Universiteit Leiden
The Netherlands

Informatica & Economie

Thanks for the gold, kind stranger! Predicting the receipt of community recognition for social media comments

By Koen Hagen

Supervisors:
Frank Takes & Gerrit Jan de Bruin

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl July 22, 2021

Abstract

Reddit has made its mark online as one of the largest platforms for social interaction. The platform works through user appreciation, in the sense that more popular posts and comments get shown to more visitors. Users can grant “Reddit gildings” to posts and comments they feel deserve extra recognition. This gives the comment a golden hue and badge, while the author receives “Reddit Premium”. This thesis aims to predict whether or not a comment receives Reddit gildings through machine learning using the comment’s properties. Our data contains all publicly available comments on all posts from November to December 2018. For each comment, a set of 24 features is gathered. The results show that there is some predictability with an AUC of 0.73. The features that have a lot of influence on our machine learning model are the comment length based features. The number of comments on average per post on the subreddit is also of high importance. The most important feature however, is the average karma of the author. The results are helpful to scholars to further explore the influence of variables on the likelihood of gaining Reddit gildings and for Reddit users seeking to gain a high social status on the Reddit platform.

Contents

1 Introduction

1.1 Context

Reddit is one of the largest social media platforms, ranking #19 in global internet engagement according to Alexa[1]. Reddit is especially well used in the United States and United Kingdom, ranking seventh and third respectively. Reddit functions as an enormous forum with a virtually unlimited number of smaller sub-forums (called subreddits). Any user can create a new subreddit and start filling it with content. Users can follow a set of subreddits, which then will appear in their home feed. Subreddits can range from only a couple to almost 32 million followers. These subreddits become their own community or "bubble", allowing people with vastly different ideas and opinions to coincide on Reddit. Every subreddit has its own set of rules and themes, which the users of that subreddit will have to abide by. Subjects of subreddits can range from serious political discussion to lighthearted sharing of cute cat pictures and everything in between. The highest-ranking posts on any of the subreddits get collected on the front page of Reddit. It is considered a remarkable feat to get a post on there. So much so that [15] shows that topics discussed on the front page often significantly increase the Wikipedia pageviews on that respective topic.

Reddit does not create any content by itself, but its users add all content on Reddit. Users can do this by creating posts on a specific subreddit. Posts typically contain either an image or a text; or redirect one to another site through a URL. The hierarchical structure of Reddit is shown in Figure 1.

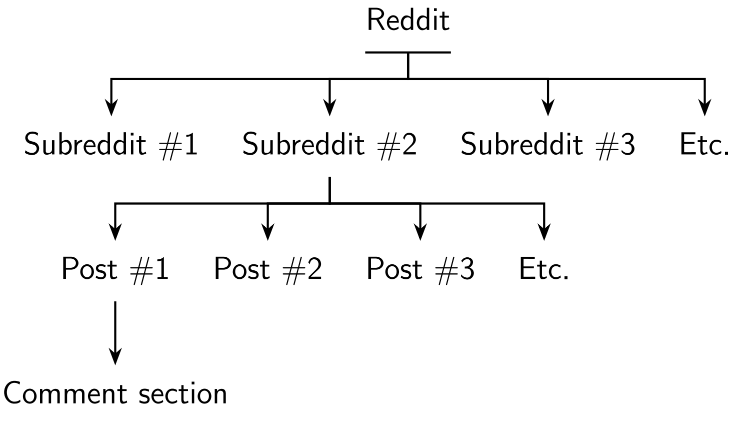

Figure 1: A schematic hierarchical structure of Reddit

Users of Reddit can comment on posts in subreddits. This includes posts

[1] `https://www.alexa.com/siteinfo/reddit.com` *requested on April 14th 2021*

that, for instance, are merely a link to another website. Comments on posts are structured as shown in Figure 2. Comments can respond to the post itself or to another comment, in which case it will position itself below and right of the comment it is responding to. When a comment is too far down the hierarchical structure, it will "collapse", making it not directly visible when viewing the comments. This will give more room for other comments on the post. Users can collapse any comment or open any collapsed comments by themselves via the user interface. In Figure 2, collapsed comments are shown in grey with a dotted border.

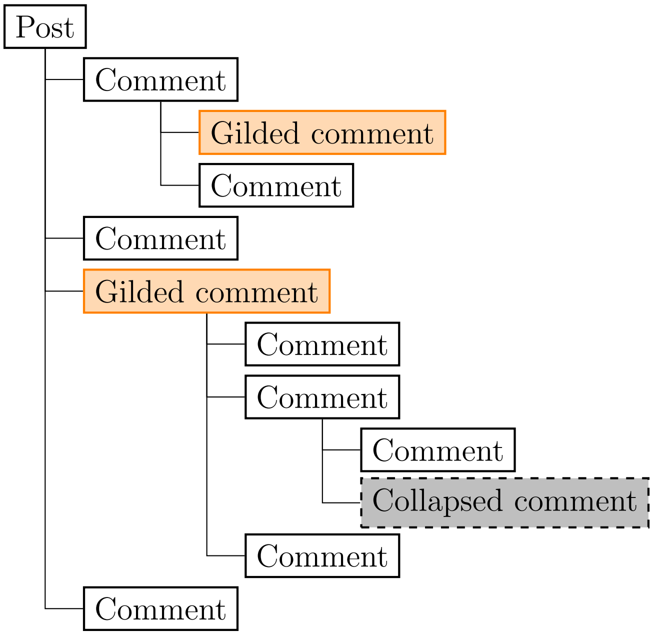

Figure 2: Typical comment section of a post on Reddit.

Any post or comment on the site can be awarded so-called Reddit gildings, making it a "gilded comment". In Figure 2, comments awarded with Reddit gildings are shown in orange. Getting gildings on a comment distinguishes that comment from others. A deeper look into the anatomy of Reddit can be found in [13]

1.2 Goal

Our goal is to obtain a clearer grasp on the distribution of user appreciation. There are multiple ways to measure user appreciation; It can be measured by looking at the "karma" scores and through the Reddit gildings mentioned earlier. Karma is defined by the number of upvotes minus downvotes, which are Reddit's equivalent of likes and dislikes. Karma determines the prevalence of the comment or post. More upvoted comments float to the top of the feed.

Obtaining karma and Reddit gildings have become goals in and of themselves for Reddit users. This is clear from the amount of research that has been done and external resources that have been created for this specific purpose. Works by [14], [8] and [3] show us that these motivators have a clear influence on user behaviour. Post schedulers, like websites Cronnit and Delayforreddit, are a common resource for gaining karma and gildings. These post schedulers keep track of which time and date are the best to post on specific subreddits and allow the user to schedule their posts at such a time. Website Laterforreddit provides the most common words and combination of words in the title of popular posts per subreddit. The user's profile page shows the number of obtained karma and whether the user has been awarded Reddit gildings.

Across the years, this striving towards as much karma as possible has become so prevalent on Reddit that it has gotten its own set of expressions and slang. For instance, "karma farming" is sometimes used to refer to people that generally do not contribute to the conversation and repost content purely for the sake of gaining as much karma as possible.

