Which Social Media Platform?

Juan Pablo De Rasis, Aziz Gülen, Brantley Vose, Yang Yang

TEAM OSU MATH

Idea

Question: Do users tend to speak differently when posting to different social media platforms?

- Collect posts from various platforms
- Compute features that capture the mood of a post
- Train models that predict which platform a given post was on based only on these features

Data Sources

- Twitter
  - Sentiment140 dataset
  - Twitter stream: random sample of tweets from one day
- Instagram
  - Kaggle
- YouTube
  - Perchance’s random YouTube video generator (100 videos).
  - Egbert Bouman’s YouTube comment scraper (<351 comments per video).
- Reddit
  - Kaggle

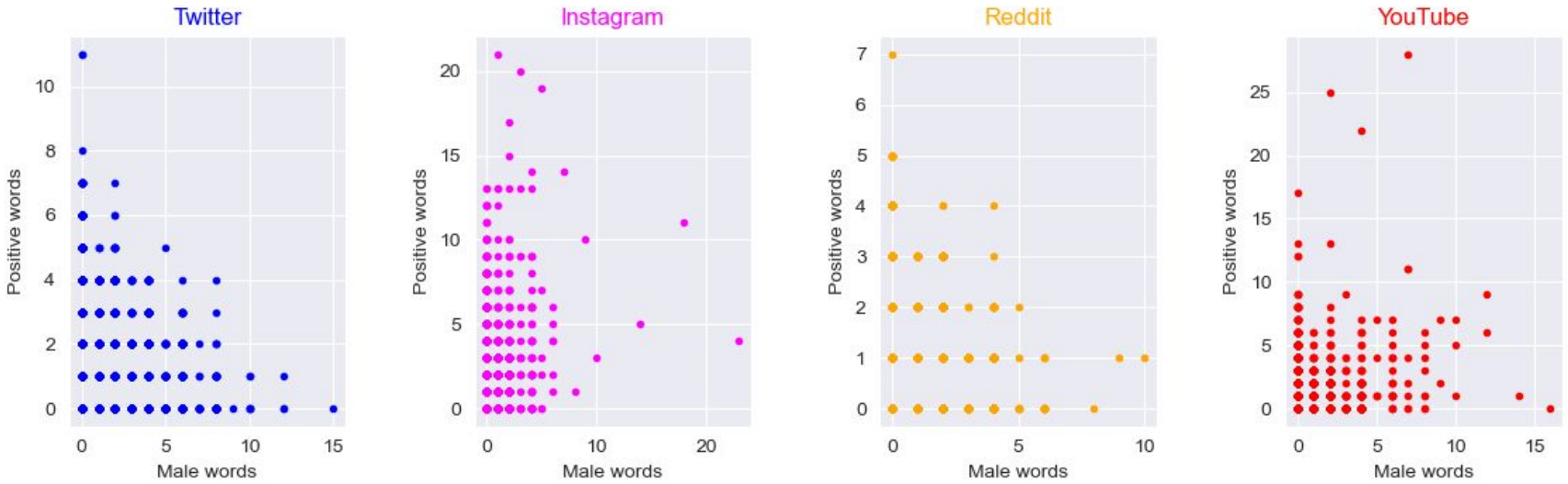
Features

- Word Count
- Positive Words
- Negative Words
  - Word lists from *Mining and Summarizing Customer Reviews* by Minqing Hu and Bing Liu
- Male Words
- Female Words
  - Word lists taken from the “Jailbreak the Patriarchy” chrome extension by Danielle Sucher
- Afinn Score
- Mood Intensity : (Pos. + Neg.)/(Word Count)
- Positivity: (Pos. - Neg.) / (Word Count)

EDA

Two interesting Observations:

- Twitter (balanced mood), others (positive mood)
- Instagram's plot on male vs positive has a different behavior from the other three (much higher slope).



