Analyzing User Engagement and Predicting Popularity on UW Subreddit Based on Post Titles

Jay Kuo, Jonic Zhu, Juntong Wu, Manqi Jiao

December 11, 2023

1 Motivation

We are investigating the topic modeling of Reddit posts using natural language processing techniques. Reddit is one of the most popular and active social media platforms with millions of users posting and commenting on a wide range of topics. Topic modeling is a popular technique in natural language processing that can automatically identify latent topics or themes from a large corpus of text data, without the need for manual labeling or annotation.

With the increasing influence of social media platforms, there is a growing concern about the spread of misinformation, hate speech, and other harmful content. By analyzing Reddit data using natural language processing techniques, we can identify and monitor the prevalence of these problematic topics and subtopics. Furthermore, we can also detect patterns of behavior, such as coordinated campaigns, that may be used to manipulate or influence public opinion. This research can, therefore, contribute to the development of ethical solutions and policies for addressing these issues and promoting responsible use of social media platforms. By identifying and addressing these ethical problems, we can create a safer and more inclusive online environment for all users.

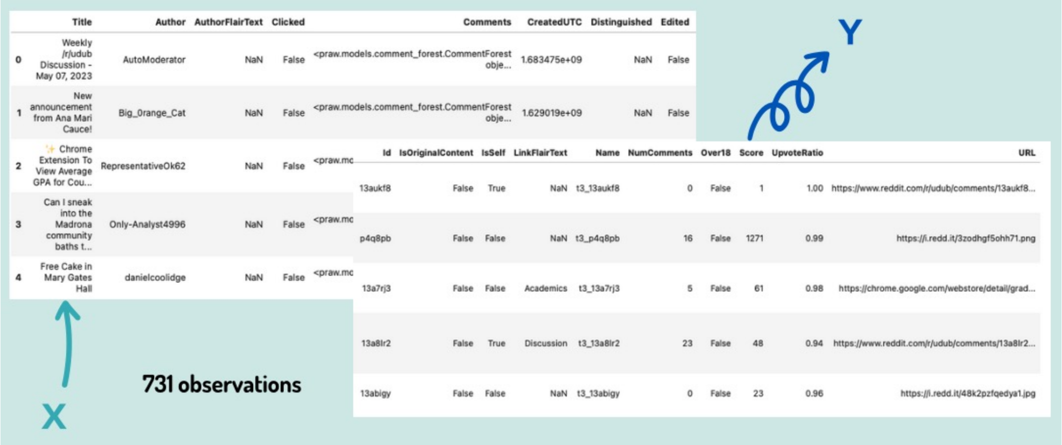
2 Concise problem definition

The objective of this study is to conduct a comprehensive exploration of the UW subreddit, focusing on topics, user characteristics, and post sentiments. The primary goal is to uncover the factors that captivate users' attention, specifically investigating whether posts with higher engagement scores also yield higher comment counts.

Furthermore, the aim is to construct a robust NLP model capable of predicting post engagement based on the titles. This endeavor entails evaluating and contrasting various models, word embedding techniques, and employing sentiment analysis methods to incorporate additional features.

Moreover, this exploration aims to address ethical concerns associated with the UW subreddit, leveraging the insights gained from the analysis. Additionally, the study acknowledges the potential concerns that may arise if similar models are deployed in online forums.
In summary, the problem entails:

- Conducting a qualitative examination of topics, user characteristics, and post sentiments within the UW subreddit.
- Investigating the correlation between post engagement scores and comment counts across different topics.
- Developing an advanced NLP model to forecast post engagement by employing title-based predictions, encompassing model comparison, word embedding methods, and sentiment analysis techniques.
- Addressing ethical considerations relevant to the UW subreddit based on the findings.



	Title	Author	AuthorFlairText	Clicked	Comments	CreatedUTC	Distinguished	Edited
0	Weekly /r/udub Discussion - May 07, 2023	AutoModerator	NaN	False	<praw.models.comment_forest.CommentForest obje...	1.683475e+09	NaN	False
1	New announcement from Ana Mari Cauce!	Big_Orange_Cat	NaN	False	<praw.models.comment_forest.CommentForest obje...	1.629019e+09	NaN	False
2	✨ Chrome Extension To View Average GPA for Cou...	RepresentativeOk62	NaN	False	<praw.mc			
3	Can I sneak into the Madrona community baths t...	Only-Analyst4996	NaN	False	<praw.mc			
4	Free Cake in Mary Gates Hall	danielcoolidge	NaN	False	<praw.mc			