While there is much literature to be found on how to obtain the most karma, like the works by [20] and [19], there is not much out there on obtaining Reddit gildings. In this thesis, we will look specifically at Reddit gildings on comments. This thesis aims to create a better understanding of gildings distribution among comments of Reddit.

We define our research question:

> **Research Question 1:** Can we predict the receipt of community recognition for user-created Reddit comments?

We will be using a machine learning model to predict the receipt of community recognition. The community recognition will be measured in the form of gilded comments. For our machine learning model to function, we need a set of features. This leads us to our sole sub-question:

> **Research Question 2:** How can a descriptive analysis of large-scale Reddit comment data contribute to generating useful features for the predictive task set out in RQ1?

1.3 Thesis overview

In Section 2, we expand on Reddit's terminology and explain some standard terms used throughout this thesis. Then in Section 3, we look for related research. After that, we define our dataset and data transformation in Section 4. Then before we get to the experiments, we describe our approach to answering the research questions in 5. In Section 6, Experiments, we explain the used machine learner to answer Research Question 1. This model could then be applied to new comments to predict if that comment is likely to get gildings or not. To create this model, we use a variety of features based on the comment. To find the features, we look at the data and describe any features that would be helpful for our classification task.

2 Background & Terminology

Section 1 introduced us to some background and terminology that may prove beneficial to understanding the research question and goal of this thesis. This section expands on the foundations laid in the first section. It focuses on explaining more Reddit terminology that would not be common knowledge to non-Reddit users or non-social media users.

- **Upvotes & Downvotes** are are a way to show appreciation to other users. All users can grant either an upvote or a downvote on any post or comment on the platform, which can then be seen by the recipient and others browsing the platform.

- **Karma** is Reddit's way of showing appreciation for a post or comment. Karma is the net sum of upvotes minus downvotes on one's comment or post. Karma is an incentive for participation and maintaining some quality among Reddit content. Karma shows up as a number in front of every comment or post on Reddit. A comment starts with one karma, as Reddit automatically makes the user upvote their own comments. Once a comment goes below a certain threshold of karma, Reddit will collapse the comment. This means users will not see the comment by default but will have to manually open the comment to view its contents. In general, karma is useless, though some subreddits require a minimum amount of karma to be able to post or comment on said subreddit or make specific requests, like material or monetary requests, et cetera.

- **Gildings** are ways to distinguish a post or comment from the others in the thread. They are used as a form of "super upvote" on the site. Gildings come in three different types; Reddit Silver, Reddit Gold, and Reddit Platinum. Each has its different costs and benefits for the receiver. Originally only one version of gildings existed, Reddit Gold. On October 2018 Reddit implemented Reddit Silver and Reddit Platinum.

- **Reddit Premium** can be bought or given to a user through gildings. Premium grants the user access to the exclusive subreddit r/lounge and gives its users ad-free browsing on Reddit. Users with Reddit Premium get coins for each month of Premium, which can be used to purchase more gildings. It also shows a premium badge on their profile page. Reddit Premium costs $44,99 per year or $5,99 per month or can be obtained through the receipt of gildings.

- The **OP** stands for Original Poster. When Reddit users refer to the OP, they could either refer to the creator of the post they are commenting on or refer to a poster of a comment that two or more users are responding to.
- **Reposting** is the act of posting content again that was initially found elsewhere. Reposts are typically posts or comments that were initially valued positively, so they are posted again by someone else or the initial poster (called the OP). Reposts can occur in various ways; They could be from the same subreddit but posted a while later; They could be initially posted on a different subreddit (called crossposting or X-posting); or they could be from another social media platform, like Facebook or Twitter.

3 Related Work

Reddit is one of the biggest social media platforms globally, which logically comes with much research about the site. Predicting the receipt of community recognition for Reddit users, specifically through Reddit gildings instead of karma, is a topic that has not been widely researched to the best of our knowledge. Of course, much research has been done closely related to our topic where we can draw inspiration from. This work is related to other works that look into Reddit and other social media.

[17] looks into the process of karma farming. The goal of gaining karma has become so large that people resort to hated tactics to gain "useless" karma. In the same vein, [12] looks at the chasing of karma by Reddit users through the scope of gaming. They look at Reddit as a series of design choices where the aim of the game is earning karma. Research is also done in predicting user behaviour through machine learning similar to what we are doing in this thesis. [7] attempts to predict retweets as a popularity measure on Twitter posts. [2] uses classification tasks to predict helpfulness votes on product reviews on amazon.com. [18] look at the prediction of likes on Facebook posts. [6] from Cornell University has looked into predicting controversiality in Reddit comments based on a set of network features.

Closer to our experiments, research has been done in predicting appreciation behaviour in the form of karma. [5] tracks the upvoting behaviour of 186 users and then tries to predict what posts and comments get upvoted. [10] researches the influence of time, titles and subreddit on post popularity. While post popularity is not used, we can draw similarities between posts and comments. [4] looks at the influence of age and gender on karma success on Reddit. Works by [20], [19] [11] and [9] predict comment karma through different classifiers. We use the findings from these works in our experiments to find which features can benefit our research.

4 Data

In this section we look at the data which is used throughout these experiments. The data is collected through Reddit's API[2] by Pushshift[3]. Subsection 4.1 looks at which data is included, and the distinctions made for which data to include in our work. Subsection 4.2 looks at the data and the fields included in our data.

4.1 Data Scope

The data contains all comments placed on Reddit over a period from November 1st, 2018, till December 31st, 2018. The data includes public and restricted subreddits but no private subreddits. Public subreddits are subreddits where everyone can read and comment on posts. Restricted subreddits allow all users to read posts but only allow a select few people to comment and post. Private subreddits only allow some people to enter the subreddit and read its contents. Thus there is no means of accessing that data as the Reddit API does not release this data.

No other distinction is made with the data. The dataset contains any comment on any post on any subreddit. This includes NSFW (Not Safe For Work) content, posts that, for example, have an erotic nature. No distinction is made between languages. In total, the data contains 112,346,556 comments by 7,317,245 users across 158,869 subreddits. The dataset does not contain any information on the post or subreddit the comment was posted in. This includes features like the contents of the post, the amount of karma of the post, the subscriber count of the subreddit, et cetera.

[2] `https://www.reddit.com/dev/api/` *requested on 21th May 2021*

[3] `https://files.pushshift.io/reddit/comments/` *requested on 21th May 2021*

4.2 Data Description

A breakdown of the data obtained from the Reddit API is given in Table 1, containing a description and an example for each attribute. The table shows the attributes that will be used throughout this work. Some features are directly copied from the raw data, and some are adapted to be more effective for our experiments.

When a user is deleted (for violating guidelines, or else), their comments will remain in the dataset. However, any user-based data would be lost. This includes username, account age, et cetera.

5 Approach

In this section we look at how we will be attempting to answer the research questions set in Section 1.2. Section 5.1 explains how we will approach our primary research question, while Section 5.2 explains our approach for the sub-question.

