Creating a Text Classifier for English from Different English Speaking Countries

John Speaks
jspeaks2@illinois.edu

Abstract

This paper explores the development of a text classifier for English tweets originating from six different English-speaking countries. Utilizing logistic regression and varying n-gram tokenization (unigram, bigram, and trigram models), the study investigates the impact of different methods of tokenization on the accuracy of the classifier. The training corpus consists of over one million English tweets, with a focus on classifying tweets based on country of origin. Evaluation metrics, including precision, recall, and F-score, reveal varying performance across the different n-gram models. The results indicate that the unigram model, despite its simplicity, achieves high accuracy. However, when applying the trained models to a new domain of hotel reviews, significant discrepancies in labeling are observed, suggesting potential challenges in cross-domain applicability. The study concludes by discussing the implications and potential applications of such classifiers in linguistic research and domain-specific tasks.

1 Introduction

Understanding the language variation across different regions and countries is a fundamental aspect of linguistics and natural language processing (NLP). English is a global language and exhibits diverse dialects and linguistic features across various countries. This variability is both a challenge and opportunity in tasks such as sentiment analysis and text classification, where accurately identifying the origin of text can provide valuable insights.

This paper focuses on the development of a text classifier capable of distinguishing English tweets originating from six different English-speaking countries: Australia, Ireland, New Zealand, South Africa, the United Kingdom, and the United States. The primary objective is to investigate the effectiveness of various linguistic features, namely n-gram tokenization, in classifying tweets based on their country of origin.

Logistic regression, a supervised learning algorithm, was used to build the text classifiers. By utilizing different n-gram models, including unigram, bigram, and trigram, the paper explores how the granularity of linguistic features influences the classifier's performance. Specifically, it is examined how the presence or absence of specific n-grams contributes to the classification accuracy across different countries.

Furthermore, the application of the trained classifier was expanded to a new domain by examining its effectiveness in labeling hotel reviews obtained from TripAdvisor. This cross-domain evaluation allows us to assess the classifier's adaptability and performance in classifying text from diverse sources.

Overall, this study contributes to the understanding of language variation in English tweets from different countries. The findings have implications for various NLP tasks, including geolocation-based content analysis, sentiment analysis across regions, and linguistic research on national dialects.

2 Training the Models

The classifiers in this paper were trained using logistic regression. Logistic regression in scikit-learn (sklearn) is a supervised learning algorithm used for classification tasks. In this process, three models (word unigram, bigram, and trigram models) were created to explore how different n-gram tokenization affects the accuracy and usability of the model. In the context of language n-gram analysis, logistic regression was used to classify text documents into different categories based on the presence or absence of specific n-grams.

Logistic regression uses a sigmoid activation function to drive all values to 0 or 1. For this reason, it is typically used for two-class classification. The classifiers in this paper are six-class classifiers, meaning the sklearn logistic regression

function is actually implementing a softmax function to compute the probability distribution over all classes. The softmax function takes a vector of raw scores and converts them into probabilities. This is known as softmax regression or multinomial logistic regression.

The training corpus for these models was a collection of over one million English tweets tagged with their country of origin. Tweets have a limited size of 280 characters. For this reason, going into training the model, it was expected that tweets would not have a large enough set of words to be able to identify a specific dialect by word occurrences, and instead tweets would be easier classified using larger N-grams that would identify regional phrases. For training, 180 thousand tweets were used while 20 thousand were reserved for testing. For each model, a vocabulary of 10,000 tokens was used where each token was an n-gram where n was one, two, or three depending on the model. Tweets were randomly sampled in equal proportion from all of the six countries in the dataset.

To prepare the data, a segment of the corpus was first imported into a Jupyter notebook file and then turned into a pandas dataframe with appropriate column titles and no invalid rows. Due to a lack of computing resources, the entire dataset (>1.2 million items) could not be loaded and for that reason only 200 thousand items were loaded. Next, the data was vectorized for training using the sklearn CountVectorizer. After a transformation to a vector, the data was fed into the model for training, and then the resulting model was tested against the 20 thousand reserved tweets to generate standard metrics like precision, recall, and f-score. Finally, the model was saved for later use on additional corpora.

More details about implementation and the python code can be found in the GitHub repository corresponding to this project[1].