Id	IsOriginalContent	IsSelf	LinkFlairText	Name	NumComments	Over18	Score	UpvoteRatio	URL
13aukf8	False	True	NaN	t3_13aukf8	0	False	1	1.00	https://www.reddit.com/r/udub/comments/13aukf8...
p4q8pb	False	False	NaN	t3_p4q8pb	16	False	1271	0.99	https://i.redd.it/3zodhgf5ohh71.png
13a7rj3	False	False	Academics	t3_13a7rj3	5	False	61	0.98	https://chrome.google.com/webstore/detail/grad...
13a8lr2	False	True	Discussion	t3_13a8lr2	23	False	48	0.94	https://www.reddit.com/r/udub/comments/13a8lr2...
13abigy	False	False	NaN	t3_13abigy	0	False	23	0.96	https://i.redd.it/48k2pzfqedya1.jpg

Figure 1: An example of what we collected.

- Identifying and evaluating potential concerns associated with the implementation of similar models in online forums.

3 Related works

- Predicting User Interaction on Social Media Using Machine Learning

Researchers suggest utilizing machine learning models to forecast user engagement metrics for advertising articles, including share count, comment count, and remark sentiment. Understanding which models are most effective at forecasting user involvement can help marketers avoid wasting money by presenting less effective advertising, and let them know which ads will work best on the platform.

- Using Natural Language Processing to Identify Effective Influencers

The framework uses natural language processing to record the mental traits and standard variables of influencers in order to make influencer profiles and predict the best influencers for given campaigns. The suggested framework can help deal with fraudulent involvement, and it is tied with a set of data that links the social media actions of influencers to purchases made by customers.

- Experiment: Can Post Timing Improve Your Instagram Engagement?

The article highlights the usage of social media management tools like Hootsuite to plan posts at the best times and stresses the need for testing and monitoring post timing to enhance Instagram engagement. The necessity of harmonizing posting time with other engagement factors is demonstrated as a significant concern for optimizing post engagement.

4 Data

We obtain data from Reddit's API (Application Programming Interface) with PRAW library. To scrape Reddit's data, we followed a tutorial of this video on YouTube. Subsequently, we used a Jupyter notebook to convert the collected information into CSV files. The data we gathered consists of 732 rows extracted from the University of Washington subreddit. The CSV file containing this data can be accessed through this link.

5 Approach

5.1 Model approach

- Metrics: Natural language processing (NLP) can address our research questions by analyzing and extracting meaningful information from text data, such as post titles and comments, to gain insights and make predictions related to engagement, feedback, topic attraction, user behavior, and ethical concerns.

- Evaluation: The evaluation metric used in our score-predicting regression model is the root mean squared error (RMSE). The metrics assess the accuracy and performance of the regression model in predicting the continuous target variable.

5.2 Model selection

We implemented the following models for our project and generates the results shown at figure 2.

- Random forest model on CountVectorizer and deep sentence embedding
- Keras neural network on CountVectorizer and deep sentence embedding
- Sentiment analysis and ensemble learning method

6 Experiments

In this experiment, we aimed to predict scores based on post titles using two different models: a random forest and a neural network. Additionally, we explored the effectiveness of two word embedding methods, namely count vectorizer and deep sentence embedder, to convert the textual data into numerical vectors for analysis.

To assess the performance of the models, we conducted comprehensive evaluations, considering various performance metrics and statistical measures. This rigorous evaluation process ensured that the models effectively captured the underlying patterns and relationships present in the data.

To validate the robustness of the models, we employed independent test sets to assess their predictive capabilities on unseen data. This validation step provided valuable insights into the generalization and accuracy of the models beyond the training data.

Furthermore, we incorporated sentiment analysis into our analysis to explore the sentiment expressed in the post titles. By considering the sentiment, we aimed to understand its impact on the prediction of scores and uncover any relationships between sentiment and engagement.

To enhance our understanding of the models' behavior and the relationships between variables, we utilized visualizations. These visual representations provided a clearer and more interpretable depiction of the results, aiding in the interpretation and communication of our findings.

By employing these structured, principled, and rigorous approaches, we aimed to develop accurate prediction models and gain deeper insights into the relationships between post titles and scores.

7 Results

- Root Mean Square Error: We have achieved 2.36 for Root Mean Square Error with sentiment analysis and 2.45 for Root Mean Square Error with sentiment analysis & bagging.
- Qualitative result: We have made a regression plot to show the relationship between the top 20 posts with the highest comment counts and indicate a weak positive relationship between the number of comments and the score.
- Word cloud: We created a word cloud that represents the prominent themes and topics discussed. The word cloud's size corresponds to the frequency of each word in the subreddit data, providing insights into the interests of the UW community.