5.1 Machine Learning Model

To answer our primary research question, we implement a machine learning algorithm to see if we can predict Reddit gildings on comments. Specifically, we choose to implement our classifier as a regression. To know how features influence the classifier, we have a look at the importance of every feature. Therefore, we choose a regression analysis because it allows us to gather the importance of the features. Importance is gathered through examining the model's coefficients.

While comments can get multiple gildings, we consider gilding to be a binary variable as 0.9% of comments only get gilded once. We choose to combine all types of gildings; gold, platinum and silver. This way, we neglect the differences in value and cost between the different gilding types. We set gilding to true if any of gildings.gid_1, gildings.gid_2 and gildings.gid_3 is true. This will be our response variable for our classification task. To accommodate for the binary response variable, we implement our classifier as a logistic regression. Then we define a set of features used to predict our response variable, 'gilded'. We use the 'lbfgs'-solver as it is the standard for logistic regressions and works for larger datasets such as ours. We also set the maximum number of iterations of the built-in optimisation algorithm to 1000. The logistic regression is implemented with scikit-learn [16] in Python. In the next section, we define a set of features that we will use.

Attribute	**Type**	**Example**	**Description**
`author_fullname`	string	t2_q2w17e	ID of the author of the comment
`subreddit_id`	string	t5_2qh1i	ID of the subreddit the comment is posted in.
`score`	int	4782	Karma received by the comment.
`link_id`	string	t3_9l9lc9	ID of the post that the comment is placed on.
`parent_id`	string	t1_e8r5v8y	If the comment is a reaction to another comment this will refer to that comment. If the comment is not a reaction to another comment this value will refer to the ID of the post that the comment is placed on, similar to the 'link_id' value.
`body`	string	"Thank you for the gold, kind stranger!"	Contents of the comment.
`gildings.gid_1`	int	3	Amount of Reddit Silver received by the comment.
`gildings.gid_2`	int	5	Amount of Reddit Gold received by the comment.
`gildings.gid_3`	int	0	Amount of Reddit Platinum received by the comment.
`created_utc`	int	1538685945	Time and date of when the comment was placed.
`is_submitter`	bool	true	Whether or not the author of the comment is also the submitter of the original post the comment is placed on.
`edited`	int/bool	1538693495	Time and date of the last edit. If the comment has not been edited, this will be set to false.
`stickied`	bool	false	Whether or not the comment has been stickied. This is generally done by moderators of the subreddit and causes the post to automatically be put on to the top.
`subreddit_type`	string	"restricted"	States whether the subreddit is set to 'public' or 'restricted'.
`author_created_utc`	int	1524912351	Time and date of the user creation. This will be set to `null`, if the account has been deleted.

Table 1: Attribute descriptions

5.2 Features

In section 6.1 we try to answer our sub-question by looking at our data to see if a set of useful features can be derived. The features are listed with an explanation of what the feature is and an approach of how the data point is reached using the variables stated in Table 1. We only gather features that are available at the time of the comment's submission. Information from after a comment has been posted could influence the likelihood of a comment getting gold and thus potentially leak information. Leaking can appear when the data used is partially based on or correlated to the target variable that we are trying to predict. The classifier could unfairly take advantage out of that data, which would result in a more accurate prediction than that is fair. An example, for instance, would be when we have two features, user gildings and user comment count. Then when the classifier looks at a comment and sees that the author has one gilding and only one comment posted, it could deduce that the comment it is currently looking at must be the comment that is also gilded. We also consider the results to be more interesting when features are limited to the time of posting. To limit the scope of this thesis, decidedly no lexical features are collected based on the content of the comment. Though, the content of the comment was still used to collect the comment length.

5.2.1 Feature categories

We define our final set of features. These features can be subdivided into categories: comment-based, subreddit-based, user-based, and post-based features. For each feature, we give an explanation and state our means of calculation in the approach, given the variables as stated in Table 1. Features were gathered through a mixture of descriptive analysis in Section 6.1, looking into previous research and using our own speculation about what features would have some importance to come up with features. In total, we collected 24 unique features for each comment.

Table 2: Features for the machine learning model

№	Feature	Explanation	Approach
–	***Comment-based*** *features are unique for every comment. Most features we use are comment-based. We define the following subreddit-based features for our project:*		

Table 2: Features for the machine learning model

№	Feature	Explanation	Approach
1	Time of day	Reddit has different visitation numbers depending on what time of day it is. A big portion of Reddit users is American, so we expect to see a peak in page views when Americans are awake. Americans constitute 46 per cent of all app installs[4]. This increase in page views might translate to more gildings on comments. On the flip side, it could also cause more comments, thus fewer gildings per comment. The demographic could also influence spending behaviours.	–We transform `created_utc` to a full date and then split off the hours. Data ranges from 0 to 23, depending on the hour of the day.
2	Day of the week	People typically have more free time on the weekend than they would be able to spend on Reddit. Though depending on the subreddit, people could also be using Reddit during work or even for work.	We transform `created_utc` to a day of the week. Data ranges from 0 to 6, depending on the hour of the day. 0 is Monday, 1 is Tuesday, et cetera.
3	Day of the month	Other than the time of the day and day of the week, we also define the day of the month. We also try the day of the month because people potentially give out more money when they got their pay for the month. Holidays may also influence subreddits like r/secretsanta and r/adventofcode. We do not have the month of the year as a feature, as our dataset only contains two months.	We transform `created_utc` to a day of the month. Data ranges from zero to 30, depending on the hour of the day.

[4] `https://sensortower.com/blog/reddit-app-install-record` *requested on April 14th 2021*

Table 2: Features for the machine learning model

№	Feature	Explanation	Approach
4	Is submitter	This is a boolean variable that is true when the author of the comment is also the submitter of the original post. Comments where the author is the same as the submitter have a unique distinction. Depending on which platform the user uses, they typically have a blue microphone icon or 'OP' next to the author's name, and the author may have their name in light blue.	A boolean variable. This feature is directly taken from `is_submitter`.
5	Comment length	The length of the comment in characters. Longer comments typically require more effort, which may be gilded faster. We see that some subreddits with minimum character counts have the highest gild-per-comment ratio.	An integer variable. We calculate the comment length by getting the number of characters of `body`.
6	Stickied	Stickied comments automatically show up at the top of a comment section. Stickied comments have their own icon and are green. On one hand, they are seen by more users browsing Reddit and could therefore get more gildings. On the other hand, gilding them will not get them more recognition as they are already at the top, which could cause that people might be less keen on gildings them.	A boolean variable. This feature is directly taken from `stickied`.
7	URL count	Comments can contain a link to an external site. These can be embedded into the text, which would then appear as blue. Reddit users may appreciate it when comments contain external sources of information. Having links in a comment typically shows that extra care has been taken, which could be rewarded more often.	An integer counter for the amount of times 'http:' or 'https:' appears inside `body`.

