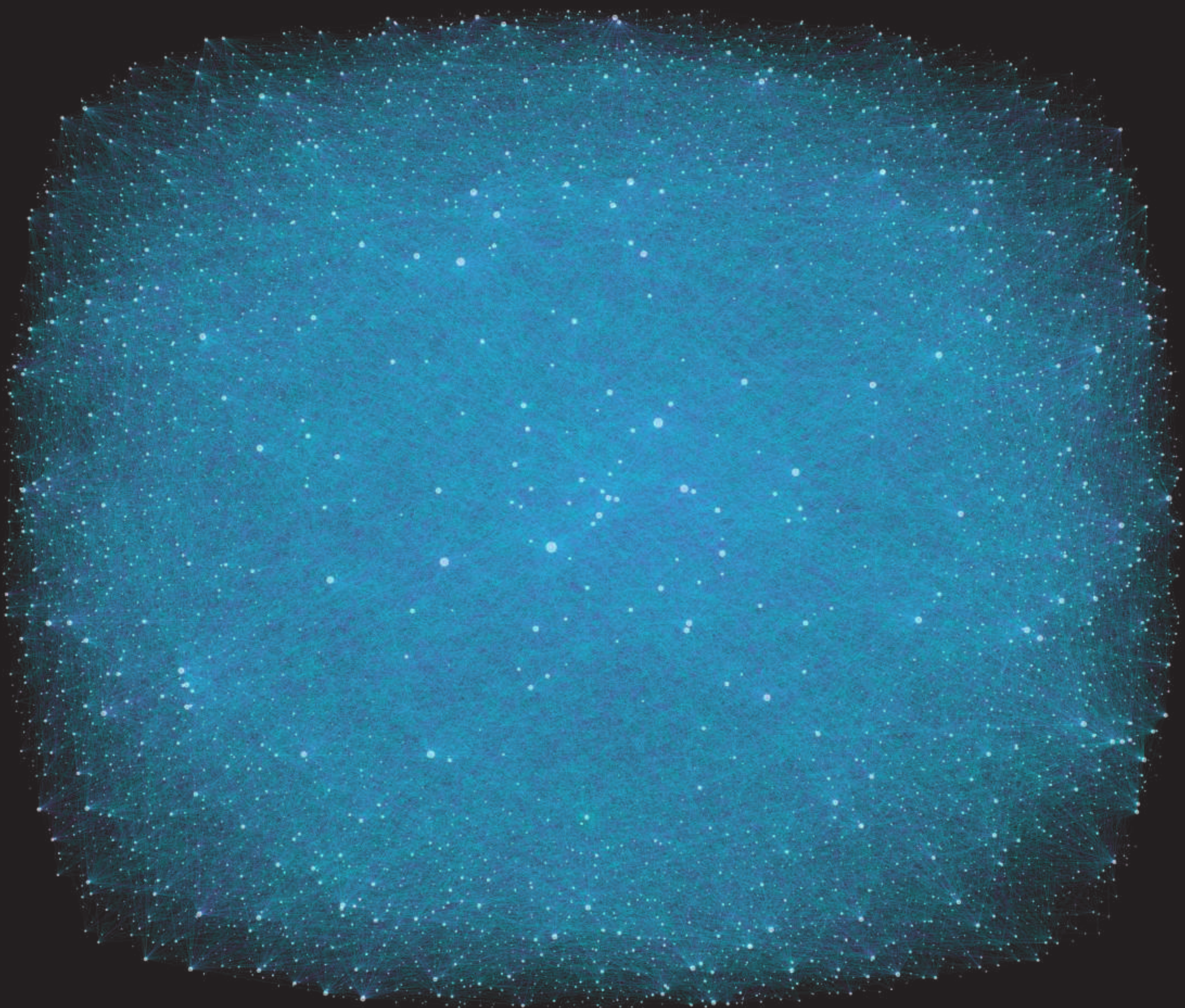
DTU Compute
Department of Applied Mathematics and Computer Science

Explaining Successful Reddit Posts

Predicting Awards from Textual and Network Features

Master Thesis

Lasse Pelle Skytte Hansen



Explaining Successful Reddit Posts
Predicting Awards from Textual and Network Features

Master Thesis
March, 2022

By
Lasse Pelle Skytte Hansen

DTU, Department of Applied Mathematics and Computer Science, Building 324, 2800 Kgs. Lyngby Denmark - www.compute.dtu.dk

Approval

This thesis has been prepared at the Department of Applied Mathematics and Computer Science, at the Technical University of Denmark, DTU, in partial fulfillment for the degree Master of Science in Human-Centered Artificial Intelligence, MSc Eng.

This thesis is credited 35 ETCS and was conducted during the period from the 12th September 2022 to the 10th March 2023. It was supervised by Sune Lehmann Jørgensen & Laura Alessandretti.

It is assumed that the reader has a basic knowledge in the areas of statistics, machine learning, and computer science.

Lasse Pelle Skytte Hansen - s154446

Lasse

Signature

10. March 2023

Date

Abstract

This thesis explores how the acknowledgment of individuals can be quantitatively measured by exploiting the Reddit award system. The thesis primarily focuses on data from the subreddit Wallstreetbets, where we analyzed the social dynamics of the community using network science and natural language processing. We developed a deep learning model that classifies awarded posts, using network features and the transformer model ELECTRA. Finally, we explained some of the underlying social mechanisms within the community, revealing a ”rich get richer” effect for the award system, where the probability of receiving an award increases prior to the amount previously awarded posts. We also investigate the impact of the network size on individuals’ identity within the community, revealing that larger networks have different social dynamics, making it more challenging for individuals to establish identity. Overall, this research provides insights into how reputation is formed and maintained within online communities and the factors that contribute to it.

Acknowledgements

A huge thanks to my supervisor Laura Alessandretti for the guidance throughout the whole process!

A special thanks to my fellow students at G. A. Hagemanns Kollegium for the motivational talks and input.

Contents

CHAPTER 1

Introduction

1.1 Motivation

Since the origins of social network analysis, there has been an interest in identifying the most influential individuals of a social network. Researchers have proposed several measures to quantify the relevance of individuals, such as the ”centrality” of a node in the network. However, it is still unclear if these traditional measures and methods within social science can effectively explain the importance of an individual. The online social platform Reddit uses a system where users can give others recognition for their contributions, through a system of ”awards” it provides a unique opportunity to quantitatively measure the acknowledgment of an individual. We will in this thesis investigate how the reputation and acknowledgment of individuals are explained through traditional network science, by exploiting the award system. We will gather data from Reddit and construct social networks from the users' interaction via posts and comments. The final aim is to develop a deep learning model that predicts if a post is awarded using network science features and the textual content created by the individual. Moreover, we will also explore some of the underlying patterns within social networks.

Understanding online social networks are essential because they have the power to influence real-world outcomes, from consumer behavior to political activism. By connecting people across the globe, these networks facilitate the rapid spread of information and ideas, which can shape public opinion, impact business decisions, and even drive social change. However, they also have the potential to perpetuate harmful stereotypes and misinformation, making it crucial to investigate and understand the underlying mechanism of how content and reputation propagate.

1.2 Reddit

Reddit is an online social network of a variety of communities where people can share and discuss interests. These communities are called subreddits. Reddit was founded in June 2005 and is the 9th most visited website in 2022 with approximately 52 million users worldwide. The platform has been involved in various social and political movements, making it a significant force in shaping online discourse and public opinion. Additionally, Reddit is a huge part of internet culture and a valuable resource for people seeking information and conversation on a diverse range of topics [1], [2].

1.2.1 The Structure of Reddit

Reddit consists of a variety of different subreddits where people can discuss and share different interests. On these subreddits, people can make posts and make comments on posts. Examples of posts can be seen in Appendix A.1. It is also possible to comment on comments, which creates a tree-like structure, see Figure 1.1 [3].

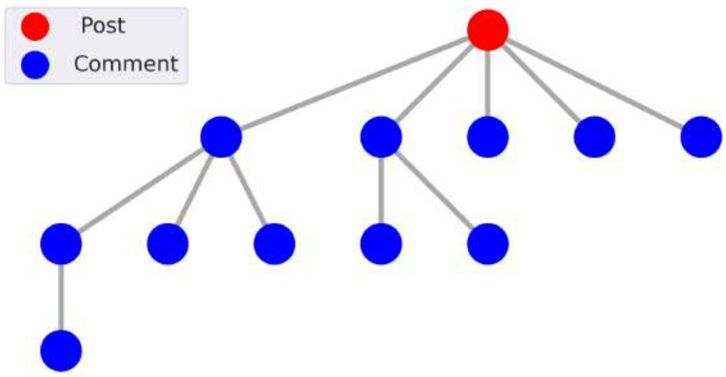


Figure 1.1: The figure shows the post/comment structure of Reddit

1.2.2 Upvotes & Awards

Reddit uses a system of upvotes/downvotes, which works similarly to other social platforms e.g. ”likes”. Posts/comments get a total score, where upvotes count as $+1$ and downvotes as -1. This system can be used for users to express their overall sentiment toward the content. Additionally, Reddit uses a system called ”karma”, which is the average of upvotes/downvotes across all of the user’s content.

Reddit also uses an award system, where users can award each other with virtual trophies for their contributions, each with its own icon and name. Awards are often used to show appreciation for posts or comments. Awards cost Reddit Coins depending on the type of award. Users can purchase Reddit Coins for real money, which can then be used to give awards or to access premium features on the site. Awards serve as a way to highlight exceptional content and to recognize the effort and expertise of users [3].

1.3 Reputation & Acknowledgment

Cambridge Dictionary uses the following definition for reputation [4].

> *"The opinion that people, in general, have about someone or something, or how much respect or admiration someone or something receives, based on past behavior or character"*

The definition of reputation emphasizes the importance of public opinion and perception in defining how individuals are perceived by others. An individual's past behavior is fundamental in determining their reputation, which can be positive or negative depending on how others perceive them.

In the context of this project, the focus is on understanding and explaining the positive reputation that is captured by Reddit awards. These awards serve as a means of acknowledgment of individuals who have made positive contributions to the Reddit community. The principles of acknowledgment and recognition are crucial in building and maintaining positive reputations, as they help to establish credibility and trustworthiness among others.

Reputation is an essential aspect of social interactions, particularly in the digital age where individuals are increasingly reliant on online platforms to connect and engage with others. By exploring the nature of positive reputation and how it is conveyed through Reddit awards, this project aims to contribute to a deeper understanding of how acknowledgment and recognition principles contribute to the construction and valuation of reputation in online communities.

1.4 Scope

We will address the following hypotheses in this thesis.

I. *On Reddit individuals with a high reputation are more likely to receive awards.*

II. *It is possible to predict if a Reddit post is awarded by using the content of the post and the position of the author in the social network.*

III. *The likelihood of receiving an award increases prior to the number of awards previously received by the author.*

IV. *It is harder for individuals on Reddit to be recognized and establish an identity in larger networks.*

In order to do this, we will gather data from different subreddits. The investigation will focus on understanding how users' reputations and acknowledgment from others are built through their interactions with others on the platform. The first step is to analyze how reputation is captured through Reddit's system of "awards". The thesis focuses primarily on the subreddit Wallstreetbets. This study will investigate how these signals of approval and recognition can be used to determine an individual's level of reputation. To better understand the social dynamics we will construct social networks from the interactions between users via posts and comments, and measure different properties of the network, as well as the overall social behavior. Additionally, a textual analysis will be conducted on individuals' content to uncover underlying patterns, using textual feature extractions and a deep learning transformer model. Underlying patterns will be investigated on how these textual and network features can be used to explain the reputation and acknowledgment of individuals. The ultimate goal of the thesis is to develop a machine-learning model that classifies if a post is awarded using both network features and the textual content written by individuals. By combining these two types of features, the model has the potential to provide insights into the complex nature of reputation on online social platforms. Additionally, we will investigate some of the underlying social mechanisms that affect Reddit's award system. This includes the social dynamics of the award system and how the likelihood of receiving an award increases prior to the number of previous awards. Furthermore, we will investigate how the identity of individuals changes for larger networks by examining the reciprocity of the edges and the number of direct mentions of individuals.

Structure

The thesis is divided into the following 8 chapters:

- Introduction
- Literature Review
- Data
- Methods
- Results - Factors of Awarded Posts
- Results - Explaining & Predicting Awarded Posts
- Discussion
- Conclusion

The thesis contains an appendix with additional figures and tables that were not included in the main text. Additionally, a GitHub repository can be found **here** which contains the source code behind the thesis.

CHAPTER 2

Literature Review

In order to gain insight and a broader understanding of the field within this thesis, we conducted a literature review. This chapter investigates relevant research within network science, online communities, and machine learning. This section presents the key findings.

Relevant literature has been found using Google Scholar and DTU findit. Some of the following keywords have been used: *”Networks”, ”Graph Theory”, ”Reputation”, ”acknowledgment”, ”Community” etc.*

2.1 Reviews

2.1.1 Network Science

The ph.d.-project *”Elites in Denmmark”* by Anton G. Larsen & Christoph Ellersgaard [5], [6] identified the Danish power elite through a social network analysis. By using network science it is possible to find individuals with high centrality. The project revealed that these individuals tend to group, forming a dense cluster of people with power in Danish society (CEOs, politicians, etc.). They also found that these individuals were highly interconnected and held multiple positions of power and influence in various organizations, making them difficult to challenge. They also observed that the Danish power elite was highly homogeneous, with a limited diversity of educational backgrounds and a strong presence of men. The results of the study have significant implications for understanding the distribution of power and influence in Danish society and for evaluating the democratic representation of its citizens.

The paper *”Identifying Social Roles in Reddit Using Network Structure”* [7] shows how network science can be used to identify social roles and individual behavior. The study reveals that Reddit roughly consists of ”answer-persons” and ”non-answer persons”. They also show that individuals change their behavior based on the subreddit they participate in. These results highlight the importance of network analysis in understanding the dynamics of online communities and how individuals interact with each other.

2.1.2 Reputation & Indentity of Individuals

The paper ”*Identity effects in Social Media*” by Sean J. Taylor et.al. [8] describes a study that examines the effects of identity cues (eg. user names, profile pictures, etc.) on how people interact with content on social media. The study conducted an experiment where the content was randomly assigned with either an ”identified” condition or an ”anonymous” condition. The experiment showed that people reacted quicker to content with identity cues than to anonymous content. The study also showed that people react according to the reputation of the content producer. Thus showing that the reputation of an individual influences how people behave and interact. The results also provide evidence to show the rich-get-rich dynamics and that inequality in social content evaluation is mediated by identity cues. Overall, the study highlights the importance of identity cues in shaping the way people perceive and interact with each other on social media.

The research described in "*Learning User Reputation on Reddit*" by Alexandre Parmentier & Robin Cohen [9] investigates the user's reputation on Reddit based on the interaction between users. By using extracted features from the interaction they build 3 machine learning models for predicting the score, sentiment, and profanity use. The data is solely based on the discussion from a parent post, where the features are extracted. The study found that certain features, such as the length of a comment, the presence of certain words, and the time of day a comment was posted, were predictive of a user's reputation score. The sentiment and profanity use models also performed well, with the sentiment model outperforming the profanity use model. The results of the study demonstrate the potential for machine learning in understanding user reputation and behavior in online communities and have applications for improving moderation and user engagement in these communities. Additionally, the findings provide insights into how to design systems that can automatically predict the reputation of users and their likelihood of engaging in negative behavior.

Furthermore, awards are used to acknowledge content that contributes to the community on Reddit and shows the potential of being a motivating factor for content creation. The paper "*How Do Peer Awards Motivate Creative Content? Experimental Evidence from Reddit*" by Gordon Burtch et.al [10] reveals that awards motivate users to create more content. They conducted an experiment where they awarded random posts over two months on Reddit. They found that peer awards induced recipients to develop longer and more frequent posts. This study indicates that awards are a big part of how users perceive each other.

2.1.3 Communities

Online communities exist everywhere on the internet and Reddit is definitely not an exception. The paper "*Loyalty in Online Communities*" by William L. Hamilton et.al. [11] investigates the components within online communities. The research is based on exploring a large set of Reddit communities. Loyalty is an essential part of a community. Every user has an opportunity to engage with a variety of different communities, especially on a platform like Reddit due to the number of different subreddits. To retain a sense of community the user must return to the community and engage with other users regularly. Loyal communities have a denser user-to-user interaction and are reciprocal where interaction goes both directions. This suggests that loyal communities have a more cohesive network and are less fragmented in smaller subgroups. The paper also reveals that loyal users tend to engage with more esoteric or less popular content. In this way, the community may develop new ideas and content collectively. This can be seen in different examples e.g. the subreddit wallstreetbets that was the staging ground for a populist uprising against the financial establishment, by collectively increasing the stock price of the struggling company, Gamestop and nearly bankrupting institutional investors who had shorted its stock in the process [12]. To accomplish this the subreddit required a loyal community that collectively followed the same movement.

2.2 Summary

Examining various research studies that have been conducted on communities provides a good basis for further investigation. The *”Elites in Denmark”* project found that a cluster of highly interconnected individuals with power and influence in Danish society exist and that they are homogeneous. This shows that network science can be used to identify central individuals and understand the behavior of communities. Furthermore, the study *”Identifying Social Roles in Reddit”* shows that individuals tend to change their behavior based on the subreddit they participate in. This shows that reputation stays within the community, this can also be seen in the study *”Identity Effects in Social Media”* which shows the importance of identity cues in how people perceive and interact with each other on social media. If a user is known and has a reputation within a community they tend to get more reactions and keep the same level of reputation. For an online community to form and for individuals to obtain reputation users must continuously engage with each other, as described in the study *”Loyalty in Online Communities”*, which shows that loyal communities are more cohesive and tend to engage with more esoteric content. Furthermore, the paper *”Learning User Reputation on Reddit”* demonstrated the potential of machine learning in understanding user reputation and behavior in online communities. The study also showed that other features such as textual features can be used to describe behavior within communities.

CHAPTER 3

Data

3.1 Gathering the Data

Reddit data is publicly available and can be gathered through different APIs (Application programming interface). Reddit provides its own API: the python Reddit API wrapper (`PRAW`), but with some limitations. `PRAW` allows requests of up to 100 items at once, with a 2 seconds delay. Another very commonly used API is the third-party Pushshift.io API Wrapper (`PSAW`), which also has some limitations.

The data has been gathered using a combination of both APIs (`PSAW`, `PRAW`). For each subreddit two datasets have been created, one containing posts and the other containing comments. For each dataset different features have been extracted. The flowing table shows the features in the dataset containing comments. The ID of the post can be used to link the two datasets.

	Description
`Author`	The name of the author
`ID`	The id of the comment
`Body`	The text of the comment
`Post_id`	The id of the post to which the comment is related to
`Parent_id`	The id of either the post or comment which the author has commented on
`Score`	The number of upvotes for the comment.
`N_awards`	The number of awards given to the comment
`Date`	The created date of the comment

Table 3.1: Comments data

The following table shows the features in the dataset containing posts.

	Description
`Author`	The name of the author
`ID`	The id of the post
`Title`	The title of the post
`Text`	The text of the post
`N_Comments`	The number of comments connected to the post
`Score`	The number of upvotes for the post
`N_awards`	The number of awards given to the post
`Date`	The created date of the post

Table 3.2: Posts data

3.2 The Data

This section provides an overview of each dataset used in the thesis, as well as a brief descriptive analysis. The datasets are collected using the method explained in section 3.1. The dataset wallsteetbets (3.2.8) is collected by my advisor Laura Alessandretti, where we additionally collected the number of awards for each post. Initially, we tried to collect and analyze data from smaller active subreddits, but throughout the project, we discovered the importance of awards as a metric in the analysis. Due to the scarcity of awards in other subreddits, the analysis presented in the thesis is primarily from wallstreetbets. However, all the data is still presented in this section.

Each subreddit has its own color which can be seen in figure 3.1. The colors occur throughout the thesis when the different subreddits are compared, but due to the large focus on Wallstreetbets, it is not consistently colored for Wallstreetbets.

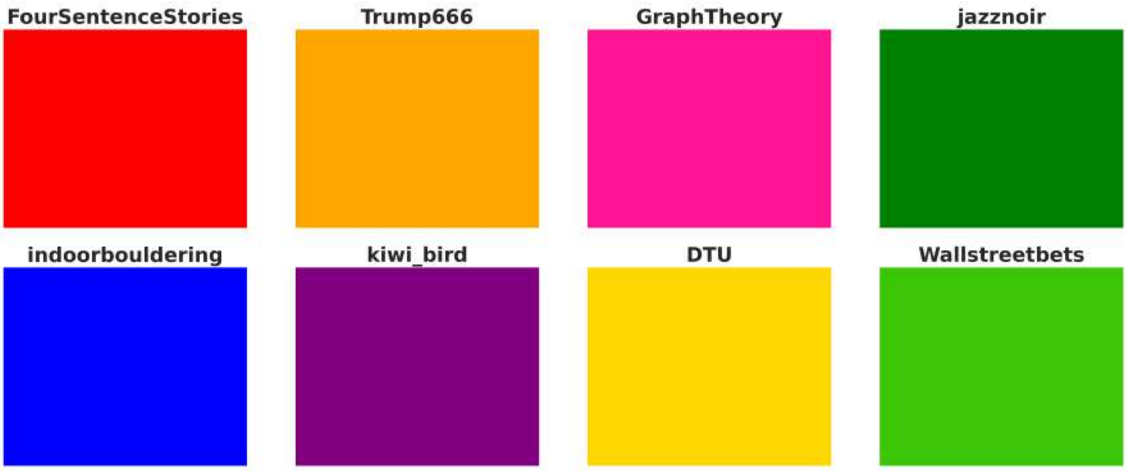


Figure 3.1: The colors indicate how each subreddit is colored.

Each dataset has been cleaned such that data with errors or empty values have been removed. We also removed bots and automatic moderators for each subreddit.

3.2.1 Jazznoir

The subreddit jazznoir is related to the sub-genre of jazz which has a melancholic feel. The dataset contains 4725 data points with 2109 comments and 2616 posts. The data is collected from 2015-01-01 to 2022-01-01. The dataset contains 1497, unique users. 8 posts are awarded which is 0.3% of all posts.

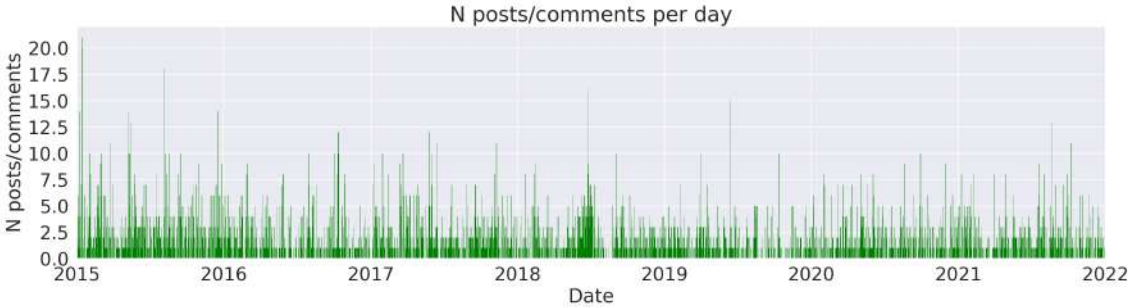
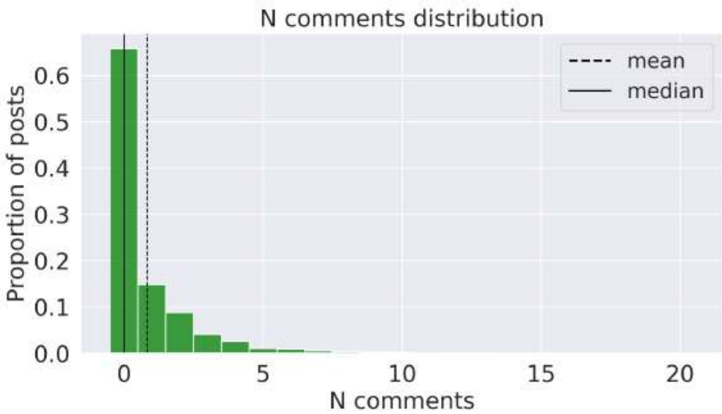


Figure 3.2: The figure shows the number of posts & comments per day with an average of 2.69 posts & comments per day.



The average number of comments for each post is 0.83, with a median of 0. The standard deviation is 1.17. The post with the most number of comments is 20.

Figure 3.3: The figure shows the distribution of comments per post

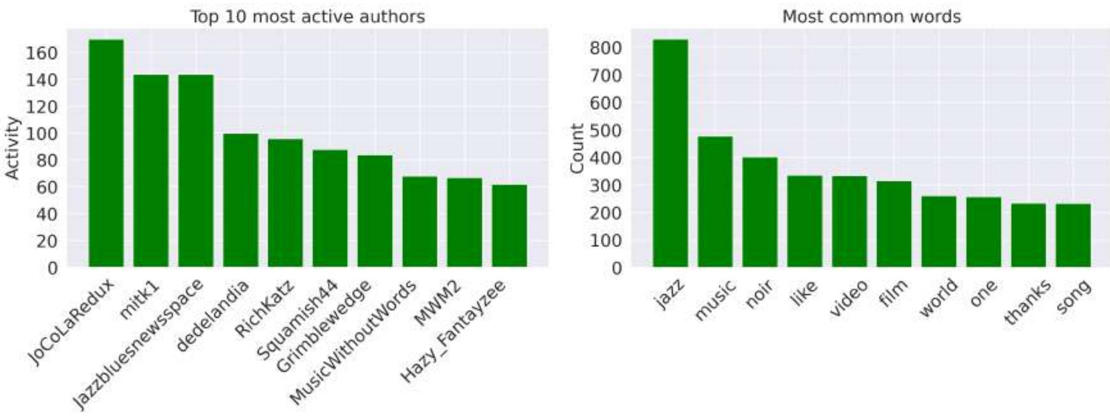


Figure 3.4: The most active users on the subreddit and the most common words used.

3.2.2 DTU

The subreddit DTU is related to the Technical University of Denmark. The dataset contains 2788 data points with 2302 comments and 486 posts. The data is collected from 2015-05-15 to 2022-01-01. The dataset contains 689, unique users. 3 posts are awarded which is 0.6% of all posts.

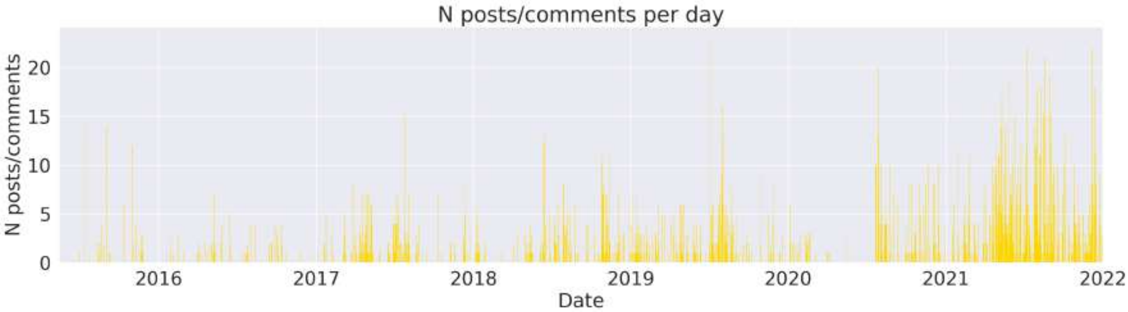


Figure 3.5: The figure shows the number of posts & comments per day with an average of 3.88 posts & comments per day.

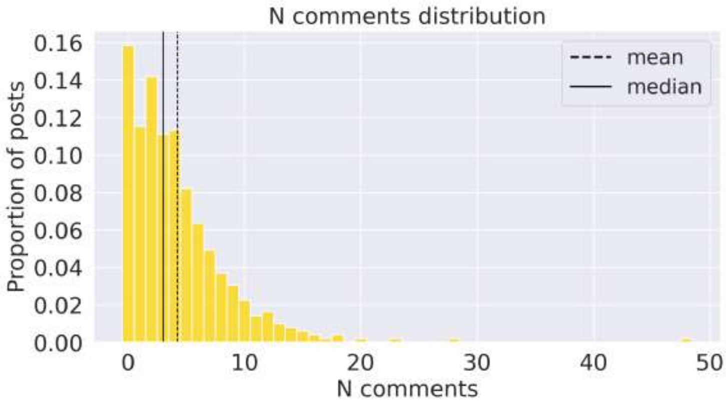
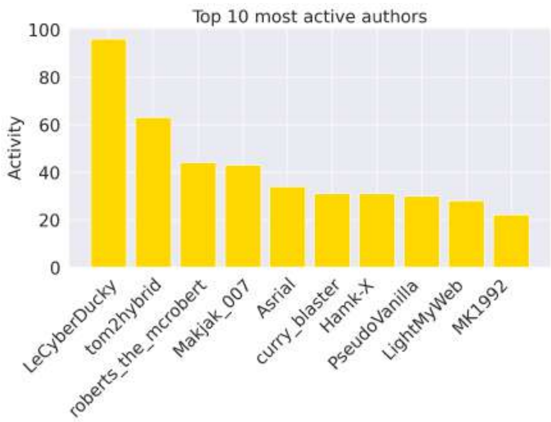


Figure 3.6: The figure shows the distribution of comments per post

The average number of comments for each post is 4.24, with a median of 3. The standard deviation is 4.42. The post with the most number of comments is 48.



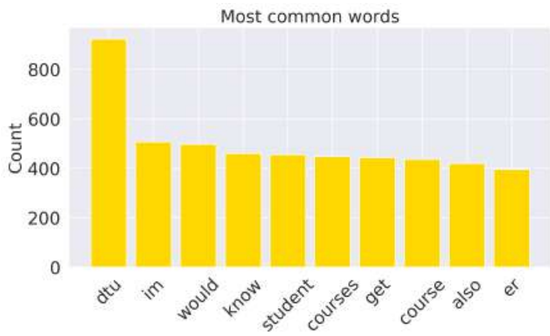


Figure 3.7: The most active users on the subreddit and the most common words used.

3.2.3 FourSentenceStories

The subreddit FourSentenceStories is a forum where users make fictive stories that are four sentences long. The dataset contains 200 data points with 79 comments and 121 posts. The data is collected from 2021-05-04 to 2022-01-01. The dataset contains 34, unique users. 5 posts are awarded which is 4.1% of all posts.

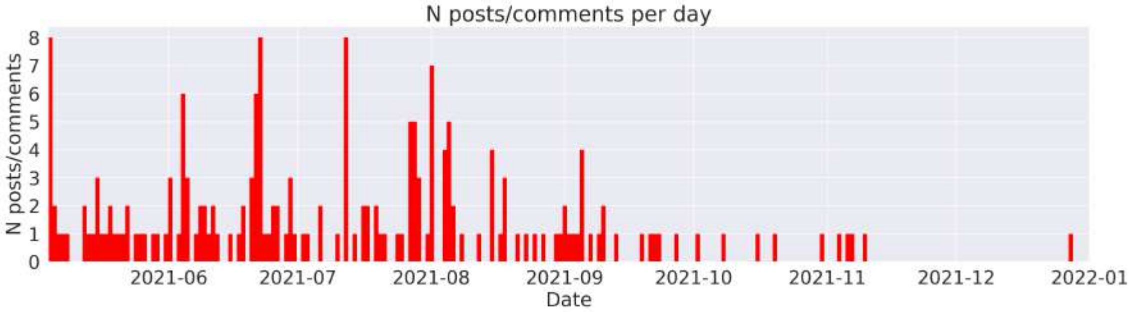


Figure 3.8: The figure shows the number of posts & comments per day with an average of 1.92 posts & comments per day.