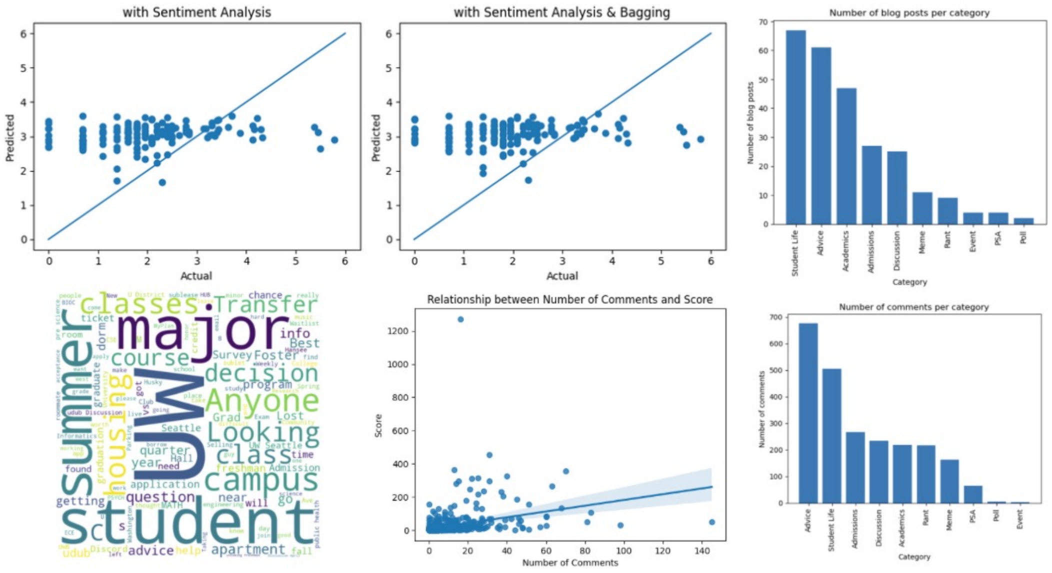


Figure 2: Analysis results.

8 Discussion of findings and insights

8.1 Model selection

While the random forest and neural network generated different distributions of predicted scores, their RMSEs are at comparable levels. As such, While model selection does have an effect on prediction results, in our case there are other potential areas of improvement for model performance, as will be discussed below.

8.2 Word embedding method

The word embedding method was studied as an option for feature engineering. The baseline was a CountVectorizer, and we also applied a deep sentence embedder in the hope to extract more semantic information, however, while the RMSE in the neural network model improved slightly when changing the embedding method to a deep sentence embedded, there was no appreciable effect for the random forest model. This result may suggest the limited number of observations we have and the brief nature of the post title alone may be inadequate for our purpose with little to no dependencies on embedding methods.

8.3 Sentiment analysis and ensemble method

To explore other ways to improve our model, quantitative sentiment analysis was used to assign a -1 to 1 value to each title as an additional feature. Separately, we also applied bagging as an ensemble method to mitigate our sample size. While both iterations lowered the RMSE compared to the baseline neural network model, the distribution of scores did not change substantially. As we will discuss in subsequent conclusion sessions, we can only marginally improve model performance by adding feature or meta-algorithm when the language embedding is inadequate or otherwise limited by the sample size.

9 Limitations and future work

9.1 Limitations

- Temporal dynamics: The research doesn't account for the timing of the post, which could significantly impact user engagement.

- Bias in Dataset: The study is confined to the UW subreddit, thus the results might not be applicable to other subreddits or online communities.
- Scope of Features: The study mainly focuses on the title of the posts and user IDs, ignoring other potential influential factors like post content, post length, presence of images/videos/links, and previous user activity.
- Subjectivity in language: The study overlooks linguistic nuances such as sarcasm, humor, or cultural references, which can alter the meaning and reception of a post.
- Lack of user-specific analysis: The study doesn't delve into understanding the patterns or behaviors of individual users who tend to create highly engaged posts.

9.2 Future work

- Incorporate Time: Future models should consider the timing of posts, given its potential influence on user engagement.
- Expand Dataset: The model can be enhanced by training it on a larger and more diverse dataset, including different subreddits, to make the findings more generalizable.
- Additional Features: Future research could include more post features like post length, media type (text, image, video, link), and user history to provide a more comprehensive understanding of engagement factors.
- User-specific analysis: Future studies could explore the behaviors and patterns of individual users who frequently generate highly engaged posts, thereby understanding what makes their posts engaging.

10 Ethical considerations

10.1 Positionality

- Privacy and Consent: It's critical to respect user privacy while collecting and analyzing data. Even though the data on subreddits is public, researchers must be cautious and considerate about how they collect, store, and use this data.
- Power Dynamics: The potential misuse of predictive models by those in power (e.g., subreddit moderators or admins) should be addressed. This could include the unfair targeting of certain users or manipulation of the platform to favor certain types of content.

