# Classifying Engaging Reddit Posts

•••

FiveThirtyEight
Dimitrios Vlahos

**What We Want to Know:**
What characteristics of a post on Reddit are most predictive of the overall interaction on the thread?

# Scraping Methodology

- Used back-up data obtained by a colleague; this data was scraped over the course of two days.
- Running personal scraper 30 minutes at a time; 10:00 AM CST and 10:30 PM CST, until obtaining an appropriate amount of samples (14K).
- Additional features in personal scrape: contents of post, and downvotes

# EDA at a Glance

- Amount of times a subreddit appeared on a thread that was gathered from 'hot' posts
- Avg. word count of successful post compared to non-successful posts.
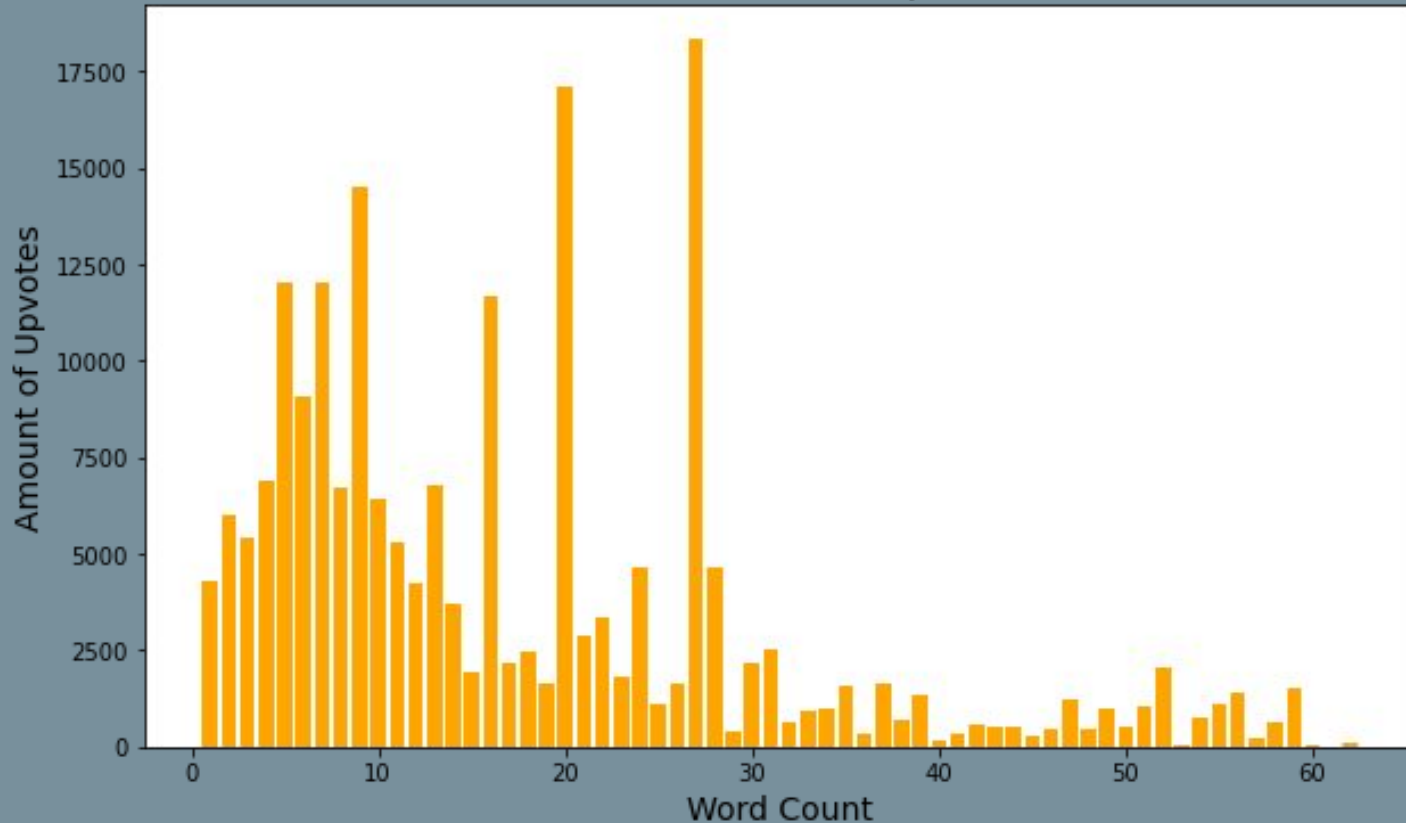
*Median Number of Comments:* **33**

*Baseline Accuracy:* **0.50**

| Subreddit | Threads in dataset |
|---|---|
| r/Superstonk | 428 |
| r/memes | 366 |
| r/antiwork | 136 |
| r/WhitePeopleTwitter | 123 |
| r/HolUp | 117 |
| r/Hololive | 114 |
| r/dankmemes | 112 |
| r/MadeMeSmile | 107 |
| r/politics | 103 |
| r/interestingasfuck | 99 |