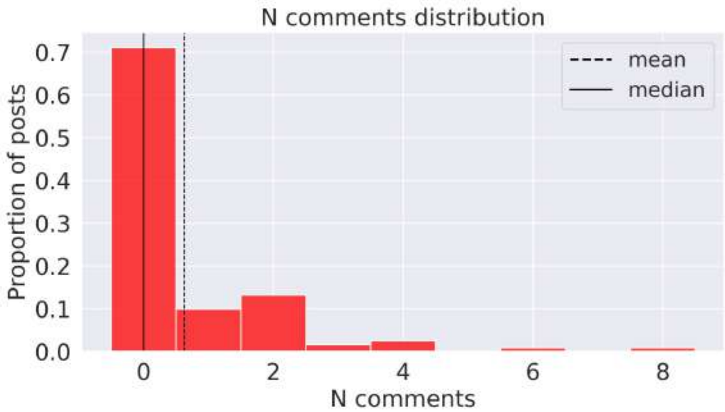
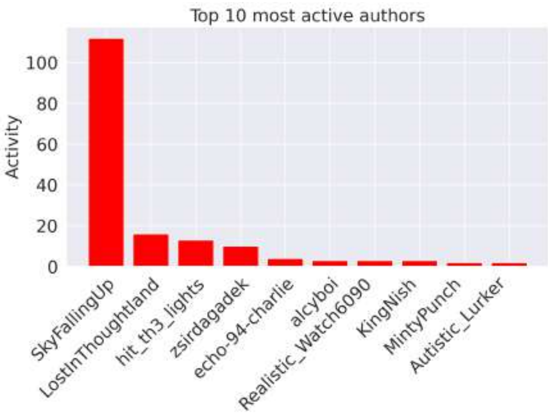


Figure 3.9: The figure shows the distribution of comments per post

The average number of comments for each post is 0.63, with a median of 0. The standard deviation is 1.27. The post with the most number of comments is 8.



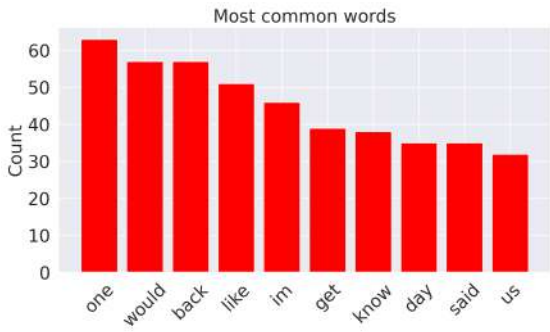


Figure 3.10: The most active users on the subreddit and the most common words used.

3.2.4 indoorbouldering

The subreddit indoorbouldering is a forum where users discuss the sport bouldering. The dataset contains 10532 data points with 8912 comments and 1620 posts. The data is collected from 2017-08-23 to 2022-01-01. The dataset contains 2529, unique users. 82 posts are awarded which is 5.1% of all posts.

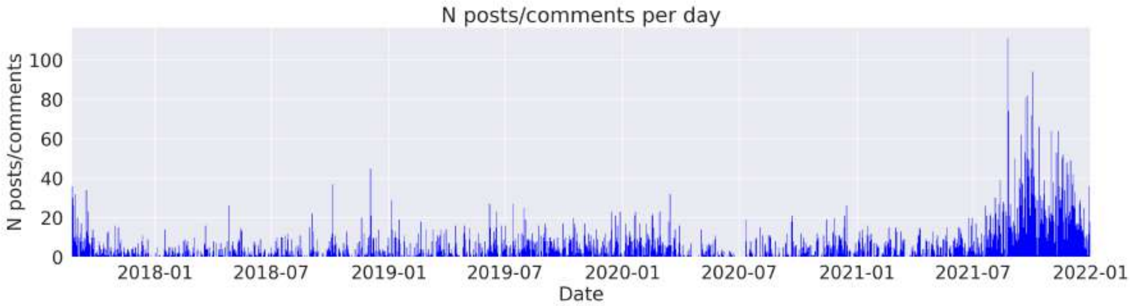


Figure 3.11: The figure shows the number of posts & comments per day with an average of 2.17 posts & comments per day.

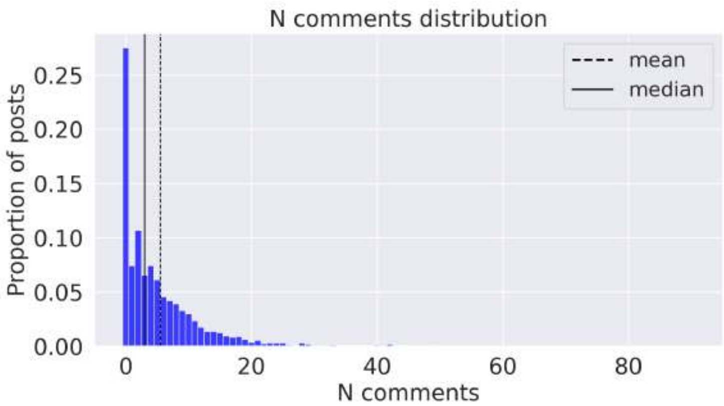


Figure 3.12: The figure shows the distribution of comments per post

The average number of comments for each post is 5.46, with a median of 3. The standard deviation is 7.63. The post with the most number of comments is 90.

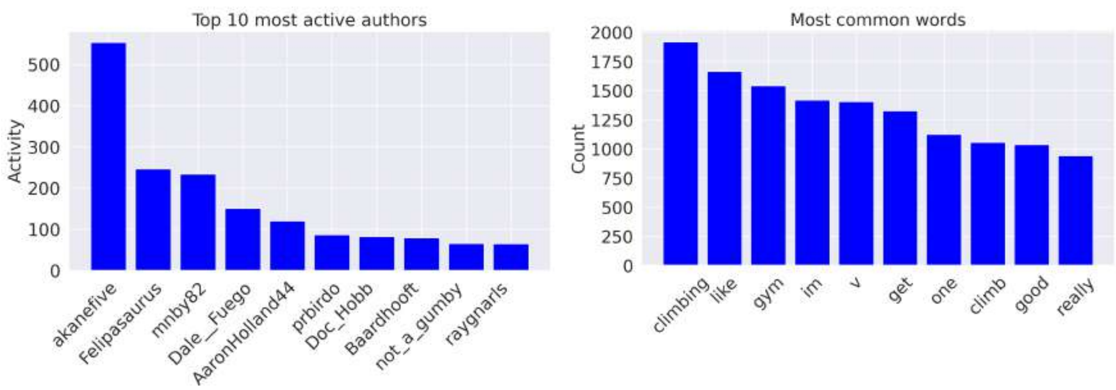


Figure 3.13: The most active users on the subreddit and the most common words used.

3.2.5 kiwi_bird

The subreddit kiwi_bird is a forum where users share pictures with the bird species kiwi. The dataset contains 862 data points with 508 comments and 354 posts. The data is collected from 2015-02-22 to 2022-01-01. The dataset contains 302, unique users. 12 posts are awarded which is 3.4% of all posts.

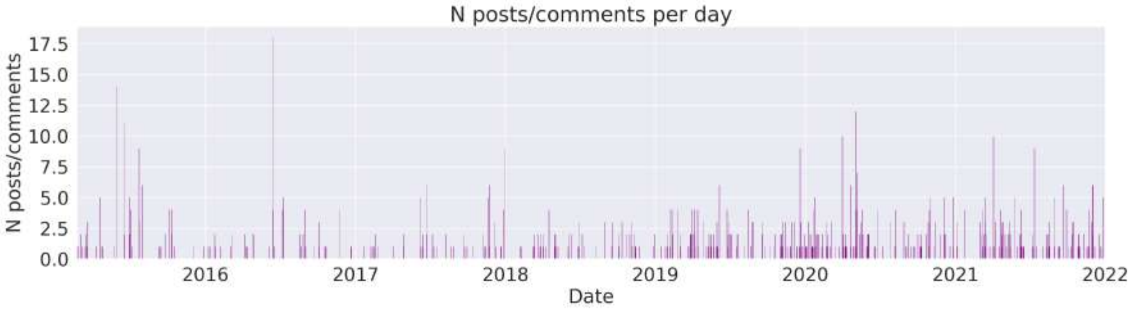


Figure 3.14: The figure shows the number of posts & comments per day with an average of 2.17 posts & comments per day.

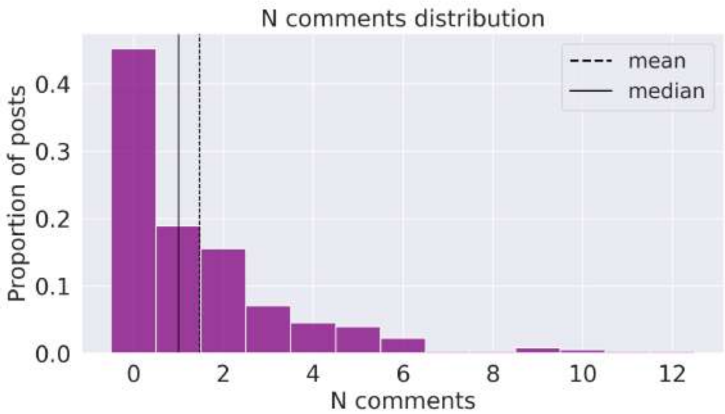


Figure 3.15: The figure shows the distribution of comments per post

The average number of comments for each post is 1.47, with a median of 1. The standard deviation is 2.04. The post with the most number of comments is 12.

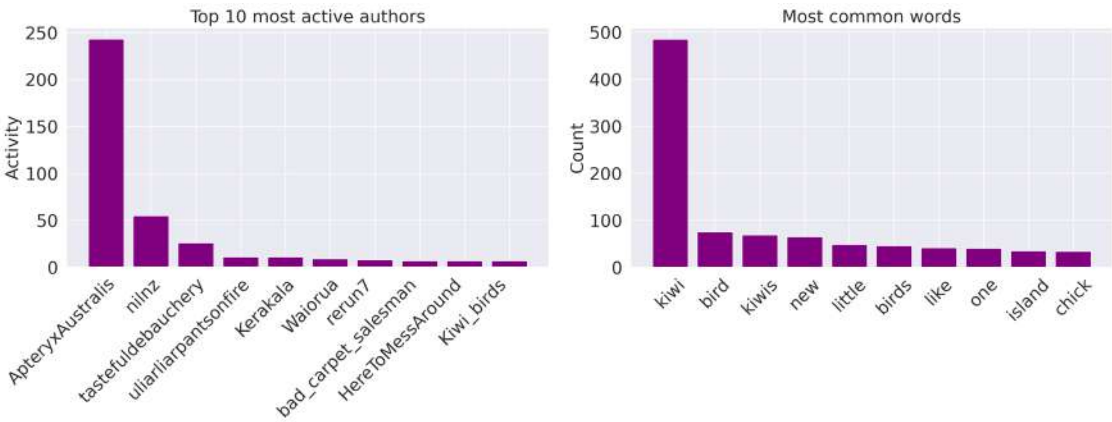


Figure 3.16: The most active users on the subreddit and the most common words used.

3.2.6 GraphTheory

The subreddit GraphTheory is a forum where users discuss graph theory and network science. The dataset contains 497 data points with 303 comments and 194 posts. The data is collected from 2015-02-17 to 2022-01-01. The dataset contains 247, unique users. None of the posts are awarded.

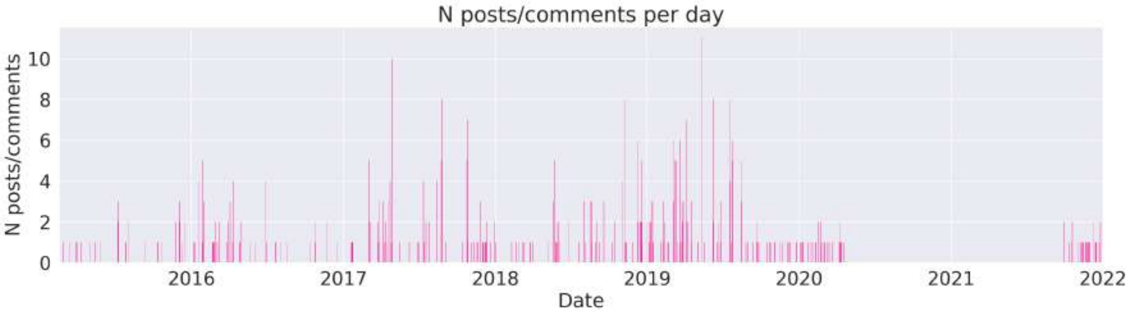


Figure 3.17: The figure shows the number of posts & comments per day with an average of 2.17 posts & comments per day.

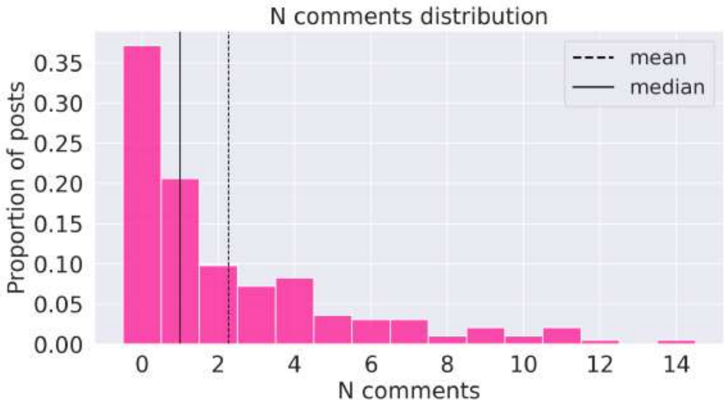


Figure 3.18: The figure shows the distribution of comments per post

The average number of comments for each post is 2.26, with a median of 1. The standard deviation is 2.94. The post with the most number of comments is 14.

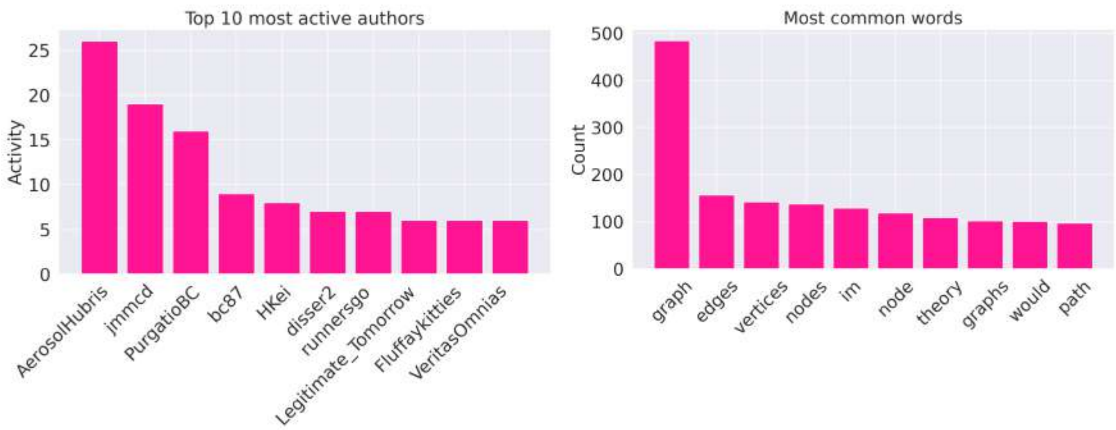


Figure 3.19: The most active users on the subreddit and the most common words used.

3.2.7 Trump666

The subreddit Trump666 is a forum where users believe that former president Donald Trump is the Antichrist. The dataset contains 6746 data points with 5616 comments and 1130 posts. The data is collected from 2015-02-22 to 2022-01-01. The dataset contains 755, unique users. 26 posts are awarded which is 2.3% of all posts.

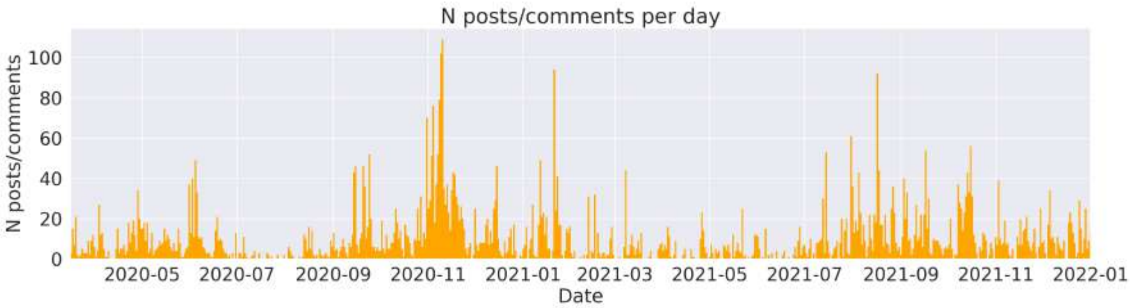


Figure 3.20: The figure shows the number of posts & comments per day with an average of 2.17 posts & comments per day.

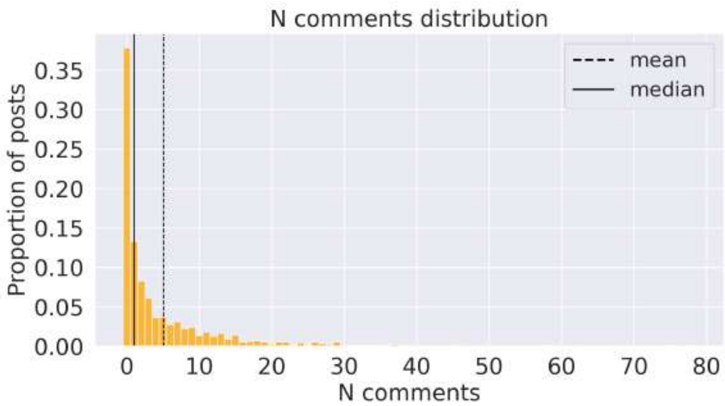


Figure 3.21: The figure shows the distribution of comments per post

The average number of comments for each post is 5.05, with a median of 1. The standard deviation is 8.96. The post with the most number of comments is 78.

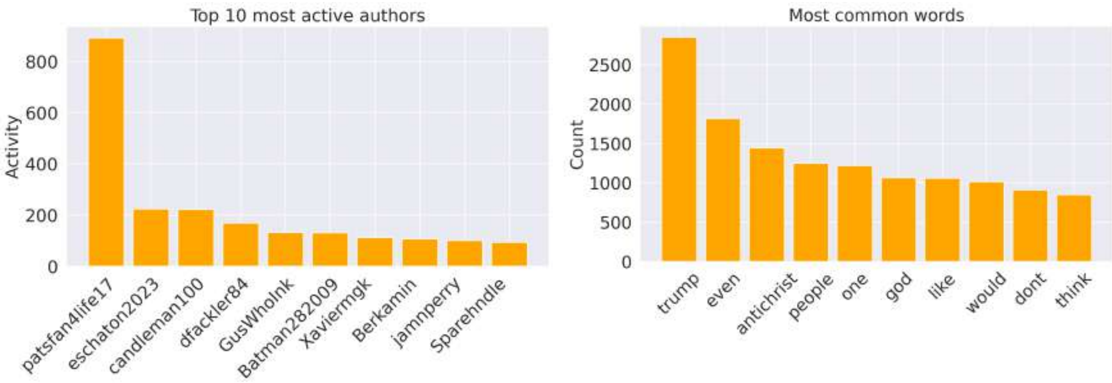


Figure 3.22: The most active users on the subreddit and the most common words used.

3.2.8 Wallstreetbets

The subreddit wallstreetbets is a forum for amateur investors who share investment ideas, strategies, and news. The dataset contains 33049869 data points with 31909222 comments and 1140647 posts. The data is collected from 2016-01-01 to 2021-02-04. The dataset contains 1323990, unique users. 28555 posts are awarded which is 2.5% of all posts.

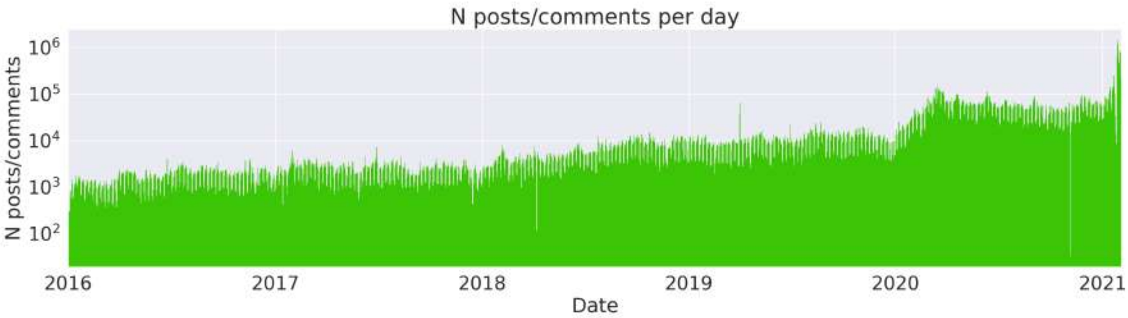
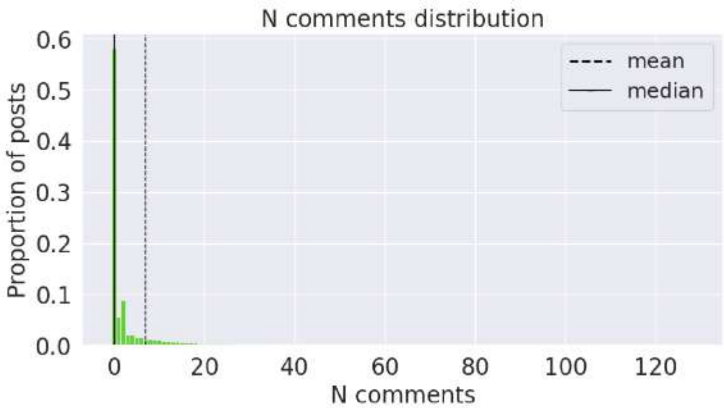


Figure 3.23: The figure shows the number of posts & comments per day with an average of 17750 posts & comments per day. Note that y-axis is log scaled



The average number of comments for each post is 6.85, with a median of 0. The standard deviation is $17.67.96$. The post with the most number of comments is 128.

Figure 3.24: The figure shows the distribution of comments per post

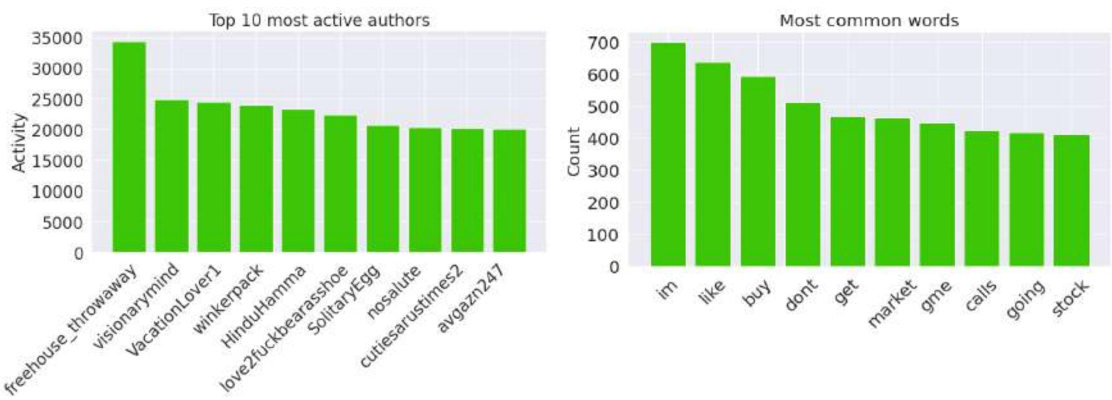


Figure 3.25: The most active users on the subreddit and the most common words used* (The most common words are computed from a sample of 8000 posts & comments, due to large amount of data)

CHAPTER 4

Methods

This chapter contains the methodology used throughout the project. This includes Network science, Natural Language Processing, Basic Statistics, Machine Learning and Transformer Models.

4.1 Data Science Tools & Practices

4.1.1 Tools

Python has been used as the main programming language for this project. A variety of libraries has been used throughout the project. `networkx` is used for the network analysis and for the final predictive deep learning model `Pytorch` has been used. Furthermore, standard libraries such as `pandas`, `numpy` etc have been used. The full list of used libraries can be seen in the `requirements.txt` file on the git repository.

4.1.2 Machine Learning Operations (MLOps)

Machine Learning Operations (MLOps) are considered the best practices for developing machine learning models and conducting data analysis. A well-organized project is key to ensuring efficient workflow and reputability [13]. Keeping track of changes in data and code version control using git ensures the reproducibility and transparency of the project. To standardize and organize the file structure the `cookie-cutter` structure has been applied. Applying the best practices of MLOps makes it easier to adjust for small changes once a promising pipeline has been established, and allows for a more agile work process. For optimizing the hyperparameters while training machine learning models, weights and biases (`wandb`) have been used, where you can monitor the model's performance through a grid search. `wandb` is a python library that enables sweeping hyperparameters and monitoring the performance of models through an online dashboard. Furthermore, cloud computation has been used to compute large computational tasks efficiently, such as training machine learning models.

4.2 Network Science

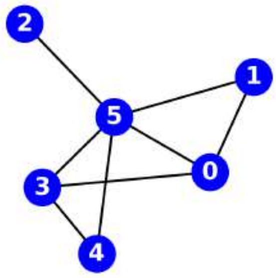
The following section presents some of the central concepts within network science.

Network science (or graph theory) is a field that studies complex networks. A network maps a complex system's distinct elements, these elements are referred to as nodes. The connections between nodes are called edges. The number of edges for a node u is called the degree $deg(u)$. Networks can either be directed or undirected. For directed networks, the edge between two connected nodes is from one node to another $i \rightarrow j$. For the undirected networks, the direction of the edge is irrelevant, thereby the direction of the edge is going both ways $i \leftrightarrow j$. Edges between nodes in networks can be assigned weights based on the frequency of interaction between the nodes, these are called weighted networks [14].

4.2.1 Adjacency Matrix

The adjacency matrix is a matrix used to represent a network. Each element in the matrix indicates whether a pair of nodes are connected (or adjacent). Each row or column in the matrix represents a node, e.g. if $A_{1,2} = 1$ then node 1 and 2 are adjacent. For simple non-weighted networks, the adjacency matrix is a $(0, 1)$-matrix. For undirected networks the matrix is symmetrical. The following plot shows the adjacency matrix and graph of an undirected network with 6 nodes [14].

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$



4.2.2 Creating Networks from Reddit Data

By using the datasets (described in Section 3.2) of posts and comments from Reddit, a network can be created. Every unique author is represented as a node in the graph. An edge is created between the two nodes if they interact by commenting on a post or comment. The graph's edges are weighted by the number of connections between the nodes. Each post/comment has a unique `post_id`/`ID` and each comment has a `Parent_id` that tells which `post_id`/`ID` the comment is connected to. Furthermore, a directed graph can also be computed.

4.2.3 Clustering Coefficient

The clustering coefficient can be used to measure which nodes tend to cluster together [15]. For unweighted, undirected graphs the local cluster coefficient is computed as:

$$C_u = \frac{2T(u)}{deg(u)(deg(u) - 1)}, \tag{4.1}$$

Where $T(u)$ is the number of triangles through the node u and $deg(u)$ is the degree of u. The average cluster coefficient for a graph can be computed by $\overline{C} = \frac{1}{n} \sum_{i=1}^{n} C_i$.

To detect if the network has a "high" clustering coefficient the average clustering can be compared to sampled clustering coefficients using a network with the same number of degrees and nodes. By using the double-edge swap algorithm [16] that randomly swaps edges in a graph while keeping the degrees of the nodes fixed, the average cluster coefficient $\overline{C}$ can be compared to the average cluster coefficient of a double-edge swap version of the same graph $\overline{C_{swap}}$. Since $\overline{C_{swap}}$ is random, the distribution of $\overline{C_{swap}}$ can be computed by sampling multiple swapped graphs. A right-tailed test can then be performed by using the cumulative distribution function $F_C(c) = \int_{-\infty}^{c} f_c(t)dt$ where f_x is the probability density function of C_{swap}. The P-value is computed by:

$$p_{value} = 1 - F_C(\overline{C}) \tag{4.2}$$

4.2.4 Reciprocal Edges

Reciprocal edges are edges that go in both directions in directed networks. The proportion of reciprocal edges R_p can be an indication of how tight the connections between the nodes in the network are. R_N denotes the number of reciprocal edges and E is the number of edges

$$R_p = \frac{R_N}{E} \tag{4.3}$$

4.2.5 Centrality

The centrality of a node can be measured in different ways. Centrality measures are a way of assigning a metric to nodes based on their position in the network [14].

Degree Centrality

The degree centrality CD_u for a node u is simply the fraction of nodes it is connected to.

$$CD_u = \frac{deg(u)}{N-1} \tag{4.4}$$

Where N is the total number of nodes and $deg(u)$ is the degree [14].

Betweenness centrality

The betweenness centrality CB_u of a node u is based on the shortest path and is a measurement of how well one node connects other nodes [17].

$$CB_u = \sum_{s \neq v \neq t} \frac{\sigma_{st}(u)}{\sigma_{st}} \tag{4.5}$$

where s is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of shortest paths that pass through v (where u is not an endpoint).

Closeness centrality

The closeness centrality CC_u is a measure of centrality based on how close the node u is to all nodes in the network [18].

$$CC_u = \frac{N-1}{\sum_x d(x,u)} \tag{4.6}$$

where $d(x,u)$ is the shortest path between the origin node u the node x.

Eigenvector centrality

The eigenvector centrality EC_u is a measure of the centrality of a node u based on its neighbors [19].

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in V} a_{v,t} x_t \tag{4.7}$$

The set of neighbors of node u is denoted as $M(u)$, and the adjacency matrix of the graph is denoted as A. The equation can be rewritten as $Ax = \lambda x$, where x is the eigenvector of A with eigenvalue λ. The eigenvector centrality can be found by finding the eigenvalues of the adjacency matrix.

4.3 Natural Language Processing

Natural language processing (NLP) is the field concerned with the interaction between human language and computer science. NLP extracts information from language and text data. [20]

4.3.1 Sentiment Analysis

Sentiment analysis is a field in NLP that extracts sentimental opinions from a text. The method transforms text into structured and quantitative measurements of sentiment. There are various different methods to compute sentiment analysis.

One of the most common sentiment analyses is VADER (Valence Aware Dictionary for sEntiment Reasoning) [21] provided by the open-source toolkit `NLTK`. VADER uses a sentiment lexicon that contains a comprehensive list of sentiment features. Each word or phrase is rated either negative, positive, or neutral. VADER also uses grammatical rules, so natural words (such as ”very”) amplify the positive or negative component of the sentence. The output of the VADER algorithm is a score of the neutral, positive, and negative sentiment, as well as a compound which is the overall score of the entire text. The compound is normalized so that the values are ranging from $(-1 : 1)$.

