Journal of Innovative Engineering and Research (JIER)
Vol.- 7, Issue -1, April 2024, pp. 18-20 (3 pages)

Language Identification at Document Level and Sentence Level in Indian Code-Mixed Social Media Text

TARJANI SEVAK, DR. GAURAV GUPTA

Department of Computer Application
Dr. A. P. J. Abdul Kalam University, Indore 452010, India
Correspond Author Email: tarjani_85@yahoo.co.in

Abstract— *Citizens are more likely to participate in conversations on numerous topics, and people voice their opinions on social media, thanks to social media's meteoric rise. Code-mixing is a popular way for people to express themselves on social networking platforms. Language detection at the document level has been considered nearly solved in some application areas, but language detectors fail in the social media context due to phenomena such as utterance internal code-switching, lexical borrowings, and phonetic typing, all of which imply that language identification in social media must be done at the word level or sentence level. Text sentiment analysis techniques usually work at a certain level, such as phrase/word, sentence, or document level. The goal of this work is to examine a solution for sentiment classification at a finer level, namely the sentence level, where the polarity of the sentence can be classified into three main categories: positive, negative, and neutral.*

Index terms: *Document Level, Word Level, Polarity, Sentence Level, Sentiment Analysis.*

I. INTRODUCTION

The ability to identify the language in which a specific segment is written is an essential precondition for any kind of automatic text processing. Many studies have been conducted to determine how to effectively conduct subjectivity text. Sentiment classification is a new branch of NLP that categorises subjectivity text as positive, neutral or negative. Sentiment classification could be done at the word, sentence, and document levels [1].

Every individual mix many languages to express their opinion, automatic detection of mixed language is very challenging. A variety of approaches are available for classifying the language at word, sentence and document levels. Sentiment analysis is now predominant approach used to extract sentiment and assessments of on-line sources.

The subjective analysis emphasizes the division of linguistic units in objective and subjective and sentiment analysis emphasizes the division of linguistic units in positive, neutral and negative. Detecting subjective and objective opinions presents a significant challenge in identifying the polarities [4].

The probability of code mixing with English language is higher in speakers whose first language is written in an Indic language since English typing makes sentence more convenient. Particularly in Indian subcontinent, this is evident. An example of code mixing is:

ચાર આ મૂવી તો મસ્ત છે . You must watch. Maza aai gayi.

This comment is written in three languages: English, Hindi and Gujarati. We adhere to the latest research on language identification for Social Media Text.

The structure of the paper is as follows: The idea of code switching and some earlier research on code mixing in social media writing are covered in the next section. Our code mixing corpus, the data itself, and the annotation procedure are all covered in Section 3. Section 4 lists the resources and tools we utilize for the language identification studies that are covered in Section 5. In Section 6, we wrap up and offer recommendations for more study on this subject..

II. RELATED WORK

Less effort has been put toward automatically identifying the language of messages with multiple dialects jumbled together. We first quickly review studies on the general properties of code mixing before getting to that subject.

Numerous scholars have examined the causes of code-mixing. Research on the practice of code-switching between Chinese and English [7][8] revealed that social incentives were not as important as linguistic ones in these highly bilingual communities when it came to causing shifts.

Das and Gamback[4] performed linguistic detections, Context in India English bengali and hindi urdu. They used SVM by using features such as n-grams with weights in linear kernels, dictionary based; minimum edit distance based on weight and word context features.

Language identification is a major issue in most of the previous works. Where the number of languages is two, in code mixed text. Some they are addressing code mixing in a single script and remaining the address code is mixed up in other scripts. The later problem is relatively easier while the complexity lies in distinguishing the tokens sharing the same script in several languages.

Sentiment analysis of multilingual Twitter data has become

increasingly important as social media platforms continue to serve as major sources of user-generated content [10]. This study explores the application of natural language processing techniques for sentiment analysis on multilingual Twitter data. The study begins by discussing the challenges posed by multilingual data, including language variation, code-switching, and cultural nuances. It highlights the need for robust techniques that can handle these challenges and accurately classify sentiment across different languages. The study explores various natural language processing techniques employed in sentiment analysis, including data pre-processing, feature extraction, and classification algorithms. It discusses methods such as tokenization, stemming, and stop-word removal for data pre-processing, as well as techniques for handling multilingual lexicons and sentiment dictionaries. Feature extraction techniques, including n-grams, statistical features, and word embeddings, are explored to capture relevant information for sentiment classification. Different classification algorithms, such as Naive Bayes, Support Vector Machines (SVM), and neural networks, are examined for sentiment analysis on multilingual Twitter data. It examines the challenges posed by languages with different structures, writing systems, and sentiment expression patterns.

We will focus here on India, a nation More than 500 spoken languages, with some 30 languages having more than 1 million speakers. There's no national language in India. 22 languages carry official status in at least parts while English and Hindi are used in the country for nation-wide communication. Language diversity frequent code mixing is caused by dialect changes and changes in the language in India. We address this problem in the context of code mixed text, with 2 separate codes Indian languages Hindi, Gujarati with English language. As a result, Indians are multilingual by adapting and necessity, and frequently changing mixed languages in the context of social media.

