiment, we employed linear SVM from Scikit-Learn as a classification technique for Natural language Identification and used Document Term Matrix with n-grams.

2.1 N-gram processing Examples of English and Punjabi

TABLE II
N-GRAM METHODS

Sl. No.	N-gram	Methods
1	Bi-gram	_P, PO,OP,PU,UL,LA,AT,TI,IO,ON,N_
2	Tri-gram	_PO,POP,OPU,PUL,ULA,LAT,ATI,TIO,ION,ON_

2.2 Algorithm Steps.

i. Digit and punctuation are removed from the text, and it is divided into individual tokens made up exclusively of letters and apostrophes. Sufficient blank space is added before and after each token.
ii. For N=1 to 5, scan each token and produce every possible N-gram. Use placements that also cover the padding blanks.

iii. To locate the N-gram counter and increase it, hash into a table. The hash table makes sure that each N-gram has its own counter by employing a traditional collision control approach.
iv. Output each N-gram along with its count when finished. In reverse order, sort the counts according to the frequency. Reverse the order of frequency such that you are left with simply the N-grams.

TABLE III
CHARACTER 2-GRAM MODEL

Sl. No.	Language	Numbers of Tokens
1	English	1767141
2	Punjabi	52376

TABLE IV
MIXED SENTENCES

Sl. No.	Data Name	Total words
1	Punjabi	52376
2	Mixed sentences of English and Punjabi	1746

From the above Table III we can see the total numbers of bi-gram of both English and Punjabi languages and Table IV shows the total mixed sentences used for training the dataset using N-gram models.

V. LANGUAGE IDENTIFICATION MODEL FLOWCHART

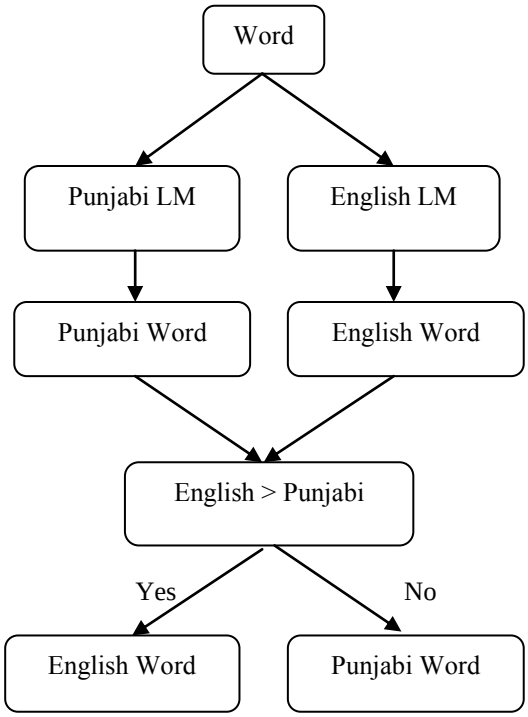

Fig. 1 Word level Language identification of English and Punjabi language.

We can see from Figure 1 above how the word identification of English and Punjabi languages has been categorised. It assesses the degree to which the scores of the two languages may be translated. The result will be English words if the English score is higher than the Punjabi score; otherwise, the words will be identified as Punjabi terms.

V. FEW CODE-MIXING AND CODE-SWITCHING EXAMPLES OF ENGLISH AND KOKBOROK

1.Mai/PU aj/PU hair/EN cut/EN karvaea/PU

English Translation: Today I had my hair cut

2. Meri/PU marriage/EN nu/PU two/EN years/EN ho/PU gye/PU

English Translation: It's two years already, that I got married

3. Mai/PU early/EN morning/PU uthda/PU han/PU
English translation: I get up early in the morning.

4. Mera/PU ghar/PU is/EN in/EN village/EN
English Translation: Mine House is in village

5. Mere/PU ghar/PU wich/PU eight/EN members/EN ne/PU
English Translation: I my house we have eight members

6. Mai/PU aaj/PU rice/EN khade/PU
English Translation: I eat rice today.

VI. RESULT

Using n-gram and dictionary-based features, three separate classifiers were used to train the model. The code-mixed corpus yielded various accuracy levels for word level evaluation. Table V provides information on the model's accuracy in relation to two classifiers, as well as shows each variable's accuracy and errors. We have also checked the recall, F1 and Root mean square of both the classifiers.