4.4 Statistics

Fundamental concepts of statistics are not included in this section.

4.4.1 Pearson correlation coefficient

The Pearson correlation ρ is a measure of linear correlation between two sets of variables. The coefficient is computed by the ratio between the covariance cov and the standard deviation σ between the two sets of data. The covariance is normalized, such that the correlation coefficient is between -1 and 1

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{4.8}$$

4.4.2 Performance Metrics

For binary classification tasks, different metrics can be used to compare and evaluate the performance of a predictive model.

Recall

The recall is defined as the proportion of actual positives that are correctly classified.

$$recall = \frac{TP}{TP + FN} \tag{4.9}$$

Where TP denotes the number of true positives and FN denotes the number of false positive predictions.

Precision

Precision is defined as the proportion of positive identifications that are actually correct.

$$precision = \frac{TP}{TP + TN} \tag{4.10}$$

Where TP denotes the number of true positives and FP denotes the number of false positive predictions.

F1-score
The F1-score is the harmonic mean of the recall and precision defined by the following equation.

$$F1 = 2\frac{precision * recall}{precision + recall} \tag{4.11}$$

4.5 Machine Learning

The following section describes the key methods of Machine Learning, used in the thesis. Fundamental concepts within machine learning and deep learning are not included but can be found here [22],[23].

4.5.1 Transformer Models

Transformer models is a deep learning model that was first introduced in 2017 in "*Attention is all you need*" by Vaswani et al. [24]. A Transformer model is built to handle sequential data, similar to Recurrent Neural Networks (RNN). But it does not handle data in a specific order and is therefore a popular choice for NLP tasks since it detects the contextual input of each word. To do this a transformer model uses a self-attention mechanism, that devotes more attention to smaller specific parts of the data. Transformer models do this by looking at the contextual relationship of the inputs. The architecture has shown huge potential and is used in many models such as Bert, ELECTRA, and GPT3.

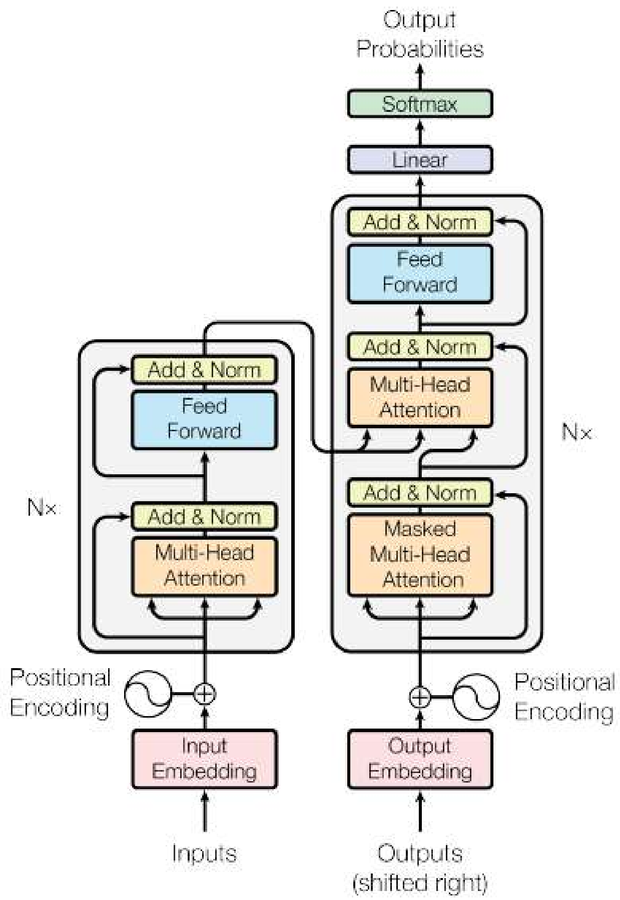


Figure 4.1: The architecture of transformer models from 'Attention Is All You Need' by Vaswani et al.[24]

The architecture of transformers consists of an encoder and a decoder. The encoder generates encodings that contain information about which parts of the inputs are relevant to each other, while the decoder decodes the encodings with the common incorporated contextual information.

Figure 4.1 shows the encoder on the left and the decoder on the right. These modules can be stacked multiple times. The input and output are passed through an embedding layer which maps them to an N-dimensional vector representation, where each word is represented as the index in the vector. Furthermore, the relative position of each word in the sentence is also represented in a vector. A transformer model uses a so-called attention-mechanism to understand the input embedding.

The attention mechanism can be described by the following equation. Where Q (query) is denoted as the vector representation of one word in the sequence. K (key) is denoted as all words in the sequence and lastly, V (value) is denoted as the value.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.12}$$

The dot product of the query Q and the key K is divided by $\sqrt{d_k}$, where d_k is denoted as the dimension of K. A softmax function is applied to distribute the values from 0 to 1. This can be simplified by the following equation.

$$a = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{4.13}$$

Where a denotes the attention-weights, which describes how each word Q is influenced by the other words K. This is a simplification that makes the attention-mechanism easier to conceptually understand.

ELECTRA
The model was proposed in the paper ”ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators” [25]. The model consists of 2 separate transformer models; the generator and the discriminator. The generator replaces tokens in the sequence and the discriminator tries to identify the masked tokens. The overall structure can be seen in the following figure.

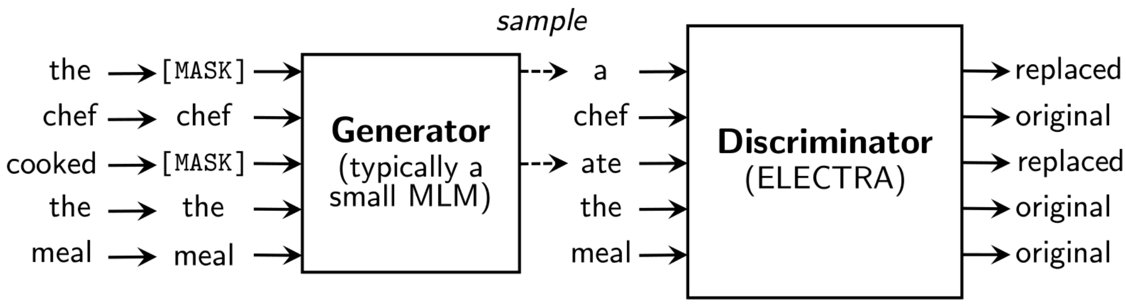
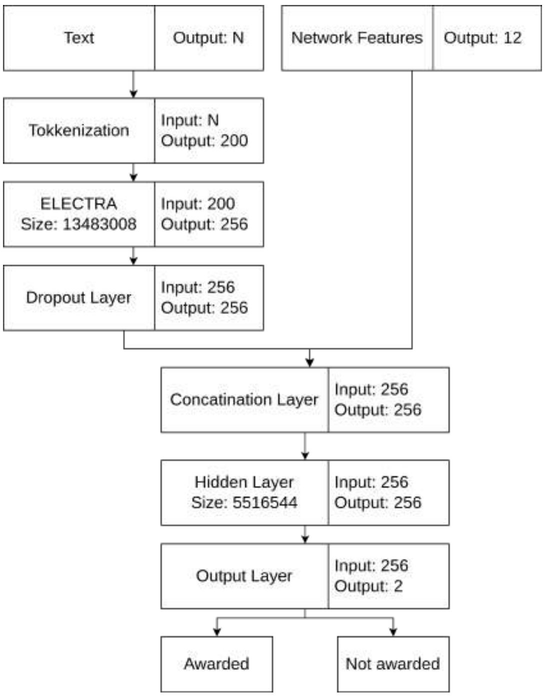


Figure 4.2: From the paper ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators [25].

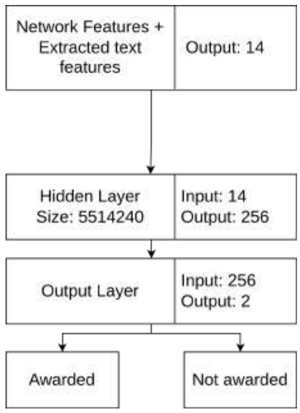
ELECTRA can be used for next-sentence prediction, but can also be used for classification tasks, where the model gives an output of a dense layer of size 256, this is also referred to as classification sequence tokens or CLS-tokens. These embeddings contain information, which can be used for a variety of classification tasks. However, the individual CLS token does not contain information that can conceptually be understood.

When using the ELECTRA model in this thesis project we used a pre-trained model with initial weights since training an ELECTRA model would require a large amount of computing. Therefore the initial weights are transferred and the last layer of ELECTRA is trained to optimize the specific classification task, this is also known as last-layer-training.

The Deep Learning Models
We tested and developed two different deep learning models to predict whether a post would be awarded or not. Model (b) uses network features and extracted textual features with a hidden layer consisting of 7 linear layers, totaling 5514240 parameters, with a ReLu activation function between each layer. The model definition is available in Appendix B.1. Model (a) uses the ELECTRA transformer 4.5.1 to handle the textual input, tokenizing the raw text before feeding it into the transformer, which had 13483008 parameters. A dropout layer is used to prevent overfitting of the textual input before concatenating the output of the ELECTRA with the network features. A hidden layer with 6 layers and 5516544 parameters is then used for the concatenated layer. The code snippet for this model is provided in Appendix B.2. Both models can be seen in the following Figure 4.3.



(a) Neural Network + ELECTRA



(b) Neural Network

Figure 4.3: Figures (a) and (b) shows the overall structure of the deep learning models used to classify if a post is awarded or not.

4.5.2 Random Forrest

Tree-based models partition the feature space into subsets and fit a simple model to each subset by recursively splitting the feature space [23]. Consider the response variable Y and the continuous input variable X. A decision tree splits the feature X by, for example, $X > c$, where c is a constant that divides the feature space into a subset so that most variance is explained in Y. Decision trees can vary in depth, which partitions the feature space into different subsets. By doing this, tree-based models can learn irregular patterns that linear models cannot learn (e.g. logistic regression). Decision trees are conceptually very simple, and one of the key advantages of decision tree models is their interpretability. They are also very versatile and can be used for categorical, discrete, and continuous input variables for regression and classification tasks.

Random forest [26] is an ensemble of decision trees that uses bootstrap aggregation also called bagging. Given a training set of the input data $X = x_1, x_2..., x_n$ and the response $Y = y_1, y_2..., y_n$ bagging is repeated B times by sampling the data with replacement. A decision tree is then fitted to each sampled data in the bag B.

After training, the prediction can be computed by taking the average across all the decision-trees output $f_b(x\prime)$ or the majority vote for classification tasks.

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b\left(x'\right) \tag{4.14}$$

Single decision trees often tend to overfit and have low bias and very high variance, but by using an ensemble of decision trees and averaging multiple trees' output with bagging it is possible to reduce the overall variance.

4.5.3 Permutation feature importance

Permutation importance is a method used to inspect the behavior of a model and the importance of individual features. The method randomly shuffles all values of a single feature, the model is then fitted to the corrupted data. By doing this any potential relationship between the target and the feature is removed. An accuracy metric of the original model and the corrupted model is then compared. The following equation shows the permutation importance i_j, where j denotes the feature. s is the score (eg. F1 score) of the original model, K denotes the number of reputations, and $s_{k,j}$ denotes the score of the corrupted model.

$$i_j = s - \frac{1}{K} \sum_{k=1}^{K} s_{k,j} \tag{4.15}$$

Permutation feature importance can be used to evaluate the predictive power of a feature in a specific model and does not necessarily explain the intrinsic predictive value of a feature by itself, but its importance in a particular model. However, it can give an indication of the overall behavior of a feature.

4.5.4 Principal Component Analysis

Principal component analysis, or PCA, is a technique to reduce the number of dimensions of an observation, such that most information is retained and represented in a lower dimensional space. PCA was first introduced in 1901 by Karl Pearson [27].

PCA is a technique that discovers the principal components of a dataset. These components are a set of orthogonal linear transformations of the original variables that convert the data into a new coordinate system, where they capture the highest variance in the dataset. The initial principal component describes the most significant variance, followed by the second component, and so on. The singular value decomposition (SVD) is used to compute the principal components. [22]

CHAPTER 5

Results - Factors of Awarded Posts

This chapter presents the result of the explorative analysis. In Section 5.1 we reveal why we are using Reddit's award system as a metric for acknowledgment and why we are focusing on the subreddit Wallstreetbets. In Section 5.4 & 5.3, we present the network and textual features for Wallstreetbets, as well as an analysis of these features.

5.1 Defining Acknowledgment on Reddit

Reputation is explained in section 1.3 as the general opinion of someone. Reddit uses a system of either scores or awards where users can acknowledge an individual's content and show their appreciation (Section 1.2). The following section shows the key results from the study on how reputation is defined for a user on Reddit. The result show why we are using awards as a metric of acknowledgment and additionally why we are using wallstreetbets as the main source of data.

5.1.1 Awards versus Scores

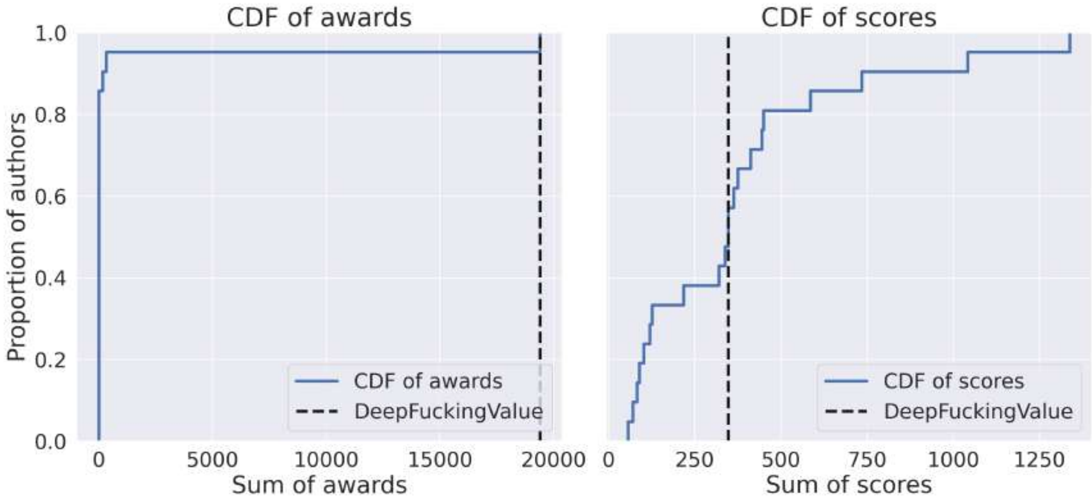
By examining users with a known positive reputation we are able to compare the score and the awards given to those individuals. The users 1RONYMAN, ControlTheNarrative, and DeepFuckingValue have been studied [28].

On the subreddit Wallstreetbets the user 1RONYMAN exploited a popular stock trading platform, where the user created a risk-free trade using a method called ”boxspread” on the trading platform Robinhood in 2019. This gained the user a huge amount of reputation [28] [29].

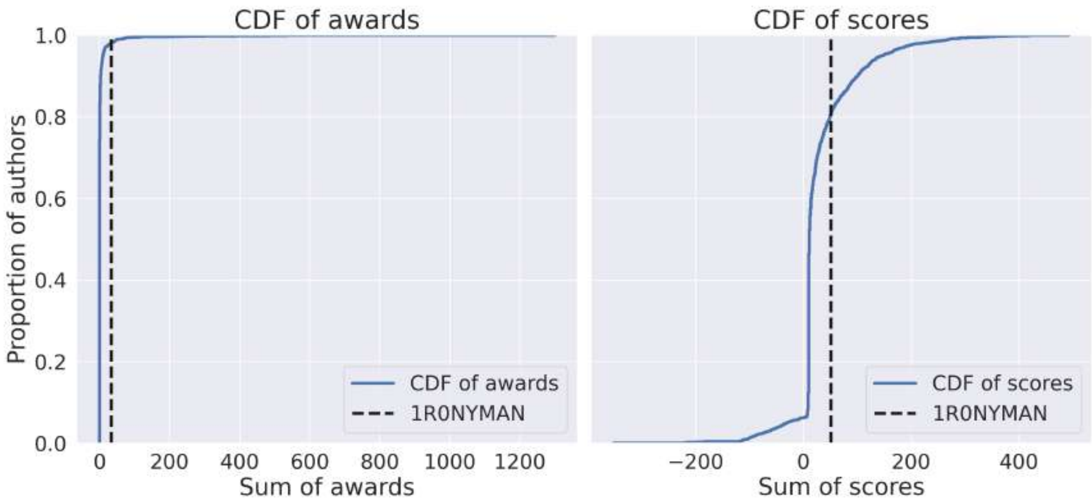
Another user known as ControlTheNarrative found a bug on the trading platform Robinhood, where the user could leverage the stocks of a ratio of $25:1$. This was later referred to as the ”free money cheat code” in the community of Wallstreetbets. The user shared the story on the subreddit and gained a lot of reputation [28] [30].

The user DeepFuckingValue is a highly known user within Wallstreetbets and is considered one of the founders of the GameStop short squeeze [31]. He first posted about the stock in September 2019 and is known in the community as a person with a very high reputation [28] [32].

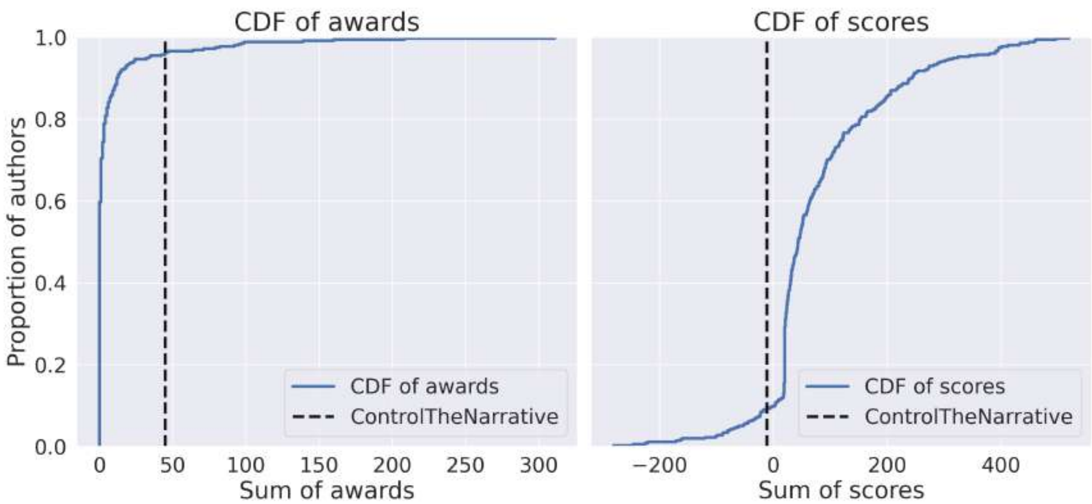
In order to investigate if these users have more awards and larger scores than the population mean. We computed the Cumulative Density Function (CDF) for the sum of scores and awards for users with the same activity (umber of posts) as DeepFuckingValue, 1R0NYMAN, and ControlTheNarrative. The CDFs can then be compared to the individual's sum of awards and scores.



(a) DeepFuckingValue



(b) 1RONYMAN



(c) ControlTheNarrative

Figure 5.1: Awards and Score comparison for highly reputed users. The blue lines indicate the CDF for awards and scores for users with the same activity level. The user's sum of awards and scores is denoted with a black line.

As seen in the examples (5.1a 5.1b 5.1c) the user has more awards relative to other users with the same activity level, while the score of the user is relatively low compared to other users.

To determine if the awards or scores are significantly higher than the population means, a one-sample t-test is performed by: $p = \frac{\sum_{i=1}^{n}[x_i > X]}{n}$. The following table shows the p-values.

	p-value, Awards	p-value, Score
1R0NYMAN	0.0396	0.1881
DeepFuckingValue	0.0	0.4286
ControlTheNarative	0.0366	0.9070

Table 5.1

For all users, the p-value for the sum of awards is below the significance level of 0.05, meaning that the users have a significantly higher sum of awards than the population means. Furthermore, the p-value for the sum of scores is above the significance level. This indicates that the number of awards is a better indicator of an individual's reputation on Reddit, assuming the users have a high reputation. This is further elaborated and discussed in Section 7.1.

5.1.2 Awards

The following section shows the result of awards as the chosen metric for acknowledgment. The following figure shows the number of awards for each subreddit used in this project.

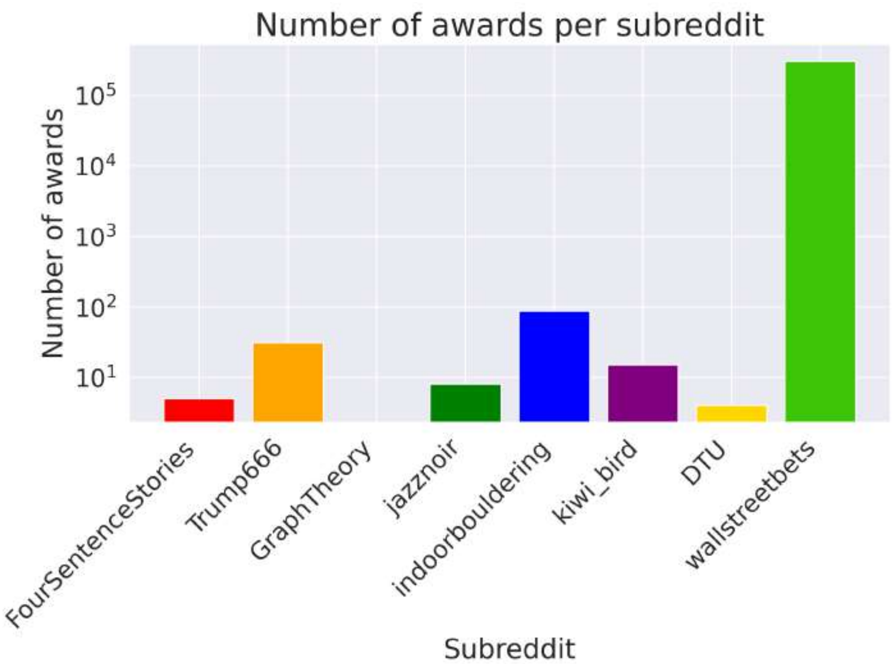


Figure 5.2: Sum of awards on subreddits with the y-axis log scaled. It can be seen that the subreddit wallstreetbets contains a larger amount of awarded posts than the other subreddits.

The figure shows that wallstreetbets has a large number of awarded posts, compared to the other posts. Wallstreetbet contains a total of 304817 awards, while the second most awarded subreddit indoorboulding only contains 88 awards. Due to the lack of awarded posts in other subreddits, wallstreetbets is used to investigate reputation. The following figure shows the distribution of the sum of awards for each user on wallstreetbets.

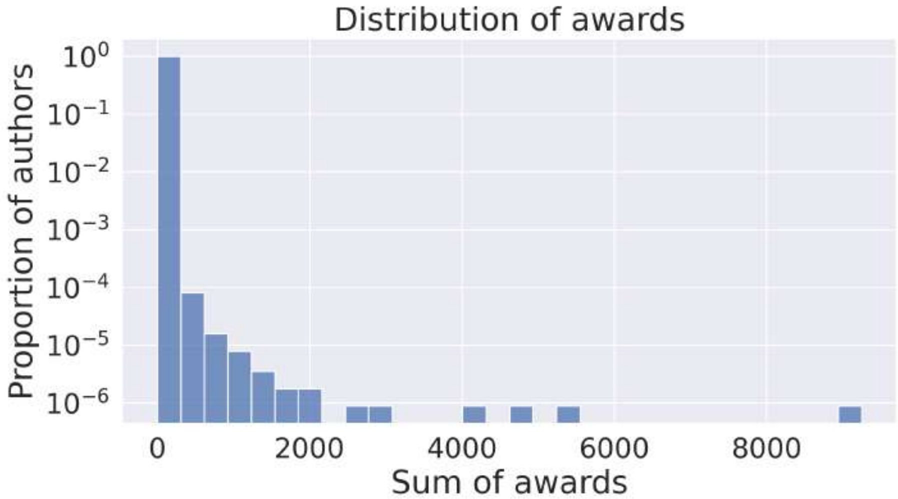


Figure 5.3: The distribution of the sum of awards for each user on wallstreetbets with the y-axis log scaled.

It can be seen that the majority of posts do not receive awards, hence making awards a rare commodity. This exclusivity makes receiving an award a notable achievement and can be seen as a symbol of recognition for the hard work and dedication put into creating the post. Only 4% of the authors from wallstreetbets received an award.

Due to the scarcity of awards on other subreddits, wallstreetbets is further investigated. For the final predictive model awarded posts are used as the dependent variable. Due to the scarcity and wide distribution of awards seen in Figure 5.3 the aim of the predictive model is to classify if a post/author receives an award. A new nominal variable called `awarded` is introduced to indicate whether a post has received an award or not.

5.2 The Effect of Reputation

The following section contains results of how reputed users may have an impact. More specifically how the user DeepFuckingValue affected the GameStop (GME) stock. Keith Gill known by the username DeepFuckingValue is seen as one of the driving factors in the GameStop short squeeze of January 2021 [31].

The figure shows the number of times the stock is mentioned each day in posts on wallstreetbets. By doing this we can measure the "trend" of the stock inside the subreddit. The Gamestop stock price is also plotted. Furthermore, each post from DeepFuckingValue is annotated by a dotted line, and a bar illustrates the number of awards given to the post.

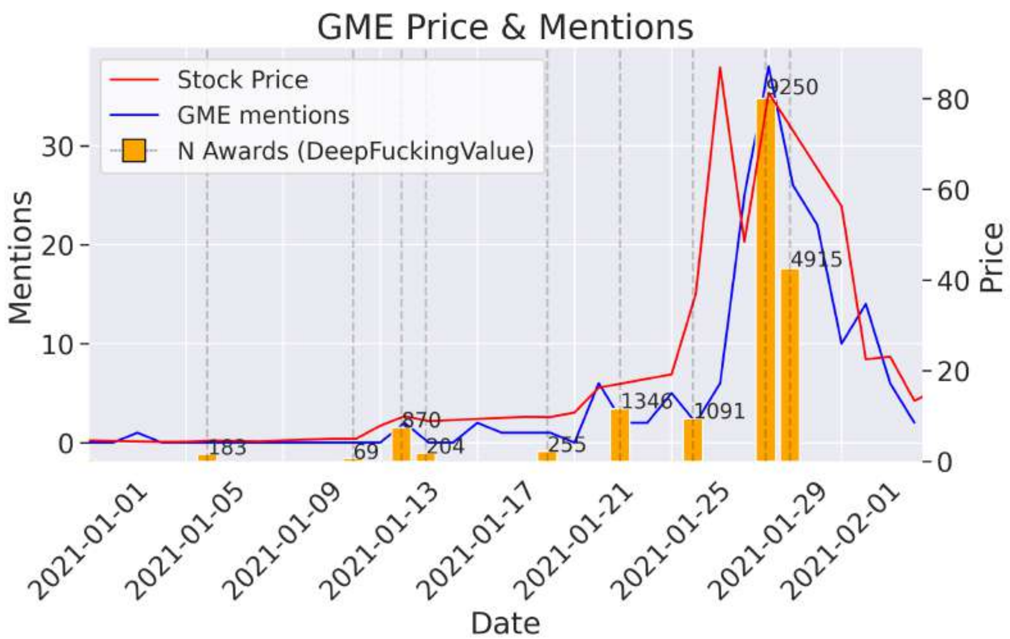


Figure 5.4: Illustration of the effect of a person with a high reputation. The red and blue lines show the stock price and the number of times GME is mentioned respectively. The dotted vertical lines show each time DeepFuckingValue made a post about GME. Additionally, the number of awards for each post is shown by the orange bars, which are annotated by the number of awards.

On the evening of 2021-01-11, DeepFuckingValue made a post (Appendix A.1.1) showing the stocks he bought in GameStop, and 3 days later 2021-01-14 the stock had increased 100.15%. DeepFuckingValue made a similar post the 2021-01-25 and 2 days later the stock increased by 352%.

This can be further investigated by comparing the return on investment (ROI) of the GameStop stock after DeepFuckingValue made a post to other randomly sampled dates. The ROI is calculated by:

$$ROI = \frac{(profit - cost)}{cost} \tag{5.1}$$

First, the same number of dates as the number of posts by DeepFuckingValue is randomly sampled from the same period as DeepFuckingValue posted about GameStop. Then the ROI three days after these random dates is computed. The mean of the ROIs is then computed. This is repeated $N = 2000$ times and then compared to the mean of ROIs for the original dates from the posts by DeepFuckingValue.

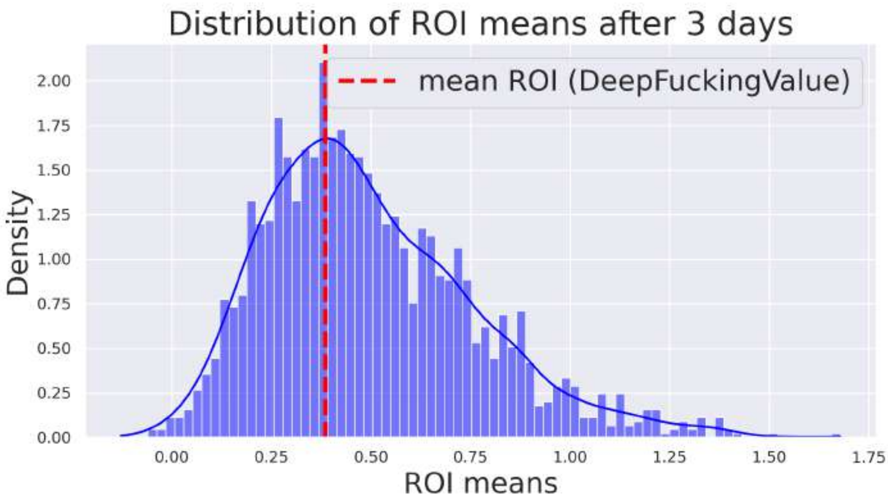


Figure 5.5: The figure shows the distribution from 2000 samples of mean ROIs of GameStop from random dates after 3 days. The red line shows the mean of ROIs for the original dates after 3 days.

The distribution of mean ROIs from random dates follows a normal distribution. It appears that the mean of ROIs for the dates of posts from DeepFuckingValue after 3 days is not higher than the population means. Additionally, we computed the p-value of the mean as being significantly higher than the population means by performing a one-sample t-test:

$$p_{value} = \frac{\sum_{i=1}^{n} [x_i > X]}{n} = 0.616 \tag{5.2}$$

This shows that means is not significantly higher than the population means, which indicates that DeepFuckingValue did not affect the stock price of GameStop. Nonetheless, it cannot be inferred that DeepFuckingValue had no impact. Other factors may have come into play, such as external events or news that may have influenced the stock price. Therefore, while the data suggest that DeepFuckingValue's posts did not have a significant impact on GameStop's stock price, it is important to consider that the user started a movement in the community and gained a large following, which in turn may have influenced the behavior of other investors in the market. DeepFuckingValue's posts may have contributed to increased trading activity in GameStop's stock. It is also worth noting that the impact of DeepFuckingValue's posts is difficult to quantify, as the stock market is a complex system with many variables at play.