Table 2: Features for the machine learning model

№	Feature	Explanation	Approach
8	Bolded	Whether the comment contains bold formatting for the full or part of the comment. Using bold in the comment shows the extra effort made by the author, which could potentially increase the chance of gildings.	We check whether `body` contains any words or sentences surrounded by two asterisks.[5]
9	Italicized	Whether the comment contains italics formatting for the full or part of the comment.	We check whether `body` contains any words or sentences surrounded by asterisks on either side, but not two asterisks.[6]
10	Direct comment	A boolean variable that is true when the comment is posted directly to the post and false when the comment is a response to another comment.	To see if a comment is direct, we check if `parent_id` is starts with 't3_'. It would start with 't1_' if the comment is a response.
11	Length above average	Length of the comment in characters minus the average length of a comment on that subreddit. The usefulness of comment length might depend on the subreddit the comment is posted. Some subreddits generally have longer comments than others, so that long comments would be less exceptional. Having the comment length divided by the average comment length would potentially combat this.	To calculate the length above average, we first calculate the average length of a comment on each subreddit, defined by `subreddit_id`. This is done by counting the number of characters in `body`. Once we have gathered that, we subtract that by the length of the current comment, which is the same as our other feature `comment length`. If the length is below average, the resulting feature score would be negative.
12	Time since first comment	The time is in milliseconds after the first comments on the same post has been placed. If the comment is the first comment of the post, this will be set to zero. This feature is used as an alternative for a "time since post creation" feature. We do not have any information in our dataset about the post itself.	For each post, we note the time as stated in `created_utc` of the first comment on the post. When we collect the features we subtract the `created_utc` by this `created_utc` of the first comment of the post that we saved.

[5] The regex used is `(\*\*.*?\*\*)` .

[6] The regex used is `(^|[^*\\\\])\*[^*]+\*([^*\\\\]|$)` .

Table 2: Features for the machine learning model

№	Feature	Explanation	Approach
–	***User-based*** *features are unique for every user. So these features would show the same for any comment by a user. The idea behind user-based features is that comments between users can deliver differences in quality. We define the following user-based features for our project:*		
13	Account age	This shows the author's age in years at the time of the latest comment in our dataset. The account creation date could influence the results if we assume that older users generally have more experience on the platform and have a better understanding of the ins and outs of Reddit, thus being able to potentially obtain more gildings.	We transform `account_created_utc` to the year the user account was created. Then we subtract that from 2018 as that is the year of the latest date we have in our dataset. When the user account has been deleted at the time of collecting the data, we do not have this information. For these instances, we set the account age to zero.
14	User deleted	A boolean variable that states whether or not the account of the author of the comment has been deleted or not.	We check whether `author_created_utc` is equal to `null`. If it is, we set this feature to True, else to False.
15	Total karma	The total amount of karma gained by the user across the time frame. We detract the karma of the current comment from the total karma not to leak information accidentally. Because, karma is closely related to gildings. Thus if we know the karma of a comment, it could hamper the integrity of the model. This feature is set to zero in case a user is deleted.	An integer variable. The total karma is calculated by accruing all `score` values for each author. Authors are defined by `author_fullname`.
16	User comment count	The total number of comments made by the user within our time frame. This feature is set to 0 in case a user is deleted.	An integer variable. The comment count is calculated by increasing a variable by one for every comment from an author. Authors are defined by `author_fullname`.

Table 2: Features for the machine learning model

№	Feature	Explanation	Approach
17	Average karma per user	The average karma per comment by a user. Total karma divided by comment count. Suppose a user has a higher average karma score. In that case, it will increase the chance of a newer comment also getting more karma. This feature is set to 0 in case a user is deleted.	An integer variable. The average karma is calculated by dividing the `total karma` by the `user comment count`, both of which are other features we use and are stated above. To prevent data leakage, we remove the karma of the comment from the average score.
–	***Subreddit-based*** *features are unique for every subreddit. Some subreddits get more gilded on than others, so having some subreddit-based features could be relevant. So these features would show the same for any comment by a subreddit. We only have data on subreddits that is based on comments. So we can not use features like subreddit subscriber count, subreddit age, et cetera. We define the following subreddit-based features for our project:*		
18	Public	A boolean variable that states whether the subreddit that the comment is posted on is public or restricted. Restricted subreddits only allow whitelisted users to comment and post on them, but they do allow all users to visit the subreddit and view its contents. They generally have fewer comments, which could increase the gilding count per comment. Private subreddits are not included in our dataset as they only allow access to the subreddit for specific users.	Checks whether `subreddit_type` is equal to "public". If it is public, the variable is set to true.
19	Subreddit post count	Total number of posts on the subreddit within the time frame.	An integer variable. The `post count` is calculated by increasing a variable by one for every post on a subreddit. Posts are defined by `link_id` and subreddits by `subreddit_id`.

Table 2: Features for the machine learning model

№	Feature	Explanation	Approach
20	Subreddit comment count	The total number of comments across all posts on the subreddit within the time frame.	An integer variable. The `post count` is calculated by increasing a variable by one for every comment on a subreddit. Subreddits are defined by `subreddit_id`.
21	Comments per post	Average number of comments left on each post on the subreddit.	An integer variable. The `comments per post` is calculated by dividing the `post count` by the `subreddit comment count`, both of which are other features we use and are stated above.
22	Average karma per comment	Total karma gained in the subreddit divided by the total number of comments. Suppose comments in a subreddit get more karma than others in the same subreddit. In that case, they might also have a bigger chance of receiving gildings.	An integer variable. First, the total karma is calculated by accruing the `score` of all comments on the subreddit. Then we divide that by the `subreddit comment count`, which we also use as a feature. To prevent data leakage, we remove the karma of the comment from the average score.
23	Average karma of posts	Total karma gained in the subreddit divided by the total amount of posts.	An integer variable. First, the total karma is calculated by accruing the `score` of all comments on the subreddit. Then we divide that by the `subreddit post count`, which we also use as a feature.

– ***Post-based*** *features are unique for every post. These features would show the same for any comment on that post. The reasoning behind these features is that individual posts gain more acknowledgements than other posts by getting more karma, thus reaching higher on Reddit's pages and even on the front page or other top pages of Reddit. Our data does not contain actual information about the post itself, such as the post's contents or its creation date. We define the following post-based features for our project:*

Table 2: Features for the machine learning model

№	Feature	Explanation	Approach
23	Post comment count	The total number of comments on the post of the comment. This feature attempts to encapsulate post popularity.	The `post comment count` is calculated by increasing a variable by one for every comment on each post. Posts are defined by `link_id`.
24	Average karma	The average karma gained by a comment on a post. This feature attempts to encapsulate post popularity.	First the total karma is calculated by accruing the `score` of all comments for each `link_id`. Then we divide that by the `post count`, which we also use as a feature. To prevent data leakage, we remove the karma of the comment from the average score.

6 Experiments

The experiments are split up over our research questions. We will start by answering research question 2 in Section 6.1 and then our main research question in 6.2.