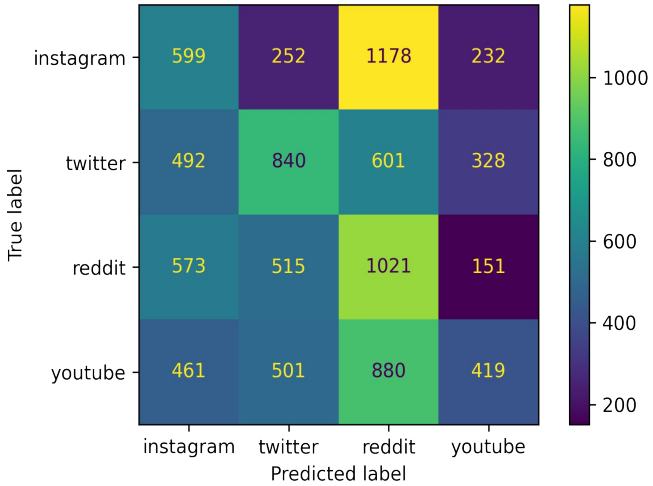
Models

We trained three kinds of models:

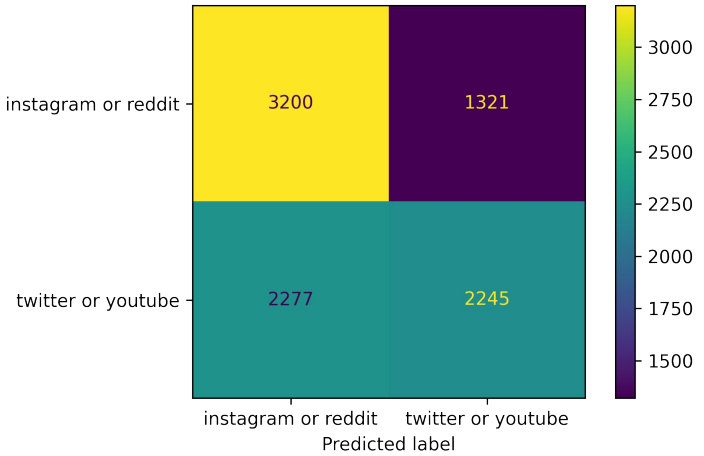
- Logistic regression
- Decision Tree
- Random Forest

Logistic Regression

Four class classification accuracy: 0.32

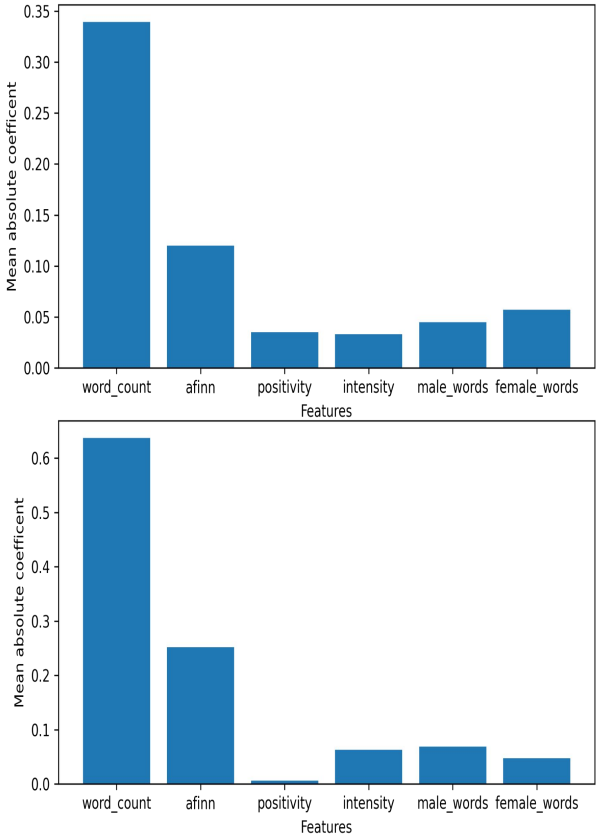


Two class classification accuracy: 0.60



Logistic Regression

- Word count was the most important feature in both the four-class (top) and two-class (bottom) cases, followed by afinn