3 Evaluating the Models

Each of the uni-, bi-, and tri-gram models were evaluated against testing data and their full evaluations are in the A appendix (figures 3-5). For each language, there was no significant variation in the recall and accuracy scores. This indicates that none of the models have a significant issue with being perpetually put in another category, as in that case an imbalance would be expected. F-score data was observed across the six English speaking countries in the dataset across each model as seen in Figure 1.

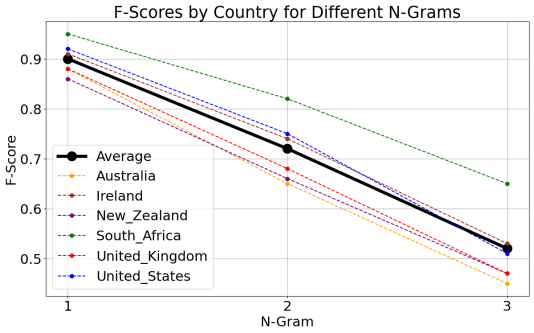

Figure 1: F-score by language for each model

It can be observed that across all six countries, the F-score decreased as the lengths of N-grams used for training decreased. Looking at the average across all countries, the F-score for the unigram model was high at 0.90, the bigram lower at 0.72, and the trigram lowest with a score of 0.52. All the models have shown a reasonable level of learning considering a random model trained on equally distributed data across six classes would be expected to have an F-score of 0.17. As seen in figure 1, all of the countries in the model experienced a similar decay as the training N-grams increased. This shows that no countries benefitted from a particular model.

Using these standard metrics, it could then be assumed that the best model trained on the data is the unigram model. At the same time, we should be skeptical of this result. Since all of the data is representative of English, the vocabulary should be fairly similar across all countries, and the limited length of a tweet would not allow for a large representative distribution of words by country dialect. The results, instead, show the opposite, where using the occurrence of only individual words are the tweets being classified.

A likely reason that only a vocabulary of individual words was needed for a most accurate model is that users from different countries are more likely to use Twitter for different topics and write tweets in different registers. For example, a user in the United Kingdom is probably much more likely to tweet about issues in British politics or events in London than someone from New Zealand would be. In this sense, the unigram classifier takes an approach of topic classification instead of using

[1] https://github.com/JTSIV1/English-Speaking-Country-Text-Classification-Model

linguistic features. Bi- and tri-gram models likely use more language-based patterns of which words occur together in different dialects to make their classifications.

4 Using the Models in a New Domain

There are many possible uses of a model that classifies English by country in the linguistic domain. For example, because so much internet data is anonymized, the geographic origin of a particular piece of content is often obscured or difficult to recover. A successful classifier like the one introduced in this paper could fix that issue, and allow for research to link country of origin to other features online.

An example pursued in this paper is adding country labels to a corpus of hotel reviews from TripAdvisor. Hotel revies can be a great asset to linguists because they are public and thus easy to find, often long enough to include a good sample of an individual's writing while also being short enough to not be a storage burden, and they come with a built-in sentiment label which is the rating they attach to the hotel with their review. A research question that could be pursued with a corpus of country labeled hotel reviews is "do English speakers in different countries express dissatisfaction in a way different enough to be demonstrable and identifiable by a classifier?" With such a research question, a linguist could begin to explore how different dialects of English go about expressing satisfaction/dissatisfaction.

The corpus chosen for the example labeling is one of 20.5 thousand hotel reviews from trip advisor. For each review in the corpus there is the content of the review and the rating the user gave as a singular integer. To label the reviews, the models previously trained on the Twitter data were loaded, and the hotel reviews were vectorized in the same manner as the tweets. The vectorized reviews were fed into the stored model and a prediction was generated. These predictions were then appended to the dataframe and saved. For this paper, this was done once for each trained model (uni-, bi-, and trigram). The code for the application of the model is included in the previously mentioned GitHub repository[2].

When labeling the reviews, completely different distributions of country labels occurred for each model as seen in figure 2.

