CMPUT 501 Project Proposal: Sentiment Analysis for Code-Mixed Social Media Text (SemEval 2020 Task 9)

Arnob Mallik and Arif Hasnat
Department of Computing Science
University of Alberta

1 Introduction

Code-Mixing (CM) is defined as the embedding of linguistic units such as phrases, words, and morphemes of one language into an utterance of another language (Raghavi et al., 2015). People are using their native language ”mixed” with English more and more frequently on online platforms. It helps in speeding-up communication and allows wider variety of expression due to which it has become a popular mode of communication on Facebook and Twitter.

The goal of our project is to design a model that assigns sentiment labels (positive, negative to neutral) to English-Hindi and English-Spanish mixed tweets.

2 Motivation

On Social media sites such as Twitter and Facebook, people post their opinions, comments, suggestions in a free form, resulting in large volume of text data available for interpretation. Analyzing the sentiments of these texts is of huge importance as it allows us to gain an overview of the wider public opinion behind certain topics. It can also be an essential part of market research as businesses generally rely upon the beliefs/views of the people about their products or services.

3 Dataset

For our project, we will primarily use the dataset provided by the ”SentiMix” task of SemEval 2020. The dataset has sentiment labels and also language labels at word level. We will also look into the dataset provided by (Prabhu et al., 2016) for English-Hindi mixed texts.

4 Related Work and Roadmap

As we are working on mixed-language datasets, we will need a suitable method for language identification at word level such as the ones of (Das and Gambäck, 2014) and (Barman et al., 2014). We will look into the methodologies of previous attempts of sentiment analysis of code-mixed texts (Mäntylä et al., 2018). (Patra et al., 2018) presents a overview of a previous task of sentiment analysis on English-Hindi mixed texts where classifiers such as SVM, Naïve Bayes, were applied on unigram, bigram and n- gram features for sentiment classification.

Also, we will study successful traditional sentiment analysis approaches such as the one of (Nakov et al., 2016) to understand the basics of the task at hand. We will look into different deep learning techniques (Zhang et al., 2018) for traditional sentiment analysis and examine how they work on our dataset.

5 Evaluation Metrics

We will measure the efficiency of our tool in terms of Precision, Recall and F-measure.

References

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*. pages 13–23.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*. pages 378–387.

Mika V Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review* 27:16–32.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-

2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*. pages 1–18.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task@ icon-2017. *arXiv preprint arXiv:1803.06745* .

Ameya Prabhu, Aditya Joshi, Manish Shrivastava, and Vasudeva Varma. 2016. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. *arXiv preprint arXiv:1611.00472* .

Khyathi Chandu Raghavi, Manoj Kumar Chinnakotla, and Manish Shrivastava. 2015. Answer ka type kya he?: Learning to classify questions in code-mixed language. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, pages 853–858.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8(4):e1253.