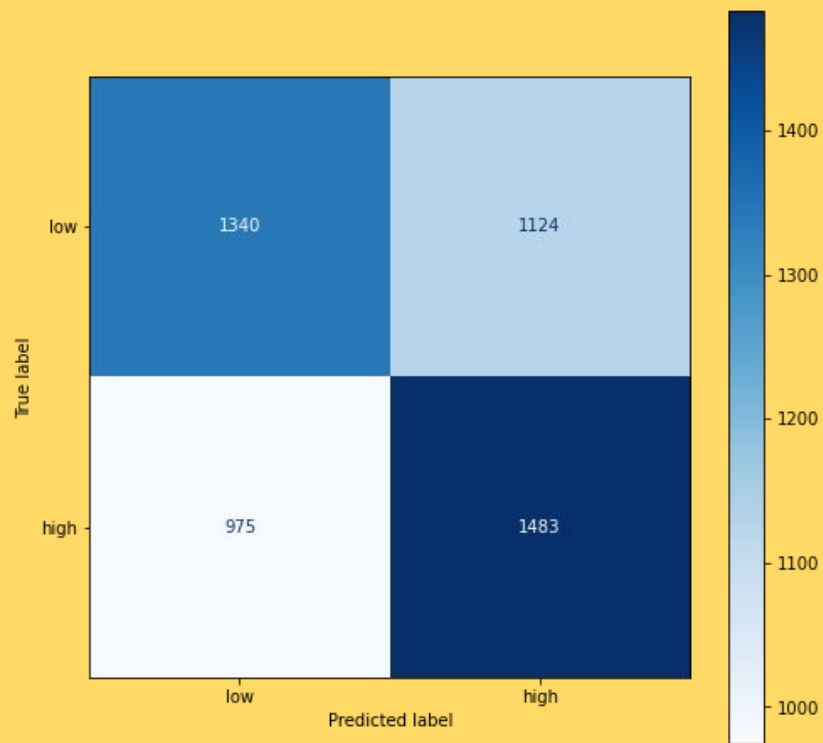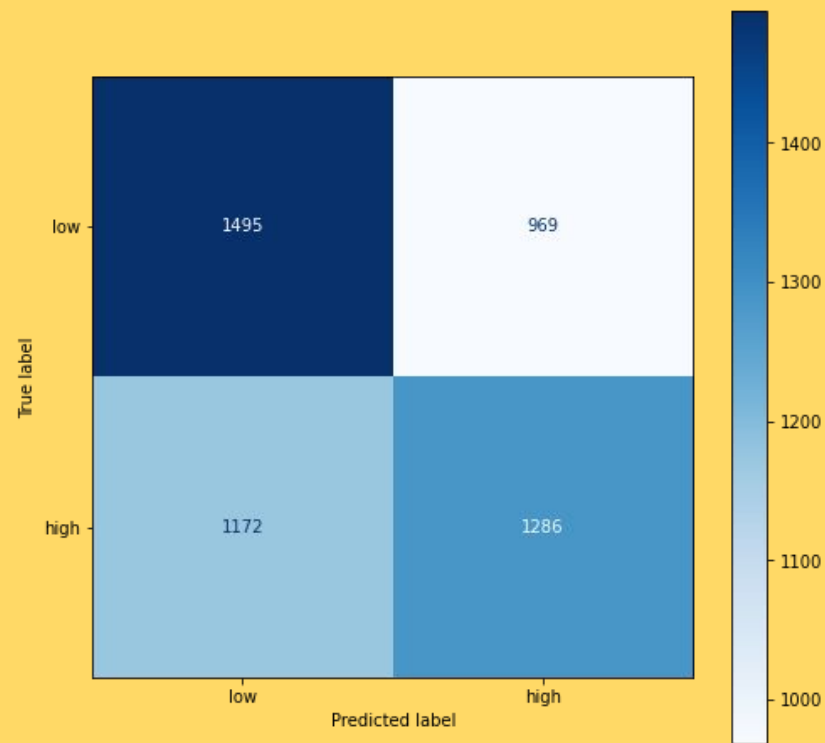| Successful | Avg. Word Count |
|---|---|
| No | 8.36 |
| Yes | 10.93 |

Number of Words in Title /w Upvotes on Thread

In general, from the bar graph there is a condensing of observations at a lower word count.

# Model Performance

| Metrics | Random Forest | Log. Regression |
|---------|---------------|-----------------|
| Test Score | **0.573** | 0.565 |
| Accuracy | **0.573** | 0.565 |
| Sensitivity | **0.603** | 0.523 |
| Precision | 0.568 | **0.570** |

**Random Forest Model**

**Logical Regression Model**

# Recommendations

- Post in one of the reddit threads listed above with a title that asks a question to get the most engagement through comments on the thread
- Keep word count in your title at 15 words or less.
- Check with our trained Random Forest Model to predict whether a potential post will be successful based on the wording used in the title of the thread.