Universitetet
i Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study programme/specialisation:
Computer Science

Spring / ~~#######~~semester, 20.19.

Open/~~#########~~

Author:
Fredrik Wigsnes

....Fredrik Wigsnes....
(signature of author)

Programme coordinator:

Supervisor(s):
Vinay Jayarama Setty

Title of master's thesis:

Predicting popularity of Reddit posts using machine learning

Credits:

Keywords:
Machine learning

Number of pages: ....56....

+ supplemental material/other: ....0....

Stavanger,....15/06/2019....
date/year

Title page for Master's Thesis
Faculty of Science and Technology

University of Stavanger

Faculty of Science and Technology
Department of Electrical Engineering and Computer Science

Predicting popularity of Reddit posts using machine learning

Master's Thesis in Computer Science
by

Fredrik Wigsnes

Internal Supervisors

Vinay Jayarama Setty

June 14, 2019

“Success is not final, failure is not fatal: it is the courage to continue that counts.”

-Winston Churchill

Abstract

Using data from the social network Reddit, we see if there are ways to predict if a submission will gain popularity, going into detail in components of a Reddit post, and try to determine if it will be successful. Analyzing the data, we see multiple factors that have impacts on what can help achieve a successful post. Timing is a big one, but mostly initial interactions with the submissions can make or break it, which can often be exploited by users upvoting content with fake accounts. Then using machine learning to see if it is possible to predict how many upvotes a submission will achieve. Random forest model achieved the best results with a mean absolute error close to 500, but it changes based on which subreddit. The findings in this thesis can help show the weakness in social media networks, relying on a minority of users the control of what will get popular, and show what elements have an impact on the number of upvotes.

Acknowledgements

I would like to thank Vinay Setty for helping me get through this master's thesis, and guiding me back on track every time I have diverged.

Contents

Abbreviations

TF-IDF	**T**erm **F**requency-**I**ndex **D**ocument **F**requency
SD	**S**tandard **D**eviations
MSE	**M**ean **S**quared **E**rror
RMSE	**R**oot **M**ean **S**quared **E**rror
RMSProp	**R**oot **M**ean **S**quare **P**ropagation

Chapter 1

Introduction

In today's society, the internet is where people spend most of their time, reading news, watching YouTube videos, or socializing with friends and stranger on social networks. On average, internet users are spending more than 2 hours on social networks each day[1] and making it the perfect arena for businesses and companies to acquire peoples time and attention. About two-thirds of Americans get parts of their news from social media[2], with four in ten getting their news from Facebook.

Receiving news from a social network would not have been problematic if it was not for fake news and algorithms keeping users on the site as long as long as possible. Old ideas are resurfacing and manifesting them self in groups of people because of it. Ideas like the flat earth movement or that vaccine's cause autism, have both been growing in the last years.

A big reason these ideas are resurfacing are the algorithms that were created to increase the time spent on social networks. The algorithm was intended to deliver similar content for example if someone is a dog person and like looking at pictures of dogs, they will show more pictures of dogs and fewer pictures of cats to keep them on the site. When this same algorithm work on news aggregation based on ones believes, these ideas that are hard to disprove can manifest them self in groups of people and will create echo chambers[3]. Moreover, it gets even worse with fake news networks exploiting this algorithm, and generating stories that will fit the narrative of individual believes, generating revenue from advertising.

1.1 Manipulation

As mentioned above, there are fake news networks that take advantage of the algorithms and try to make as much profit as possible from it. However, it is not just businesses that try to take advantage of the platform. Countries like China and Russia are experts in using social media and the internet as a tool to control their narrative and people, even taking it as far as disrupting other countries elections, creating chaos and disarray.

Manipulations, in today's society, should not be looked upon lightly. It exists all over the interwebs with fake accounts distributing fake news, and creating conflict to stall progress. It is a weapon meant to disorganize and create conflict and will continue to grow and be used as a weapon in the future to come. NATO's Secretary general Jens Stoltenberg said this in a speech to the US Congress:

> "We have only just seen the beginning of the threats in cyberspace. Artificial intelligence, quantum computing, and big data could change the nature of conflict more fundamentally than the Industrial Revolution."[4]
>
> –Jens Stoltenberg

Misinformation is happening all over, and users have to be more cautious when they read content on the internet as ever before.