5.3 Network Features

The following section presents the result of the network analysis of wallstreetbets. This includes results from how the networks were created using discrete-time dynamic graphs, as well as an analysis of the features extracted from the networks. The distinctions between the features for awarded and not awarded posts are analyzed to identify the factors contributing to receiving an award.

5.3.1 Discrete-Time Dynamic Networks

We created discrete-time dynamic networks for wallstreetbets. These networks are a static snapshot of 10 weeks in the period before a post was created, this creates a network for each individual post. By doing this we are able to analyze the network before a post is created and the behavior of the individual who made the post in that 10-week period. The data is balanced such that it contains an equal amount of awarded and not awarded posts, this is done by downsampling the majority class not awarded posts.

The following plot shows the number of nodes for each discrete-time dynamic network, created for each individual post.

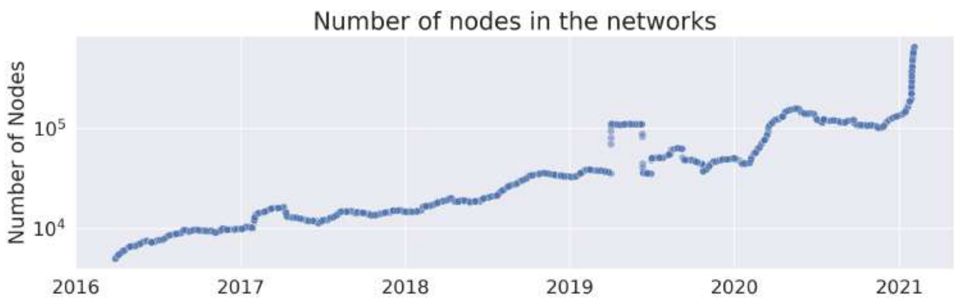


Figure 5.6: The number of nodes for each network by date. The y-axis is log-scaled

The following figure illustrates the number of awards given to the authors of the posts from where a network is created.

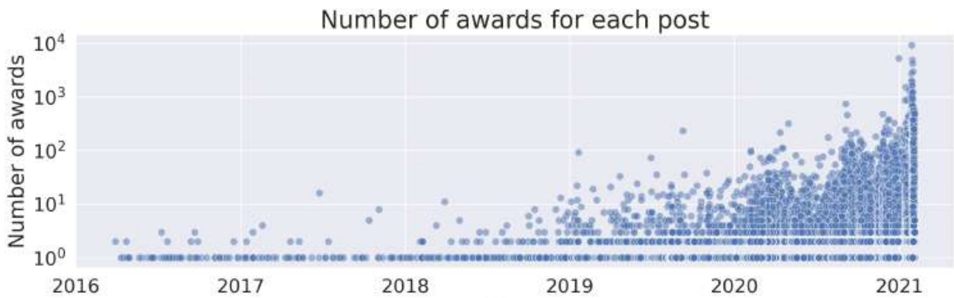


Figure 5.7: The number of awards given to each post. The y-axis is log-scaled

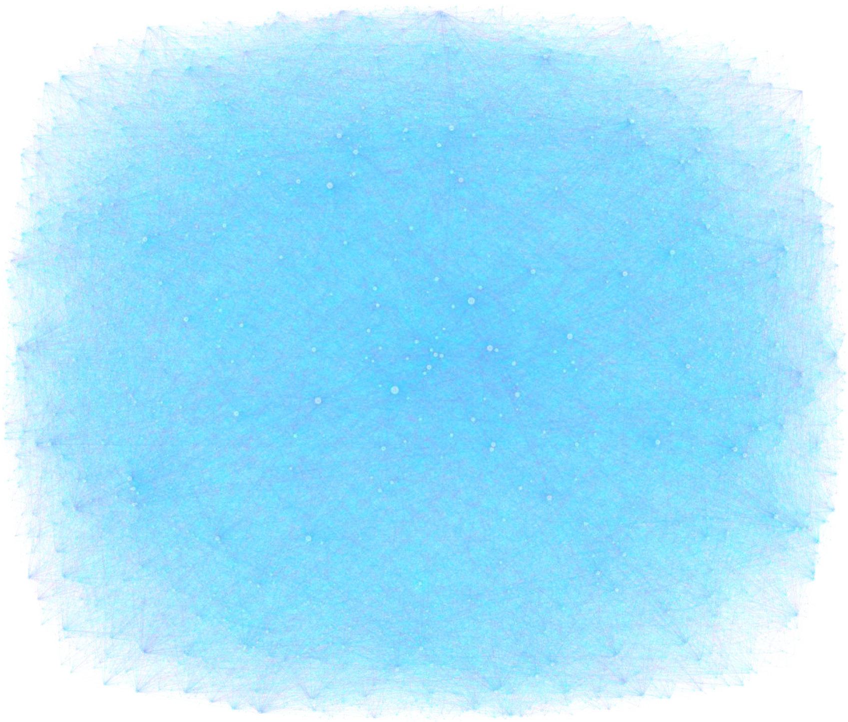


Figure 5.8: An example of a discrete-time dynamic graph (2020-12-25 to 2021-03-25) of the first graph created. The edges are colored by the reciprocity and the size of the node is defined by the number of degrees. Only nodes from the biggest coherent part of the network are included.

By visualizing networks it is often possible to get a better understanding of the social dynamics of a community. However, due to the size of this network, it is hard to comprehend all the information. We visualized this primarily because it is it's aesthetically pleasing, but also to see how complex the network is.

5.3.2 Network Features

The following section contains an analysis of the networks from the discrete-time dynamic networks from wallstreetbets. For each network, we extracted and computed a variety of features that are used to explain and understand the network. This includes overall network features and features from the authors of the posts, by which the discrete-time dynamic networks were created. It is important to note that the interactions from the post from which each network where created is not included, since the network is created in the time period before the post where made. By doing this the features are not biased toward the posts, but instead, explain the social dynamic of the author within the network.

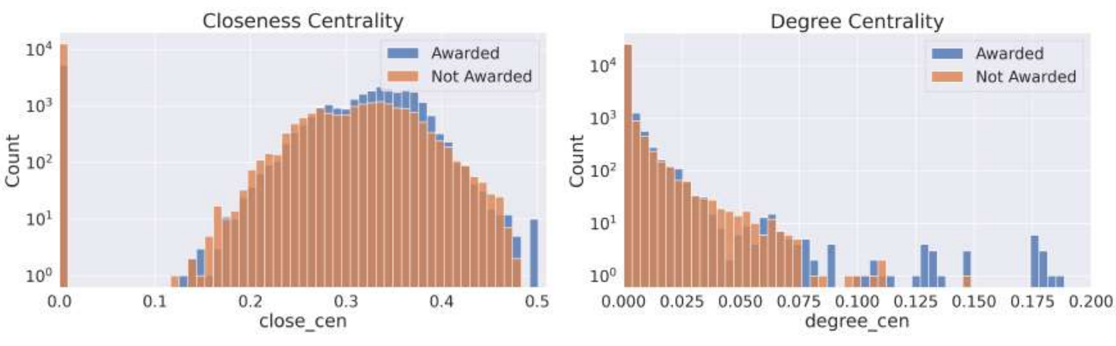
Feature	Description
edges	The number of edges in the network
nodes	The number of nodes in the network
degree	The degree of the author of the post
degree centrality	The degree centrality of the author of the post (Section 4.2.5)
closeness centrality	The closeness centrality of the node (Section 4.2.5
in-degree	The in-degree of the author of the post in the directed network
out-degree	The out-degree of the author of the post in the directed network
fraction of reciprocal	The fraction of reciprocal edges of the whole network (Section 4.3)
reciprocal edges of the author	The sum of reciprocal edges of the author of the post (Section 4.3)
reciprocaledges	The total sum of reciprocal edges (Section 4.3)
activity	The number of posts & comments by the author of the post in the time interval
mentions	The number of times the author of the post has been mentioned in other posts & comments.

Table 5.2: The network features extracted from the discrete-dynamics graphs for each post

The following sections contain an analysis of some of the key features. Distributions of all features can be seen in Appendix A.4.

Closeness centrality & degree centrality

We computed the closeness centrality and the degree centrality of the author of the post. The following figure shows the distribution of each post/author.

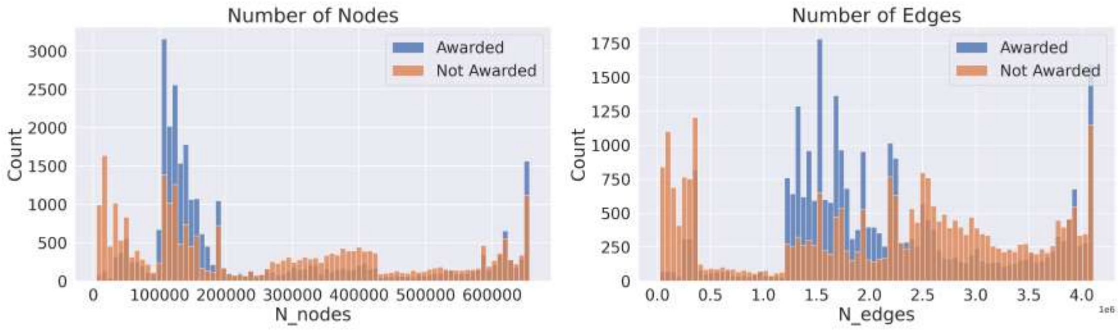


(a) (b)

Figure 5.9: The closeness centrality (a) and the degree centrality (b) colored by awarded or not awarded.

A visual interpretation of the Figure 5.9 (a) suggests that the closeness centrality for both authors that received and did not receive an award follows a similar normal distribution. This indicates closeness centrality does not explain much of the variance for awards. However, few awarded authors have a high closeness centrality around 0.5. Additionally, it is hard to distinguish awarded users by observing the degree centrality. But some awarded users have a high degree centrality. It is important to note that the y-axis is log scaled on both figures, so the smaller values appear proportionally larger.

Number of Nodes & Edges
We computed the total number of edges and nodes for each network, which can be seen in the following figure.

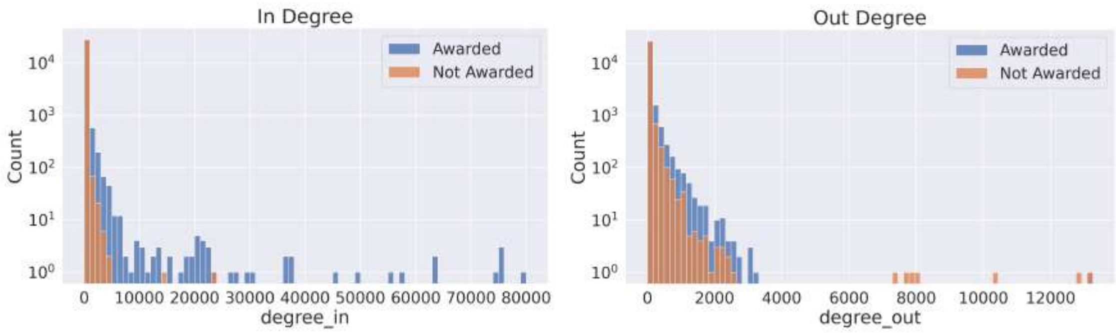


(a) (b)

Figure 5.10: Distribution on the number of nodes (a) and the number of edges (b) colored by awarded or not awarded.

It is hard to distinguish the distribution of the number of nodes and edges for networks created for awarded and not awarded posts. However, it appears that it is more likely to receive an award if you make a post when the network has a size of around 150000 nodes. Another explanation could be that more awards were given at a certain point in time. This can be seen in appendix A.8, which shows the number of awards given by date.

In and out Degree
We computed the in-degree, the number of edges directed toward the author of the post, which is how many times other users engage with the author. Additionally, we also computed the out-degree in the 10-week period, which is the number of edges going away from the author of the post. The out and in-degree can be seen in the following figure.



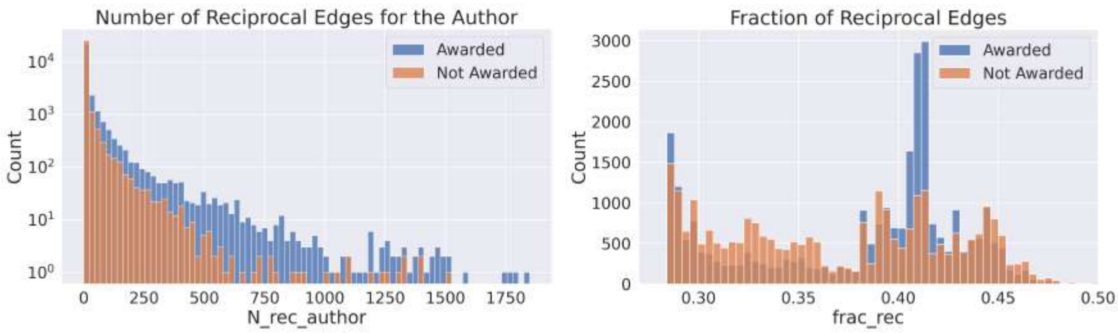
(a) (b)

Figure 5.11: Distribution of the in-degree (a) and the out-degree (b) colored by awarded or not awarded. Note that the y-axis is log scaled for both figures

The figure shows that awarded users appear to have a larger in-degree, meaning that users are more likely to comment on the author's content. Additionally, awarded users also appear to have a higher out-degree, meaning that they are more likely to comment on other users' content. This suggests that awarded users are not only popular within the network, but also active and engaged members of the community. Additionally, few not awarded authors have a large out-degree, this could potentially be "bots" or similar.

Reciprocal Edges

Reciprocal edges are edges where interaction occurs in both directions between the nodes, such that both users have engaged with each other. Figure 5.12 shows the reciprocal edges for the author of the post and also the fraction of the total number of reciprocal edges in the network, which we computed by $\frac{Nnodes}{Nreciprocaledges}$. The fraction of reciprocal edges can be used as a measure of how engaged and connected the community generally is (see Section 4.3).



(a) (b)

Figure 5.12: The number of reciprocal edges for the author of the post (a) and the fraction of the total number of reciprocal edges in the network. (b) The distributions are colored by awarded or not awarded. Note that the y-axis of figure (b) is log-scaled

The distribution of reciprocal edges for the author of the post indicates that awarded users appear to be more active and engaged in the community. The figure of the fraction of reciprocal edges suggests that posts are more likely to get awarded when the fraction of reciprocal edges is around 0.40 however, this could also be because more awards were given at a certain point in time. This can be seen in Appendix A.15 where the fraction of reciprocal edges are plotted by date.

Mentions
Reddit uses a system where users can mention or ”tag” other users in their content by writing ”/u/” followed by the specific username. We computed the number of times other users mentioned the author of the post in the network.

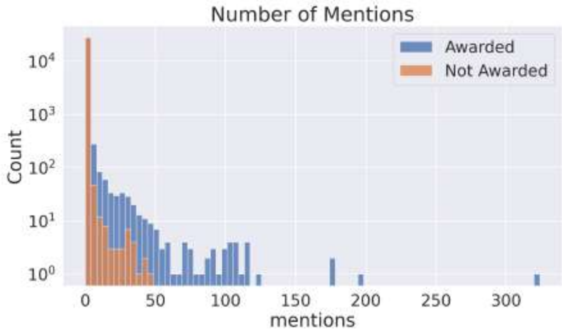


Figure 5.13: The distribution of mentions of the user, colored by awarded or not awarded posts. Note that the y-axis is log-scaled.

The figure shows that an individual is more likely to receive an award if the person is previously mentioned by other users. This indicates users are more likely to be acknowledged if they already are known within the community.

5.4 Textual Features

The following section shows the key result from the textual analysis of the posts from wallstreetbets. This includes textual feature extraction such as the VADER sentiment algorithm. Additionally, the embeddings from ELECTRA are also presented.

5.4.1 Sentiment

Using the VADER algorithm explained in Section 4.3.1 it is possible to extract the sentiment compound from posts. We conducted a sentiment analysis of awarded and not awarded content for posts on wallstreetbets. The following figure shows the sentiment compound from the VADER algorithm.

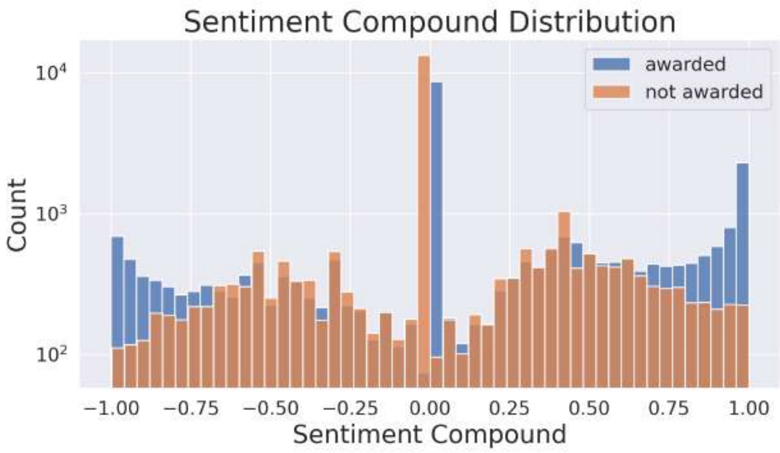


Figure 5.14: The figure shows the sentiment compound from the VADER algorithm for Awarded or not awarded posts. Note that the y-axis is log-scaled.

The figure shows that the awarded posts appeared to be more negative and positive, while the not awarded posts were more neutral in tone. This could indicate that awarded content is more subjective, which tends to evoke stronger emotional responses, whether positive or negative, while more objective content tends to be more neutral in tone. An example of a positive awarded post can be seen in Appendix A.1.

5.4.2 Text Length

The following section shows the result of the text length analysis. We counted the character length of posts on wallstreetbets. The following figure shows the text length for awarded and not awarded posts.

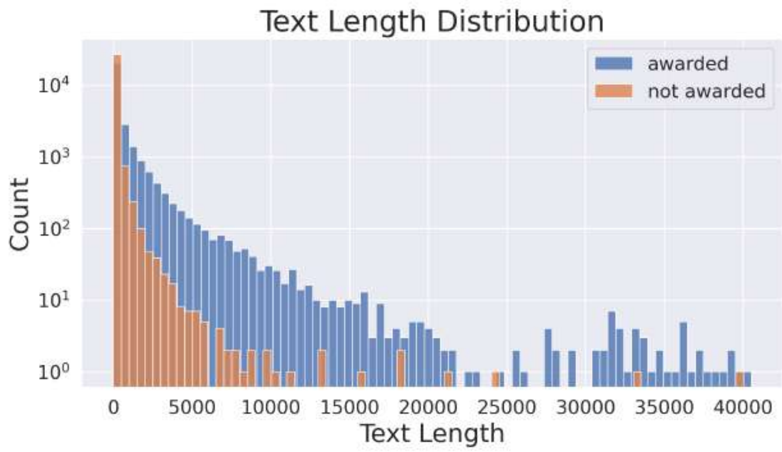


Figure 5.15: The text length distribution of awarded and not awarded posts. Note that the y-axis is log scaled and therefore the longer text lengths appear proportional bigger, in reality, the majority of the post has a short text length.

The figure shows that awarded posts appear to be longer. The average text length of an awarded post is 794.61 characters and 137.6 for posts without awards. One possible explanation for this relationship could be that larger content contains more detailed and nuanced information. In relation to wallstreetbets, awarded content could contain more detailed information about e.g. a stock, which could be valuable to other users and is therefore acknowledged with awards. Examples of texts can be seen in Appendix A.1.

5.4.3 ELECTRA Embeddings

The transformer model ELECTRA explained in Section 4.5.1 can be used to extract information from the text. ELECTRA transforms the raw text into classification sequence tokens (CLS tokens), which retain information used for classification tasks. Using ELECTRA we extract the CLS tokens from different posts. ELECTRA returns 256 CLS tokens for a single text input.

Using the 256 outputs of ELECTRA we reduced the number of dimensions by using PCA 4.5.4. This allows us to visualize the CLS-tokens. The following figure shows the two first principal components of the CLS tokens from posts randomly sampled from wallstreetbets. The dots are colored by awarded and not awarded posts.

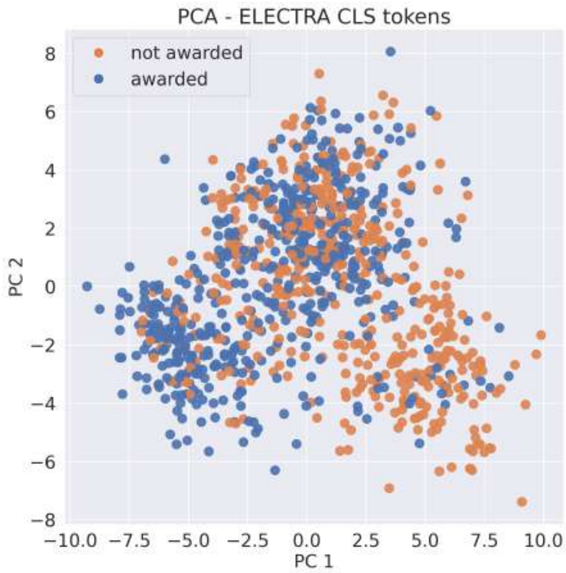
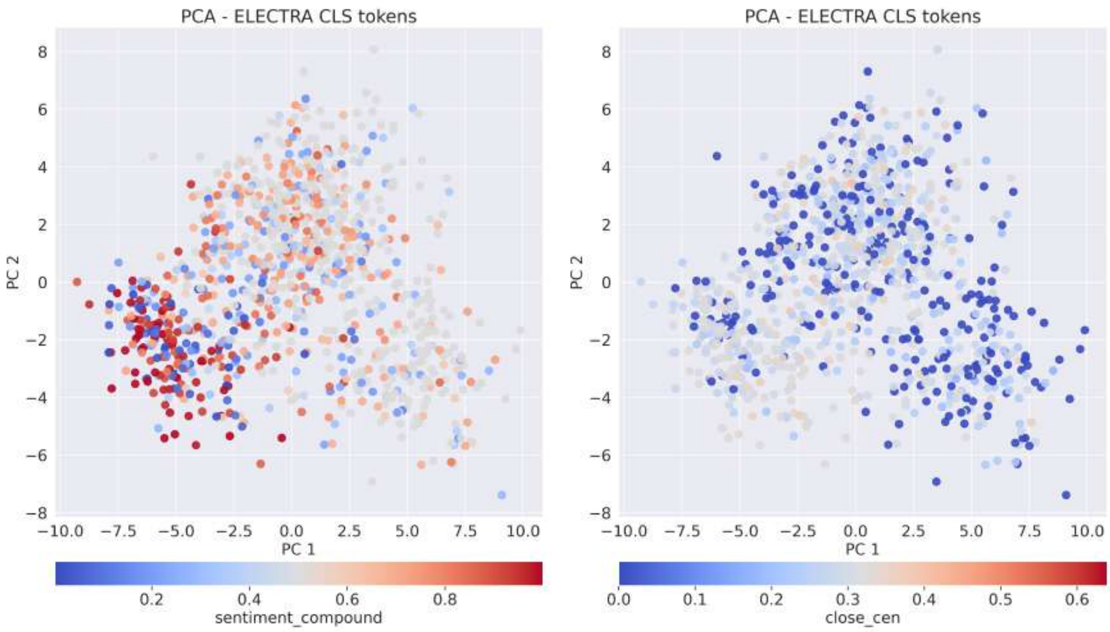


Figure 5.16: Awarded and not awarded posts from wallsteetbets

Figure 5.17: Principal components of CLS tokens colored by awarded and not awarded posts from Wallstreetbets.

It can be seen in Figure 5.17 that some of the ELECTRA CLS-tokens from awarded posts tend to cluster together. This shows that the ELECTRA model is able to extract valuable information from the raw text. However, it's important to note that using the ELECTRA model does have its limitations. One of the key challenges is that the model is a black box, which means that we cannot interpret the behavior of the model.

Furthermore, we examined the relationship between other features and the CLS-tokens. The following figure shows an example of the sentiment compound and the closeness centrality.



(a) VADER sentiment compound

(b) Closeness Centrality

Figure 5.18: Principal components of CLS tokens colored by the sentiment compound of Figure (a) and the closeness centrality on Figure (b)

It can be seen in Figure 5.18 (a) that ELECTRA captures similar information from the text, as the VADER sentiment analysis does. A cluster of posts with a positive sentiment can be seen on the left and neutral posts can be seen on the right.

The posts have been colored by the closeness centrality in Figure 5.18 (b), where it appear to be no clusters. This indicates that there is no correlation between how users express themself in text and how they react in the network. This can also be seen in similar figures with the other features in Appendix A.2.

5.5 Correlations

In order to gain an overview of the in-between correlation of the dependent variable awarded, but also the correlation in-between the features a correlation matrix has been computed using the Pearson correlation (Section 4.4.1).

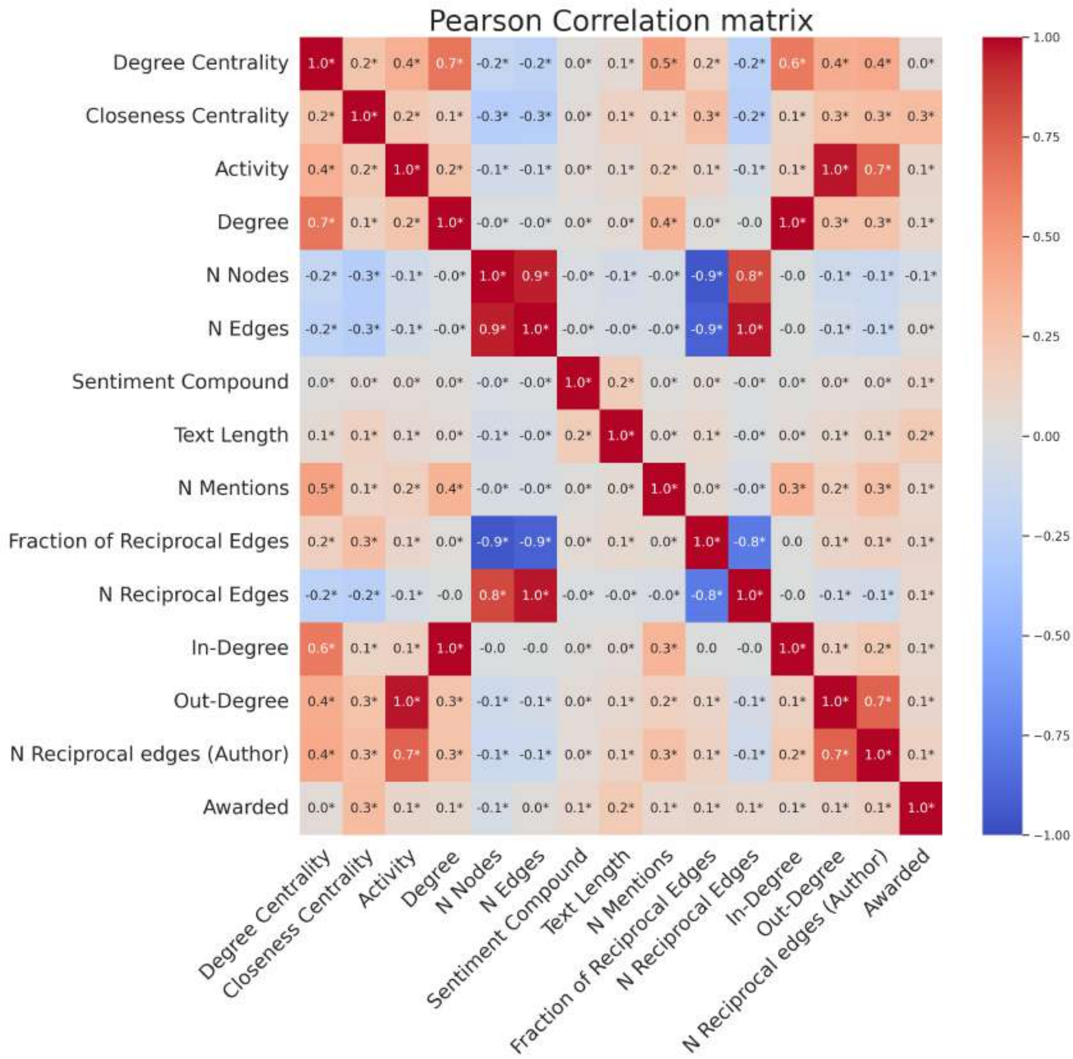


Figure 5.19: The figure shows the Pearson correlation matrix 4.4.1 for each feature. The correlation is colored and the value is shown. Correlations with a p-value below the significance level of 0.05 has been denoted with a "$*$".

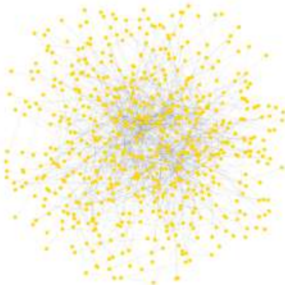
Closeness centrality and text length are the strongest correlated features for the dependent variable awarded with a Pearson correlation of 0.3 and 0.2. However, this is still a relatively weak correlation. Additionally, the matrix shows a strong correlation between some of the independent variables, such as the fraction of reciprocal edges and the number of reciprocal edges or the number of nodes and edges. This is to be expected, given the inherent relationships among these features. It is important to note that the Pearson correlation is a measure of a linear relationship, and can therefore not capture a non-linear relationship.

5.6 Other Subreddits

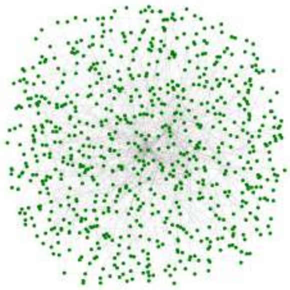
The following section presents an examination of the smaller subreddits that were excluded from the predictive model. Our original intention was to analyze a broad range of smaller subreddits and use them for explaining and predicting reputation. However, this proved to be unfeasible due to the limited number of awards given in smaller subreddits. This section provides a brief overview of the subreddits.

5.6.1 Network Analysis

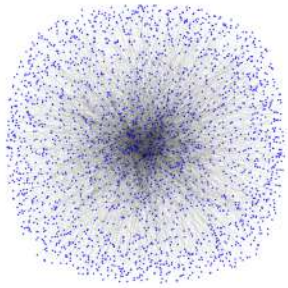
For each subreddit (Section 3.2) a network has been created using the method explained in Section 4.2.2. The networks are colored by the respective subreddit color 3.1 and plotted using the Fruchterman-Reingold force-directed algorithm (spring layout) that treats edges as springs holding nodes close while treating nodes as repelling objects. By doing this high degree nodes tend to cluster together.



(a) DTU

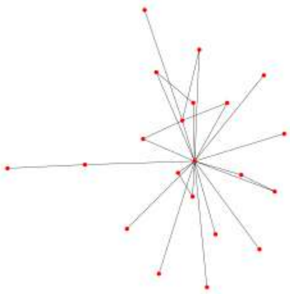


(b) Jazznoir

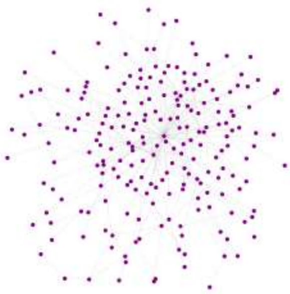


(c) Indoorbouldering

(d) Trump666



(e) FourSentenceStories



(f) Kiwi_bird

(g) GraphTheory

Figure 5.20: The figure shows the biggest coherent component from each subreddit.

