Predicting Tweet Popularity through a Feature-based Approach

Group 12:
Mojan Majid and Huayin Luo

What type of tweets are receiving the most engagement?

1. Engagement with content
2. **"Fake news", misinformation**

1) Motivation

It is an interesting and important question to determine what type of tweets are receiving the most engagement and popularity, from multiple viewpoints

Motivation - Existing Work

Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 168 (2020) 123–128
Procedia Computer Science
www.elsevier.com/locate/procedia

Complex Adaptive Systems Conference with Theme:
Leveraging AI and Machine Learning for Societal Challenges, CAS 2019

Prediction of Likes and Retweets Using Text Information Retrieval

Ishita Daga[a]*, Anchal Gupta*[a], Raj Vardhan[a], Partha Mukherjee[a]

[a]Pennsylvania State University, Great Valley, 30 E. Swedesford Road, Malvern, PA – 19355, USA

AStA Advances in Statistical Analysis (2021) 105:335–352
https://doi.org/10.1007/s10182-021-00390-z

ORIGINAL PAPER

Check for updates

Predicting the popularity of tweets using internal and external knowledge: an empirical Bayes type approach

Wai Hong Tan[1] · Feng Chen[2]

Received: 16 November 2019 / Accepted: 8 February 2021 / Published online: 26 February 2021

Predicting popularity of user content is a perplexing problem that captivates many.

- **Limited text length**

  280 char (*as of Feb 2023, 4000 char for Twitter Blue users)

- **Popularity is complex & ambiguous**

  Most practical approaches take into account time-based data

2) Data Description

Reward Variables?

1. favourite count
2. retweet count
3. creation time
4. tweet text
5. user screen name
6. user status count
7. user followers count

Features

We used Cait's dataset that had 10,000 unique tweets.This data was extracted using the Twitter Developer API.

#CDNPOLI, no retweets (avoid duplicate tweets)
Limitations: truncated text

Data Description - New Variables

We implemented some additional variables based on the existing variables:

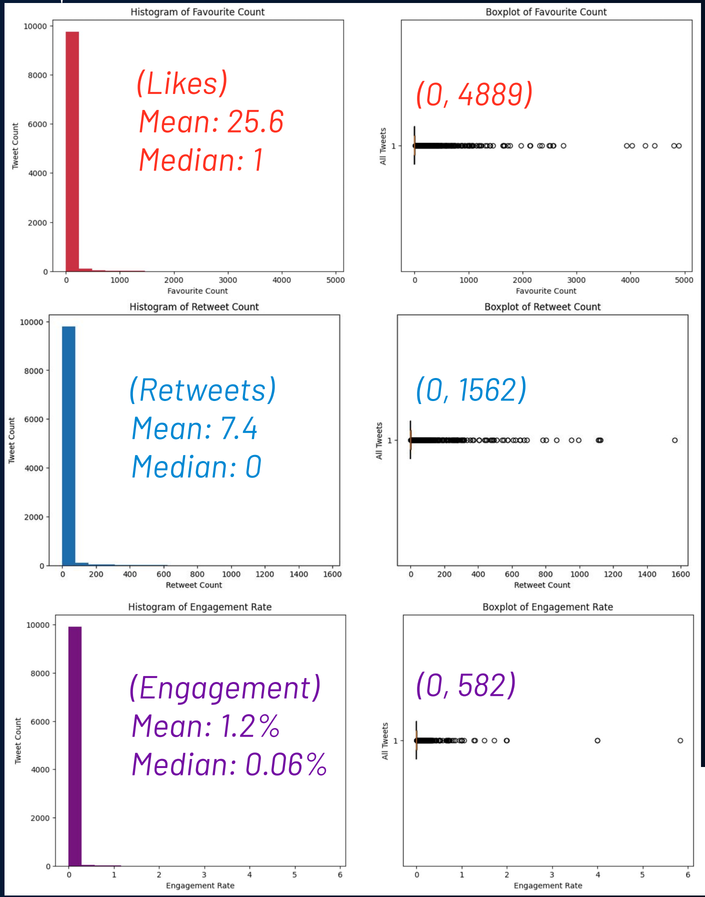
Variable Name	***Data Type***
Engagement Rate	Continuous (from 0 to 1) $\frac{100(\#\text{ likes} + \#\text{ retweets})}{\#\text{ followers}}$
Engagement Score	Binary (1 if engagement rate >0.01, 0 otherwise)
Low Engagement Score	Binary (1 if engagement rate >0.001, 0 otherwise)
Time of Day Created	Categorical (Morning, Afternoon, Night, Evening) *Limitation: Created Time (UTC), we divide based on EST*
Sentiment Score	Continuous (from -1 to 1)