Renee DiResta is the Director of Research at New Knowledge and a Mozilla Fellow in Media, Misinformation, and Trust. She studies the effect of misinformation on social media, how it is curated, and how it is used as propaganda to hurt the bottom line. She has spent two years looking at data from popular social media networks and found out that there only needs to be a handful of users to game the system. It most often is done to generate user interactions and get ad revenue from it, but it also is propaganda, trolls, terrorist organizations, and state intelligence services that are actively trying to manipulate people. What they found out is that these entities actively roam the big five social networks, Facebook, Twitter, Instagram, Youtube, and Reddit. These are the social networks most people are on, so it makes it easy for the attackers to only have to deal with these sites. In 2016 Renee and her team saw a lot of fake accounts emerging during the presidential election, and most often they are there only to facilitate polarisation.

Fake account is a big problem on all social networks, Facebook deletes over 1 million fake accounts a day.[5] There are sites that one can go and purchase accounts for specific use cases ranging in age and account activity. One can be so specific and get a ten-year-old account of a 30 year something Swedish female with an extensive network of friends; it is possible to request such an account from these vendors.[6]

Sebastian Bay who is a Senior Expert at NATO Strategic Communications Center of Excellence wrote a report called The black market for social media manipulation[7] where it goes into detail how these vendors that sell fake account works and how they use it. One way is what they call meta manipulation or artificial lift, where they inflate the view count to trick the algorithm of social media platforms into making their content trend on that platform. Reddit is profoundly affected by meta manipulation, and most often, it is corporations trying to reach users for gain in revenue, there is even a subreddit all about catching these obvious submissions called HailCorporate.

1.2 Motivation

A lot of what was mentioned above should instantly sound an alarm and make people understand that this is a big problem that will only become more problematic as chatbots, user tools, and automation gets better. Knowing that organizations and intelligence services are actively working on how to manipulate social network users, is a big motivator to figure out how to do this before them, and make it public so that it can either be patched or learned to spot and not be taken advantage of by these individuals.

1.3 Problem Definition

This paper will take a look at the social network site called Reddit, and see if there are ways to take advantage and find patterns that can lead to a higher chance at a successful submission. The reason Reddit was selected to gather data from is that it is conveniently broken into what they call subreddits. Subreddits are groups of niche subjects examples such as news, gaming, and politics. Making it simple to differentiate what type of content the submission is, and then use the upvotes a post receives to measure its popularity. Starting by looking at Reddit as a whole and later dive into single subreddits like news and try to predict with the use of machine learning.

1.4 Use Cases

Being able to find out if a news article will be popular is worth a lot to a news agency, but as mentioned above, there are people out there ready to abuse the system. The best use case is to combat the abuse and help with spreading news from all different viewpoints to fix the problems of echo chambers. Also, help find flaws that can be exploited and fix them.

1.5 Challenges

When displaying news articles to users it would be most desirable that it was ranked by quality and not popularity, but the way reddit works popularity is king and it is the users that decide what will be at the top. Popularity is often not the same as quality, and popularity is not as easy to predict as it can change based on recent events. The news cycle has become constant and it often seems like there is no time to dwell on the recent news.

1.6 Outline

Start by going trough related works and see if there are similarities and information that can help with understanding how reddit and social media sites work. Begin to gather data from reddit and structuring it to make it easy to use. Before trying any machine learning look at the data and see if there is interesting patterns and useful elements that can be used in machine learning. Start with some simple models of machine learning and see if it is possible to predict any values, and later try more advanced forms of machine learning algorithms like neural networks.

Chapter 2

Related Work

There are a lot of similar works that dive into how Reddit works, and what makes a popular post reach the front page. The paper called "Popularity and Quality in Social News Aggregators: A Study of Reddit and Hacker News"[8] goes into detail how popularity and quality are not always in agreement. It is a big difference in predicting popularity vs. quality. What might be popular today is most likely not as popular tomorrow. It changes every day based on what is happening in the media, and or what is trending at the moment. Posts that are better quality is not always the ones that make it to the top, and lesser quality posts will. On youtube, creators are trying to predict how the trending tab works. If a video reaches trending, it can receive about 100 times more views than what it would have received otherwise. Most content on social media is short-lived. To predict if a youtube video will hit the trending tab, one would not look at how many views it got during its first week but look at the first hour of posting. It is not similar to the stock market where it keeps going; most of the views of a video are made the first couple of days.