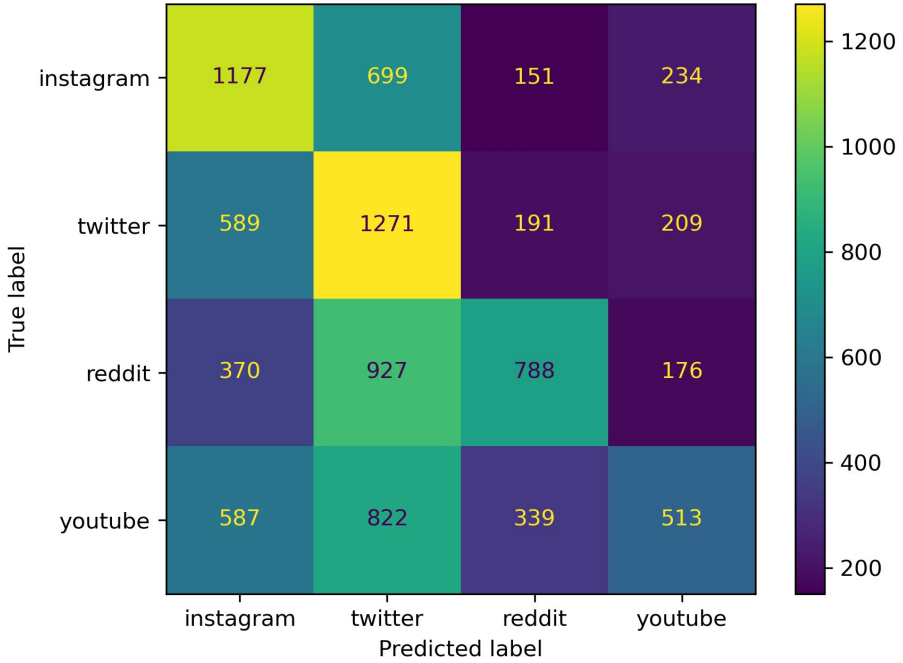


Decision Trees

Best decision tree model had max_depth = 8

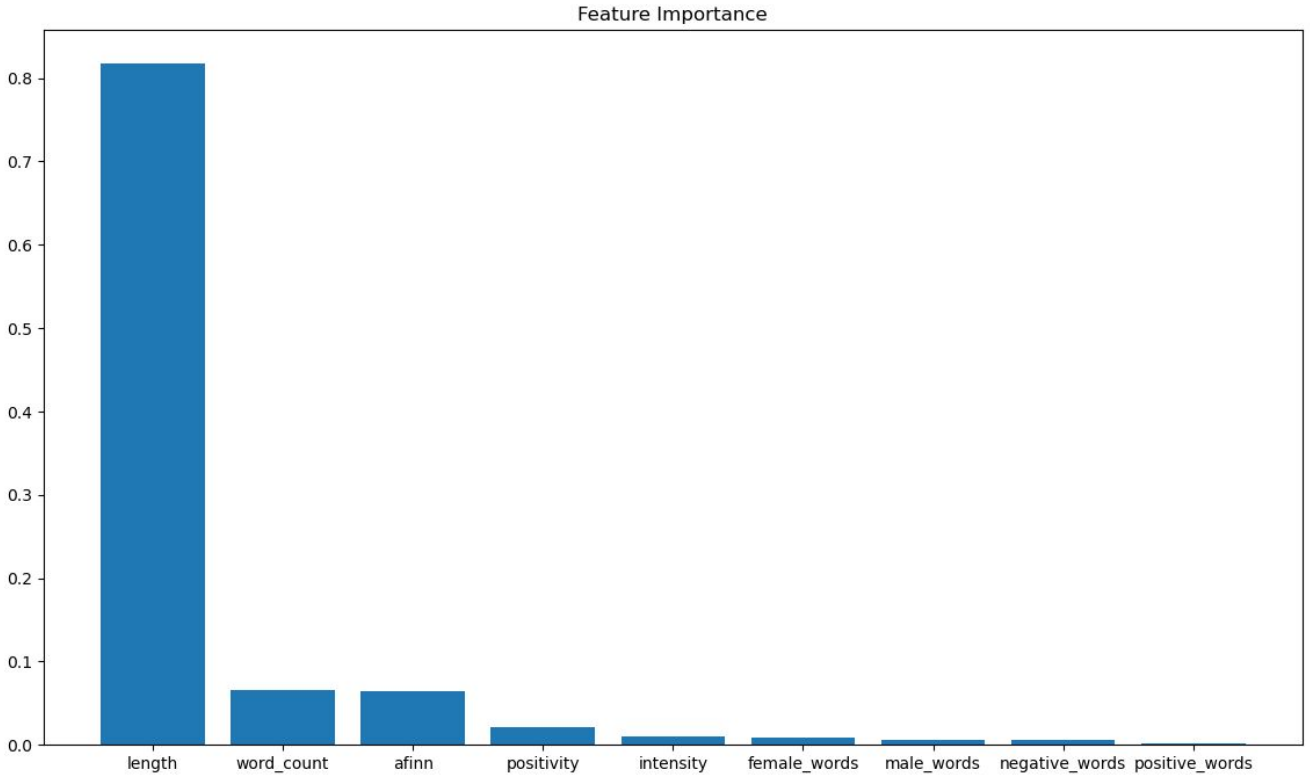
It was able to learn 140 character limit for twitter posts.