Features

Limitations:
(1) Range of values
(2) Distribution of values

3) EDA

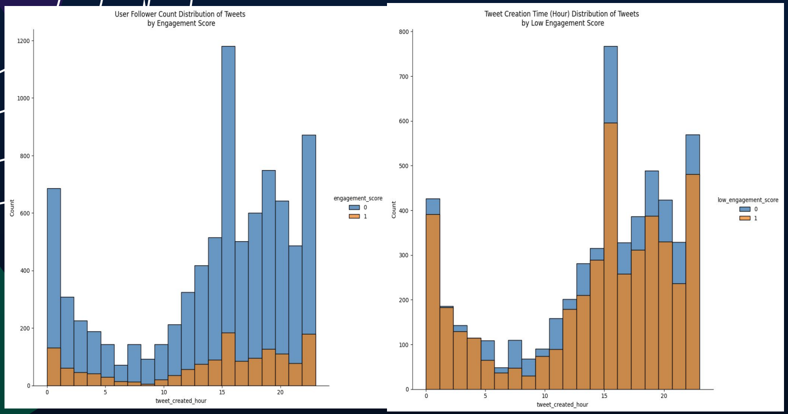
Some exploratory graphs & tables



$$\frac{100(\#\ \text{likes} + \#\ \text{retweets})}{\#\ \text{followers}}$$

[Figure 1: Distribution of Favourite Count, Retweet Count, & Engagement Rate]

EDA - More Visualizations



Not Significant: Tweet Creation Time, by Engagement Scores

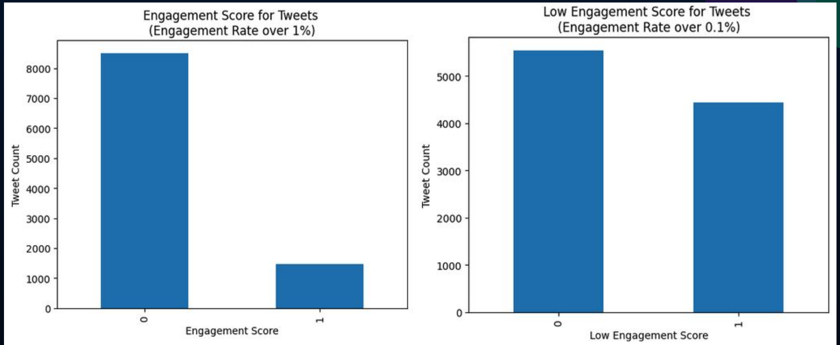
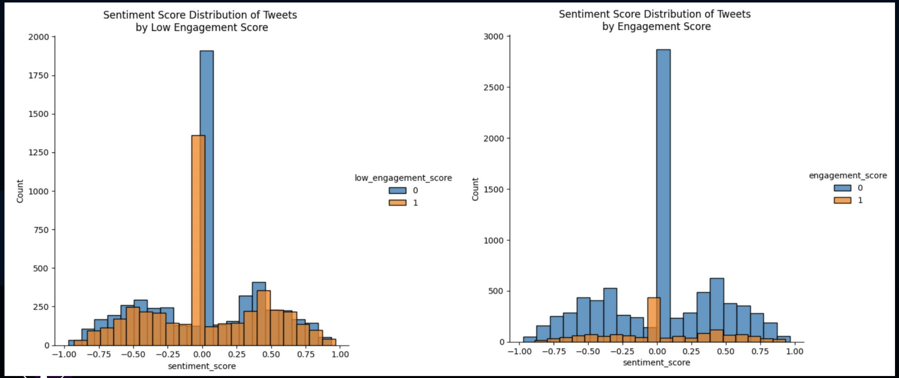


Figure 2: Engagement Score & Low Engagement Score Distribution



Significant: Sentiment Score, by Engagement Scores

We looked at the difference in the distribution of variables, by engagement score, to identify potential features.

4) Our Model

1002 Features

1. Sentiment Score for tweet [-1, 1]
2. Tweet Text Vectors (1000 words)

NLTK VADER Pretrained Sentiment Model

Preprocessing + TF-IDF

+

1. Username length

=

Logistic Regression

Outcome

Engagement Score
Low Engagement Score

Binary: 0 or 1

We used a two step approach, with Text Features for Non-Text Prediction

5) Results

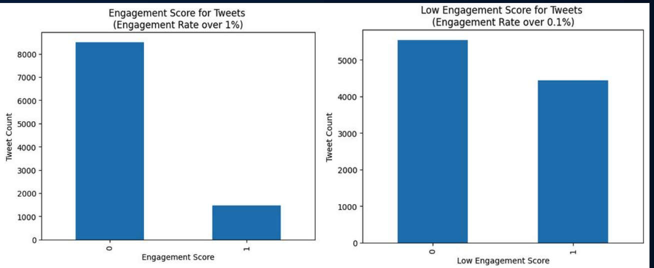
Engagement Score
Train Accuracy: 84.7%
Test Accuracy: 85.7%

Low Engagement Score
Train Accuracy: 67.4%
Test Accuracy: 67.0%

Engagement Score: model performs similarly to a "naive" model.

