Predicting Popularity: A Social Media Experiment

Bryce Anthony and Hannah Kanjian

Davidson College

Abstract

Social Media is a constant influence on our lives and can be very powerful when it comes to marketing. The goal of our project is to create a model that can predict the popularity of a given post in terms of the number of likes a post receives and total audience engagement with the post.

Introduction

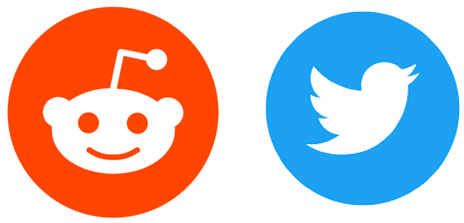

Figure 1:Reddit and Twitter Logos

Starting with a Kaggle dataset containing data from “Reddit Top 1000 Posts,” we analyze how the title of a Reddit post and time the post was created affect the total number of likes and overall engagement a post will receive.

To train a model with this data, we needed to do a great deal of preprocessing. We used a classic approach to Natural Language Processing (NLP) of creating Word Embeddings (3) in addition to analyzing the sentiment of each post. This additional data derived from each post was then used to create several models to predict the popularity of a post.

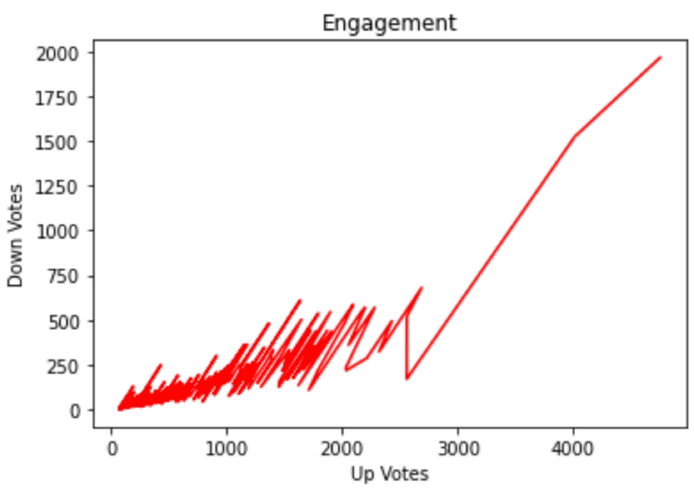

Figure 2:Visualizing Engagement through Up and Down Votes, we note the relation between the two forms of engagement.

Experiments & Results

Using the "Reddit Top $1,000$ Posts" data set from Kaggle, we started by training our model on one data set (r/Shower Thoughts), and then expanded our data to include many different topics.

For both the Shower Thoughts data set and more general data set We experimented with the three models listed below:

- Logistic Regression
- Random Forest Classifier
- SVC (Support Vector Classifier)

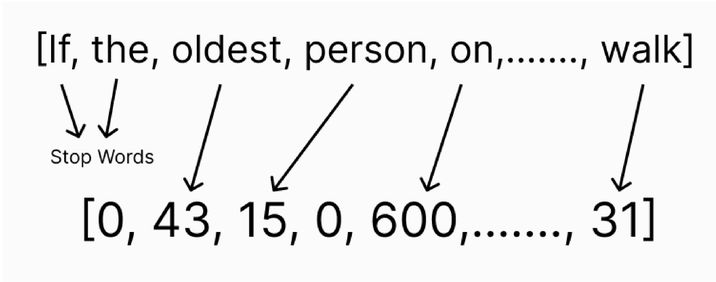

Figure 3:Basic Structure of Word Embedding

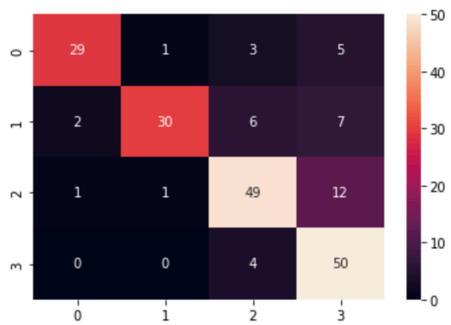

Figure 4:Heat Map of Total Engagement Predictions from the Random Forest Classifier using word embeddings

Model	WE	WE + SA	Test
Logistic Regression	20.5%	23.5%	25.5%
Random Forest Classifier	66.5%	71.9%	77.5%
Support Vector Classifier	20.5%	23.5%	25.5%

Figure 5:Accuracy for **Up-Vote Predictions**

Model	WE	WE + SA	Test
Logistic Regression	25.6%	21.8%	-
Random Forest Classifier	70%	69.4%	69%
Support Vector Classifier	25.6%	21.9%	-

Figure 6:Accuracy for **Total Engagement Predictions**

Conclusion

- We used Machine learning to predict the popularity of a post based on it's text content and time of creation
- This worked best for predicting Up-votes/likes using a Random Forest Classifier that took in word embeddings of the titles in combination the sentiments of the title from the post (77.5% accuracy).
- For predicting total engagement, the best model was again a Random Forest Classifier using only word embeddings (69% accuracy).
- Both models were over than 2x more accurate than the dummy model and highlight the idea that there may be objective features that can be used to predict the popularity of a post.
- The inability of the model to more accurately predict the popularity of a post emphasizes the complexity of the human mind when it comes to interpreting information. Although we found it possible to predict a post's performance to a certain degree.
- In terms of further work, we believe there is a lot to be learned from a deep learning perspective. We would continue by running the word embedding through a deep learning model.

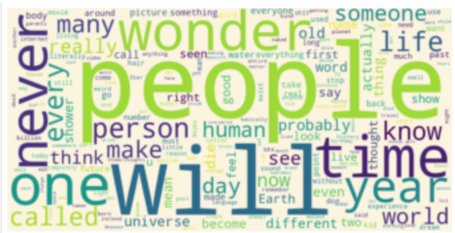

Figure 7:A Word Cloud of the the most frequent words in the dataset

Broader Impacts

Since social media is such a large part of our day to day lives, marketing has a huge impact on how we think and what we buy.

Our model has the capacity to positively influence the community, however it can also be used negatively. All of which is subjective to the content and message of what we put online.

In addition, this model could lead to a lack of diversity in online content were it to be used on a larger scale.

References

[1] Ishita Daga, Anchal Gupta, Raj Vardhan, and Partha Mukherjee.
Prediction of likes and retweets using text information retrieval.
Procedia Computer Science, 168:123--128, 2020.
“Complex Adaptive Systems”Malvern, PennsylvaniaNovember 13-15, 2019.

[2] Yonghui Dai and Tao Wang.
Prediction of customer engagement behaviour response to marketing posts based on machine learning.
Connection Science, 33(4):891--910, 2021.

[3] Bei Yu, Miao Chen, and Linchi Kwok.
Toward predicting popularity of social marketing messages.
pages 317--324, 03 2011.

DAVIDSON