TABLE V

ALGORITHMS ACCURACY AND ERROR

Sl.No.	Methods	Accuracy	errors
1	Dictionary Based	48	52
2	Character n-gram	88	12

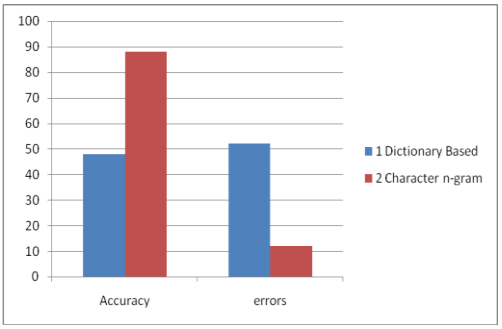

Fig. 2 Dictionary based and N-gram chart.

VII. OBSERVATIONS

An example of Englishization is the blending of languages like English and Punjabi. It is a sign of Indianization, nativization, and acculturation of English language when PunjabiEnglish mix is used to transmit socio-cultural ideas, beliefs, and conceptions. The data also includes information on the attachment of inflection and reduplication of specific words and phrases. English nouns are typically inflected with a Punjabi plural suffix. There are also been seen few words which are taken as borrow words from both the sentences. Sometime the speakers was actually those borrow words to address the speakers. This creates ambiguity because there was a dearth of structured English-Punjabi code mixed data, the size of the corpus used in this was quite small. Additionally, some of the words in Hindi or English occurs to be the same in the Punjabi language.

VIII. CONCLUSION AND FUTURE SCOPE

With the increased usage of social media, where culture as well as the language are more frequently intermingled, the scope and the dimension of the language processing have greatly advanced. This study on the identification of code-mixing in the setting of social media messages was provided in this publication. Language identification is difficult since social media postings often use vocabulary from multiple languages. Therefore, word level identification is required for this process. In this study, the corpus was exclusively created using Whatsapp groups, manually collecting the common conversation mixed sentences, Facebook postings that combined mixed English and Punjabi language. This model can be used in the future to test different languages and social media text formats, including tweets. And there might be additional characteristics that improve accuracy. Additionally, although the focus of this study has been on the use of code-mixing in romanized social media in India messages, there are other instances in which it could occur, such as when English is used alongside Unicode and romanized Indian text.

In the future we are thinking to explore much more deeper to identify the word level language of English-Punjabi code-mixing and code-switching sentences by using different machine learning approaches like SVM, RM,HMM, NN, etc. This paper will surely help the others researcher to understand how to explore much deeper in the area of Natural Language Processing in Indian languages especially the Punjabi language.

REFERENCES

[1] Grefenstette, Gregory. "Comparing two language identification schemes." Proceedings of JADT. Vol. 95. 1995.

[2] Uchoi, Enjula and Lenin Laitonjam. “An Unsupervised Word Level Language Identification of English and Kokborok Code-Mixed and Code-Switched Sentences.” *Journal of emerging technologies and innovative research* (2020):

[3] Combrinck, H. P., and Elizabeth C. Botha. "Automatic language identification: Resisting complexity." *South African Computer Journal* 2001.27 (2001): 18-26.

[4] Takcı, Hidayet, and İbrahim Soğukpınar. "Centroid-based language identification using letter feature set." *Computational Linguistics and Intelligent Text Processing: 5th International Conference, CICLing 2004 Seoul, Korea, February 15-21, 2004 Proceedings 5*. Springer Berlin Heidelberg, 2004.

[5] Manning, Christopher, and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[6] P. R. Rao and S. L. Devi, “Cmee-il: Code mix entity extraction in indian languages from social media text@ fire 2016-an overview.” in FIRE (Working Notes), 2016, pp. 289–295.

[7] A. Selamat and N. Akosu, “Word-length algorithm for language identifi cation of under-resourced languages,” Journal of King Saud University Computer and Information Sciences, vol. 28, no. 4, pp. 457–469, 2016

[8] D. Nguyen and A. S. Dogru ˘ oz, “Word level language identification ¨ in online multilingual communication,” in Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 857–862.

[9] R. V. Kumar, M. A. Kumar, and K. Soman, “Amritacen nlp@ fire 2015 language identification for indian languages in social media text.” in FIRE Workshops, 2015, pp. 26–28

[10] H. Jhamtani, S. K. Bhogi, and V. Raychoudhury, “Word-level language identification in bi-lingual code-switched texts,” in Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing, 2014.

[11] T. Vatanen, J. J. Vayrynen, and S. Virpioja, “Language identification of ¨ short text segments with n-gram models.” in LREC, 2010.