The following table shows an overview of the networks of different subreddits.

subreddit	period	A	N	E	$\overline{W}$	W_μ	$\overline{D}$	D_μ	$\overline{C}$	$\overline{A}$	$\bar{t}$	R	$R_{\%}$	P_C
FourSentenceStories	2021-05-04 - 2021-12-28	200	34	28	2.679	1.0	1.647	1.0	0.275	5.824	19	15	0.429	0.023
Trump666	2020-03-16 - 2022-01-01	6746	755	1628	2.653	1.0	4.313	2.0	0.203	7.910	53	1048	0.493	0.000
GraphTheory	2015-02-17 - 2021-12-26	497	247	163	1.718	1.0	1.320	1.0	0.016	2.012	59	124	0.564	0.286
jazznoir	2015-01-01 - 2022-01-01	4725	1497	1139	1.474	1.0	1.522	1.0	0.010	2.986	122	640	0.453	0.016
indoorbouldering	2017-08-23 - 2022-01-01	10532	2529	4866	1.665	1.0	3.848	2.0	0.062	3.967	112	3487	0.532	0.000
DTU	2015-05-15 - 2022-01-01	2788	689	1196	1.805	1.0	3.472	2.0	0.064	4.046	136	1014	0.600	0.000
kiwi_bird	2015-02-21 - 2021-12-27	862	302	301	1.595	1.0	1.993	1.0	0.048	2.778	65	128	0.358	0.085

Table 5.3: A (Activity) is the total sum of posts and comments on the subreddit. $\overline{A}$ is the average activity across all authors. N is the number of nodes/authors. E is the number of edges. $\overline{W}$ is the average weight of the edges. $\overline{\Delta t}$ is the average time period of the users. R is the number of reciprocal edges and R_P is the proportion of reciprocal edges (Equation 4.3). $\overline{D}$ is the average degree. P_C is the p-value of the average cluster coefficient being different from the double-edged swapped graphs (Equation 4.2).

5.6.2 Textual Analysis

This section shows the result of the textual analysis of the different subreddits and that each subreddit uses a distinct way of communicating.

Using the VADER 4.3.1 sentiment analysis we computed the sentiment compound of content from different subreddits. Furthermore, completely neutral sentiment scores of 0 have been removed for better interpretation and readability. The following figure shows the distribution of each subreddit.

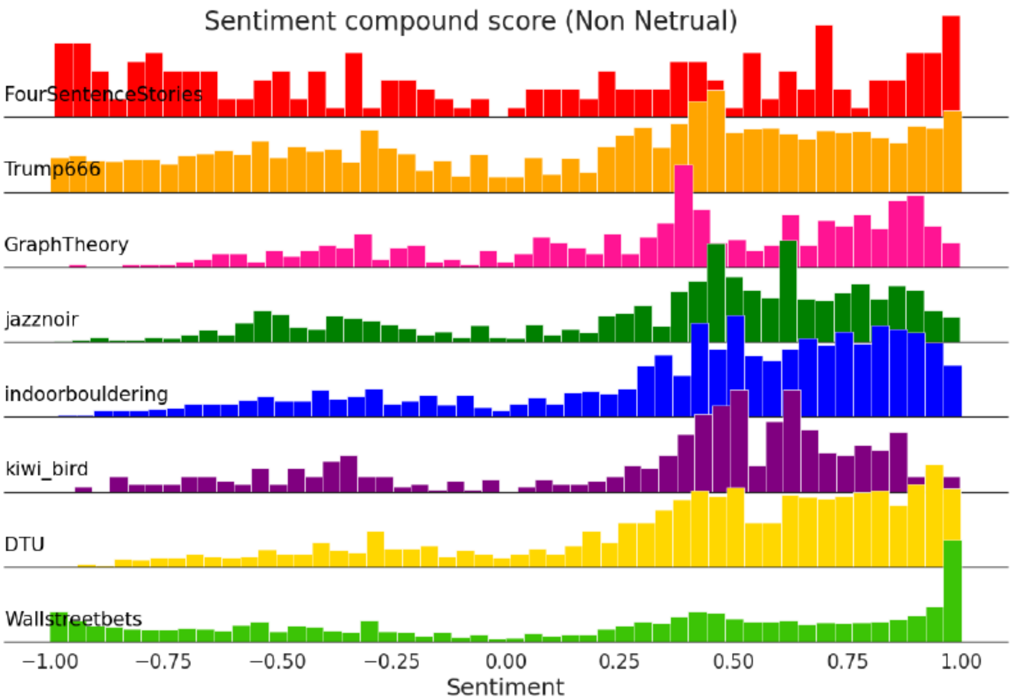


Figure 5.21: The figure shows the sentiment compound from the VADER algorithm for each subreddit. It shows the distribution, with the y-axis being the density for each subreddit

It can be seen that each subreddit varies in sentiment. The subreddits FourSentenceStories and Trump666 appear to use a more negative tone compared to other subreddits.

It can also be seen in Figure 5.21 that many of the subreddits appear to follow a bimodal distribution.

Additionally, we computed the ELECTRA CLS-tokens for posts from the different subreddits. Each subreddit is sampled an equal amount of times. The dimensionality of the CLS-tokens is reduced by using PCA. The following figure shows the result of this.

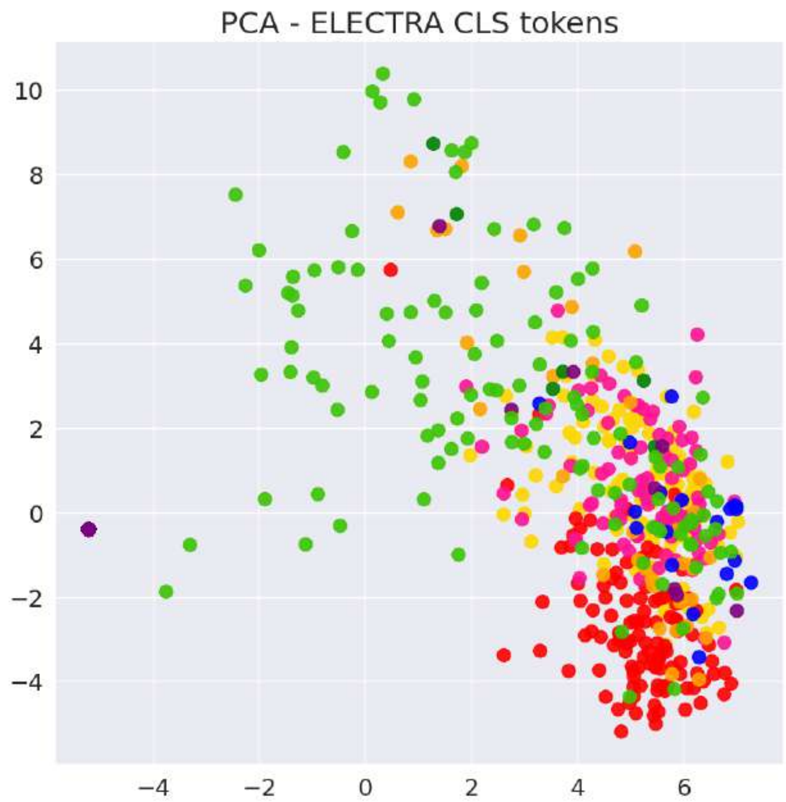


Figure 5.22: The principal components of the ELECTRA CLS-tokens colored by the different subreddits

The figure shows that some of the principal components from the different subreddits tend to cluster together. This indicates that subreddits use distinct ways of communicating. This can also be seen in Figure 5.21, which shows a variation in the tone used in posts.

CHAPTER 6

Results - Explaining & Predicting Awarded Posts

This chapter contains a predictive analysis of how an individual obtains an award. In Section 6.1 we present the predictive model used to classify if a post gets awarded, using the network and textual features. In Section 6.2 we explored the importance of identity and how the social dynamics change for a larger network. Additionally, in Section 6.3 we show how the probability of receiving an award increases if the user is already awarded.

6.1 Predicting Awarded Posts

The aim of the predictive model is to classify if a post gets an award. The dependent variable of the model is therefore the nominal variable `awarded`. The following section shows the results of the different models used to classify if a post is awarded or not. Due to the lack of awards given in other subreddits, only data from wallstreetbets is included in the predictive model.

6.1.1 The Data

The dataset used for the model is based on discrete-time dynamic graphs taken in the interval from before a post was created. This creates a static snapshot of the graph. The time interval of the snapshot is 10 weeks. For each post/network, different features have been extracted.

Feature	Description
N edges	The number of edges in the graph
N nodes	The number of nodes in the graph
text	The title and text of the post concatenated
text length	The length of the text
sentiment compound	The VADER sentiment compound of the text (Section 4.3.1)
degree	The degree of the author
degree centrality	The degree centrality of the author (Section 4.2.5)
closeness centrality	The closeness centrality of the node(Section 4.2.5
in-degree	The in-degree of the author in the directed graph
out-degree	The out-degree of the author in the directed graph
Fraction of Reciprocal Edges	The fraction of reciprocal edges of the whole network (Section 4.3)
N Reciprocal Edges (author)	The sum of reciprocal edges of the author (Section 4.3)
N Reciprocal Edges	The total sum of reciprocal edges (Section 4.3)
activity	The number of posts & comments in the time interval by the author
mentions	The number of times the author has been mentioned in other posts & comments.

Table 6.1: Independent variables.

The original data is imbalanced with more posts being not awarded. This is corrected by downsampling the majority class not awarded posts, such that the models do not overfit the majority class.

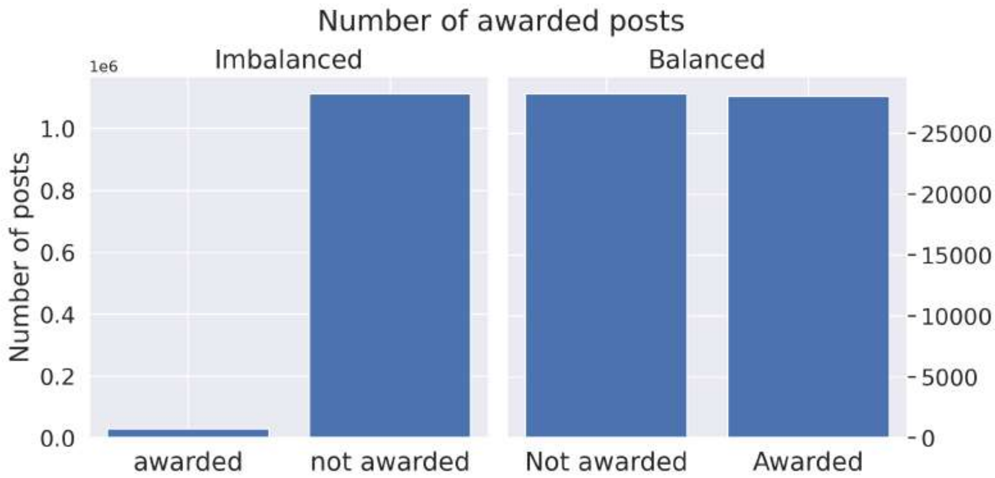


Figure 6.1: The data has been balanced by downsampling the majority class. The final data contains 56278 points.

The data is furthermore normalized and divided into training, test, and evaluation dataset with a size of 80%, 10%, and 10% respectively.

To investigate if it is possible to classify awarded posts, we conducted a preliminary analysis of the features. The following figure shows the principal components of the features used to classify awarded posts, where each dot is colored by awarded content.

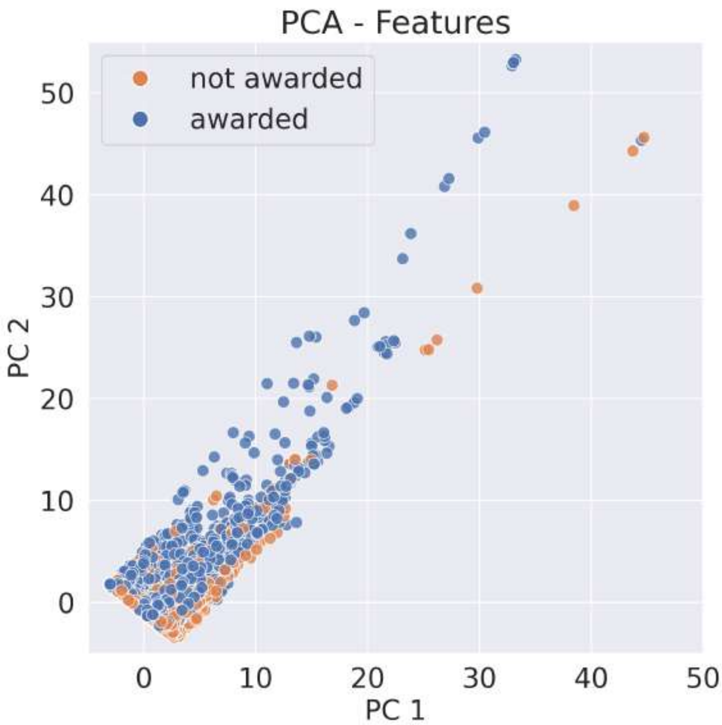


Figure 6.2: The first to principal components from the PCA using the features. This shows some clusters of awarded posts, which indicates that a classification task is possible.

6.1.2 Random Forrest

Due to the simplicity and interpretability random forest is a good baseline. Random forest obtained an F1 score of 74.82% of the test set. Since a Random Forrest model does not handle sequential data the raw text has been excluded. Instead features from the text have been extracted and the sentiment compound & the text length have been included in the model (see Table 6.1).

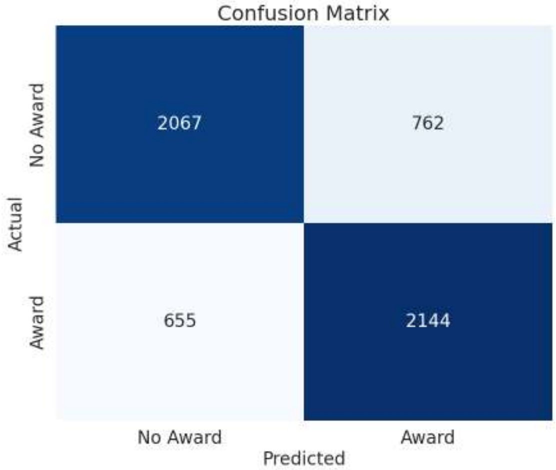


Figure 6.3: Confusion Matrix

	Precision	Recall	F1-score	Support
not awarded	0.7594	0.7306	0.7447	2829
awarded	0.7378	0.7660	0.7516	2799
accuracy			0.7482	5628
macro avg	0.7486	0.7483	0.7482	5628
weighted avg	0.7486	0.7482	0.7482	5628

Table 6.2

6.1.3 Neural network

The model uses a neural network which is further explained in Section 4.5.1. The model uses the same features as the Random Forrest model, where features have been extracted from the text. The model has been trained using the cluster computer HPC on DTU. The model has been optimized using a hyperparameter sweep which can be seen in appendix A.5.2. The model obtained an F1 score of 74.13%.

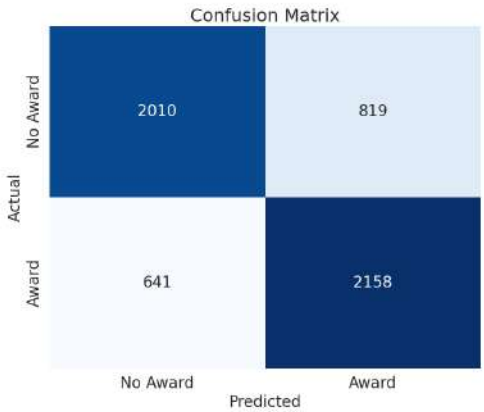


Figure 6.4: Confusion Matrix

	Precision	Recall	F1-score	Support
not awarded	0.7574	0.7144	0.7343	2829
awarded	0.7270	0.7688	0.7474	2799
accuracy			0.7415	5628
macro avg	0.7423	0.7416	0.7413	5628
weighted avg	0.7423	0.7415	0.7413	5628

Table 6.3

6.1.4 Neural Network + ELECTRA

The model is a combination of the transformer model ELECTRA and a neural network. The model is further described in Section 4.5.1. The model has been trained using the cluster computer HPC on DTU. The model has been optimized using a hyperparameter sweep which can be seen in appendix A.5.2. The model obtained an F1 score of 78.13%.

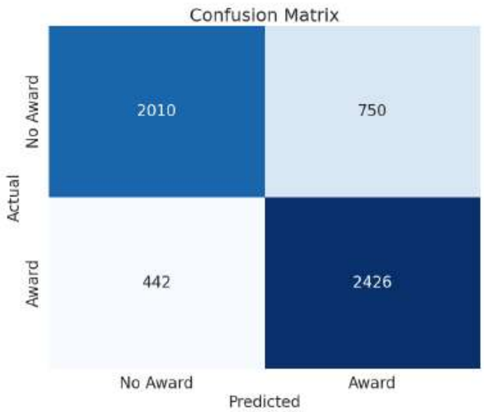


Figure 6.5: Confusion Matrix

	Precision	Recall	F1-score	Support
not awarded	0.8130	0.7217	0.7217	2829
awarded	0.7583	0.8403	0.7647	2799
accuracy			0.7821	5628
macro avg	0.7857	0.7810	0.7810	5628
weighted avg	0.7852	0.7822	0.7813	5628

Table 6.4

6.1.5 Overview

The following table shows an overview of the models used to classify awarded. Every model has been trained on the same training data and tested using the same test data. It shows the weighted average F1-score, Recall, and Precision.

Model	Features	F1-score (%)	Recall (%)	Precesion (%)
Random Forrest	Network Features + Extracted text features	74.82	74.82	74.83
Logistic Regression	Network Features + Extracted text features	70.47	70.52	70.69
K Nearest Neighbor	Network Features + Extracted text features	70.50	70.50	70.52
Nerual Network	Network Features + Extracted text features	74.13	74.15	74.23
Nerual Network + Electra model	Network Features + raw text	**78.13**	78.22	78.52

Table 6.5: Performance metrics (Section 4.4.2)

The neural network with the ELECTRA transformer model performed best with an F1-score of 78.13%. This model was the only one to use the raw text as input instead of extracted text features. This suggests that the extracted text features sentiment compound and the text length is not enough to capture all information retained in the text. By using the ELECTRA model we were able to capture more information and underlying patterns from the text that would normally not be captured by a traditional feature-based approach.

6.1.6 Feature Importance

To inspect the performance of the model we computed the permutation feature importance for the Random Forrest model. Due to computational limitations, it was not possible to compute the permutation importance of the Deep Learning ELECTRA Model, since it requires training numerous times. As explained in Section 4.5.3 the permutation importance is computed for each feature where the model is trained K times for each feature. The following figure shows the permutation feature importance using the F1-score as the metric s for the Random Forrest model.

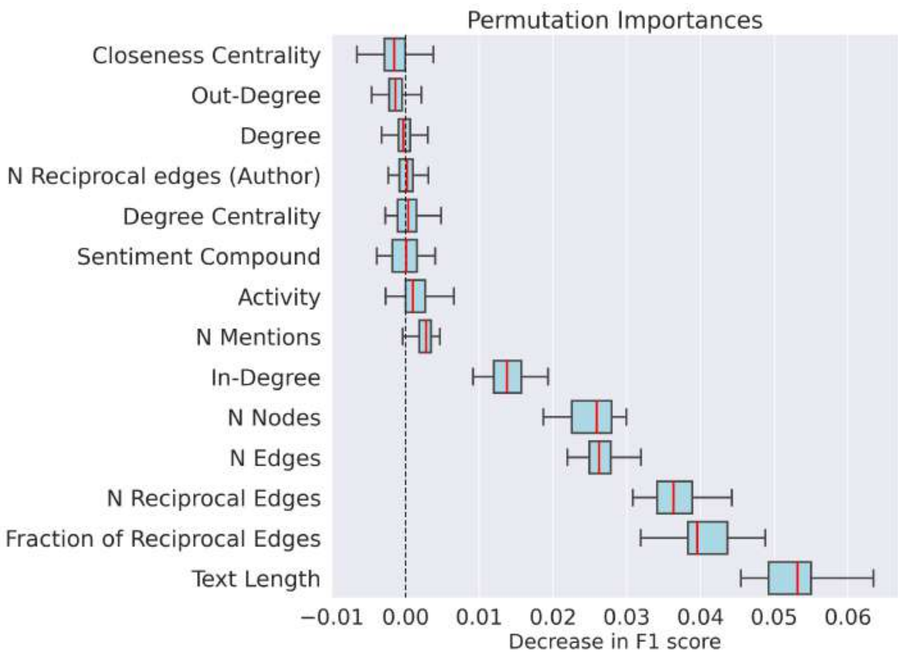


Figure 6.6: The Figure shows the decrease in F1 score of each feature, using the permutation feature importance for the random forest model

It is important to note that the permutation feature importance does not necessarily explain the intrinsic predictive value of a feature, but instead the predictive power of the features in the specific model the Random Forrest model. This means that permutation feature importance could be different for the Deep Learning ELECTRA model. However, it can be seen that when the text length is corrupted the F1 score decreases, this could indicate that users who write longer text are more likely to get an award, this suggests that they create informative text that contributes to the community. Additionally, network features such as the number of reciprocal edges, the fraction of reciprocal edges, and the in-degree and the number of edges and nodes also decrease the F1 Score when corrupted. This suggests that the network features also contribute to the explanation of individuals' reputations. This is further discussed in Section 7.2 & 7.3.

6.2 Individuals Identity in Communities

The following section shows the result of the study we conducted on the importance of identity in communities. As stated in Section 2.1 and in the paper *"Identity Effects in Social Media"* an individual's identity is a crucial part of a user's reputation within a community. For a user to have a true and accurate reputation, they must be well-known and recognizable within the community.

The subreddit wallstreetbets gained widespread popularity at the end of 2020, resulting in a large influx of new members joining the community. This section shows the result of the investigation conducted on how community size affects the identity of an individual. Using the discrete-time dynamic graphs (Section 5.3.1) we analyzed the relationship between various features and the size of the wallstreetbets networks. The summary of these results is furthermore discussed in Section 7.6

6.2.1 Recirpocal Edges

The total number of reciprocal edges explained in section 4.3, can be seen as a metric of how well the individual nodes are connected and how dense the community is. The following figures illustrate the relation between the total number of reciprocal edges and the network size.

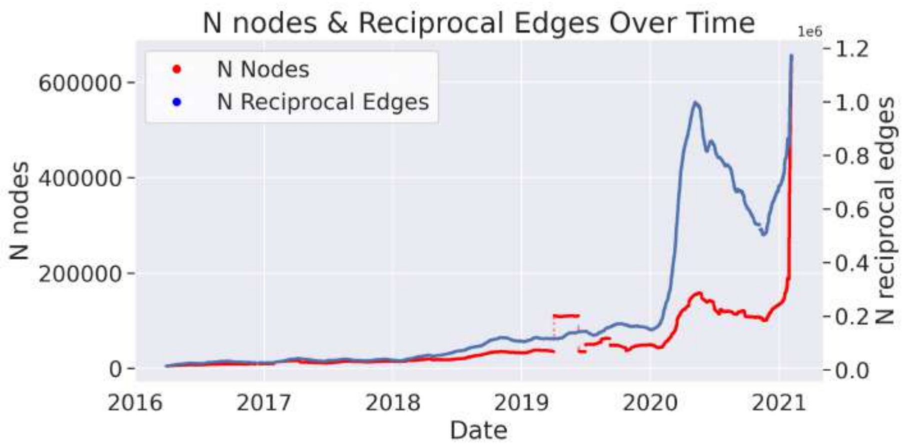


Figure 6.7: The total number of reciprocal edges and nodes by date. The red line shows the number of nodes for each network. The blue line shows the total number of reciprocal edges of the same network.

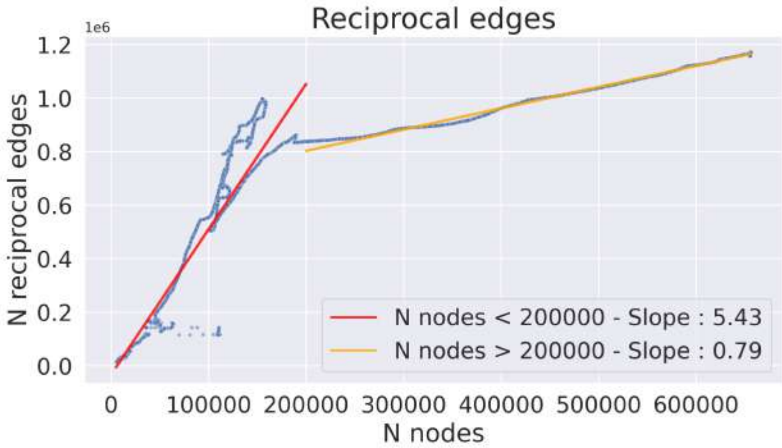


Figure 6.8: The total number of reciprocal edges and nodes. The lines show a linear regression for N nodes < 200000 & > 200000

We performed a linear regression for the networks with N nodes < 200000 & > 200000 It can be seen that the slope for $Nnodes < 200000$ is 5.43 meaning that when a node is added to the network the number of reciprocal edges is increased by 5.43. For $Nnodes > 200000$ the slope is 0.79 meaning that when a node is added the number of reciprocal edges is increased by 0.79. This suggests that when the community gets too big it losses its ability to form strong connections between nodes.

6.2.2 Mentions

Reddit uses a system where users can mention individuals in the text by writing "/u/" followed by a username. We computed the total number of mentions, which can be used as a measure of how well users know each other within the community. The sum of mentions is compared to the network size, N Nodes.

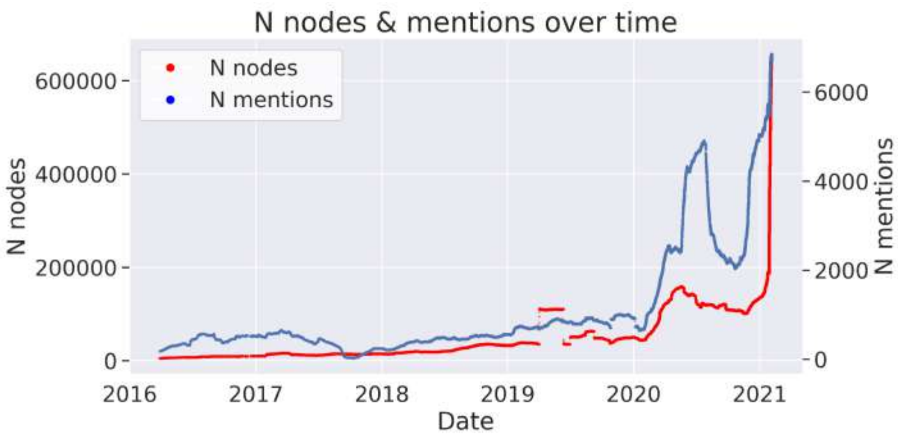


Figure 6.9: The total number of mentions and nodes by date. The red line shows the number of nodes for each network. The blue line shows the number of mentions of a user in the same period

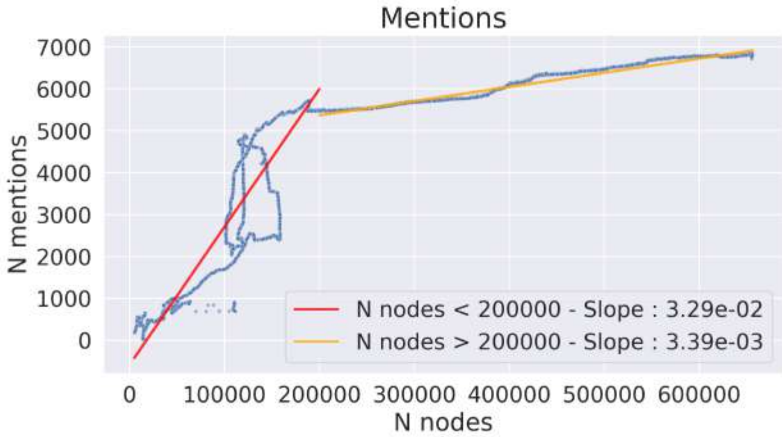


Figure 6.10: The total number of mentions and nodes. The lines show a linear regression for N nodes < 200000 & > 200000

We performed a linear regression for the networks with N nodes < 200000 & > 200000. It can be seen that the slope for $Nnodes < 200000$ is $3.29e^{-2}$ meaning that when a node is added to the network the number of mentions is increased by $3.29e^{-2}$. For $Nnodes > 200000$ the slope is $3.39e^{-3}$ meaning that when a node is added the number of reciprocal edges is increased by 3.39^{-3}. This illustrates that when the community reaches a certain size it becomes more difficult for individuals to establish strong connections and increase their mention count.

6.2.3 Permutation Importance

To further investigate how the community changes based on the network size, two random forest models 4.5.2 have been computed using data from different time periods. Additionally, the permutation importance has been computed for both models. 4.4.2. The data has been divided such that each set contains the same amount of awarded posts, additionally, each dataset has been balanced. The first set is from the start of the period when the network was smaller and the second is from the end of the period. The random forest model has been used due to computational simplicity.

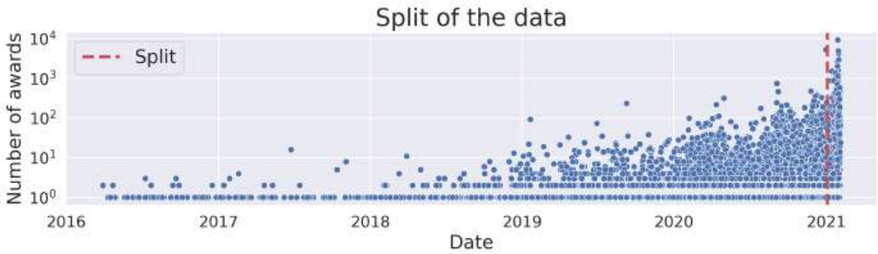


Figure 6.11: The red line indicates the point in time the data is divided. It can be seen that the majority of the awarded post is at the end of the period, hence splitting the data towards the end such that the two sets are the same size.

Each dataset contains 24876 datapoints with a 20% test set each. The data is divided from 2016-03-28 to 2021-01-03 & 2021-01-03 to 2021-02-04, they obtained a F1 score of 0.75 & 0.74 respectively.