10.2 Sociotechnical Considerations

- Bias and Fairness: There is a risk of algorithmic bias in predictive models, where the model's predictions may be biased against certain groups of users due to the inherent biases in the training data. Mitigation strategies should be in place to ensure fairness.
- Transparency and Accountability: The workings of predictive models should be transparent to avoid misuse or misunderstanding. Users should have the right to understand how their data is being used and how predictions about user engagement are made.

10.3 Philosophical considerations

- Autonomy: The autonomy of users to post without undue influence or manipulation from predictive models should be respected.
- Justice: It is important to consider the distribution of benefits and burdens when using these predictive models. For instance, if certain types of posts are consistently favored, it could lead to an unfair distribution of attention and engagement.

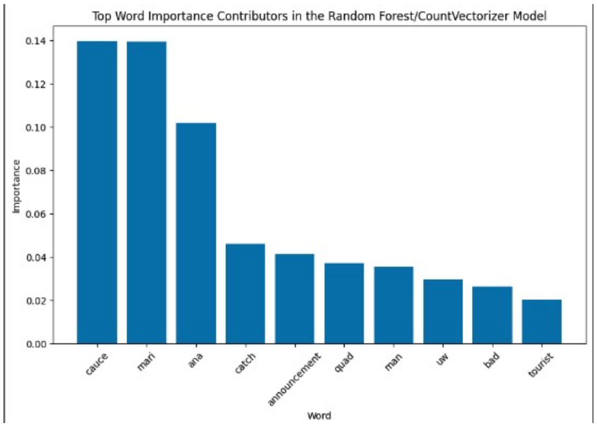


Figure 3: Feature (word) importance in the random forest CountVectorizer model.

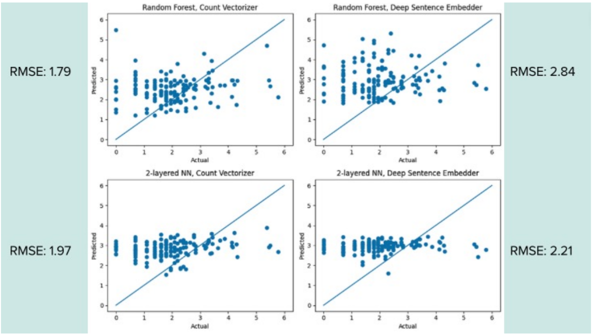


Figure 4: RMSE results

11 Conclusion

With a combination of qualitative and quantitative analysis, we were able to demonstrate three major themes around the UW campus subreddit and engagement analysts at large:

11.1 Qualitative theme has a major impact on post engagement

Repeating themes such as campus life, academics, and administrative support come up as major predictors for discussion engagement. While our model performance demonstrated such correlation is nowhere close to definitive, the insight provides substantial opportunity for future feature engineering based on pre-defined categories, which may also have a bigger impact on the potential use of the model.

11.2 Categorical post sentiments are directly correlated to higher engagement

With quantitative analysis attributing varying importance to certain words as demonstrated below, we see reoccurring themes that agree well with the thematic observations above. At the same time, we also observe certain categories of post features, such as emotional ones, are better predictors of post engagement quantitatively.

11.3 Model selection and feature engineering is critical to predictive analytics accuracy

In our machine learning models, word embedding methods had a much smaller impact on the RMSE compared to model selection. While constrained by the number of data points available, we were able to exercise engineering other aspects of the model, including ensemble methods for performance improvements. Future iterations should focus on feature selection while keeping in mind that reliance on mere post titles is likely to be insufficient.

12 References

Crowe, C. (2018). Predicting User Interaction on Social Media using Machine Learning (Doctoral dissertation, University of Nebraska at Omaha).

Fang, X., Wang, T. (2022). Using Natural Language Processing to Identify Effective Influencers. International Journal of Market Research, 64(5), 611-629.

Experiment: Can Post Timing Improve Your Instagram Engagement? https://blog.hootsuite.com/experiment-post-timing-instagram-engagement/

The python reddit api wrapper. PRAW. (n.d.). https://praw.readthedocs.io/en/stable/

YouTube. (2021, March 14). How to scrape reddit automatically label data for NLP projects — reddit API tutorial. YouTube. https://www.youtube.com/watch?v=8VZhog5C3bUt=625s

What is Cython? - cython basics. CallMiner. (n.d.). https://callminer.com/blog/a-breakdown-of-cython-basics

R/UDUB. reddit. (n.d.). https://www.reddit.com/r/udub/