Predicting Tweet Specificity Using Machine Learning Algorithms

Maanasa Darisi
mvd456
mdarisi074@utexas.e

Misty Peng
mp46528
misty.peng@utexas.edu

Abstract

The idea of tweet specificity can be characterized by the level of details included in the work. This includes the concept of whether or not the entirety behind the text's intention can be comprehended just by the text on its own, and the notion that a more specific tweet would need less background knowledge to be understood. Using a machine learning algorithm to develop predictions of how specific a line of text is could prove useful for many functions, including summarization and analyzing speech to determine its effectiveness. The idea that a text is more specific indicates that less background knowledge is needed to fully understand it, making it a clearer statement.

1 Related Work

A 2019 paper titled *Predicting and Analyzing Language Specificity in Social Media* posts also aimed to quantify the level of specificity in a dataset of tweets. They rated each tweet on a scale between 1 and 5 and used their results to analyze the social and mental health factors that associate with language specificity.

2 Data

The dataset that we are using is from the GitHub page of Yang Zhong, an author of the paper referenced in the previous section. The dataset itself contains over 7,200 tweets and each tweet is followed by a score between 1 and 5, with 1 indicating the tweet is more general and 5, more specific.

Specificity is based on how much background context is needed to to understand the complete intent behind the tweet. If less context is needed, then the tweet is more specific. These scores are given on a continuous basis rather than a categorical one. This dataset previously contained more demographic information such as the Twitter user's age and level of education as it came from another study that aimed to use those tweets to predict a user's political ideology removed in the version of the dataset that we will be using.

3 Methodology

We will be using two different supervised machine-learning techniques to predict the specificity of these tweets. The first will be a linear regression model, as our predictions will be on a continuous scale. The GitHub account from which we are pulling our data includes two other files where the data is already portioned off into two separate datasets. Using those, we can train our model on one and test it on the other. The features that we will be using to make our predictions include the length of the tweets, part-of-speech tagging, word frequencies, as well as the number of named entities. All of these features have some capability of influencing how much detail is in each tweet. After we have created our model and run it on the testing data, we will be able to evaluate the model using error rates, R^2, and the Pearson correlation.

The second method that we will be using to make our predictions is bag-of-words and TF-IDF. Here, we can calculate specificity from scores of individual words. These scores will come from the individual words and can then be normalized to fit between 1 and 5 so that these scores can be compared to the original output as well the output from the linear regression model. Afterward we can evaluate the model using the percentage error from the original score.

4 Team Structure

As we are using two different methods for predicting the tweet specificity, we have divided the work amongst us so that each partner tackles one algorithm. Maanasa will be completing the linear regression model and the components attached to that, such as part of speech tagging. Misty will be working on Bag-of-Words and TF-IDF scores. Evaluations for both algorithms will be done by each respective partner and together analyzed to draw conclusions.

5 Results

5.1 TF-IDF Scores

The TF-IDF model calculates term frequency–inverse document frequency as opposed to the Bag-of-Words model originally calculated. Using this algorithm, each unique word occurring in the series of tweets is assigned a TF-IDF value. From these, the average score for each tweet is calculated and compared to the original annotation. Terms like hashtags or <USER> and <URL> have inconsistent TF-IDF scores due to how unique or too common they are, respectively. Due to these inconsistencies, it was important for us to see how their individual calculated TF-IDF scores affect tweet scoring and if we could fix that by setting certain scores for these terms.

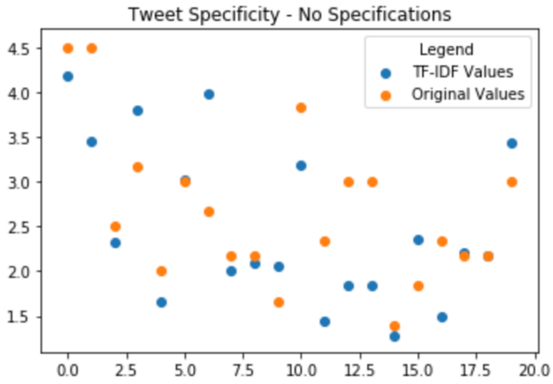

Figure 1. Original values in orange and in blue the new calculations of TF-IDF, not considering special cases like hashtags and mentions of other users and websites.

Figure 1 depicts tweets scores calculated with only TF-IDF word scores. The total distance between the scores of all the tweets is 4996.0316 and the average is 0.8748 for these scores.

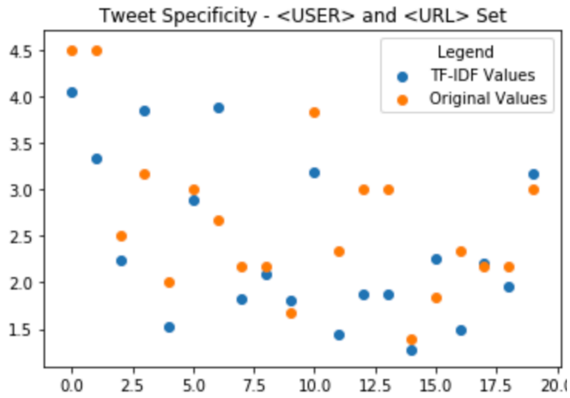

Figure 2. Original values in orange and in blue are the TF-IDF scores of tweets using set values for hashtags and mentions of users and websites instead of their calculated TF-IDF values.

Figure 2 depicts a graph of the same twenty sentences with the plot points in orange being the same as Figure 1 while the blue plot points represent scores that were calculated with a set value for hashtag appearances and a set value for mentions of other users and website URLs. The total difference for this set is 4799.5245 and the average became 0.840. We set the values for hashtag appearances at 8 and mentions at 9 for specificity.