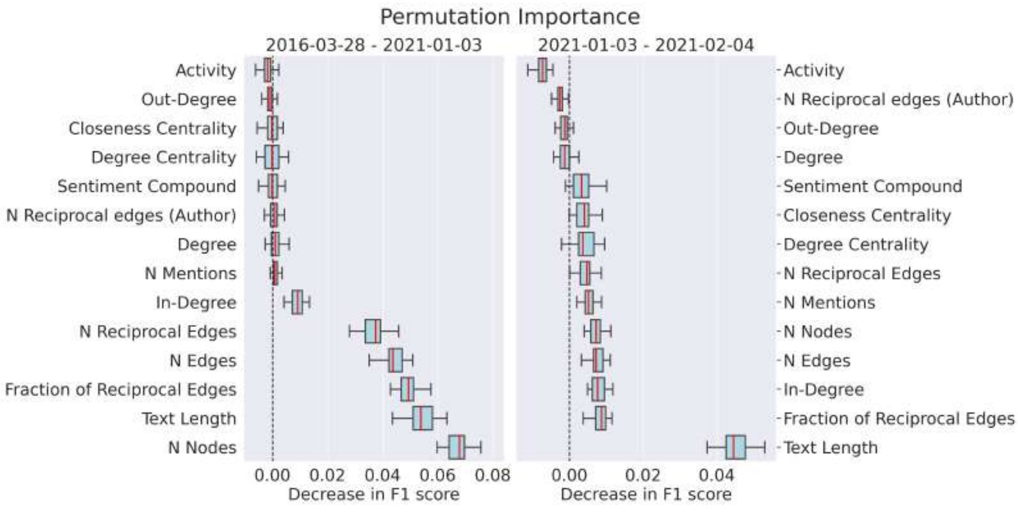


Figure 6.12: Permation Importance for the early and late period of wallstreetbets. write more here

It can be seen that features (6.1) such as the fraction of reciprocal edges, number of nodes, number of edges, and the number of reciprocal edges contribute to higher accuracy in the model for the early period. While these features do not contribute so much to the model for the late period. Additionally, the text length decreases the accuracy for both models when it is corrupted. The results of the permutation importances suggest that network features have a larger predictive power in smaller communities since primarily only the text length appears to have an effect on the later model.

6.3 Awards Increases Awards

This section seeks to investigate how the probability of receiving an award increases if the user is already awarded. As previously mentioned, awards are a rare commodity, with only 4% of the authors from wallstreetbets receiving an award. These awards are expected to be given to users that have produced high-quality, informative, entertaining, or otherwise valuable content for the community. However, there are indications that individuals receive more awards if they are previously awarded. This could suggest a ”rich get richer” effect [33], which refers to the phenomenon where those who are already successful are more likely to continue to be successful. This concept is often used in the context of economics, but can also be applied to social systems such as awards on Reddit communities. We computed the percentage of awards given to the top n% awarded users of the awarded users. The following figure shows the cumulative rate of the total awards given to the top awarded users.

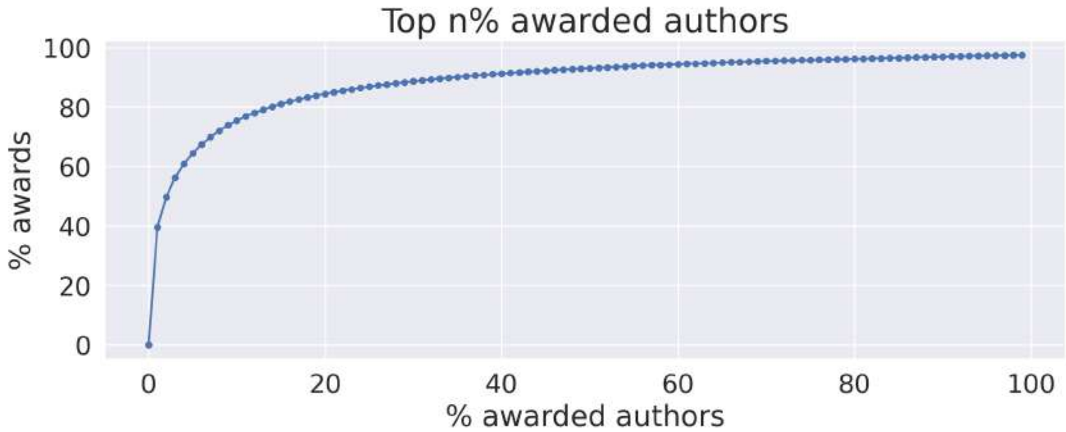


Figure 6.13: The cumulative distribution of the number of awards given to the awarded users.

It can be seen that the majority of awards are given to a small proportion of the awarded users. The top 1% awarded users have received 40% of the total awards. This suggests that the award system follows the ”rich get richer” phenomenon. Users who are already well-known or have established a reputation are more likely to receive awards, which enhances their visibility and popularity, leading to even more awards. This creates a feedback loop where the rich get richer and the poor get poorer.

To further investigate the ”rich get richer” effect we computed the probability of receiving an award prior to the individual receiving K awards. The conditional probability for receiving an award is computed for $P(A|K = 0), P(A|K = 1)...P(A|K = n)$. This is done by considering the sequence of awards received by a user e.g. `Not awarded, Awarded, Not awarded, Awarded` and then computing the cumulative number of awards in the sequence e.g. $0, 1, 1, 2$. The probability prior to the amount previously awarded post is then computed for each K (awarded post) by counting the number of occurrences of K in the sequence. The mean of conditional probability for each K is then computed for multiple individuals. Furthermore, this can be compared with a null model where the sequence of awards is randomized. This is done by randomly sampling the value awarded from users with the same activity level (number of posts) and from the same date. For each model, a linear regression line is fitted to data and the 95% confidence interval for that regression

The following figure shows the conditional probability for each K for the true model and the null model.

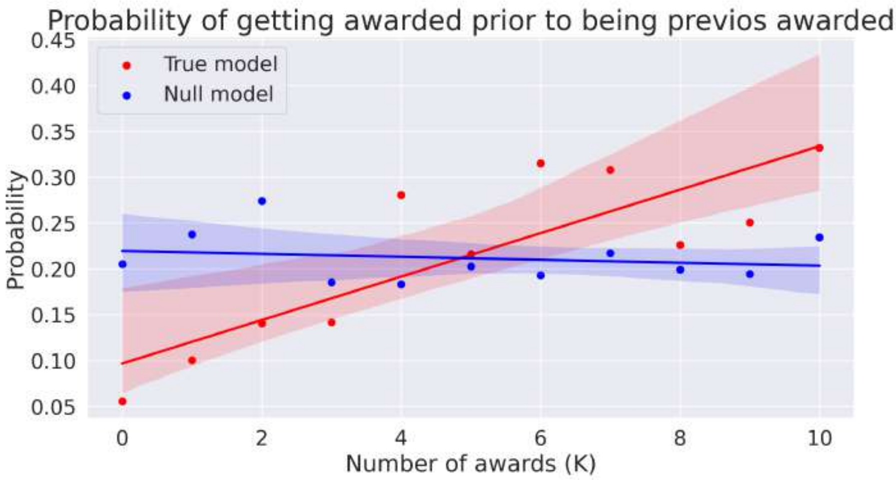


Figure 6.14: The conditional probability prior to being previously awarded K times. The red line shows the true model and the blue line shows the null model where the sequence of awards is randomized. The error bars indicate the 95% confidence intervals of the regression lines.

It can be seen that for the true model the probability $P(A|K)$ increases for larger K, while for the null model, the probability appears flat. This suggests that the award system follows the ”rich-get-rich” phenomenon where it is more likely for already awarded users to receive additional awards.

6.3.1 Effect of the network size

To investigate how this concept is affected by the community size, we computed the % of awards given to the top n% awarded individuals for different periods of the data. As previously stated wallstreetbets had a large increase in members at the end of 2020. We divided the data into 4 equal parts and computed the cumulative distribution of the number of awards given to the awarded users in each time period.

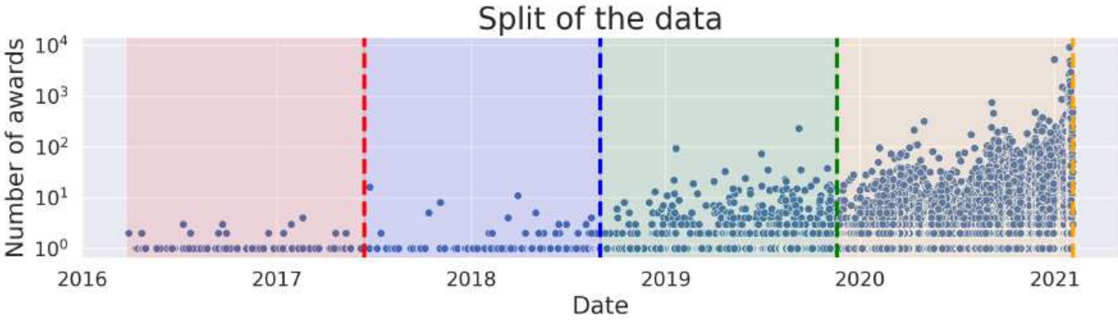


Figure 6.15: The Figure shows the data divided into 4 equal parts by date, where each part is colored

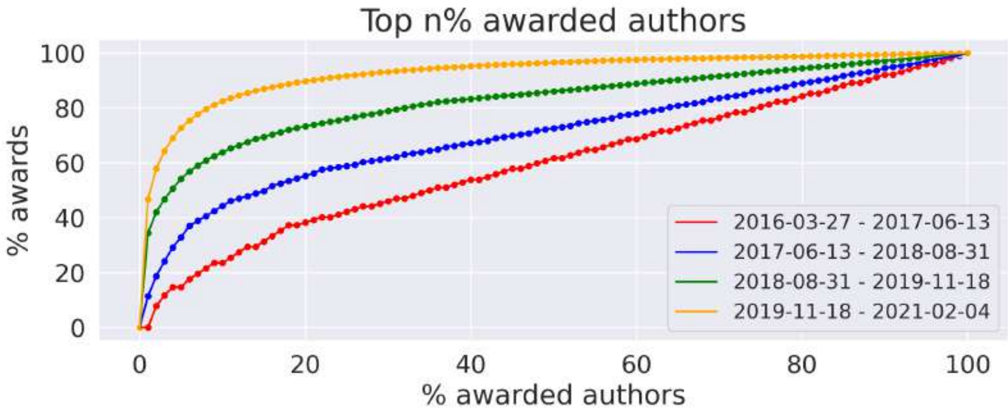


Figure 6.16: The cumulative distribution of the number of awards given to the awarded users for each different time period. Each color denotes the time period shown in Figure 6.15

Figure 6.16 shows that the effect increases in the later periods when the community gets bigger. The following table shows the % of the total awards for each period, received by the top 1 % awarded individuals in that period.

Period	% awards received by the top 1%
2016-03-27 - 2017-06-13	4%
2017-06-13 - 2018-08-31	11%
2018-08-31 - 2019-11-18	35%
2019-11-18 - 2021-02-04	47%

Table 6.6: % of the total awards for each period received by the top 1 % awarded individuals in that period.

The increase in this phenomenon could be due to many different factors. The self-reinforcing feedback loop, where individuals who receive more awards become more visible and influential, could potentially be bigger in larger communities. Another factor could be that as the community grows larger, the competition for reputation can become more intense, making it more difficult for new or less well-known individuals to gain a reputation. This is further discussed in Section 7.5.

CHAPTER 7

Discussion

This chapter contains a discussion of the results. The results presented in chapters 5 & 6 reveal that the hypotheses stated in Section 1.4 can be supported. Hypothesis (I) is discussed in Section 7.1 and hypothesis (II) is discussed in Sections 7.2, 7.3 and 7.4. Additionally, hypothesis (III) is discussed in 7.5. Finally, Section 7.6 discusses hypothesis (IV).

7.1 Awards as a Metric of Reputation & Acknowledgment

In Section 5.1 we investigated the use of awards as a metric of acknowledgment and recognition of individuals on wallstreetbets. The following section discusses the key results.

Users with a know high reputation, appear to have more awards compared to the population means, which can be seen in Table 5.1. This suggests that these highly reputable users have been recognized through awards for their contributions to the community. It can be assumed that these users are in fact highly reputed, by the existence of papers and articles about them, since it is not common for every Reddit user to have articles written about them [28] [32] [29] [30].

Figure 5.3 shows that the majority of posts on Reddit do not receive an award. The scarcity of awards helps to maintain the value and prestige of the award system, as it demonstrates that only the best content is being recognized. This can motivate users to continue creating high-quality content, leading to a self-reinforcing cycle of improvement and recognition. This sacristy is also an issue since the community needs a big enough size for users to receive awards, this can be seen in Figure 5.2 which shows that the subreddit wallstreetbets has a substantially large amount of awards relative to other subreddits. Initially, we tried to investigate and focus on smaller subreddits since it is expected to have a better community as described in Section 7.6 & 6.2, but due to the scarcity of awards, it was hard to investigate and find a measurement for acknowledgment in these datasets.

One of the key limitations of the analysis is the collection of the data. The dataset of awards was collected using the python library `PRAW`, which collects the data from Reddit's own API. This means that the number of awards for each post is computed at the time of the scrapping. Ideally, this should be a fixed factor where the number of awards is collected at a specified period after the post where created. This is unfortunately not possible since Reddit does not save the history of awards for each post. The library `PSAW` is also not preferable since it collects data from posts at an unknown point in time and then saves it in `PSAW`'s own database.

In Section 1.2 the "karma" system on Reddit is described as an average of scores across all content of a user. While it could potentially be used as a metric, it is not considered, since it can be earned across a wide range of subreddits, making it difficult to evaluate within a single community.

7.2 Network Science

We analyzed the network of wallstreetbets using discrete-time dynamic graphs. When using discrete-time dynamic graphs we analyzed the 10 weeks before the creation of a post. Social dynamics often change and are very volatile, so by analyzing the network before the point in time of the post we are predicting, we can get a more accurate representation of the social dynamics captured by the network features. If we used the entire period of the data, the network and the network features would capture dynamics that would not be relevant to the individual we are trying to predict. Additionally, using the entire period would result in over-representing the users who had a consistently high level of activity throughout the entire duration, leading to an inaccurate portrayal of the network's true dynamics. Another advantage of discrete-time dynamic graphs is their computational simplicity. It was proven to be easier to divide the data into smaller chunks of data, instead of computing the features for a network created by the entire period.

Centrality measures such as the degree or closeness centrality are often used to quantify the relevance of individuals in a social network. However, a visual interpretation of the distribution of the centrality measures for awarded and not awarded posts (Figure 5.9) does not suggest that these measures are a significant good representation of individuals' importance. This can also be seen in Figure 6.6 which shows the permutation feature importance of the Random Forrest, it can be seen that the corruption of the closeness nor betweenness centrality affects the performance of the random model. This is of course assuming that awards are a true measure of acknowledgment. Due to computational limitations, other centrality measures such as the betweenness and eigenvector centrality are not computed, these measures could potentially show a more promising understanding of the reputation of an individual.

Other features did however show a promising potential of explaining how individuals receive an award. The visual interpretation of the distributions of the in-degree for awarded and non-awarded posts (Figure A.29) and the permutation importance from the random forest model (Figure 6.6) suggests that it is more likely to be awarded if you have a high in-degree. A similar trend can be seen in the distribution of the number of mentions of the individual (Figure 5.13), which also indicates that highly mentioned individuals have a higher likelihood of being awarded. The results of this indicate that awarded individuals have an established identity within the community.

7.3 Text

In Section 5.4 we showed the results from the analysis of the textual content from the post. This included the sentiment and the text length, which were extracted from the text. We also investigated the embeddings from the ELECTRA model.

The sentiment analysis was conducted using the VADER algorithm. The results in Figure 5.14 showed that awarded posts appeared more negative and positive, compared to not awarded posts. This indicates that awarded post is more polarizing and that the content might be more subjective, which tends to evoke stronger emotional responses. It is also important to consider that the VADER algorithm may not accurately capture the nuances and complexities of human language, particularly when it comes to sarcasm, irony, and other forms of figurative language.

We also extracted the text length from the posts. The result seen in Figure 5.15 shows that awarded posts appear to be longer. This indicates that awarded post contains more detailed information that is valuable to the community. This could be detailed information about stocks or investment ideas. It is possible that this pattern only is true for wallstreetbets, as other subreddits cover different topics, where the same level of detail is not appreciated.

The textual content can potentially retain a large amount of information and methods such as text length and sentiment analysis may not extract sufficient information, therefore we used the ELECTRA transformer model. Using the outputs from ELECTRA, we visualized the 256 CLS-tokens, by reducing the dimensions using PCA (see Section 5.4.3). The result in Figure 5.18 shows the Principal components colored by awarded post, where it can be seen that ELECTRA explains some of the variations of the awarded post. This applies that the textual features retain predictive power. Additionally, the overview of the models 6.5 shows that ELECTRA contributes to higher accuracy when comparing the two deep learning models with the extracted features and with the ELECTRA.

The limitation of using ELECTRA is the lack of explainability of textual features. Since ELECTRA contains more than $10e^6$ parameters it is hard to explain the behavior of the model.

7.4 Predicting Reputation

As previously mentioned the initial idea was to investigate many smaller subreddits, which should be included in the final model, such that reputation is more generally explained instead of primarily within wallstreetbets. However, due to the lack of awards in other subreddits only data from wallstreetbets were included in the development of the final model. The following sections address the challenges related to the final predictive model.

7.4.1 Data Bias

As seen in Section 3.2.8 wallstreetbets gained a large influx of users toward the end of 2021. This results in more people being able to award other users, which could affect the likelihood of being awarded based on the period the individual made a post. The dataset used for the models is balanced such that it contains the same amount of awarded and not awarded posts. However, the percentage of awarded posts by date which can be seen in Appendix A.30 shows that the dataset is imbalanced by the date. The figure shows a higher percentage of awarded posts in the later periods of the data. Network features such as the number of nodes and edges might be biased toward the end since there are more nodes in this period. The figure in Appendix A.24 showing the number of nodes for the awarded and not awarded posts does not however show any clear correlations. It is also important to note that the models capture the changes in social dynamics that occur when the network size increases and are able to predict these changes with an F1 score of 78.13%.

7.4.2 The Deep Learning Models

The ELECTRA transformer model uses attention-mechanism, described in 4.5.1, which captures the contextual relationship of the input and devotes more ”attention” to smaller specific parts of the input. The transformer model has been used to obtain information retained in the text in the form of CLS-tokens, these tokens have been combined with the network features to recognize the underlying relationship between network features and CLS-tokens (see Section 4.5.1). However, the model does not capture any relations between network features and raw text input. Ideally, an attention mechanism should be applied to a combination of network features and text as the input. By doing this it could potentially find the contextual relationships between a network's features and specific words.

The attention mechanism described by Equation 4.12 in Section 4.5.1, does however not handle continuous variables (network features) as input but uses a vector representing, where each word/input is represented as the index in the vector. A proposal for this could be to use Graph Neural Networks (GNN) as input instead of the network features. This is further discussed in Section 7.7. This would require training and developing a model from scratch, which would require a substantial amount of data and computational power, which was not possible within the constraints of this project.

7.4.3 Training the Deep Learning Models

Both deep learning models have been trained successfully, which can be seen in Appendix A.5.1 & A.5.1. The figures show that the training and validation loss lowers and converges. The training and validation accuracy also increases and converges after enough epochs. To obtain the optimal hyperparameters `wandb` have been used to "sweep" all the hyperparameters through a Bayes search that finds the optimum faster than a random grid search. More details on this method can be found in Section 4.1.2. The results from the search can be seen in Appendix A.5.2, which shows that both models performed better with smaller batch sizes and more epochs.

Through the development of the models, many different versions have been tested. A code snippet for the final models can be seen in Appendix B.1 & B.2.

7.4.4 Excluded Features

Some network features have not been included in the data used for predicting reputation, this is due to the computational complexity. The betweenness centrality algorithm has a relatively high running time for big networks compared to other centrality measures since it needs to compute the total number of shortest paths (Section 4.2.5). The eigenvector centrality is also computationally complex since it computes the eigenvector for the adjacency matrix A, which has the same dimension as the number of nodes (Section 4.2.5). However, it is to be expected that the centrality measures are highly correlated with each other.

7.5 Rich Get Richer Effect

Section 6.3 shows how the probability of receiving an award increases prior to the more awards you have, which can be seen in Figure 6.14. Figure 6.13 also shows that the top 1 % of the most awarded individuals received 40 % of all the awards. This could indicate a "rich get richer" effect [33], where successful users are more likely to continue to be successful. The concept of the "rich get richer" is often used within economics. However, the same concept can be applied to other systems. In social sciences, capital is often divided into 4 distinct forms: economic, symbolic, cultural, and social [34]. Where symbolic capital is referred to as honor, prestige, or recognition. Reddits' award system provides a unique opportunity to measure the symbolic capital of an individual quantitatively. We used this to study the "rich get richer" effect of awards, which suggests that symbolic capital in the form of awards behaves similarly to economic capital.

This phenomenon is also seen within network science. This is known as a scale-free network where the distribution of degrees follows a power law meaning that a few nodes have a very high number of connections, while most nodes have only a few connections [14]. A similarity can be seen between the concept of scale-free networks and the distribution of awards. The degree can be seen as a measure of social capital, while the award system can be seen as a measure of symbolic capital.

This concept was further investigated for different network sizes of wallstreetbets. The inequality of the distribution of awards increases when the Wallstreetbets network size increases (Figure 6.16). This indicates that there are different mechanisms driving the award systems for larger networks. The results suggest that when the network is small every user that creates valuable content gets acknowledged through awards. For bigger networks, it's primarily individuals with an established reputation that gets acknowledged for their work. For larger networks, reputed individuals who receive more awards become more visible and influential leading to a self-reinforcing feedback loop, where they keep receiving more awards. When the network is too extensive it also becomes harder for less well-known individuals to establish an identity within the community and gain a reputation. This can also be seen in Section 6.2 where the social dynamics of larger networks change and it becomes difficult for individuals to establish strong connections. This can be seen in Figure 6.10 and Figure 6.8 which show the decrease in mentions and reciprocal edges for larger networks.

However, it is important to note that the phenomenon is not necessarily caused by a ”rich get richer” effect, but it could be that a few users simply produce high-quality content and are therefore recognized more by the community through awards.

When we conducted the experiments we purposely did not control for the activity level of each individual, since it mimics real-world economics. Individuals can ”work hard” and accumulate more wealth or success in the ”real world”, which would cause inequality in the distribution of wealth. However, the distribution should not be that unequally distributed, despite some people being highly active. We adapted this idea to the experiment, such that it mimics the dynamics of economics.

7.6 Identity Importance for Reputation in Communities

In Section 6.2 we investigated how the individuals' identity is affected by the size of the network. This was done by examining the total number of mentions and the total number of reciprocal edges in the network.

The two slopes for the regression computed between the total number of nodes and reciprocal edges, which can be seen in Figure 6.8, indicate that it is harder to form stronger connections in bigger networks. It may be that in smaller networks, individuals are more likely to form connections with others who have similar interests or goals, while in larger networks, connections may be driven more by convenience or proximity.

The regression computed for the total number of nodes and mentions, which can be seen in Figure 6.10 also indicates that there are different mechanisms driving the social dynamics of bigger networks. It suggests that in large networks, it becomes increasingly challenging for new nodes to establish a strong presence and build their reputation. This could potentially be because smaller communities are stronger and the identity of individuals is lost when a network reaches a certain size. It becomes harder for individuals to establish an identity within the community since there are too many people.

Furthermore, we also computed the permutation importance of a Random Forrest model (Figure 6.12) which indicates similar findings. Where features such as the fraction of reciprocal edges, number of nodes, number of edges, and the number of reciprocal edges contribute to higher accuracy in the model for the early period with smaller networks.

7.7 Further Work

This section contains discussions and ideas about potential further work, that was out of the constraints of this thesis. This includes the improvement of the data, model, and the use of Graph Neural Networks (GNN).

7.7.1 Data

Due to the constraints of the thesis, only data from wallstreetbets have been used for developing the model. Initially smaller subreddits were analyzed, but due to the lack of awarded posts which is used as the dependent variable, these subreddits have not been included in the model. Ideally, data from many subreddits should be used in the development, such that the model explains individuals' reputations more generally, instead of primarily on wallsteetbets.

Furthermore, it could be interesting to include textual analysis of the interaction between users, and more specifically the interaction between users and the individual that is being predicted. This will capture some of the social dynamics between the nodes, that are not captured through the network analysis. It could be expected that users interact differently towards a user with a high reputation. An example of this can be seen in Appendix A.1.1 where a user comments on the individual DeepFuckingValue. It can be seen that the user expresses themself in a manner that would not be expected for a user without a reputation.

7.7.2 Model Improvements

It could be interesting to control for the date discussed in Section 7.4.1 since the result suggests a bias towards the date a post is created. Additionally, the result in Section 6.3 shows the likelihood of receiving an award increases prior to the number of previous awards by the user. This feature could be included in the model since it shows promising predictive power.

7.7.3 Graph Neural Networks

A future possibility for further improvement could be to explore the possibilities within Graph Neural Networks.

Graph Neural Networks (GNN) are Neural Networks that are capable of dealing and working with graph-structured information or data. A GNN model takes networks/graphs as input and can learn information retained in the network, this can be used for a variety of machine-learning tasks. GNN models take the adjacency matrix as input.

A variation of GNN that could be ideal for this project is Temporal Graph Networks (TGNs) explained in the paper "Temporal Graph Networks for Deep Learning on Dynamic Graphs" by Emanuele Rossi et al [35]. TGNs handle dynamic graphs represented as sequences of timed events. Which could have huge potential for predicting awarded posts.

Another approach is Graph Attention Networks (GATs) [36] described in the paper "Graph Attention Networks" by Petar Veličković et al. GATs use an attention-mechanism similar to transformer models, where the model finds contextual relationships between nodes. This works very similarly to the attention-mechanism in a transformer model since the GAT takes the adjacency matrix as input which is a $(0, 1)$matrix, this is similar to the word embedding in the transformer model which is also a $(0, 1)$ vector representation of each word in the sequence.

When using a GNN model you do not have to extract features for the networks, instead, the GNN model finds information retained in the network. When using a deep learning model instead of feature extracting you lose the ability to interpret the learned features and their relationship to the dependent variable.

CHAPTER 8

Conclusion

In this thesis, we explored how reputation & acknowledgment through awards can be explained by using network science and natural language processing on the subreddit Wallstreetbets. The results support the hypotheses stated in Section 1.4.

Hypothesis I
Using Reddit's award system we were able to measure an individual's acknowledgment and recognition. This was investigated by examining individuals with a known high reputation on the subreddit Wallstreetbets. These users had a significantly high amount of awards compared to users with the same activity level. The study revealed that Reddit's award system is a unique opportunity to quantify how an individual is perceived.

Hypothesis II
In order to investigate how individuals obtain an award for their posts, we used discrete time dynamic networks. For each post, we computed a network of a 10-week period before a post was created. This enables us to extract different network features in the time before the post. To extract information retained in the text we used the ELECTRA transformer. Using the network features and the ELECTRA transformer model we developed a deep learning model that can classify if a post gets an award or not. The model performed well with an F1-score of 78.13%. The model is a combination of various network features and textual content, which indicates the complex nature of acknowledgment within communities.

Hypothesis III
We explored the social dynamics of acknowledgment through awards. We studied the distribution of awards across the awarded users, which shows that a small group of people receive the majority of all the awards. The top 1 % of the most awarded individuals received 40 % of all the awards on Wallstreetbets. In addition, we showed that the probability of receiving an award increases prior to the more awards you have. This indicates a "rich get richer" dynamic.

Hypothesis IV
We investigated the impact of network size on individuals' identity and reputation within the wallstreetbets community. By analyzing the reciprocity and number of mentions, we found that larger networks have different social dynamics. The results indicated that there are relatively fewer reciprocal edges and number of mentions in larger networks, suggesting that it is more challenging for individuals to establish an identity within larger communities. Our analysis also revealed similar findings in the permutation feature importance computed for models of both larger and smaller networks.

Bibliography

[1] Elizabeth Dubois and Grant Blank. “The echo chamber is overstated: the moderating effect of political interest and diverse media”. In: *Information, Communication & Society* 21.5 (2018), pp. 729–745. DOI: 10.1080/1369118X.2018.1428656. eprint: https://doi.org/10.1080/1369118X.2018.1428656. URL: https://doi.org/10.1080/1369118X.2018.1428656.

[2] Yonghwan Kim, Ji Eun Lee, and Eun-Ju Park. “Motivations for using the social networking site Reddit”. In: *Telematics and Informatics* 34.7 (2017), pp. 1294–1304.

[3] Alexey Medvedev, Renaud Lambiotte, and Jean-Charles Delvenne. “The Anatomy of Reddit: An Overview of Academic Research”. In: May 2019, pp. 183–204. ISBN: 978-981-13-6240-8. DOI: 10.1007/978-3-030-14683-2_9.

[4] Rationality. *The Cambridge Dictionary of Philosophy*. Cambridge University Press, 1999.

[5] Christoph Ellersgaard. *PhD Thesis - Elites in Denmark: Power Elites and Ruling Classes in a Welfare State*. June 2016. DOI: 10.13140/RG.2.1.2176.9840.

[6] Anton Grau Larsen. “Elites in Denmark: Identifying the elite”. English. PhD thesis. 2015. ISBN: 9788776119096.

[7] Cody L. Buntain and Jennifer Golbeck. “Identifying social roles in reddit using network structure”. In: *Proceedings of the 23rd International Conference on World Wide Web* (2014).

[8] Sean Taylor et al. “Identity effects in social media”. In: *Nature Human Behaviour* (Nov. 2022), pp. 1–11. DOI: 10.1038/s41562-022-01459-8.

[9] Alexandre Parmentier and Robin Cohen. “Learning User Reputation on Reddit”. In: *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. 2019, pp. 242–247.

[10] Gordon Burtch et al. “How Do Peer Awards Motivate Creative Content? Experimental Evidence from Reddit”. In: *Management Science* 68.5 (2022), pp. 3488–3506. DOI: 10.1287/mnsc.2021.4040. eprint: https://doi.org/10.1287/mnsc.2021.4040. URL: https://doi.org/10.1287/mnsc.2021.4040.

[11] William L. Hamilton et al. *Loyalty in Online Communities*. 2017. DOI: 10.48550/ARXIV.1703.03386. URL: https://arxiv.org/abs/1703.03386.

[12] Abhinav Anand and Jalaj Pathak. “The role of Reddit in the GameStop short squeeze”. In: *Economics Letters* 211 (2022), p. 110249. ISSN: 0165-1765. DOI: https://doi.org/10.1016/j.econlet.2021.110249. URL: https://www.sciencedirect.com/science/article/pii/S0165176521004808.

[13] Nicki Skafte Detlefsen. *Machine Learning Operations*. https://github.com/SkafteNicki/dtu_mlops. 2022.

[14] Albert-László Barabási and Márton Pósfai. *Network science*. Cambridge: Cambridge University Press, 2016. ISBN: 9781107076266 1107076269. URL: http://barabasi.com/networksciencebook/.

[15] Jari Saramäki et al. “Generalizations of the clustering coefficient to weighted complex networks”. In: *Phys. Rev. E* 75 (2 Feb. 2007), p. 027105. DOI: 10.1103/PhysRevE.75.027105. URL: https://link.aps.org/doi/10.1103/PhysRevE.75.027105.

[16] Ron Milo et al. “Network motifs: simple building blocks of complex networks”. In: *Science* 298.5594 (2002), pp. 824–827.

[17] Linton C. Freeman. “A Set of Measures of Centrality Based on Betweenness”. In: *Sociometry* 40.1 (1977), pp. 35–41. ISSN: 00380431. URL: http://www.jstor.org/stable/3033543 (visited on 12/12/2022).

[18] Linton C. Freeman. “Centrality in social networks conceptual clarification”. In: *Social Networks* 1.3 (1978), pp. 215–239. ISSN: 0378-8733. DOI: https://doi.org/10.1016/0378-8733(78)90021-7. URL: https://www.sciencedirect.com/science/article/pii/0378873378900217.

