Zero Shot Sentiment Analysis on Tweets in Any Language

Anonymous ACL submission

1 Introduction

Linguists have identified more than 6500 languages in active use (Anderson, 2010), over 100 of which are in active use on Twitter (Hong et al., 2011). Unfortunately, the natural language processing (NLP) community has focused primarily on understanding English language text.

In this article, we will build a sentiment analysis (Zhao et al., 2016) convolutional neural network classification model, pretrained on the geolocation problem, by training it on a set of sentiment-labeled tweets in a popular language.

We hypothesize that this trained model can zero shot learn (Xian et al., 2017) to evaluate sentiment in any target language by training on unlabeled tweets in that language.

Since training a model requires a large dataset of sentiment-labeled tweets, it is not possible to validate accuracy on *low resource languages* where the required dataset does not exist by definition. Therefore, we will test our hypothesis on target languages with existing sentiment-labeled tweets. Since we are not using the target language's labels during training, we expect our sentiment classification model to also work on low resource languages where no such labels exist.

2 Background

2.1 Sentiment Dataset

We will use a publicly available sentiment-labeled dataset (Mozeti et al., 2016) containing 1.6 million tweets in 13 languages, listed in Figure 1. Since all existing sentiment classifiers require labeled data, this allows us to train models on many languages besides English, which accounts for 42% of tweets. However, there are still hundreds of languages popular on Twitter with insufficient amounts of sentiment-labeled data. Morever, the

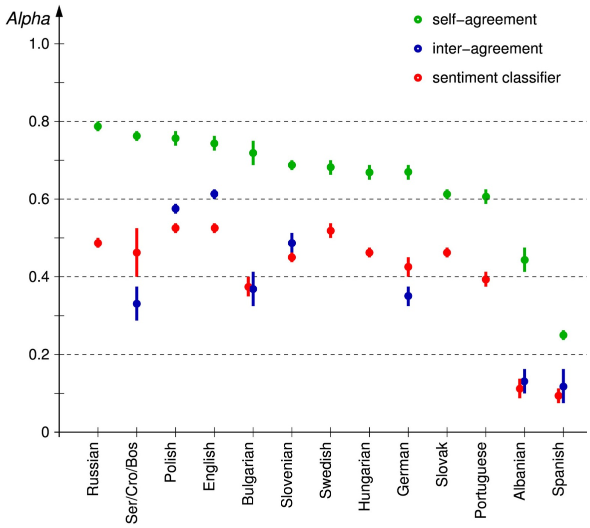

Figure 1: The dataset generated had low levels of *self-agreement* between an annotator's earlier and later evaluations, and *inter-agreement* between different annotators.

imprecision in the dataset serves as a rough upper bound for the accuracy of a sentiment classifier trained on it. In an effort to avoid using imprecise data, we are incentivized to explore using precise data from other languages.

2.2 UnicodeCNN

The *Unicode Convolutional Neural Network* (UnicodeCNN) (Author, 2019) is a state of the art model for tweet geolocation which predicts the exact GPS coordinates of any tweet in any language at any location.

Inspired by the *Character-Level Convolutional Neural Network* (CLCNN) (Zhang et al., 2015), which classifies English text by using 70 characters commonly found in English text, it operates at the character (sub-word) level. Whereas the CLCNN and other NLP networks only work on a specific language, the UnicodeCNN is able to support all languages by ecoding all characters into

