Project Proposal – Language classifiers

Tianjin Ji
tji33@gatech.edu

Project Overview

In this project, we are interested in using machine learning techniques to build language classifiers to identify the language English for each tweet from the UMass Global English on Twitter Dataset. Natural Language processing is a challenging field and social media languages are informal and hard to identify, which makes it hard to correctly classify the tweet. However, language identifier for social media data such as tweets could be the first step in processing and understanding messy and unstructured data.

Data description

The UMass Global English on Twitter Dataset can be downloaded from Kaggle: https://www.kaggle.com/datasets/rtatman/the-umass-global-english-on-twitter-dataset. In this dataset, we have 10,502 annotated tweets from 130 countries. For each row of the data, the columns of our interest are Tweet, Country(where the tweet is sent), Definitely English, Ambiguous, and Definitely not English. The Definitely English, Ambiguous, and Definitely not English columns have previously annotated binary value of 0 and 1 and are our response variables. On the other hand, the Country and Tweet columns has raw text strings and are our predictors. The dataset has several other columns such as Tweet ID, Date(of the tweet), Automatically Generated Tweets, Code-switched, and Ambiguous due to Named Entities. In this project, we will not study these columns because we want to focus on the prediction of whether a tweet is English or not.

Research Question and methodology

The question we want to address is whether a given tweet belongs to Definitely English, Ambiguous or Definitely not English category. The categories are mutually exclusive as a tweet cannot be both Definitely English and Ambiguous at the same time. This question can help us identify whether a tweet is English or not.
In this project, we will use classification methods such as KNN, SVM, Logistics Regression and other methods to determine of a given tweet is English, Ambiguous or definitely not English. The dataset is first divided into 80/20

training and testing split. Since both of our predictors (Country and tweets) are strings, we need to preprocess these into numerical values to fit into models. We will use a numerical encoder to convert “Country” column into numerical values. For the tweets data, we will use TfidfVectorizer from Python’s sklearn package to convert the text data into TF-IDF form. The TF-IDF form is a way of getting the relative frequency of the words in a document, and the result is stored in a sparse matrix. Then we fit the models to the preprocessed data and perform cross validation to find the best tuning parameter. After finding the optimal model, we will compare the performance different models using test data. Accuracy rate and confusion matrix can be used to evaluate and compare the models.