Social media also experience what is called popularity bias. Popularity bias is when a post has become so popular that it catches a wave and continues to grow in popularity while shadowing for other posts. Often other posts that are of better quality. When a post reaches high popularity often it will be shown to even more users and gain even more popularity, as for in Reddit's case if a post reaches status hot, it will be shown to about 50 times more users, and keep growing.

Even though popularity not always equal quality, the paper concludes that Reddit and Hacker news are impressively good at promoting quality posts. Most of the popular post was of higher quality. The same conclusion does this paper make: "Popularity Dynamics and Intrinsic Quality in Reddit and Hacker News"[8].

Another study called "Random Voting Effects in Social-Digital Spaces: A case study of Reddit Post Submissions"[9] looks at the effect on early voting and how it affects the submission later on. By randomly upvoting and downvoting new submissions and seeing if the average posts that received an upvote compared to a downvote were different. The result shows that a positive treatment increased the final rating of submission by 11.02% on average. It also increased the probability to receive a high rating ($\geq$2000) by 24.6%. A negative treatment also had a negative effect of 5.15% on average.

Having such a significant increase in upvotes by positively alter the submission early on is hugely beneficial to a manipulator, and Reddit is highly susceptible to this as described in the introduction.

A paper called "How does the front page of the Internet behave? Readability, emoticon use, and links on Reddit"[10] goes into detail on a variety of indicators, text readability, emoticon usage, and domain linking. They found out that the popular subreddits behave very differently from each other. There is a variety of range with word complexity and emoticons. Some communities limit the domains that are allowed to use, and others only allow text posts. A post that is successful in the subreddit "news" will not be popular in other subreddits like "funny". Reddit is made up of thousands of communities with their own rules and norms. To be able to predict if a post will be successful, one can not look at Reddit as a whole but match a post with specific subreddits.

Another paper that takes a look at initial comments and if it can predict how well a post will do. The paper is called "Predicting Reddit Post Popularity Via Initial Commentary"[11] and takes the first ten comments of a Reddit post and tries to predict its popularity. With 2000 training examples and 220 testing examples, they try different algorithms with a baseline of 0 upvotes. They initially tried regression models, but the majority of Reddit posts never surpasses the ten comments, so they went with trying to classify if a post was above a certain threshold. By using features based on the comments and trying different algorithms, they get a result of up to 89% accuracy. With a baseline of 73% there is a small increase in performance.

The scientific paper called "Widespread Underprovision on Reddit"[12] found out that 50% of the popular posts had been posted before and ignored. One would assume there were some external elements at play like bad timing or bad luck, maybe the title of the post was better; unfortunately, they do not go into details about that.

Chapter 3

Data collecting

The data collection chapter is split into three parts. First, go through how the submissions were collected from Reddit. The second part goes into the collection of user information. Then finally present the values that are present in the dataset that will be used for machine learning.

3.1 Submissions

Reddit is a collection of multiple subreddits all with their theme and users. A couple of subreddits was selected to get a proper distribution of different types of content. The main objective was to predict the popularity of news articles, but at the same time, see if this could work on different parts of Reddit. The subreddits that were selected are:

- art
- AskOuija
- AskReddit
- atheism
- confession
- Documentaries
- food
- Futurology
- gaming
- GetMotivated
- investing
- legaladvice
- LifeProTips
- MachineLearning
- movies
- news
- norge
- personalfinance
- PewdiepieSubmissions
- politics
- privacy
- programming
- psychology
- Python
- reddevils
- science
- Showerthoughts
- space
- sport
- technology
- The_Donald
- The_Muller
- tifu
- todayilearned
- videos
- worldnews

On Reddit, there is the option of structuring the content into seven categories. Best, hot, new, rising, controversial, top, and gilded. "Best" is the main category and show content according to Reddit's algorithm designed to keep the user on their site for longer. "Hot" is all about showing the most recently popular upvoted posts. "New" will show submissions that have just been posted in the last couple of minutes. "Rising" are posts that have high interaction early on. "Top" is the most upvoted post in a specific period, and then there is "gilded" which show the posts that have received the most medals (silver, gold, and platinum) which are a token of appreciation one can give to other users for a small fee which supports the site.