[19] Phillip Bonacich. “Power and Centrality: A Family of Measures”. In: *American Journal of Sociology* 92.5 (1987), pp. 1170–1182. DOI: 10.1086/228631. eprint: https://doi.org/10.1086/228631. URL: https://doi.org/10.1086/228631.

[20] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. “Natural language processing: an introduction”. In: *Journal of the American Medical Informatics Association* 18.5 (Sept. 2011), pp. 544–551. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2011-000464. eprint: https://academic.oup.com/jamia/article-pdf/18/5/544/5962687/18-5-544.pdf. URL: https://doi.org/10.1136/amiajnl-2011-000464.

[21] C. Hutto and Eric Gilbert. “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1 (May 2014), pp. 216–225. DOI: 10.1609/icwsm.v8i1.14550. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

[22] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738.

[23] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[24] Ashish Vaswani et al. *Attention Is All You Need*. 2017. DOI: 10.48550/ARXIV.1706.03762. URL: https://arxiv.org/abs/1706.03762.

[25] Kevin Clark et al. *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators*. 2020. DOI: 10.48550/ARXIV.2003.10555. URL: https://arxiv.org/abs/2003.10555.

[26] Tin Kam Ho. “Random decision forests”. In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[27] Karl Pearson F.R.S. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. DOI: 10.1080/14786440109462720.

[28] *Spectator - who are the reddit traders*. https://www.spectator.co.uk/article/who-are-the-reddit-traders/. Accessed: 14-02-2023.

[29] *Marketwatch - 1R0NYMAN*. https://www.marketwatch.com/story/trader-says-he-has-no-money-at-risk-then-promptly-loses-almost-2000-2019-01-22. Accessed: 14-02-2023.

[30] *CNBC - ControlTheNarrative*. https://www.cnbc.com/2020/06/18/young-trader-dies-by-suicide-after-thinking-he-racked-up-big-losses-on-robinhood.html. Accessed: 14-02-2023.

[31] Abhinav Anand and Jalaj Pathak. “The role of Reddit in the GameStop short squeeze”. In: *Economics Letters* 211 (2022), p. 110249. ISSN: 0165-1765. DOI: https://doi.org/10.1016/j.econlet.2021.110249. URL: https://www.sciencedirect.com/science/article/pii/S0165176521004808.

[32] *The Guardian - DeepFuckingValue*. https://www.theguardian.com/technology/2022/nov/04/how-i-lost-1m-during-the-pandemic. Accessed: 14-02-2023.

[33] J. Reiman and P. Leighton. *The Rich Get Richer and the Poor Get Prison: Ideology, Class, and Criminal Justice: Eleventh Edition*. Nov. 2016, pp. 1–258. DOI: 10.4324/9781315639079.
[34] P. Bourdieu and R. Nice. *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press, 1984. ISBN: 9780674212770.
[35] Emanuele Rossi et al. *Temporal Graph Networks for Deep Learning on Dynamic Graphs*. 2020. DOI: 10.48550/ARXIV.2006.10637. URL: https://arxiv.org/abs/2006.10637.
[36] Petar Veličković et al. *Graph Attention Networks*. 2017. DOI: 10.48550/ARXIV.1710.10903. URL: https://arxiv.org/abs/1710.10903.

CHAPTER A

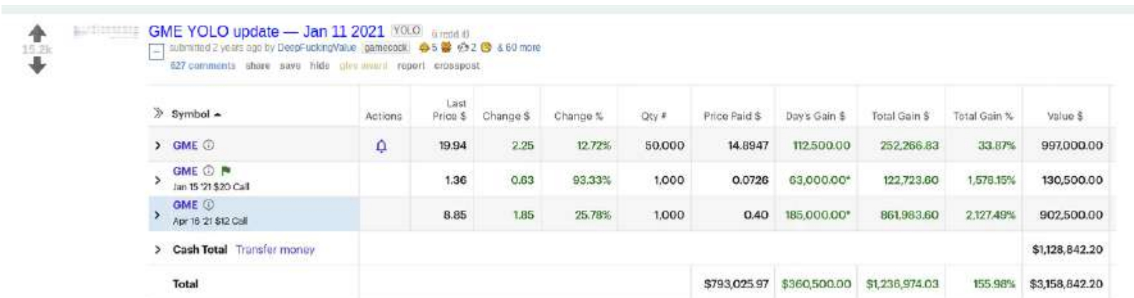
Appendix

A.1 Text Examples

The following example is from an awarded post from wallstreetbets with a VADER sentiment score text of 0.9543 and a text length of 632.

> This is probably going to be a huge week Let's show them how much we really like the stock, let's destroy that 1000, let's zoom past 5000 and then let's get out of this universe, going to the moon was plans of the past, we're far too good for that, let's explore in deep space. Love the sight of fear in the mornings, watching the news every day seeing they're trying to make us crumble, do they not know diamond is too strong to shatter (I ain't no scientist so if I'm wrong forgive me) Good luck this week people, keep them □□□□ and remember, together apes strong □□□□□□□□□□□□ Also not financial advice, I just really like the stock.

A.1.1 DeepFuckingValue



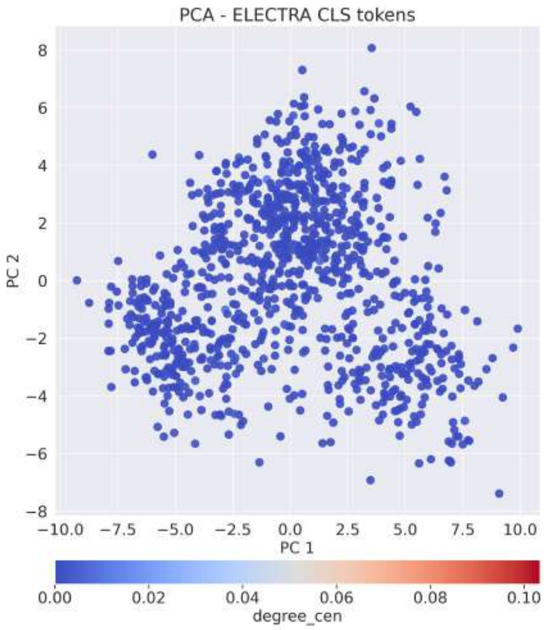
Symbol	Actions	Last Price $	Change $	Change %	Qty #	Price Paid $	Day's Gain $	Total Gain $	Total Gain %	Value $
GME		19.94	2.25	12.72%	50,000	14.8947	112,500.00	252,266.83	33.87%	997,000.00
GME Jan 15 '21 $20 Call		1.36	0.63	93.33%	1,000	0.0726	63,000.00*	122,723.60	1,578.15%	130,500.00
GME Apr 16 '21 $12 Call		8.85	1.85	25.78%	1,000	0.40	185,000.00*	861,983.60	2,127.49%	902,500.00
Cash Total Transfer money										$1,128,842.20
Total						$793,025.97	$360,500.00	$1,236,974.03	155.98%	$3,158,842.20

Figure A.1: A post from DeepFuckingValue

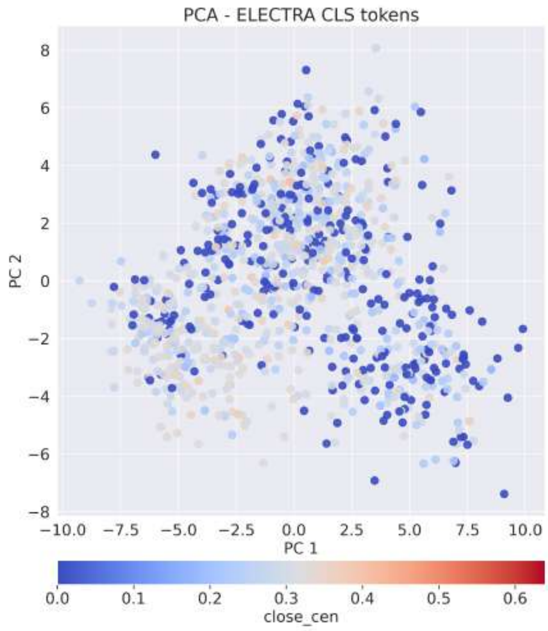
Figure A.2: A comment on the post from DeepFuckingValue seen in Figure A.1.1

The post and comments can be seen **here**.

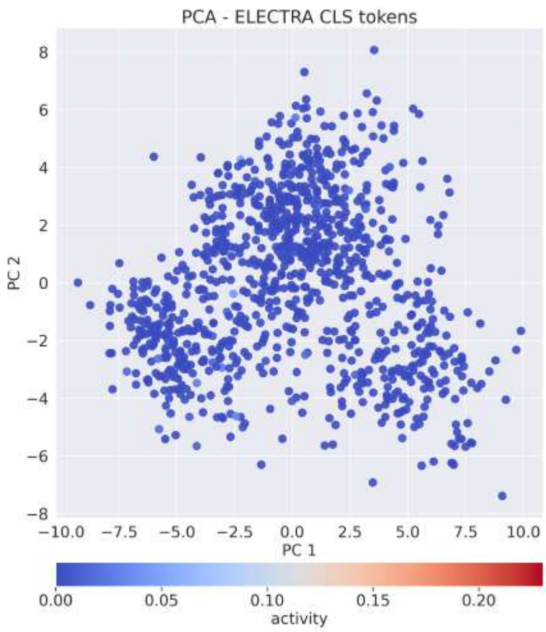
A.2 ELECTRA - PCA



(a) Degree Centrality

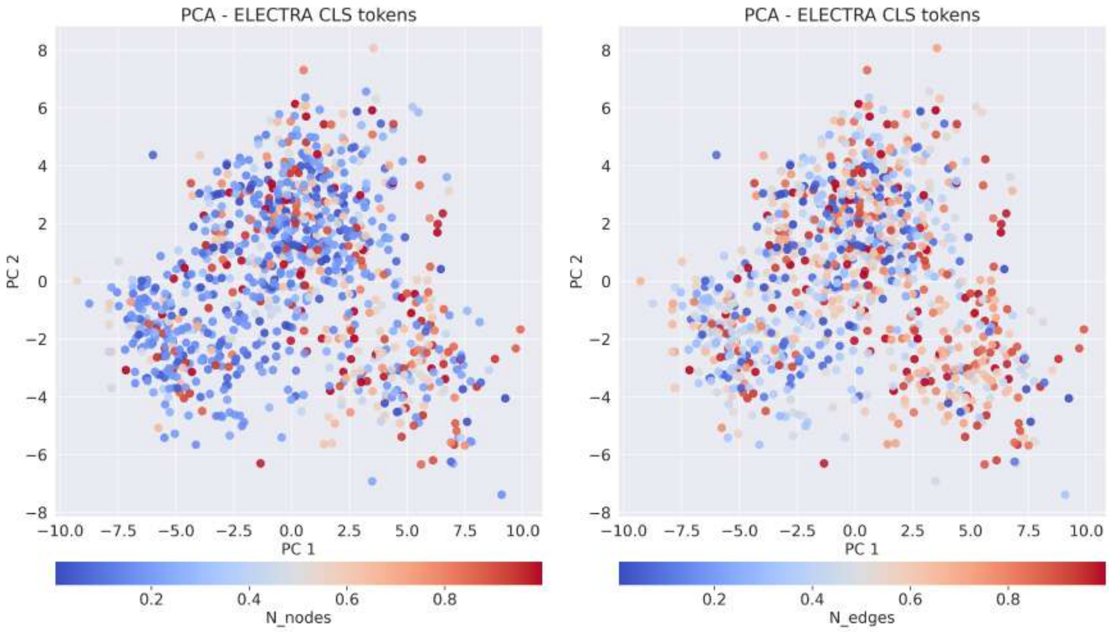


(b) Closeness Centrality



(c) Activity

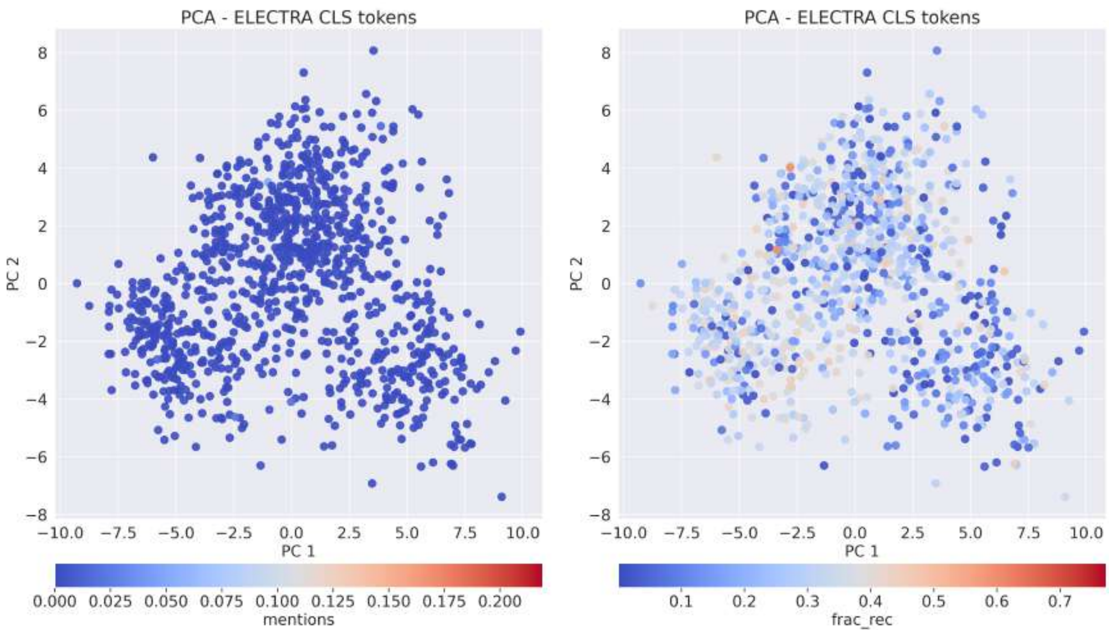
Figure A.3: Principal components of CLS tokens



(a) N Nodes

(b) N Edges

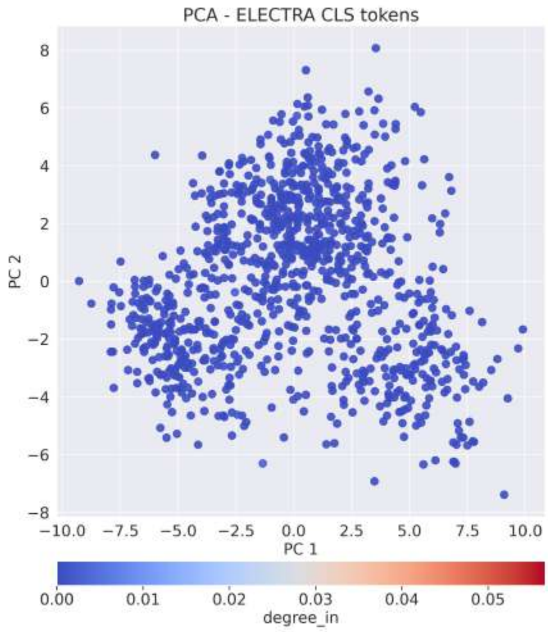
Figure A.4: Principal components of CLS tokens



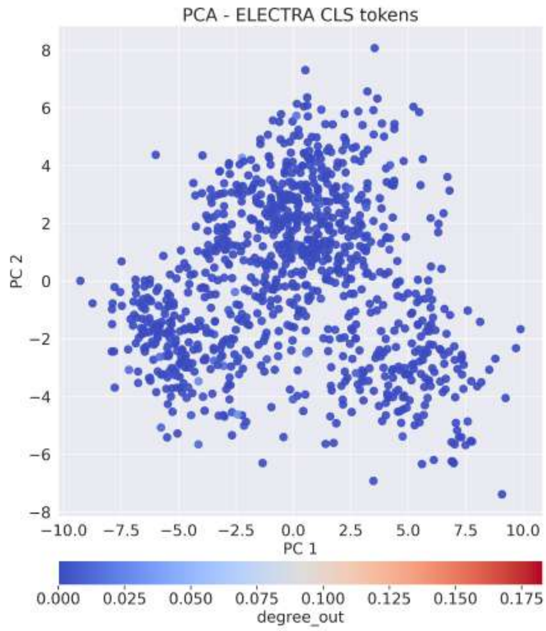
(a) Author Mentions

(b) Fraction of reciprocal edges

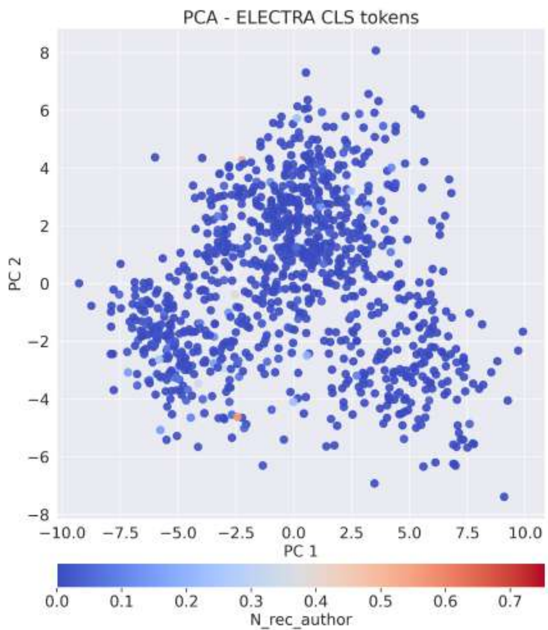
Figure A.5: Principal components of CLS tokens