Accuracy: ~41%



Decision Tree Feature Importance

Length and word count together account for over 88% of feature importance.

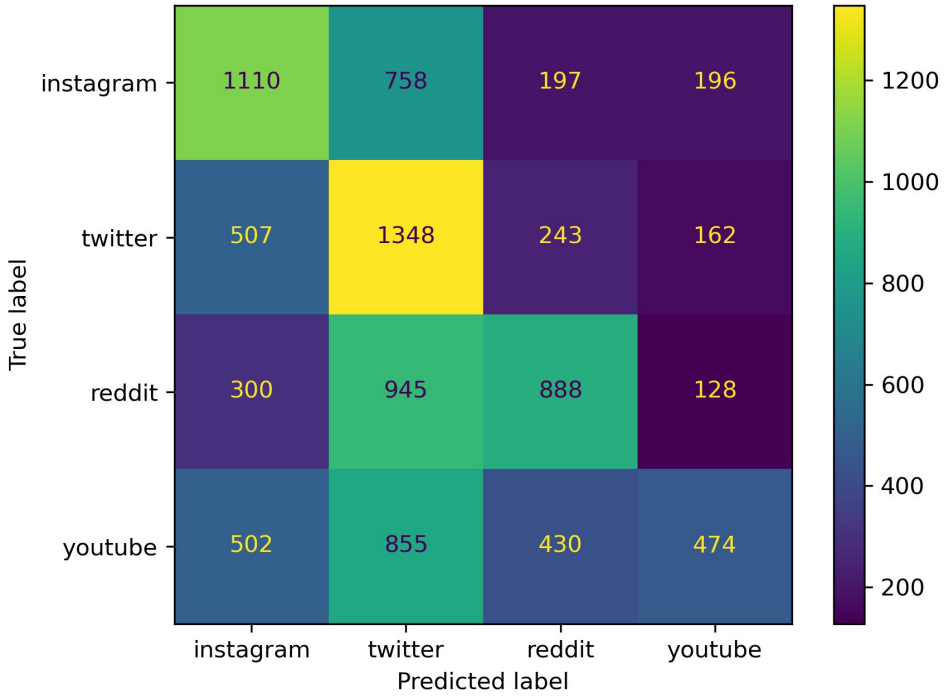


Random Forest

Trained a random forest with

- 150 estimators
- max_depth = 10

Accuracy: ~42%



Conclusions

- Length of post was the strongest feature in every model
- Additional features did not significantly contribute to model performance
- What we should do next in order improve the performance of our models:
    - More Feature Engineering