6.1 Descriptive analysis

To answer our second research question as stated in Section 1.2, we need to look for patterns by analysing our data. We use three different ways of processing the large-scale Reddit comment data to look for patterns: By looking at the top comments, users and subreddits. We can potentially use that to define features in Table 2. Subsection 6.1.1 examines the top posts within our time frame. Looking at the content of these posts can help us discover why these posts are gilding in great numbers. In Subsection 6.1.2 we look at the difference between subreddits. Which subreddits get the most gildings per comment posted and discover why that is the case. Subsection 6.1.3 looks at the top percentage of users on Reddit. We try to find out who these users are and possibly find any reasons for their popularity, such as them being real-life celebrities, moderators, et cetera. In this manner, we can find any patterns or differences in their Reddit usage compared to regular users.

6.1.1 Comment

Table 3 shows the top ten comments with the gilding count. The table shows the comment's author, the number of gildings the comment received and the comment itself or a snippet of the comment. The comments are shortened to fit within the table, but for our analysis, we look at the full comments.

Author	**Gildings**	**Comment**
ThatsBushLeague	92	"You don't. You try to get gold by being sneaky and then you end up with 74 silver."
giovanniversace	78	[deleted]
scienceftw	64	"Hey guys, that's my video! I will try to hop on later and answer some questions ..."
LazyIdiotofthe88	50	"I swear to god I've seen this before"
BrandonHawes13	49	"Three guys are walking through the woods when they find a lamp. One of them picks ..."
hugthebug	48	"I'm French, living near Lyon. This is the original comment. Everyone is mistaken : ... "
iamthatis	46	"Hey all, BAM! Surprise 1.4! Been working really hard to get this release out to you ..."
_ancora	46	"Why are people acting so dense about this? Kevin Hart told some jokes where the ..."
codythisguy	37	"Very true! And they got gold and silver. Ugh."
NotEvenEvan	37	"Reddit's new award system. You can award any comment or post you want by holding ..."

Table 3: Top ten highest comments within timeframe

Looking at the contents of the most gilded comments, we see a couple of things stand out. Often, massively gilded comments share some new insight into current world events by people who have knowledge or direct involvement about the topics at hand. The comment by u/giovanniversace has been deleted from the platform. Responses to the comment suggest that the author awarded gildings to himself through another account though this can not be confirmed. Comments in Table 3 are typically reasonably long, like the comment by u/hugthebug and u/BrandonHawes13; or they are very short, like the comment by u/LazyIdiotofthe88. The longer comments all contain URLs or bold/italic text. We also notice that the poster is the same as the submitter of the original post for the comments by u/iamthatis, u/_ancora and u/codythisguy. Comments by u/ThatsBushLeague, u/codythisguy and u/NotEvenEvan directly talk about gaining Reddit gildings or offer to give away gildings themselves. Based on this, we think the following features could be of use: 4, 5, 7, 8 and 9. The numbers refer to the first column (№) of Table 2.

6.1.2 Subreddit

Table 4 shows the top 20 subreddits that have the highest gild-per-comment ratio. which is a snippet of the data we looked at. The full top 50 can be found in Appendix A. We also use the 20 most gilded subreddits, which can also be found in Table 10 in the appendix A.

Subreddit	Comments	Gilded	Gild/Comment
secretsanta	64,075	3561	0.055575
apolloapp	19,915	230	0.011549
intel	29,907	314	0.010499
Bitcoin	190,182	1719	0.009039
HighQualityGifs	25,891	162	0.006257
picrequests	10,149	61	0.006010
photoshopbattles	67,276	360	0.005351
bitcoincashSV	10,475	53	0.005060
instant_regret	104,235	463	0.004442
PhotoshopRequest	34,976	143	0.004089
AskHistorians	32,986	134	0.004062
GamePhysics	21,481	85	0.003957
denvernuggets	48,472	179	0.003693
self	47,810	166	0.003472
Dreams	11,367	39	0.003431
adventofcode	14,003	48	0.003428
WritingPrompts	98,271	325	0.003307
Detroit	17,553	55	0.003133
HumansBeingBros	113,061	348	0.003078
tifu	175,825	535	0.003043

Table 4: Subreddits with the highest gildings-per-comment ratio within our timeframe. Subreddits below 10,000 comments in our dataset have been filtered out.

Looking at the top 20 most gilded subreddits in appendix A, we see one subreddit with by far the most gildings, r/AskReddit. Interestingly, we do not see that subreddit in Table 4 for subreddits with the highest gildings-per-comment ratio. Subreddits with a high gild count do not necessarily mean there is a higher chance of obtaining gildings there. Instead, we see many subreddits that maintain a minimum comment length (r/AskHistorians, r/WritingPrompts). Comments that do not reach a certain threshold of characters would be automatically deleted. This ensures a level of quality on all comments on the subreddit. We also see some seasonally bound subreddits (r/secretsanta, r/adventofcode). We presume these would not appear as high during other months of the year. Subreddits where people respond with an image of some kind also do well (r/HighQualityGifs, r/picrequests, r/photoshopbattles, r/PhotoshopRequest). These comments could be picked out by checking for URL links in the comment. We define the following features from this in Table 2: subreddit specific features, like №18, №19, №20, №21. We also find the №3, №7 and №10. The numbers refer to the first column.

6.1.3 Users

We look at the top 50 users with the highest gildings-per-comment ratio across our timeframe. Table 5 shows a snippet of the complete list, the rest of which can be found in appendix B. We have filtered out users that posted less than ten comments in our timeframe to limit the prevalence of users that got a high ratio mostly because they only posted few comments. Because of the sheer number of users on the platform, there are bound to be some that got lucky. These users are not as valuable for analysis. A list of the most gilded users can also be found in the appendix.

Looking at the list in Table 5 and Table 12 from Appendix B, we notice many top users write longer posts on average. We also see that people on the list frequent one or a few subreddits, be it r/AskReddit, r/news, r/photoshopbattles or something else. Top users like u/ThatsBushLeague, r/Lonnbeimnech, r/gchamblee give up to date information on the news or politics. Others like u/_vargas_, u/sweet_fatal_jesus and u/GuyWithRealFacts write stories or give out facts. We also see some real-life famous people entering the top like u/bernie-sanders, u/_scienceftw_ and u/MacaulayCulkinAMA (Bernie Sanders, Mark Rober and Macaulay Culkin). Other users that gain much gilding are poem writers, like u/Poem_for_your_sprog and u/ SchnoodleDoodleDo. Some of these are difficult to turn into features as there is, for example, no hard definable aspect that makes someone a real-life celebrity. This does show that there are some users that are more recognisable than others. We can potentially find these users with some user-specific features. We define the following features from Table 2: User-specific features, like №15, №16 and №17. Another feature we can define is №5.

Author	Comments	Gildings	Gild/Comment
giovanniversace	12	81	6.75
ThatsBushLeague	23	93	4.04
Poem_for_your_sprog	97	214	2.20
vargas	22	38	1.72
Seano151	14	22	1.57
CellDood	14	20	1.42
St0pX	23	31	1.34
TheJellyTruck	11	13	1.18
Spuffdozer	12	14	1.16
sadfvliugsedfvliugsa	30	32	1.06
Anikdote	24	25	1.04
Allergics	11	10	0.90
Halon_1211	17	15	0.88
thicknatural	13	11	0.84
Cavinhuntsman	10	8	0.8
TBytemaster	10	8	0.8
dimfresh	16	12	0.75
Lonnbeimnech	12	9	0.75
GuyWithRealFacts	29	21	0.72
ShaunFosmark	19	13	0.68

Table 5: Top 20 Reddit users sorted by highest gild-per-comment ratio. Users with less that ten comments in our dataset have been filtered out.