Low Engagement Score: performs better than both random & naive baseline models

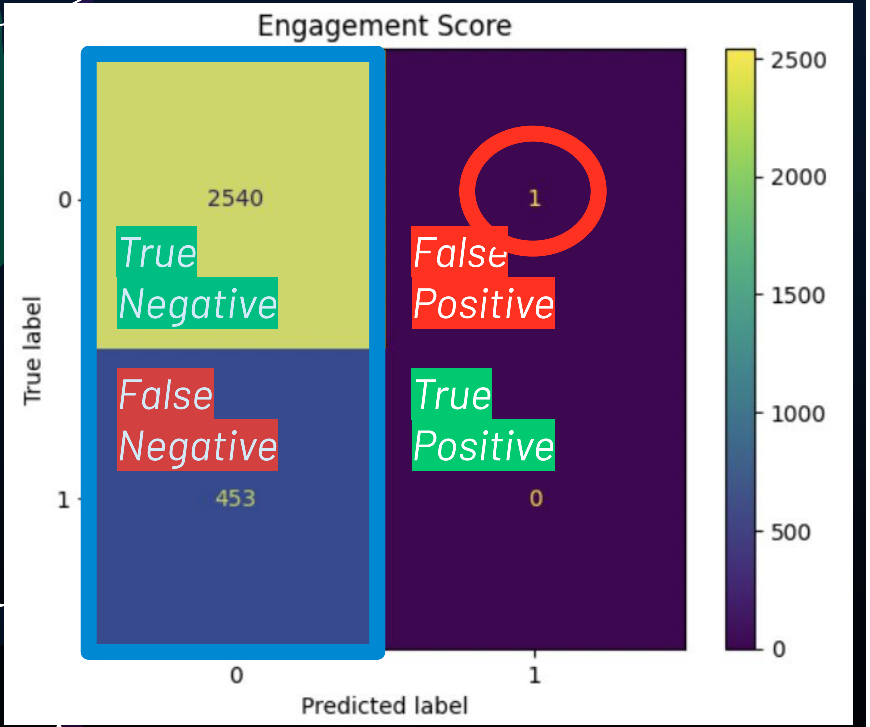
- *Difference = Balanced dataset?*

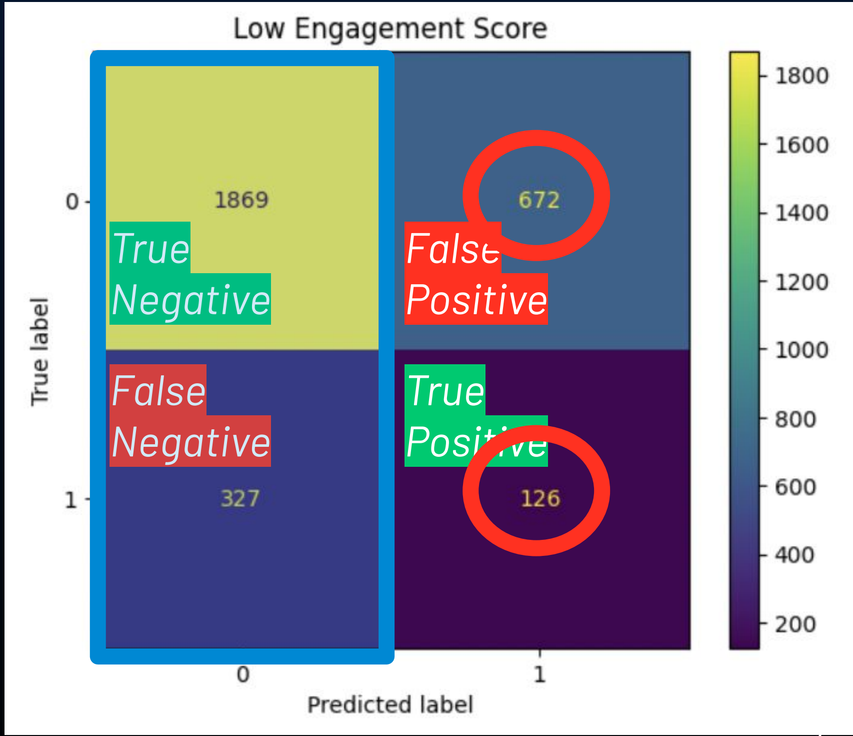


(Recall)

Results - Conclusions & Next Steps

"Naive" approach?





FNR:0.15, FPR: 1.0

FNR:0.15, FPR: 0.84

Better! Trying.

Results - Conclusions & Next Steps

Our model shows an interesting starting point for combining word features from NLP approaches with other informative variables in predicting popularity and engagement of a tweet. There are many ways in which we can improve:

(1) Full tweet text (instead of truncated)

- *More insightful NLP vectors*

(2) More balanced dataset

- *balanced amount of popular/not popular tweets*
- *more very popular/viral tweets*

(3) Other features...

Thank you!