III. DATA COLLECTION

The dataset was obtained from Gujarati, Hindi mixed Dataset, and it consists of two csv files with the file names Gujarati and Hindi. The two datasets were combined using the concat function. The dataset was then displayed using the sample function.

	lang	text
0	HM	VA VA KYA BAT HE
1	HM	Bahut khub jitu bhay
2	HM	Sab se 1 nambr
3	HM	Sachi bat he jitu bhai AAP ko call nahi kiya ...
4	HM	joks bhejo

IV. EXPERIMENT

Text normalization techniques are applied to handle variations in spelling, punctuation, or capitalization. For mixed-Indic text, normalization needs to be performed in a script-specific manner to preserve the integrity of each language involved.

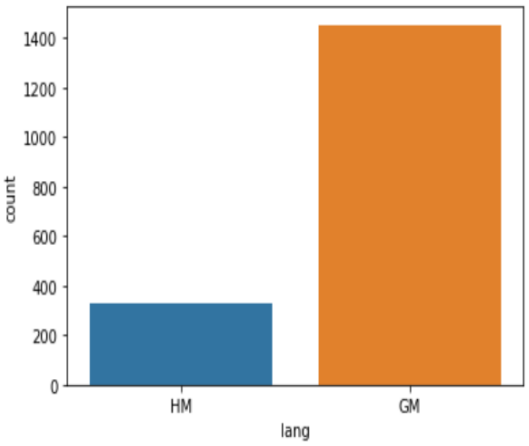
Word embeddings, such as word2vec or GloVe, are often used in sentiment analysis. For mixed-Indic social media text, it is essential to use language-specific embeddings or resources that capture the nuances of each language script.

To apply LSTM for sentiment analysis on mixed-Indic social media text, the first step is to pre-process the text data, including language-specific tokenization, normalization, and encoding. This ensures that the mixed-Indic text is prepared for input into the LSTM model.

Pre-processing involves deleting unnecessary columns from the dataset and remove punctuation or stop words that aren't necessary for analyzing the results, decreasing the review's content, and using a lemmatize to ensure correct results.

Data has been divided into an 80:20 ratio. used 20% for testing and 80% for training.

Exploratory Data Analysis (EDA) used in the data analysis process that helps us understand languages used in our dataset and characteristics of the dataset.



V. CONCLUSION

Determining the language of the provided data is the system's primary goal. Many NLP applications focus on detecting language from code-mixed data since this is a common trend among users of social media networks like Facebook and Twitter. In this work, the language recognition algorithm used for a Twitter dataset that was code-mixed in Hindi, English, and Gujarati. An initial study on language identification using code-mixing of Indian languages in social media communication has been presented by us. We provided an overview of our Twitter dataset comprising Hindi, English, and Gujarati tweets.

A machine learning-based SVM classifier is utilized for training and testing, and word- and character-based n-gram embedding vectors are employed as features. Precision, recall, F1-measure, accuracy, and other performance indicators were used to assess the system.

REFERENCES

[1]. V. S. Jagtap, Karishma Pawar 2013. Analysis of different approaches to Sentence-Level Sentiment Classification International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume 2 Issue 3, PP : 164-170.

[2]. Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. EMNLP 2014 13 (2014).

[3]. Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. In Proceedings of the 2010 Human Language Technology Conference of the North American Chapter of the Association
for Computational Linguistics, pages 229–237, Los Angeles, California, June. ACL

[4]. Amitava Das, Björn Gambäck. 2014. Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text. Proceedings of the 11th International Conference on Natural Language Processing, Pages 378–387, Goa, India, Dec. ACL

[5]. Kerstin Denecke, "Using SentiWordNet for Multilingual Sentiment Analysis", ICDE Workshop 2008.

[6]. Nagesh Bhattu Sristy, N. Satya Krishna, Vadlamani Ravi, "Language Identification in Mixed Script", ACM, New York, USA, December 8–10, 2017, Bangalore, India.

[7]. David C. S. Li. 2000. Cantonese-English code-switching research in Hong Kong: a Y2K review. World Englishes, 19(3):305–322, November.

[8]. Hong Ka San. 2009. Chinese-English code-switching in blogs by Macao young people. MSc Thesis, Applied Linguistics, University of Edinburgh, Edinburgh, Scotland, August.

[9]. Younes Samih, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer, and Thamar Solorio. 2016. Multilingual Code-switching Identification via LSTM Recurrent Neural Networks. EMNLP 2016 (2016), 50.

[10]. Goel, V., Gupta, A. K., & Kumar, N. (2018, November). Sentiment analysis of multilingual twitter data using natural language processing. In 2018 8th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 208-212). IEEE.

[11]. Shelke Rita and Singh Thakore Devendra. 2020. A novel approach for named entity recognition on Hindi language using residual bilstm network. International Journal on Natural Language Computing (IJNLC) 9, 2 (2020), 1–8.

[12]. Goel, V., Gupta, A. K., & Kumar, N. (2018, November). Sentiment analysis of multilingual twitter data using natural language processing. In 2018 8th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 208-212). IEEE.

[13]. Xin Wang, Yanquing Zhao and Guohong Fu. (2010). A Morpheme - based Method to Chinese Sentence -Level Sentiment Classification. International Journal on Asian Language Processing , 95-105.-2010