Overview

For our project, we decided to explore how users tend to speak when posting to different social media platforms. Are user sentiments on different platforms significantly different? If so, could we train a model to predict which platform a given post was written for?

To answer these questions, we decided to

- collect tweets, Instagram captions, Reddit posts and YouTube comments,
- compute features that capture the sentiment of a post,
- train a model to predict which platform a post was written for based only on these features, and finally
- interpret the results.

Our features included various measures of positive and negative emotion, emotional intensity, and references to men and women.

Results

We trained three kinds of models for the task of platform prediction: logistic regression, decision trees, and random forests. Since our class labels were balanced, we chose accuracy as our metric of success.

The logistic regression model trained to classify data into four classes has an accuracy of 32%. This model had a tendency to conflate posts from Instagram with those from Reddit, so we decided to try training a logistic regression model on only two classes: “Instagram or Reddit” and “Twitter or YouTube”. This model achieved an accuracy of 60%.

The Decision tree model we trained had an accuracy of ~41%. Interestingly, it learned the 140 character limit for tweets and was able to distinguish tweets from the others. When examining the feature importance of this decision tree, “length” and “word count” accounted for over 88% of the feature importance.

Finally, we trained a random forest model with 150 estimators and a maximum depth of 10. This produced only marginal improvements over the decision tree, bringing the accuracy up to 42%.

Conclusions

Across all models, the length of the post was the strongest platform predictor. The remaining features did not significantly contribute to the models. This suggests that Twitter, Instagram, Reddit and YouTube do not differ significantly in positivity or negativity of sentiment, or male or female references. To improve the models in the future, we could produce a wider range of features capturing more emotional sentiments or topics.