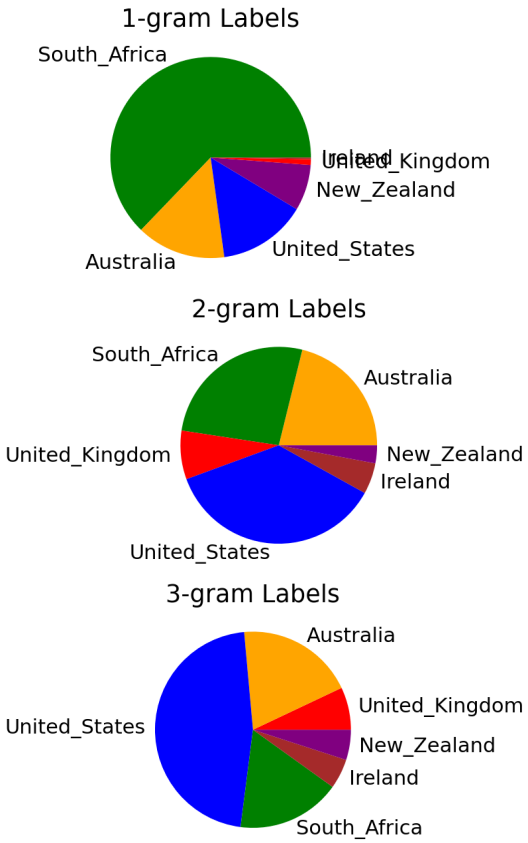

Figure 2: Proportion of Labels by Country

This indicates that there is likely an issue with the models. While the unigram model had the highest average F-score, it produced the least likely labeling of the review corpus. Considering most trip advisor users are from the United States according to the website itself, it is highly unlikely that the corpus would have such a large proportion of South African reviews and a low proportion of ones from the United States. The bigram and trigram models are both more reasonable in that sense, but there is no way to say if one is more accurate than the other since there is no ground truth in the review corpus like for the original Twitter data. In this scenario, while an accurate country labelling would enable interesting linguistic research, this model seems to be trained on a corpus too dissimilar from the one being analyzed. It is possible that the tone and register of reviews from one country may be more similar to the tone and register of a different country on Twitter. A more in-depth study would have to take place in the future to determine how registers differ on the two platforms.

These models could still be used for linguistic study, just in a more similar domain. For example,

[2] https://github.com/JTSIV1/English-Speaking-Country-Text-Classification-Model

not all Twitter data is available with a location tag. These specially trained for Twitter models could be used to label those tweets and generate a larger location tagged corpus for linguistic study.

5 Conclusions

Using vectorization with different length N-grams, it was possible to train three models to classify English tweets by country across six countries. Each model was demonstrably effective, but they each performed very differently on the test data. Each country experienced the same variation across N-grams, and no one model was a better fit for a specific country. Future studies could adapt each model to perform differently to the ones in this paper by tweaking the parameters of training for each model. When applying these models to the domain of hotel reviews, completely different labels were produced. A model generated on Twitter data may be too dissimilar to the corpus of hotel reviews and require a broader corpus to generate more consistent and accurate results. At the same time, it is impossible to say that any of the generated labels are inaccurate because there is no ground truth, just that they are different so at least two of the models are very inaccurate out of domain. These models could still be used effectively for future experiments as long as they are used to label data in a more similar domain.

A Appendix of Standard Metrics

	Precision	Recall	F1-score	Support
Australia	0.88	0.88	0.88	3350
Ireland	0.91	0.91	0.91	3386
New Zealand	0.86	0.86	0.86	3125
South Africa	0.95	0.95	0.95	3404
United Kingdom	0.88	0.88	0.88	3367
United States	0.92	0.92	0.92	3368
accuracy			0.90	20000
macro avg	0.90	0.90	0.90	20000
weighted avg	0.90	0.90	0.90	20000

Figure 3: Standard Metrics for Unigram Model

	Precision	Recall	F1-score	Support
Australia	0.65	0.63	0.64	3373
Ireland	0.74	0.72	0.73	3377
New Zealand	0.66	0.65	0.66	3160
South Africa	0.82	0.85	0.84	3341
United Kingdom	0.68	0.68	0.68	3339
United States	0.75	0.78	0.76	3410
accuracy			0.72	20000
macro avg	0.72	0.72	0.72	20000
weighted avg	0.72	0.72	0.72	20000

Figure 4: Standard Metrics for Unigram Model

	Precision	Recall	F1-score	Support
Australia	0.45	0.41	0.43	3303
Ireland	0.53	0.50	0.51	3333
New Zealand	0.47	0.44	0.45	3134
South Africa	0.65	0.69	0.67	3487
United Kingdom	0.47	0.47	0.47	3389
United States	0.51	0.59	0.55	3354
accuracy			0.52	20000
macro avg	0.51	0.52	0.52	20000
weighted avg	0.52	0.52	0.52	20000

Figure 5: Standard Metrics for Unigram Model