Stochastic Gradient Descentiment Analysis: Classifying Affect in Mixed Language Tweets

Connor Boyle
boyle128@uw.edu

Martin Horst
mmhorst@uw.edu

Nikitas Tampakis
tampakis@uw.edu

1 Introduction

Although a large percentage of the world's population is multilingual, most natural language processing (NLP) tasks focus on monolingual contexts. Our project aims to investigate how state-of-the-art machine learning approaches to sentiment analysis perform on code-mixed (i.e. multilingual) sentences. Speakers of Spanglish and Hinglish (Spanish-English code mixing and Hindi-English code-mixing respectively) naturally blend lexica and syntax across languages. Can language models from each language be combined to correctly interpret affect in tweets? Do features like emojis, punctuation, and orthographic variation also play a role in the emotion expressed in a tweet? We demonstrate that although accuracy of sentiment classification can be improved by leveraging state-of-the-art methods, there is still substantial improvement to be made in this NLP task.

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. One interesting statistic from the SemEval 2020 overview paper is that there were over twice as many submissions for Hinglish as there were for Spanglish.

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 each one of three 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). We were 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 also reached out to the original organizers, who informed us that we could submit to a new Spanglish competition on a different platform. The test set on this new platform has its labels withheld, however, many of the test instances overlap with our training data (therefore it does not have the expected guarantee against overfitting).

3 System Overview

Our baseline system extracts unigrams, bigrams, and trigrams of tokens or language-tags (e.g. ["Yo", "tengo" "hungry"] and ["lang2", "lang2", "lang1"]) which are then fed to a one-vs-all collection of linear support vector machines (SVM), which we optimize with stochastic gradient descent learning from sci-kit learn (Pedregosa et al., 2011). As input the model takes in a sparse vector of ngram counts from pre-processed multi-lingual tweet tokens; it outputs a class of negative, neutral, or positive.
For this testing round we used the "dev" set of tweets (around 2,300) as our testing set. After the system outputs its predictions, our evaluation script calculates a weighted F1 score using the labeled version of the dev set. We again used sci-kit learn's F1 calculator for this step.

4 Approach

We decided to implement the baseline classifier using unigrams, bigrams, and trigrams as our fea-

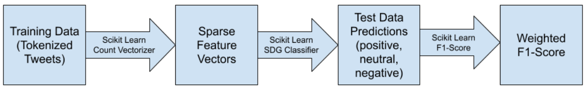
Training Data (Tokenized Tweets)
Scikit Learn Count Vectorizer
Sparse Feature Vectors
Scikit Learn SDG Classifier
Test Data Predictions (positive, neutral, negative)
Scikit Learn F1-Score
Weighted F1-Score

Figure 1: System architecture diagram

tures. Because our dataset is specific to code-mixed tweets, these are ngrams both of tweet tokens and also of language labels. From the original dataset, the language labels could be lang1, lang2, other, ne, or ambiguous. In this case, lang1 refers to English and lang2 refers to Spanish. At this time we are using no other features, however for our next deliverable we plan to extract sub word-level features such as capitalization (e.g. "SpOngEbOb cAsE"), character repetition (e.g."sigghhhhhhhh"), emoticons, and emojis; or to use embeddings that reflect these sub word-level features.

We are also considering other classifiers, such as LSTMs and other recurrent neural networks, which were popular among other participants in the original shared task, or Deep Averaging Networks, which have been shown to achieve high performance with relatively little training time.

5 Results

Results for our baseline classifier are given in table 1 below. We report two baseline systems for tweet classification that use no features to predict the sentiment of a given tweet. The first system assigns all "dev" set tweets a random sentiment classification. The second system assigns all "dev" set tweets a positive sentiment classification. Our model's weighted F1 score is reported in the third row and surpasses the two baselines. Due to our system architecture, we expect variability between trials, and so each report is an average of 5 separate runs.

System	Precision	Recall	F1
Random	0.1544	0.4551	0.4032
All Positive	0.1666	0.3333	0.3329
SGDClassifier	0.3865	0.4438	**0.5075**
Task Baseline	??	??	0.656
TBD	–	–	–

Table 1: Averaged results from five system trials

6 Discussion

One interesting aspect of our initial trials was that pre-processing the data (e.g. accounting for UNK tokens, lower-casing tweets, and standardizing encodings) did nothing to alter the F1 score. We posit that the vocabulary was not large enough to create robust ngram vectors for the SGD classifier thus improving classification.

There are many directions where we could take this work moving forward to further inform and improve our model. One idea is to analyze code-mixing features from a "primary" language of the tweet. Is a tweet following the syntax of a dominant language with a few words from a secondary language embedded? We could naïvely use the counts from the tagged tokens to assign a primary language based on the language with the highest token count, and then use this as an additional feature for the classifier. Another idea is to determine the rate of language mixing that happen in a tweet. One hypothesis is that the more "turns" a code-mixed tweet has, the higher the emotional content.

Should we create separate models for each of the language tag types, Spanish, English, named entity, or other?

Emojis are hyperpresent in tweets (Barbieri et al., 2016), and a qualitative review of the data suggests that this is the case in our dataset, If we leveraged a lexical resource such as the Emoji Sentiment Ranking from (Kralj Novak et al., 2015) this would give us additional features which could inform our classifier.

Broadly looking across the tweets in our dataset, language styles heavily vary. One of the approaches we explore may really working for some tweets, but not for others. Incorporating boosting may be a way to improve our accuracy by combining multiple model predictions into the classifier.

7 Conclusion

We are looking forward to seeing how our approach performs on the Hinglish dataset and further exploring how language/data-aware and agnostic approaches can affect the model accuracy.

References

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

Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How cosmopolitan are emojis? exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 531–535.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one*, 10(12):e0144296.

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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.