6.2 Predictive analysis

The community recognition, as stated in the research question, will be measured through gilded comments. We first gather the importance of the features we have gathered in Subsection 5.2.1. Then we determine the effectiveness of our model through different examination techniques.

6.2.1 Experimental Setup

Our dataset, as described in Section 4, is not optimal to use for machine learning models. Because the data is highly skewed towards non-gilded comments, we would need to input an enormous number of comments into our machine learning algorithm to get statistically relevant results. To allow us to process our data in a reasonable time, we filtered out all comments that received one or more gildings. The total comments are brought down to 115,951 comments by 95,005 users across 6,554 subreddits with this filter. This means that about one in 969 comments gets awarded something, showing the exclusivity of getting any comment gilded.

To train and test our model, we combine this dataset containing all gilded comments with equally as many randomly selected non-gilded comments. This brings the dataset to a total of 231,902 comments. We make a 50%/50% split in gilded and non-gilded comments so that there are enough gilded comments in the dataset for the model to analyse. The dataset is then split into 70% training data and 30% testing data.

6.2.2 Performance Metrics

To evaluate our machine learning model, we use a set of performance metrics. All performance metrics are taken from the book "Regression Analysis by Example"[1] and implemented using sklearn.metrics [16]. First, we calculate the accuracy score. The accuracy score states the percentage of predictions that our machine learning algorithm got correct. 100% would mean it predicted everything correct, and 0% would mean it predicts everything wrong. Other than the accuracy score, we also compute the F1 score. The F1 score is the weighted average between the precision score and the recall score. In our case, the precision score states: Of the number of gilded comments we predicted, how many are actually gilded? The recall score asks a slightly different question: Of the amount of actual gilded comments, how many did we actually predict?

Secondly, we generate the ROC-curve and calculate the AUC. The ROC-curve, which stands for Receiver Operating Characteristic, shows the True Positive Rate plotted against the False Positive Rate at different threshold levels. The steeper the curve (or the closer to the top left), the more accurate the predictions would be. A straight line from the bottom left to the top right is called the no-skill prediction line and would suggest completely random predictions. The Area Under the Curve, or AUC, is the area percentage that is to the right and below the line. The higher the AUC, the more accurate the machine learning algorithm.

Lastly, we generate the confusion matrix. The confusion matrix is a table that places the predictions against their actual values. This gets us the True Negative (TN), False Positive (FP), False Negative (FN) and True Positive (TP) in that order. In our case, we want as high TN and TP as possible and as low FN and TP as possible.

6.2.3 Features

Table 6 shows the importance of our features on response variable 'gilded'. We obtain the importance scores by getting the coefficients found for each input variable. The higher the absolute value is 0 (positively or negatively), the more influence that feature has on the outcome. A positive score indicates that as the predictor variable increases, the response variable also increases. On the flip side, a negative score indicates that as the predictor variable increases, the response variable decreases. The importance score means how much prevalence the algorithm should give that feature in deciding the resulting prediction. For example, an increase in the average user karma by one unit increases the likelihood of being the response variable 'gilded' by a factor of 0.00482 if the other features do not change.

The results show that only a few features have a notable influence on the machine learning model. The features; average user karma per comment, subreddit comment per post, length and length above average have the most impact on our results. Other features do not have nearly as much impact as those, with some even seemingly not having any impact at all. That does not mean that those features do not correlate whatsoever with our response variable. It could be that another feature better encompasses what we are trying to test with that feature. Hence, the model chooses to look at that other feature instead. For instance, Average user karma might diminish the importance of total user karma. The categories are also included as a column.

6.3 Results

Table 7 displays the logistic regression results through the performance metrics defined in 6.2.2. Other than the accuracy and F1 score, the precision and recall scores are also shown. The classifier has a higher precision score than a recall score. This suggests that the model is good at selecting gilded comments as opposed to selecting non-gilded comments, but also misses relatively many gilded comments. We see that all performance metrics waver around 60-70%, with the Accuracy score being the highest. A random classifier would end up with 50% accuracy.

	Score
Accuracy	0.6712
Precision	0.6941
Recall	0.6043
F1	0.6461

Table 7: Logistic regression performance

Figure 3 shows the ROC-curve of our model. The AUC is 0.730. Table 8 shows a confusion matrix for our model. From the confusion matrix, we find that our model has more false negatives than false positives. This means that the model, combined with the optimal decision boundary that the model has chosen, finds it harder to accurately predict gilded comments than non-gilded comments.

Feature	**Importance**	**Category**
user_avg_karma	0.00482	user-based
subreddit_comment/post	-0.00472	subreddit-based
length_above_avg	-0.00318	comment-based
length	-0.00193	comment-based
post_average_karma	0.00129	post-based
subreddit_avg_karma/post	0.00073	subreddit-based
monthday	-0.00038	comment-based
hour	-0.00036	comment-based
subreddit_avg_karma/comment	0.00018	subreddit-based
weekday	-0.00010	comment-based
post_comment_count	0.00004	post-based
public	-0.00003	subreddit-based
user_age	0.00002	post-based
user_karma	0.00001	user-based
is_submitter	-0.00001	comment-based
user_comment_count	-0.00001	user-based
user_deleted	-0.00001	user-based
URL_count	0.00000	comment-based
italics	0.00000	comment-based
subreddit_comment_count	0.00000	subreddit-based
time_since_first_comment	0.00000	comment-based
bold	-0.00000	comment-based
direct_comment	-0.00000	comment-based
stickied	-0.00000	comment-based
subreddit_post_count	-0.00000	subreddit-based

Table 6: Importance per feature, sorted by importance

6.4 Discussion

Both features about the length of comments have significance to the machine learning model. Interestingly, on the whole, it is a negative importance score, meaning shorter comments do better in getting gildings. This suggests that people do not want to read overly long comments and prefer short (potentially witty and straight-to-the-point) comments. It could be a negative correlation in some cases but a positive one in others. For example, funny comments are short, but informative comments are long. Subreddits where there is a high discussion rate with a lot of karma given (`subreddit_avg_karma/post`) turn out to play a significant factor. The amount of karma per user (`user_avg_karma`) is another feature that ended up with a strong importance score, higher than any of the other features we created. This would make it seem that some users know their way to a gilded comment better than others. In general, the

