Ling 573

Connor Boyle
boyle128@uw.edu

Martin Horst
mmhorst@uw.edu

Nikitas Tampakis
tampakis@uw.edu

Abstract

This is where the abstract for our reports will go.

1 Introduction

2 Task Description

Our primary task consists of classifying code-mixed Spanish-English social media (Twitter) posts with sentiment labels, which was part of SemEval 2020's Task #9: SentiMix (Patwa et al., 2020). Our adaptation task will be the analogous Hindi-English sub-task of the same SemEval 2020 task.

The training and testing data sets of each language sub-task are both comprised of tokenized mixed-language tweets, with token-level language labels and tweet-level sentiment labels. The sentiment labels are categorical and use the categories: 'positive', 'neutral', or 'negative'.

Submitted classifier models were evaluated based on their multi-class support-weighted F1 score on held-out evaluation ('test') data.

The fully-labeled Hinglish training and test data, as well as fully-labeled Spanglish training data are publicly available (on the competition website) and we have downloaded them. However, the labels are missing for the Spanglish test records. We determined that the Spanglish test data is mostly a reshuffling of other publicly-available test data (Aguilar et al., 2020). Martin was able to match at least 2,200 (out of 3,718) unlabelled SentiMix test data records to labels in the other publicly-available test data set. We have also emailed the organizers of the shared task about accessing the test data labels or otherwise indirectly accessing the test data.

3 System Overview

4 Approach

5 Results

6 Discussion

7 Conclusion

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. Lince: A centralized benchmark for linguistic code-switching evaluation.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain. Association for Computational Linguistics.