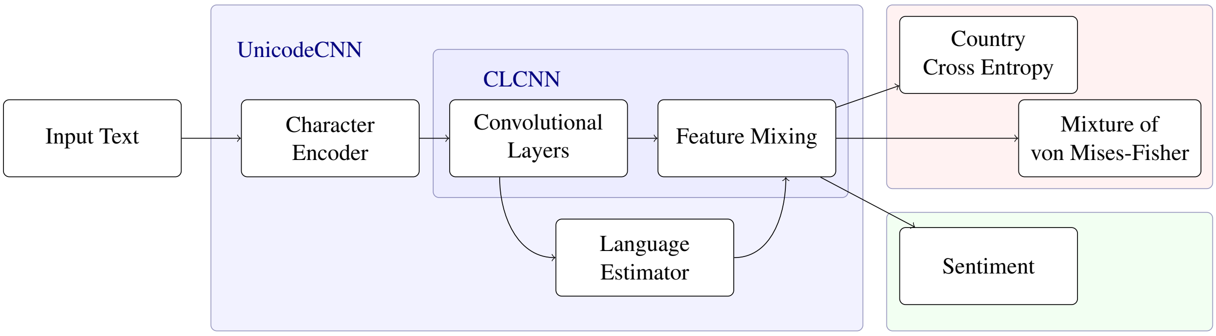

Figure 2: We replace UnicodeCNN's final *Mixture of von Mises-Fisher* (MvMF) layer with a layer of sentiment classification.

Unicode and estimating the tweet's language.

It was trained on all 900 million geo-encoded tweets sent over the course of nine months by over 3.0 million users in over 100 languages.

3 Transfer Learning

It is common to train a sentiment classifier on a set of tweets in a single language, manually labeled according to whether the emotion expressed is *negative*, *neutral*, or *positive*.

A simple baseline is a bag-of-words model which learns to predict the sentiment of a tweet in that language by remembering the number of occurrences of each word and disregarding all other features including punctuation, grammar, and the relationships among words due to their positions (Mozeti et al., 2016).

3.1 Problem

Our goal is to evaluate the sentiment of tweets in any language, trained on sentiment-labeled tweets in another language.

The challenge was that all previous sentiment classifiers worked only on a single language, and it is not feasible to generate labelled datasets for all languages of interest. The UnicodeCNN offers our model two improvements.

1. It's trained on geolocation, a similar problem for which there exists a massive amount of reliable auto-generated data.
2. While the CLCNN may be extended to work on other languages similarly sequential and atomic in nature, the UnicodeCNN can already analyze text in any language.

In this experiment, we will use a UnicodeCNN pretrained on the geolocation problem, and replace the final MvMF layer with a layer of sentiment classification as shown in Figure 2.

3.2 Experiments

We will retrain some layers of the model on a training set and evaluate on a testing set, where the languages of the two sets may or may not agree, as shown in Figure 3. We will repeat this process by varying the number of layers retrained, as well as varying the languages and sizes of the training and testing sets. We've chosen to train on English tweets because of the availability of data, and we've chosen to test on Portuguese because we expect great improvements on the accuracy of existing models. While we expect strong results due to the similarity between the two languages, we believe the determining factor is the quality and quantity of the source data.

#	Retrain	Train	Test
1	All	English	English
2	All	Portuguese	Portuguese
3	All	English	Portuguese
4	Final	English	English
5	Final	Portuguese	Portuguese
6	Final	English	Portuguese

Figure 3: Main experiments highlighted.

Then, we will repeat this experiment, replacing the Portuguese tweets with tweets in other languages listed in Figure 1, and compare against the bag-of-words models.

This would prove that an English tweet sentiment classifier can zero-shot learn to evaluate sentiment of tweets in other languages, specifically low resource languages where no sentiment-labeled data exist.

References

Stephen R Anderson. 2010. How many languages are there in the world. *Linguistic Society of America*.

Anonymous Author. 2019. Geolocating tweets in any language at any location. *Submitted*.

Lichan Hong, Gregorio Convertino, and Ed Huai hsin Chi. 2011. Language matters in twitter: A large scale study. In *ICWSM*.

Igor Mozeti, Miha Grar, and Jasmina Smailovi. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PLOS ONE*, 11(5):1–26.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning - the good, the bad and the ugly. *CoRR*, abs/1703.04394.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *CoRR*, abs/1509.01626.

Jun Zhao, Kang Liu, and Liheng Xu. 2016. Sentiment analysis: Mining opinions, sentiments, and emotions bing liu (university of illinois at chicago) cambridge university press, 2015 isbn 9781107017894. *Computational Linguistics*, 42:1–4.