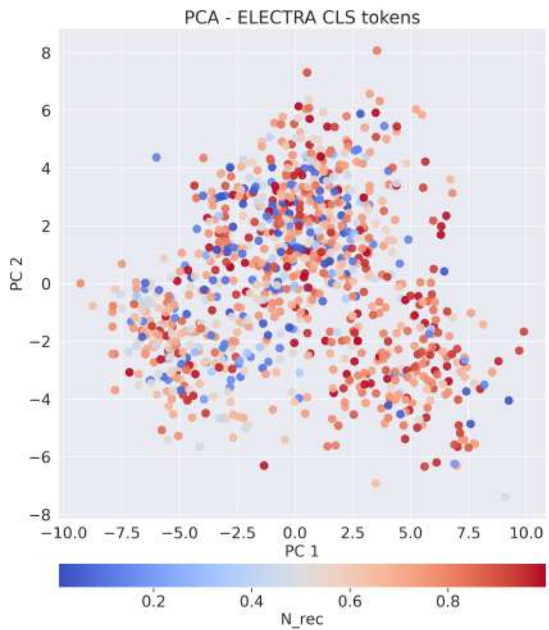
(a) In degree



(b) Out degree

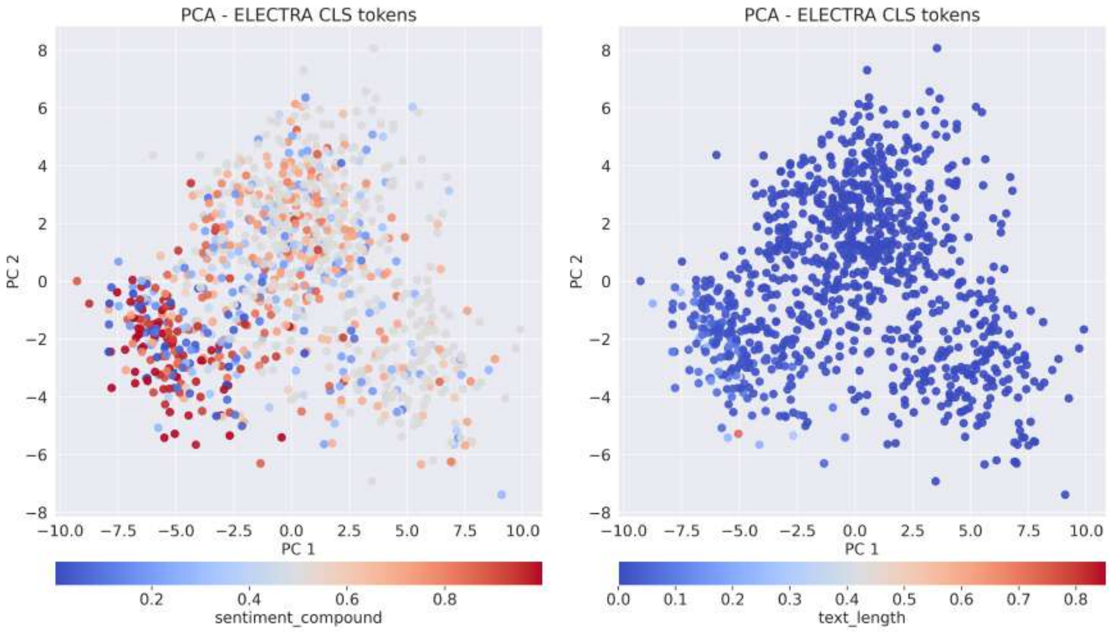


(c) N Reciprocal Edges for the author



(d) N$_reciprocaledges$

Figure A.6: Principal components of CLS tokens



(a) Sentiment compound

(b) Text Length

Figure A.7: Principal components of CLS tokens

A.3 Features by Date

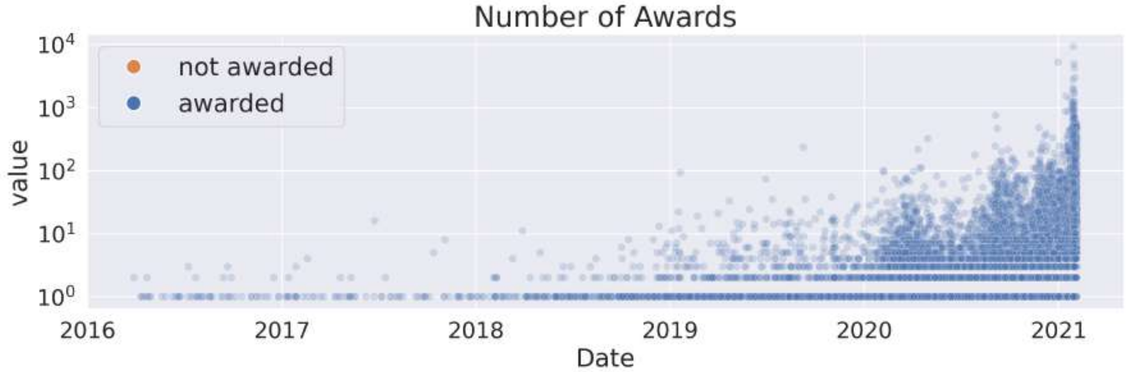


Figure A.8: The number of awards given to each author of a post by date

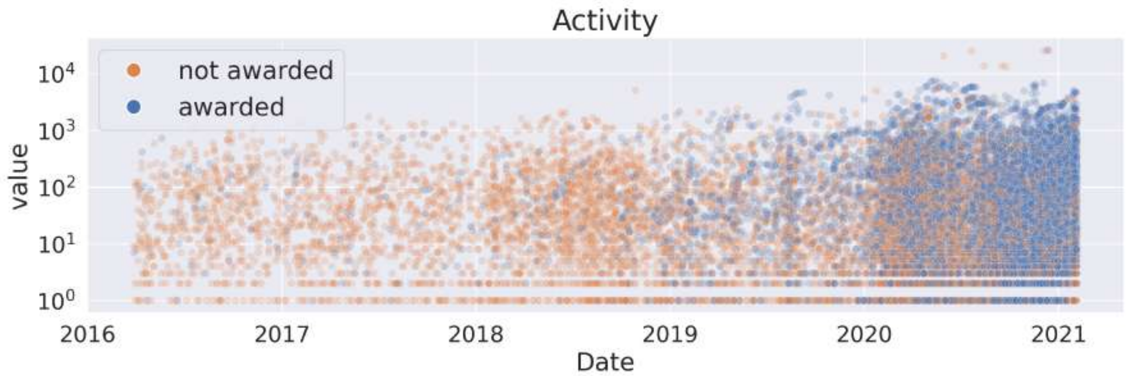


Figure A.9: The activity for each author of a post by date

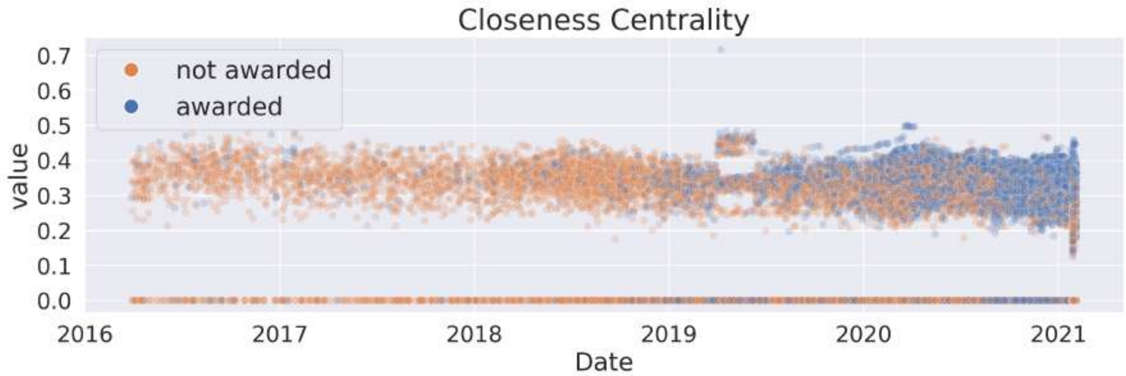


Figure A.10: The closeness centrality for each author of a post by date

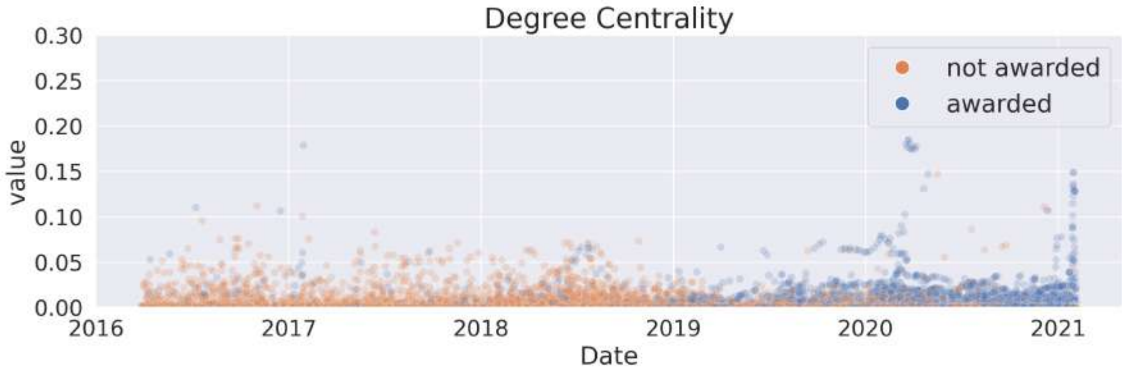


Figure A.11: The degree centrality for each author of a post by date

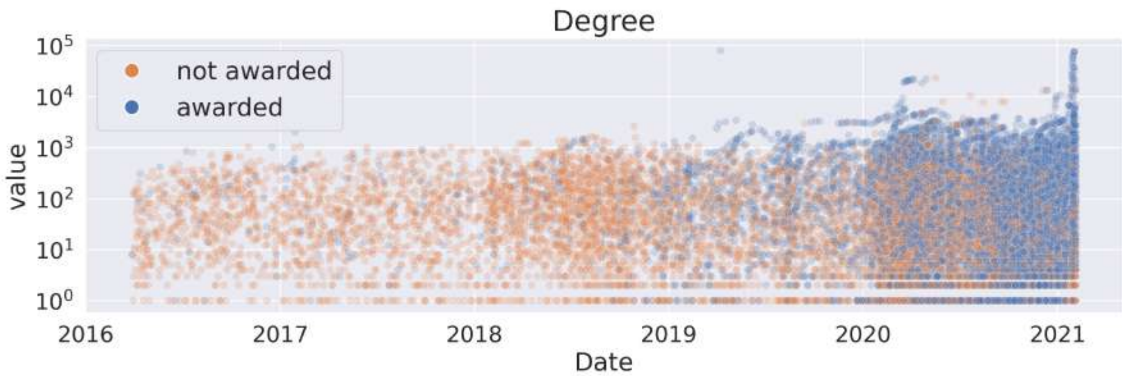


Figure A.12: The degree for each author of a post by date

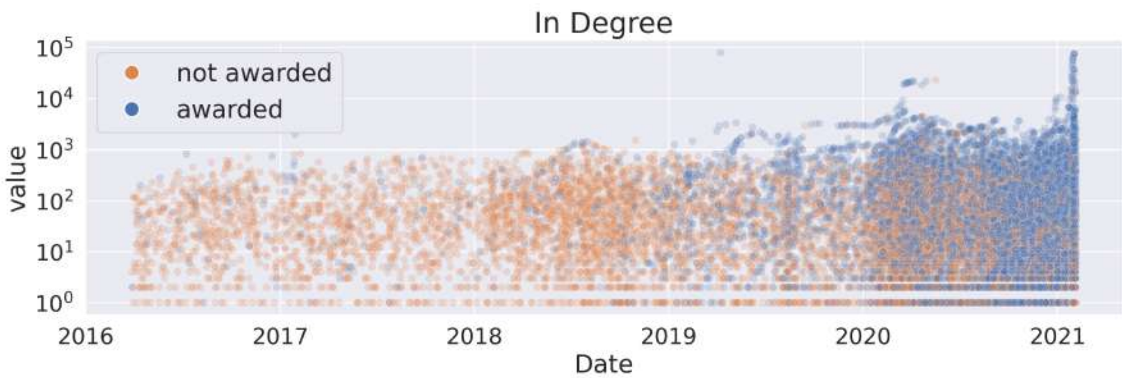


Figure A.13: The in-degree for each author of a post by date

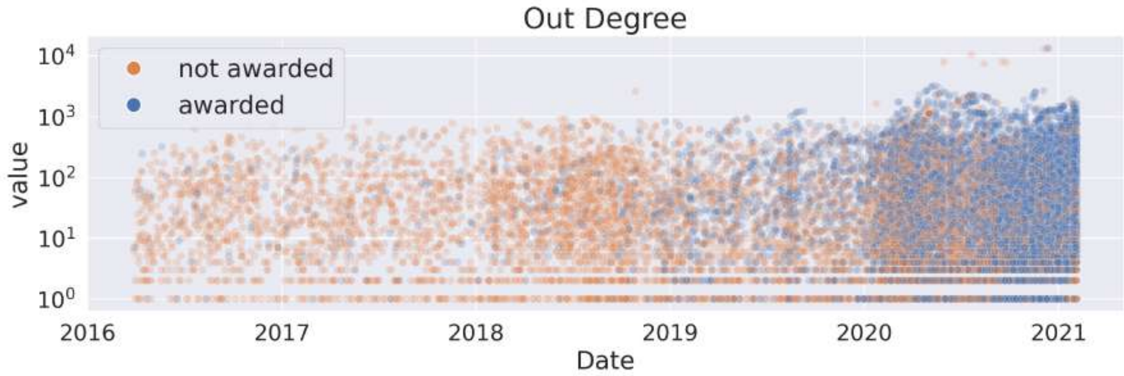


Figure A.14: The out-degree for each author of a post by date

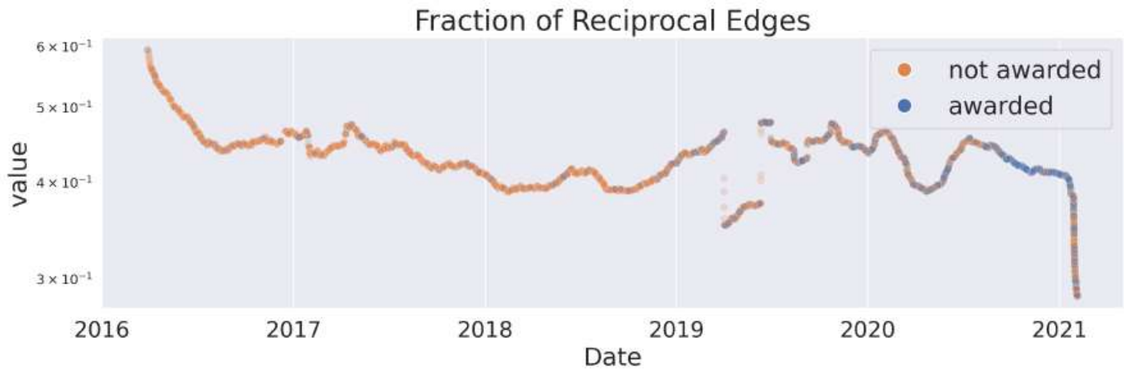


Figure A.15: The fraction of reciprocal edges

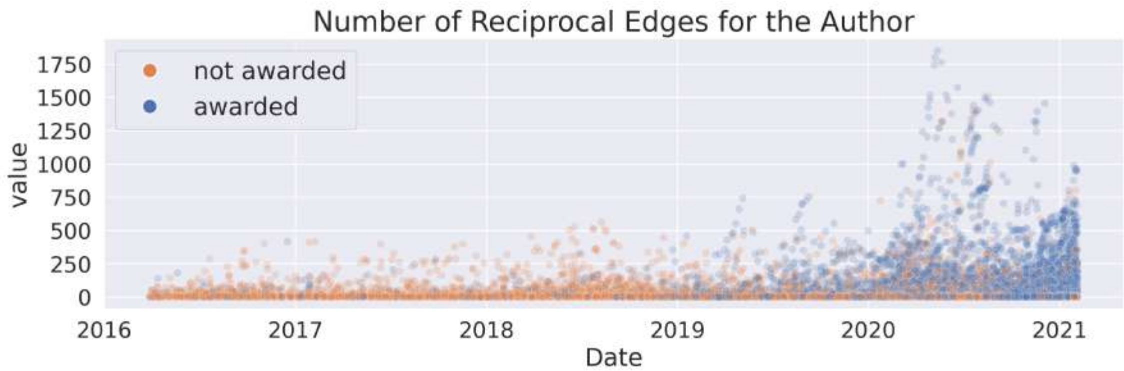


Figure A.16: The number of reciprocal edges for the author of a post

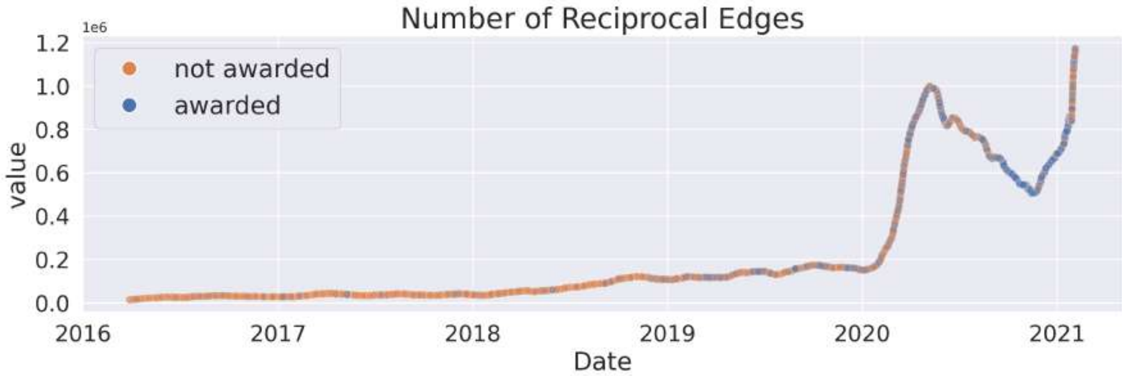


Figure A.17: The number of reciprocal edges in each network

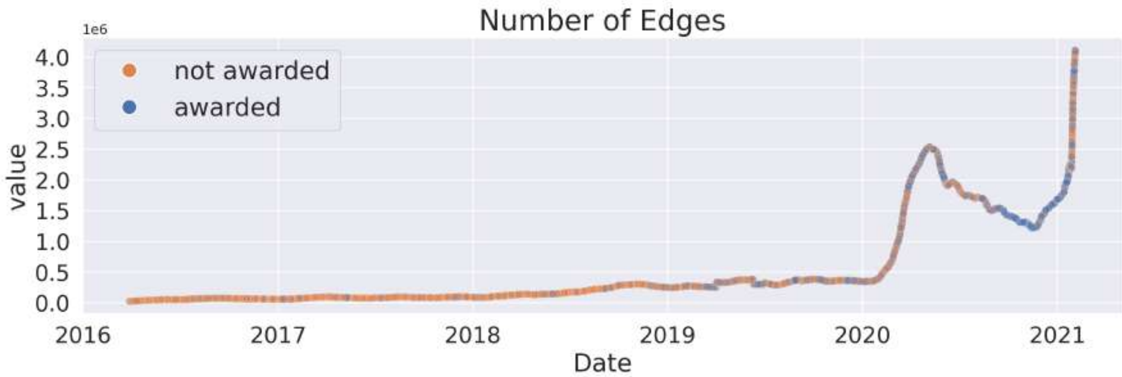


Figure A.18: The number of edges in each network

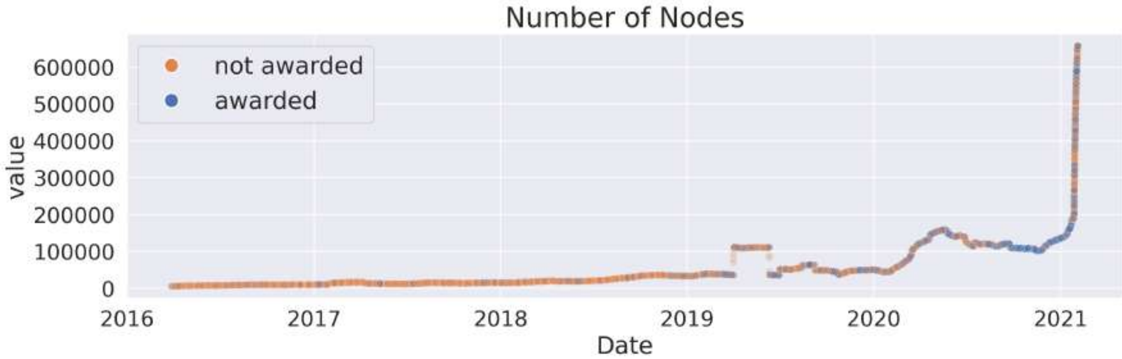


Figure A.19: The number of nodes in each network

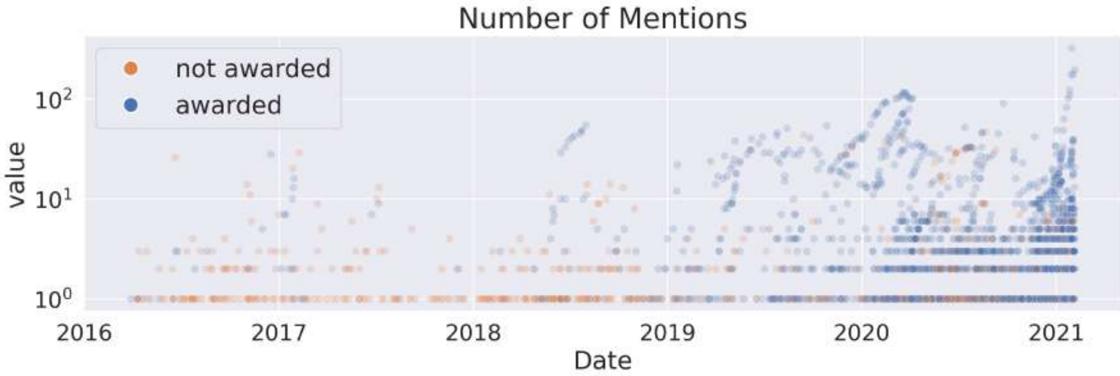


Figure A.20: The mentions of each author of a post by date

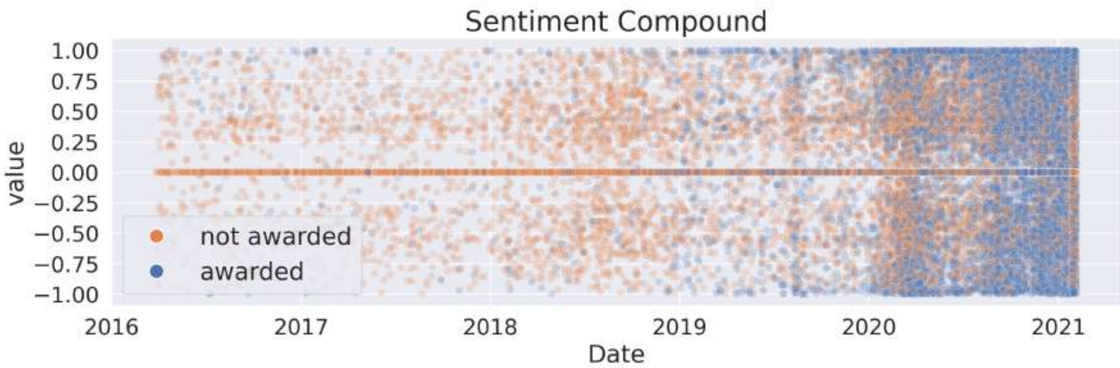


Figure A.21: The sentiment compound for each post by date

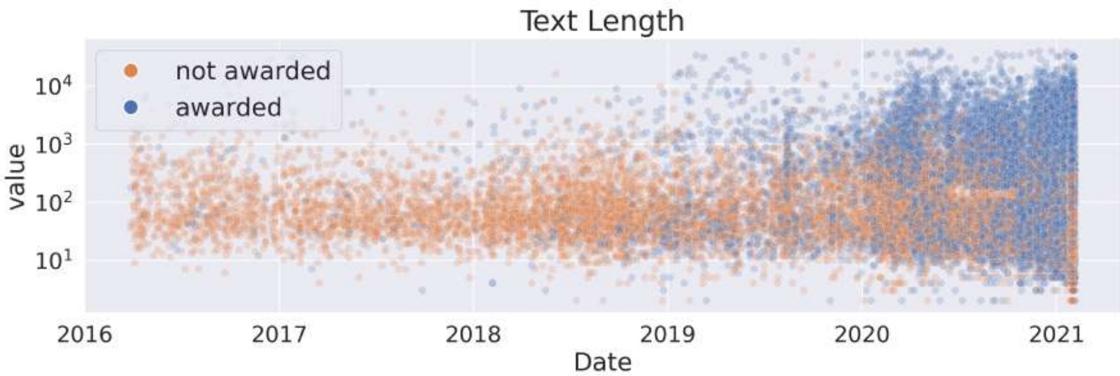
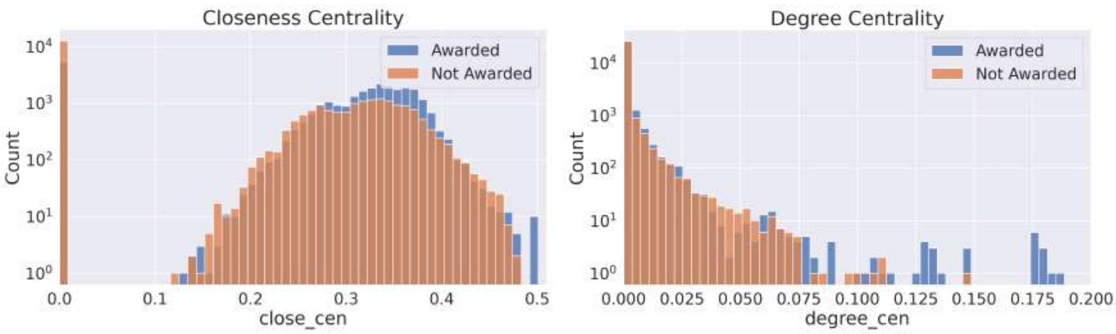


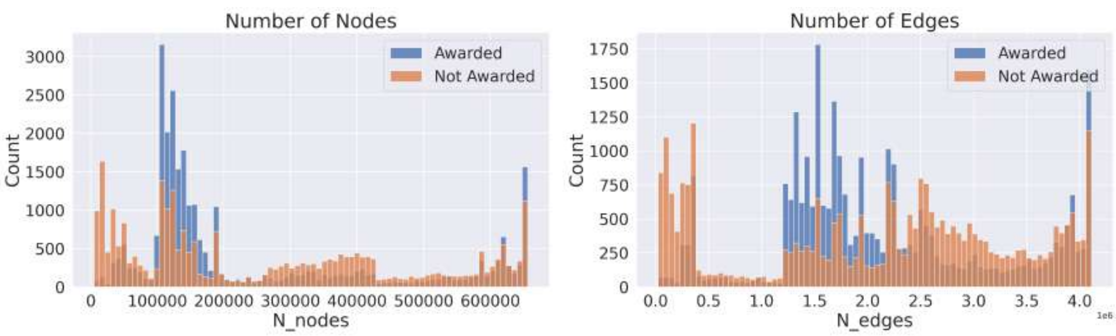
Figure A.22: The text length for each post by date

A.4 Features



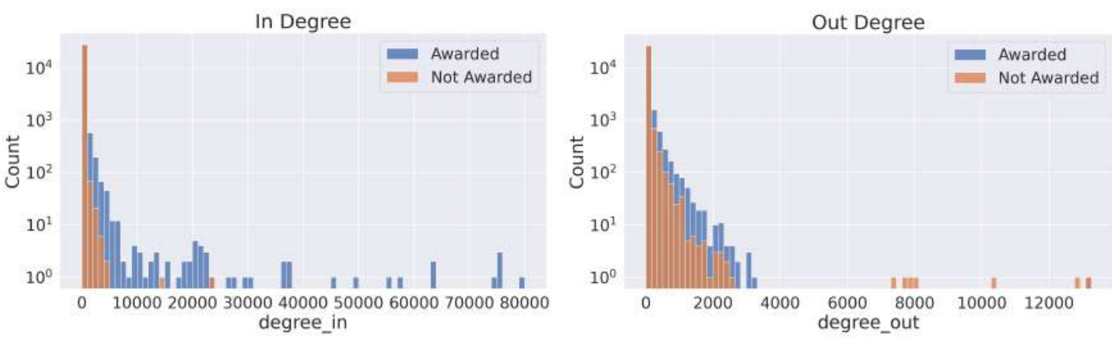
(a) (b)

Figure A.23: The closeness centrality (a) and the degree centrality (b) colored by awarded or not awarded.



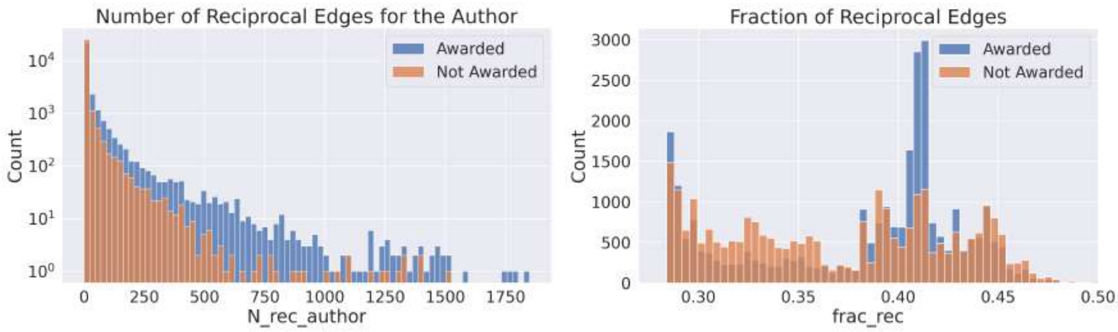
(a) (b)

Figure A.24: The number of nodes (a) and the number of edges (b) colored by awarded or not awarded.



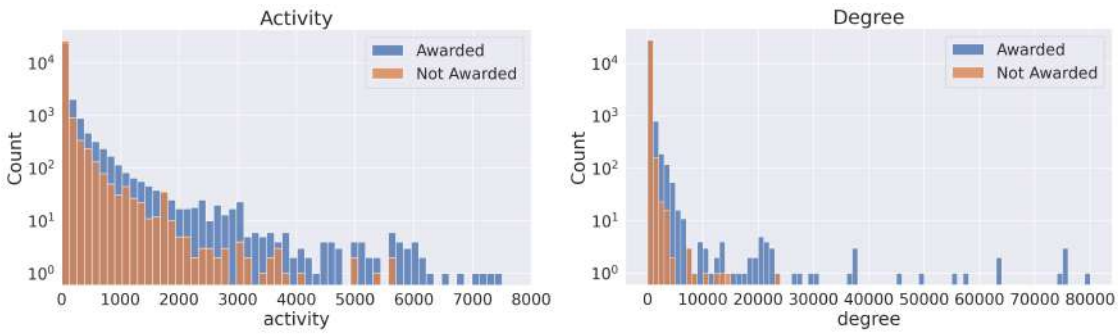
(a) (b)

Figure A.25: The number of nodes (a) and the number of edges (b) colored by awarded or not awarded.



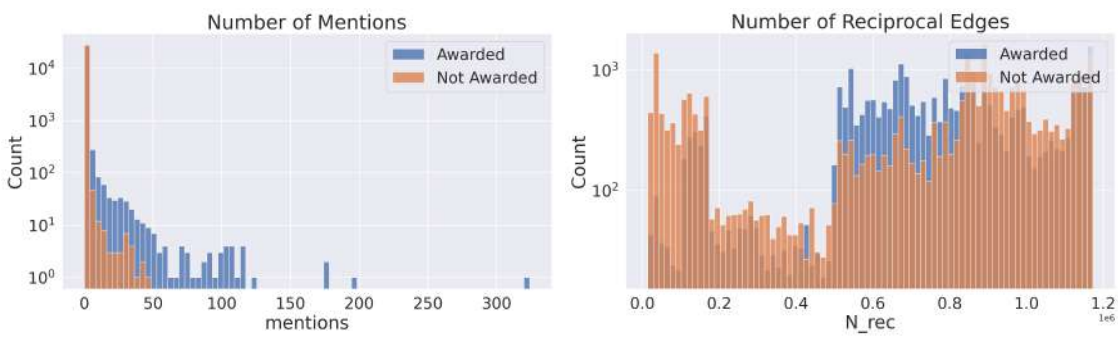
(a) (b)

Figure A.26: The number of reciprocal edges for the author of the post (a) and the fraction of the total number of reciprocal edges in the network. (b) The distributions are colored by awarded or not awarded.



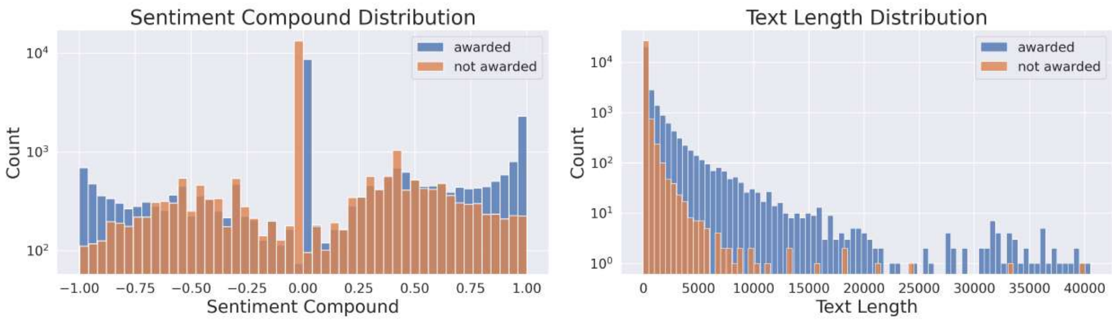
(a) (b)

Figure A.27: The activity for the author of the post (a) and the degree for the author of the post (b). The distributions are colored by awarded or not awarded.



(a) (b)

Figure A.28: The number of mentions of the author of the post (a) and the number of reciprocal edges in the network (b). The distributions are colored by awarded or not awarded.



(a) (b)

Figure A.29: The sentiment compound of the post (a) and the text length of the post (b). The distributions are colored by awarded or not awarded.

A.5 Appendix - Models

The following figure shows the percentage of awarded posts by date of the balanced data used for the model.

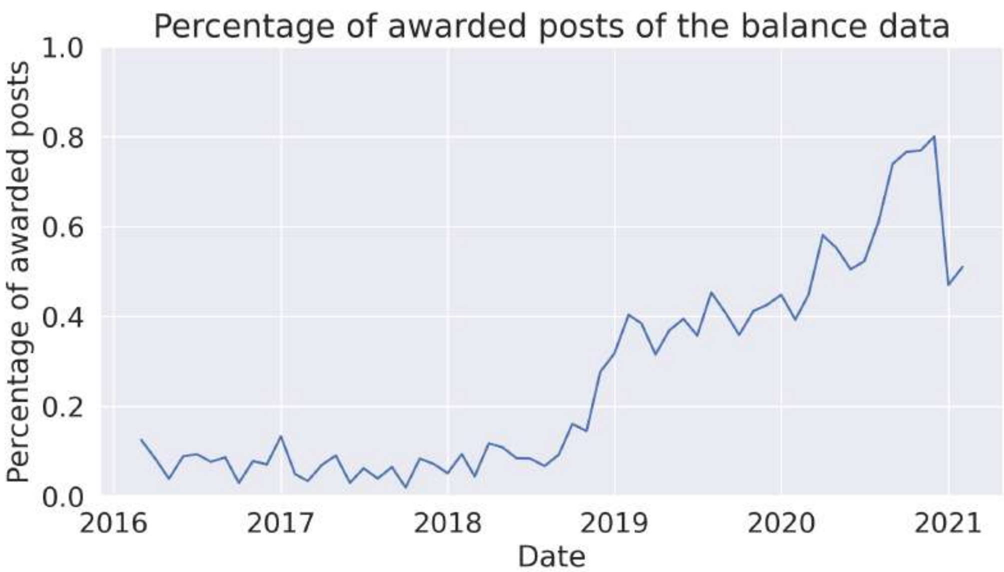


Figure A.30

A.5.1 Training

Neural Network

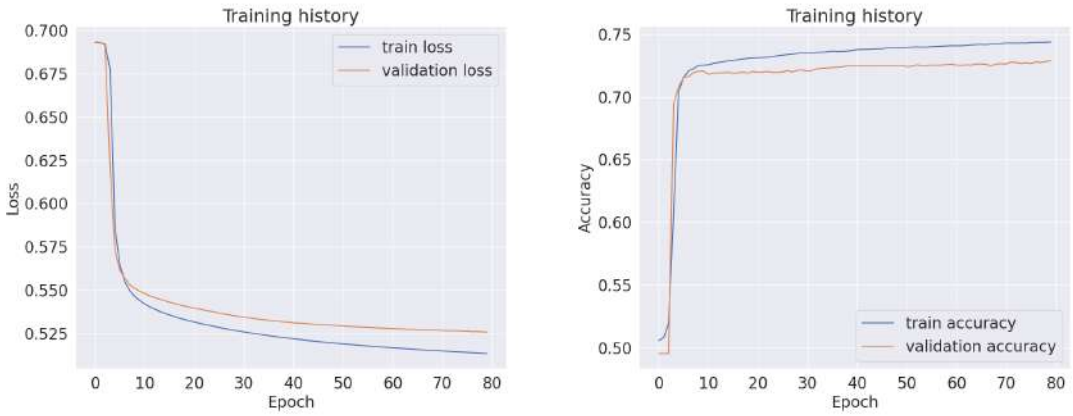


Figure A.31: Training History

Neural Network + ELECTRA

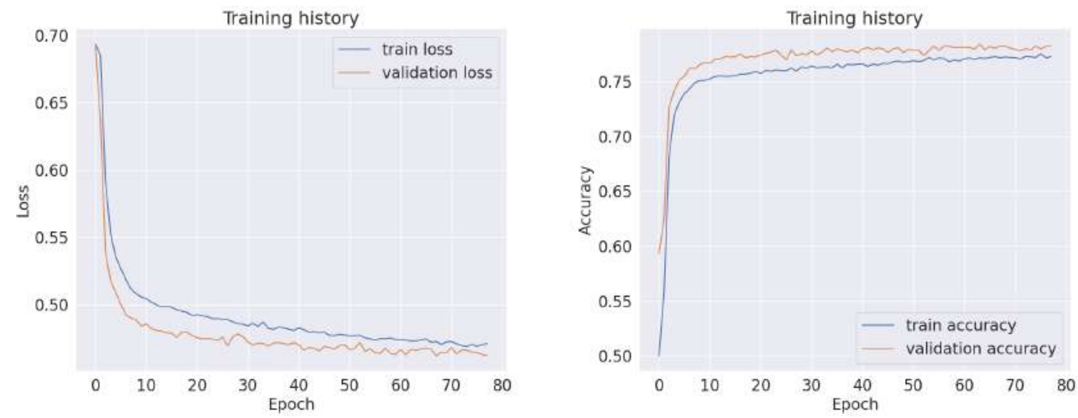


Figure A.32: Training History

A.5.2 Hyperparameter Sweeping

Neural Network

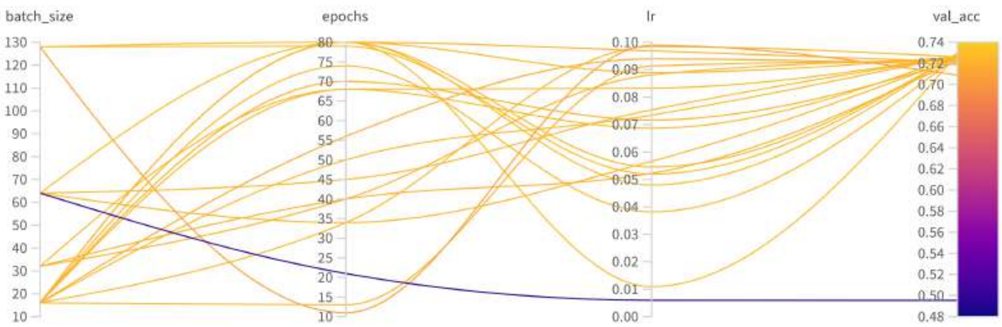


Figure A.33

Neural Network + ELECTRA

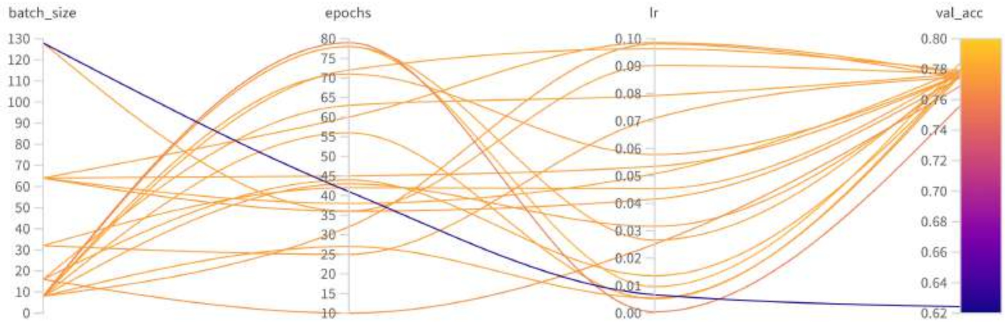


Figure A.34

CHAPTER B

Code

B.1 Model: Neural Network

```
class Classifier(nn.Module):
    def __init__(self,num_labels=2):
        super(Classifier,self).__init__()
        self.num_labels = num_labels

        # network + text features
        self.network_input = nn.Linear(in_features=14,out_features=256)
        self.dense_net2 = nn.Linear(in_features=256,out_features=512)
        self.dense_net3 = nn.Linear(in_features=512,out_features=1024)
        self.dense_net4 = nn.Linear(in_features=1024,out_features=2048)
        self.dense_net5 = nn.Linear(in_features=2048,out_features=1024)
        self.dense_net6 = nn.Linear(in_features=1024,out_features=512)
        self.dense_net7 = nn.Linear(in_features=512,out_features=256)

        # output layer
        self.out_proj = nn.Linear(256, self.num_labels)

    def forward(self,network_features=None):
        x_net = F.relu(self.network_input(network_features))
        x_net = F.relu(self.dense_net2(x_net))
        x_net = F.relu(self.dense_net3(x_net))
        x_net = F.relu(self.dense_net4(x_net))
        x_net = F.relu(self.dense_net5(x_net))
        x_net = F.relu(self.dense_net6(x_net))
        x_net = F.relu(self.dense_net7(x_net))

        logits = self.out_proj(x_net)
        return logits
```

B.2 Model: Neural Network + ELECTRA

```
class ElectraClassifier(nn.Module):
    def __init__(self,num_labels=2):
        super(ElectraClassifier, self).__init__()
        self.num_labels = num_labels

        # text features
        self.electra = ElectraModel.from_pretrained('google/electra-small-
            discriminator')
        self.dense_txt = nn.Linear(self.electra.config.hidden_size, self.
            electra.config.hidden_size)
        self.dropout_txt = nn.Dropout(self.electra.config.hidden_dropout_prob)

        # combined features
        self.dense_cat1 = nn.Linear(in_features=(256+12),out_features=512)
        self.dense_cat2 = nn.Linear(in_features=512,out_features=1024)
        self.dense_cat3 = nn.Linear(in_features=1024,out_features=2048)
```

```
        self.dense_cat6 = nn.Linear(in_features=2048,out_features=1024)
        self.dense_cat7 = nn.Linear(in_features=1024,out_features=512)
        self.dense_cat8 = nn.Linear(in_features=512,out_features=256)

        # output layer
        self.out_proj = nn.Linear(256, self.num_labels) # 2 labels

    def classifier(self,sequence_output,network_features):
        # text features
        x_txt = sequence_output[:, 0, :] #[CLS] token
        x_txt = F.relu(self.dense_txt(x_txt))
        x_txt = self.dropout_txt(x_txt)

        # combined features
        x = torch.cat((x_txt,network_features),dim=1)
        x = F.relu(self.dense_cat1(x))
        x = F.relu(self.dense_cat2(x))
        x = F.relu(self.dense_cat3(x))
        x = F.relu(self.dense_cat6(x))
        x = F.relu(self.dense_cat7(x))
        x = F.relu(self.dense_cat8(x))

        # output layer
        logits = self.out_proj(x)
        return logits

    def forward(self, input_ids=None,attention_mask=None,network_features=None
        ):
        discriminator_hidden_states = self.electra(input_ids=input_ids,
            attention_mask=attention_mask)
        sequence_output = discriminator_hidden_states[0]
        logits = self.classifier(sequence_output,network_features)

        return logits
```

Technical
University of
Denmark

www.compute.dtu.dk