These numbers (8 and 9) were specifically chosen after testing a range of scores because they produced the most accurate tweet scores. Since these scores produce the most similar overall scores as the annotation, this process shows that hashtags and mentions are relatively specific and strongly affect tweet specificity.

5.2 Linear Regression Model

The next model that was used for predicting specificity scores was Linear Regression. We started by creating the various features we would need in order for the model to create its predictions. The designed features included the length or the tweet, the number of words for each tweet, the number of named entities in each tweet, part of speech tags, as well of the frequencies of individual words in each tweet.

When looking at Pearson Correlation values for these features, tweet length had the highest with a score of 0.541. Next highest, number of words and number of named entities had scores of 0.464 and 0.379 respectively. After this, the part of speech tags, IN, NNP, NN, CD, and DT followed in score. In tenth place, word frequency sums had a correlation score of 0.202.

After creating these features, we trained the model and then fit it onto our testing data, which consisted of 1000 tweets.

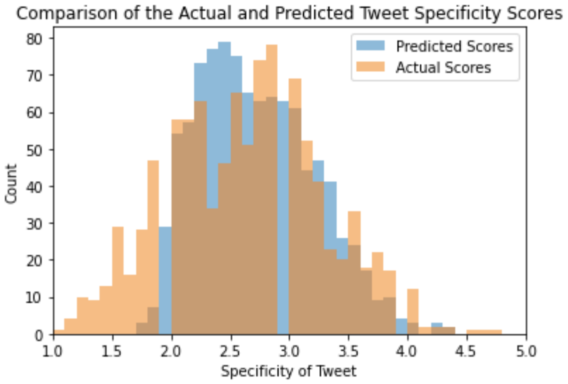

Figure 3: Distribution of predicted vs actual tweet specificity scores for testing data

As seen in Figure 3, while the actual scores of specificity comfortably ranged from 1 to 5, the predicted values strayed below a score of 2 very few times. This drawback in the model accounts for a lot of the error in the data.

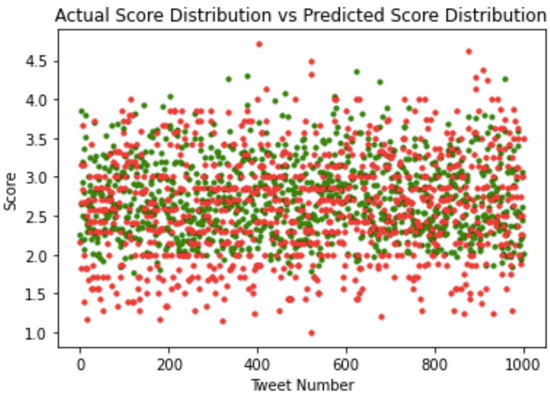

Figure 4: Distribution of predicted vs actual tweet specificity score for testing data, red = actual, green = predicted

Here, Figure 4 provides another view of the original data and the prediction values. Again, we see the predicted scores rarely drop below 2 or above 4. In order to get a clearer view of the data, we can next look at Figure 5.

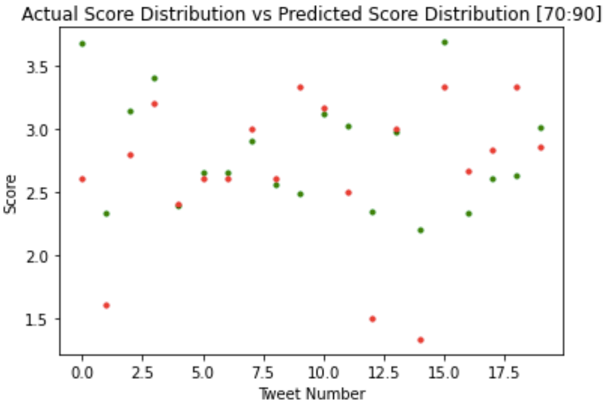

Figure 5: Distribution of predicted vs actual tweet specificity score for testing data, index 70 through 90, red = actual, green = predicted

Looking at Figure 5, where we randomly selected a narrower range to observe, we get a more distinct view of how the predictions hold up in comparison to the original scores. As we can see, while some predictions are incredibly close to the expected value, others are quite a distance apart.

In order to better understand how well the model fit, we then calculated the r^2 and the root mean square error (RMSE). The RMSE value for this linear regression model was 0.52358, which indicates that for a range of scores between 1 and 5, it was not an excellent fit. The r^2 value also lets us know the same, as it was 0.36492, which while positive, is not as close to 1 as a good fit would've indicated.

Finally, we looked at the absolute error between the actual scores and the predicted ones. The mean absolute error was 0.419 with a standard deviation of 0.315 and a median of 0.367. For comparison, when looking at solely TF-IDF for predicting values, the mean absolute error was 1.630 while the median was 0.717. When looking at the quantiles, the 25th quantile for absolute error was 0.162, and 75th quantile was 0.612. The maximum absolute error was 1.636. These numbers indicate that while our predictions were not completely accurate, the features we designed put us closer to the correct values than TF-IDF on its own. The implementation of the features we created seem to be a solid start but in order to decrease the error even more, the creation of more features that correlate greatly with the original specificity scores is necessary, and a next step may be to combine the two methods used in this project to create a better fit.

References

Gao, Y., Zhong, Y., Preoţiuc-Pietro, D., & Li, J. J. (2019). *Predicting and Analyzing Language Specificity in Social Media Posts.* Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 6415-6422.

Preot¸iuc-Pietro, D.; Liu, Y.; Hopkins, D.; and Ungar, L. H. 2017. *Beyond Binary Labels: Political Ideology Prediction of Twitter Users.* ACL, 729–740.