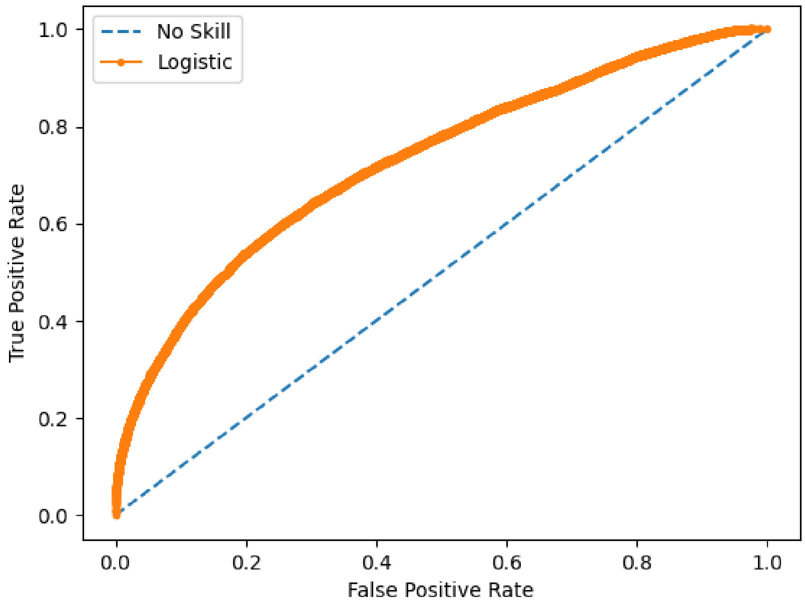

Figure 3: ROC-curve of the machine learning model. The orange line is our model, and the blue line is a hypothetical no-skill prediction line.

		True class		
		Positive	Negative	Total
Predicted class	Positive	$17,207$	$6,134$	$23,341$
	Negative	$9,116$	$13,924$	$23,040$
	Total	$26,323$	$20,058$	$46,381$

Table 8: Confusion matrix of the machine learning model

features based on karma averages tend to score highly. This seems logical as there exists a strong correlation between karma and gildings. Comments that are highly upvoted tend to get seen more thus have more chance to get gilded. Based on our dataset, gilded comments obtain about 659 karma on average, while all comments, including gilded and non-gilded comments, gain only seven karma. Some of the features that have shown no correlation may be because the result is mostly the same for most comments, except for a minority. This is the case for stickied comments, in which case comments are mostly not stickied; URL count, as most comments do not contain any URL links; bolded and italicised. Because of this, the data to make an accurate evaluation of these features by the machine learning model becomes slim. Most of these are binary values.

In a broader view, all our feature subcategories have at least some features that have relatively much importance. Our post-based features have the fewest importance. This could mean that post information has less influence on receipt of gilding than the other categories. Maybe because there are many posts on the platform, each on its own subreddit with its own community, making it harder to get a clear grasp of its features' influence.

To answer the research question, the results do show that we can predict the receipt of community recognition for user-created Reddit comments. Looking at the performance metric scores, there is still much room for improvement. The AUC is 23% higher than if one were to randomly assign a prediction to each comment about whether or not it is gilded. However, there is a limit to how accurate the prediction can be, as gildings can happen on any comment without rhyme or reason. Gold is given by a single person without any guidelines, making it impossible to predict with total accuracy. Getting gold is extremely rare, so even though a comment ticks off all the boxes for the highest chance of getting gold, it might not get it at all. The fact that the confusion matrix has a relatively high false-negative count with the optimal decision boundary, further shows that some gilded comments are tough to predict.

7 Conclusion & Outlook

In this work, we predicted the receipt of Reddit gildings on comments through machine learning. By doing so, we evaluate the predictability and check how the chances of getting gilded can be improved.

Our findings show that there is a sense of predictability in receipt of gildings as we found an accuracy score of 0.67, a F1-score of 0.65 and an AUC of

0.73. The features that have the a lot of influence on our machine learning model are the length based features. The number of comments on average per post on the subreddit is also of high importance, though the most important feature is the average karma of the author.

Some improvements can be made to better the results. Future research on this topic could look further into selecting more or better features to improve the performance and results. Most features considered are not related to the contents of the comment itself. One could also dive further into the lexical side by including more linguistic features, for example, looking for specific words and phrases or tone and sentiment. This could be done through text mining. The use of different classifiers could also potentially result in higher performance. Finally, with our limitations, we only studied comments posted across a period of two months. Results could very well differ when viewed from a more extensive period of time, like a year. For future work, it would be desirable to look at a broader time range.

There are also a few limitations that could have influenced the results. While we did our best to avoid it, there would still be a chance for data leakage to occur with the current feature set. As some features may be generated using data from the current comment that give an unfair advantage for the algorithm. All features that may be subject to this have been adapted to minimise this problem. Another limitation is the 50:50 rebalancing that was used to train and test our model. Preferably, no rebalancing would occur.

Future work could expand upon our research by look into the different types of Reddit gildings. We considered all gilding types to be of the same value, but the rewards and costs differ depending on the type of gilding. People might look for different features when rewarding a comment with Reddit platinum as opposed to Reddit silver. Next to comment data, having post data from the same time range would be a valuable expansion to this work.

References

[1] Samprit Chatterjee and Ali S Hadi. *Regression analysis by example.* John Wiley & Sons, 2015.

[2] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web*, pages 141–150, 2009.

[3] Jenny L Davis and Timothy Graham. Emotional consequences and attention rewards: the social effects of ratings on reddit. *Information, Communication & Society*, 24(5):649–666, 2021.

[4] S Craig Finlay. Age and gender in reddit commenting and success. *Journal of Information Science Theory and Practice*, 2014.

[5] Maria Glenski and Tim Weninger. Predicting user-interactions on reddit. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 609–612, 2017.

[6] Jack Hessel and Lillian Lee. Something's brewing! early prediction of controversy-causing posts from discussion features. *arXiv preprint arXiv:1904.07372*, 2019.

[7] Liangjie Hong, Ovidiu Dan, and Brian D Davison. Predicting popular messages in twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 57–58, 2011.

[8] Benjamin D Horne, Sibel Adali, and Sujoy Sikdar. Identifying the social signals that drive online discussions: A case study of reddit communities. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–9. IEEE, 2017.

[9] Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. Talking to the crowd: What do people react to in online discussions? *arXiv preprint arXiv:1507.02205*, 2015.

[10] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. What's in a name? understanding the interplay between titles, content, and communities in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, 2013.

[11] Daria Lamberson, Leo Martel, and Simon Zheng. Hacking the hivemind: Predicting comment karma on internet forums. Master's thesis, Stanford University, 2014.

[12] Adrienne Massanari. Playful participatory culture: Learning from reddit. *AoIR Selected Papers of Internet Research*, 2013.

[13] Alexey N Medvedev, Renaud Lambiotte, and Jean-Charles Delvenne. The anatomy of reddit: An overview of academic research. In *Dynamics on and of Complex Networks*, pages 183–204. Springer, 2017.

[14] Carrie Moore and Lisa Chuang. Redditors revealed: Motivational factors of the reddit community. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[15] Daniel Moyer, Samuel Carson, Thayne Dye, Richard Carson, and David Goldbaum. Determining the influence of reddit posts on wikipedia pageviews. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, 2015.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[17] Annika Richterich. 'karma, precious karma!'karmawhoring on reddit and the front page's econometrisation. *Journal of Peer Production*, 4(1), 2014.

