Making It Big on Reddit

Larry Jackelen, Professional Reddit Lurker

Problem Statement

Reddit is a a social news website and forum where content is socially curated and promoted by site members through voting[1]. According to the site, in 2019, there were approximately 430 million monthly users[2]. With this many eyeballs, getting engagement on a Reddit post can have a great impact, which leads us to our problem:

What characteristics of a post on Reddit are most predictive of the overall interaction–as measured by the number of comments.

[1]https://www.techtarget.com/searchcio/definition/Reddit
[2]https://www.theverge.com/2020/12/1/21754984/reddit-dau-daily-users-revealed

Data Collection

Post information from Reddit was accessed using the PRAW API[3] once a day for 5 consecutive days in late May 2022

The posts of interest were those in Reddit's "hot" section[4] which aggregates and sorts posts by a proprietary algorithm that looks at the voting and age of posts, among other items.

The top 2,000 posts per day were kept for analysis. Post data included:

- Title
- Subreddit and Subreddit information
- Length of time the post has been on Reddit
- Number of Comments
- Site being linked

[3]https://praw.readthedocs.io/en/stable/index.html
[4]https://www.reddit.com/r/all/

NLP Analysis

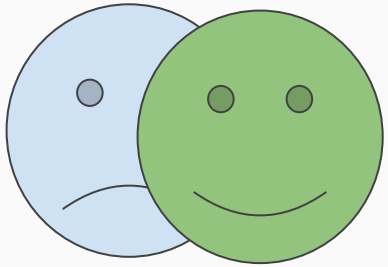
A major focus of model data preprocessing was the Natural Language Processing (NLP) of the post titles.

Techniques like stemming, lemmatization, and parts-of-speech tagging were utilized to gain insight into what hot post titles looked like.

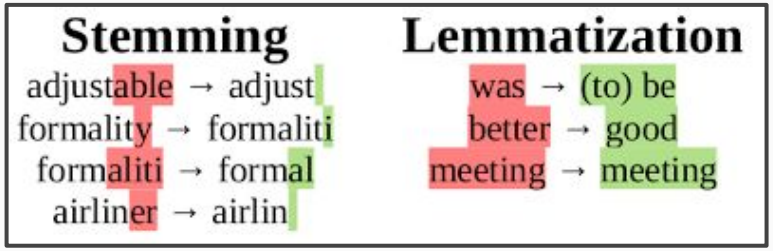
Ultimately, four dataset versions were taken to modelling:

- No NLP changes
- Stemming
- Lemmatization
- Lemmatization, then Stemming

Sentiment Analysis

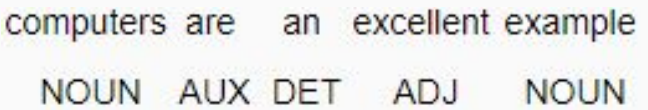


Emoji Counts



https://devopedia.org/lemmatization

Parts of Speech Tagging



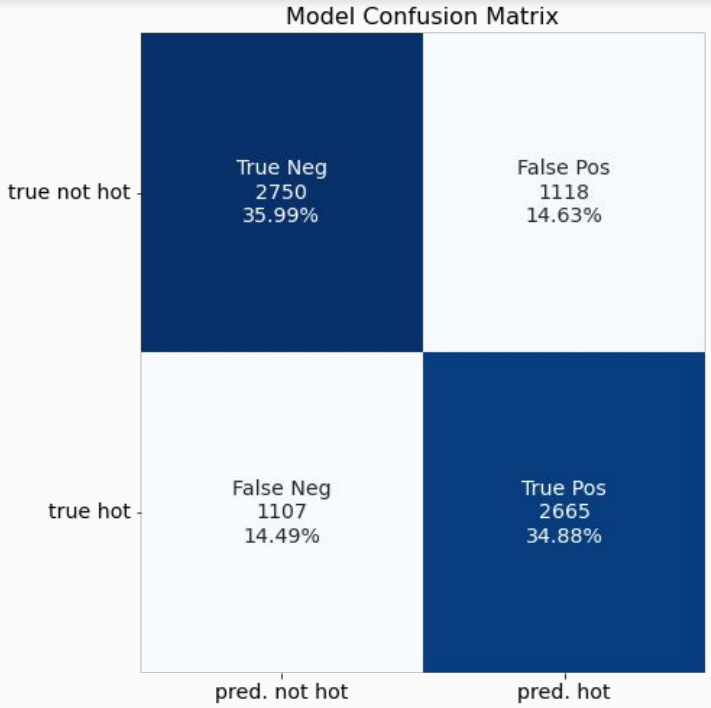
Reddit’s (Next) Top Model

The modelling process took each of the 4 dataset versions and ran them through a gauntlet of default configured models including Logistic Regression, KNN, Decision Trees, and Random Forest.

Logistic Regression and Random Forest were the top scorers, and were chosen for additional optimization

The Lemmatization dataset was selected for use.

Lasso and Grid Search techniques were employed to improve the models and … **RANDOM FOREST** was your winner! 72% +/- 2% Accuracy



Results

Interpretation of a Random Forest model is notoriously tricky, but it can point us in the direction of which variables were important within the model. Combining those importance measures and their respective distributions within the dataset, some suggestions can be made:

- **Be Patient**
  - Reddit takes the age of the post into account for Hot! Give the post around 8 hours to really rev up before making a final call on its success.
- **Aim for large subreddits**
  - More eyeballs means more possible commenters. Look 2.5M subscribers or more and pay attention to the submission guidelines and the meta of the subreddit.
- **Action words**
  - The percent of verbs in a title matters. Maintain a 10-12% ratio to keep things in a successful prose
- **Avoid Short Titles**
  - Single words or plain emojis are not proven winners
- **Reddit Media Domains are not an absolute**
  - They are convenient and useful, but not a one-size-fits-all. Select your domain with purpose.

Thank you!

Contact me:

Right Behind You
123 Your Street
New York, NY