[18] Ferran Sabate, Jasmina Berbegal-Mirabent, Antonio Cañabate, and Philipp R Lebherz. Factors influencing popularity of branded content in facebook fan pages. *European Management Journal*, 32(6):1001–1011, 2014.

[19] Jordan Segall and Alex Zamoshchin. Predicting reddit post popularity. *nd): n. pag. Stanford University*, 2012.

[20] Tim Weninger. An exploration of submissions and discussions in social news: Mining collective intelligence of reddit. *Social Network Analysis and Mining*, 4(1):173, 2014.

Appendices

A Subreddits

Table 9: Top 20 most gilded subreddits

Subreddit	Comments	Gilded	Gild/Comment
AskReddit	11,618,595	17,446	0.001502
politics	4,098,549	5081	0.001240
secretsanta	64,075	3561	0.055575
funny	2,105,966	3036	0.001442
pics	1,259,641	2238	0.001777
worldnews	1,706,878	1776	0.001040
Bitcoin	190,182	1719	0.009039
gaming	1,713,059	1688	0.000985
aww	1,010,752	1669	0.001651
news	1,377,180	1641	0.001192
nba	2,034,564	1595	0.000784
videos	803,684	1590	0.001978
todayilearned	1,046,192	1525	0.001458
gifs	732,705	1492	0.002036
nfl	2,704,621	1471	0.000544
Showerthoughts	1,090,262	895	0.000821
BlackPeopleTwitter	555,222	851	0.001533
WTF	414,798	750	0.001808
interestingasfuck	397,460	672	0.001691
mildlyinteresting	599,693	661	0.001102

Table 10: Extended table of the top 50 most gilded subreddits with a minimum of 10,000 comments per subreddit.

Subreddit	Comments	Gilded	Gild/Comment
secretsanta	64,075	3561	0.055575
apolloapp	19,915	230	0.011549
intel	29,907	314	0.010499
Bitcoin	190,182	1719	0.009039
HighQualityGifs	25,891	162	0.006257
picrequests	10,149	61	0.006010

photoshopbattles	67,276	360	0.005351
bitcoincashSV	10,475	53	0.005060
instant_regret	104,235	463	0.004442
PhotoshopRequest	34,976	143	0.004089
AskHistorians	32,986	134	0.004062
GamePhysics	21,481	85	0.003957
denvernuggets	48,472	179	0.003693
self	47,810	166	0.003472
Dreams	11,367	39	0.003431
adventofcode	14,003	48	0.003428
WritingPrompts	98,271	325	0.003307
Detroit	17,553	55	0.003133
HumansBeingBros	113,061	348	0.003078
tifu	175,825	535	0.003043
findareddit	22,727	67	0.002948
Frat	11,609	34	0.002928
IAmA	141,140	413	0.002926
DataHoarder	38,663	103	0.002664
Wetshaving	14,056	37	0.002632
NASCAR	138,430	358	0.002586
HadToHurt	20,309	51	0.002511
rickandmorty	26,789	67	0.002501
mylittlepony	14,963	37	0.002472
OutOfTheLoop	88,448	218	0.002464
jailbreak	86,366	211	0.002443
SanJose	11,659	28	0.002401
whitepeoplegifs	33,652	80	0.002377
texas	53,919	125	0.002318
reactiongifs	47,565	107	0.002249
cringe	53,502	120	0.002242
YouShouldKnow	25,612	57	0.002225
MadeMeSmile	104,171	228	0.002188
help	14,239	31	0.002177
Wellthatsucks	142,196	299	0.002102
translator	23,966	49	0.002044
gifs	732,705	1492	0.002036
WatchPeopleDieInside	116,232	230	0.001978
videos	803,684	1590	0.001978
KidsAreFuckingStupid	84,207	164	0.001947
Unexpected	160,641	311	0.001935
Romania	136,467	264	0.001934
nevertellmetheodds	30,169	58	0.001922

GetMotivated	62,808	120	0.001910

B Users

Table 11: Extended table of the top 50 most gilded users per comment with a minimum of 10 comments per user.

Author	**Comments**	**Gildings**	**Gild/Comment**
giovanniversace	12	81	6.750000
ThatsBushLeague	23	93	4.043478
Poem_for_your_sprog	97	214	2.206185
vargas	22	38	1.727272
Seano151	14	22	1.571428
CellDood	14	20	1.428571
St0pX	23	31	1.347826
TheJellyTruck	11	13	1.181818
Spuffdozer	12	14	1.166667
sadfvliugsedfvliugsa	30	32	1.066667
Anikdote	24	25	1.041667
Allergics	11	10	0.909090
Halon_1211	17	15	0.882353
thicknatural	13	11	0.846153
Cavinhuntsman	10	8	0.800000
TBytemaster	10	8	0.800000
dimfresh	16	12	0.750000
Lonnbeimnech	12	9	0.750000
GuyWithRealFacts	29	21	0.724137
ShaunFosmark	19	13	0.684210
craighamnett	12	8	0.666667
bernie-sanders	14	9	0.642857
ilikeyourjacket	11	7	0.636363
Sweet_Fetal_Jesus	11	7	0.636363
evilbarbiedoll	13	8	0.615384
Ibringturtles	36	22	0.611111
ErikThe	10	6	0.600000
slightlytense	10	6	0.600000
gchamblee	52	31	0.596153
Rasputins_Moms_Anus	48	28	0.583333
EatYourTartOut	14	8	0.571428
dcx666	20	11	0.550000
h8ed-program	11	6	0.545454

clit-eastwould	15	8	0.533333
Bloosuga	19	10	0.526315
Jficek34	27	14	0.518518
deffsight	22	11	0.500000
jacob_the_snacob	20	10	0.500000
DrWankalot	20	10	0.500000
Steak_M8	12	6	0.500000
TooManyGatsbys	12	6	0.500000
DoctorWhoodlum	10	5	0.500000
vigorous_cottage	10	5	0.500000
shawnrai	10	5	0.500000
ChristianSgt	37	18	0.486486
Andysgotgame	36	17	0.472222
July_Sandwich	15	7	0.466667
MacaulayCulkinAMA	15	7	0.466667
mossyfox	13	6	0.461538
aloofsavior	13	6	0.461538

Table 12: Top 20 most gilded users.

Author	Comments	Gildings	Gild/Comment
Poem_for_your_sprog	97	214	2.206185
SchnoodleDoodleDo	443	194	0.437923
TooShiftyForYou	3725	132	0.035436
PoppinKREAM	432	129	0.298611
ThatsBushLeague	23	93	4.043478
giovanniversace	12	81	6.750000
hugthebug	664	80	0.120481
Portarossa	699	70	0.100143
iamthatis	557	66	0.118491
AnthonyChristopher	146	57	0.390410
Sim888	1294	53	0.040958
LazyIdiotofthe88	229	50	0.218340
BrandonHawes13	412	49	0.118932
AutoModerator	2,534,494	47	0.000018
_ancora	141	46	0.326241
codythisguy	332	46	0.138554
slakmehl	1644	41	0.024939
vargas	22	38	1.727272
NotEvenEvan	108	37	0.342592