To gather a good understanding of how Reddit works and what chances one has to reach the front page. The data collected had to sample from category "new". This way, the data will follow submissions from the beginning and collect data along the way to see how a post can become popular.

Every two minutes, the data collector would go on each subreddits page and look at the submissions found in the category "new". Then every hour collecting the changes in the post.

Luckily Reddit has made it extremely easy to collect data from their site. All one has to do is append ".json" behind their URL, and the whole page is loaded in JSON format, making it easy to gather only the data without needing to parse any HTML.

The data collected on a post:

- ID
- URL
- Permalink
- Subreddit
- Subreddit Subscribers
- Title
- Author
- Upvotes
- Upvote Ratio
- Created
- Is selftext
- Selftext
- Number of comments
- Number of crossposts
- Number of Gildings
- Link to Thumbnail
- Contest Mode
- Is Original Content
- Is Reddit Media Domain
- Is Robot Indexable
- Is Video
- Locked
- Over18
- Send Replies
- Spoiler
- Stickied
- No Follow
- Category
- Content Categories
- Distinguished
- Edited
- Media
- Secure Media
- Selftext HTML
- Suggested Sort
- Pwls
- Wls
- Parent Whitelist Status
- Whitelist Status

3.2 Users

Every submission includes the username of the post author, so to get more information about the users, a new query has to be made to request this information. The reason for doing this later was because the time the query takes to finish would take to long and create problems with collecting the posts. By doing it later, it was also possible to not have the problem of requesting the same user twice. The data collected on the users are:

- Username
- Submission Karma
- Comment Karma
- Number of Comments
- Number of Submissions
- Created
- Verified Email
- Is Employee
- Is moderator
- Deleted

One annoying part about the user is that one can not acquire the total amount of comments and submissions a user has done. The query only allows up to 100 comments and submissions, so it stops at 100. Also, if the user has deleted their account in the meantime of collecting the submissions and collecting the user information, no user information can be collected.

3.3 The dataset

Some values in the dataset did not have any useful properties to them, and so some were removed, and others altered to help make them more beneficial for machine learning. The final dataset consisted of these values:

- Subreddit
- Subreddit Subscribers
- ID
- Title
- Number Of Words
- Number Of Characters
- Number Of Stopwords
- Average Word Length
- Author
- Author karma
- Author comment karma
- Author comments gt100
- Author submission gt100
- Author created
- Author verified email
- Author employee
- Author mod
- Created
- Datetime
- Hour
- Minute
- Upvotes every hour
- Downvotes every hour
- UpvoteRatio every hour
- Comments every hour
- Crossposts every hour
- Silver
- Gold

- Platinum
- Total Gilds
- Final Number Of Comments
- Final Number Of Crossposts
- Final UpvoteRatio
- Final Downvotes
- Final Upvotes

Going through the different values beginning with the subreddit of which the post originates.

Subreddit subscribers are how many subscribers the subreddit had at the time of posting.

Every submission on Reddit has an ID, and this is saved so to be able to find the submission later if needed.

All Titles are stemmed, and symbols and numbers have been removed to help remove any unneeded text. The Title is also used to get a few details on how many words, characters, stopwords, and average word length.

Then we have the Author who is the user that submitted the post. With that, we get information on the user like the number of upvotes on previous submissions and comments. If the user has more than a hundred comments or submissions, and if the user is either a mod, employee or has a verified their email address.

Created is the time in epoch of when it was submitted. Datatime is the epoch value transformed into DateTime format. Hour is the hour of posting between 0-23 and Minute is the minute of posting between 0-59.

Then we have Upvotes, Downvotes, UpvoteRatio, Comments, and Crossposts that are recorded every hour for 24 hours.

Then finally we have silver, gold, and platinum that a post receives from other users during the first 24 hours, and total gilds are just them all added together.

Chapter 4

Experimental Evaluation

Before diving into any machine learning, looking at the data can tell us a lot how Reddit works and might help later in finding out what features that is important. This chapter is divided into three parts; first looking through the dataset and visualizing the features and finding out what features to use. Then dive into some basic machine learning models and finish with some more advanced machine learning models like neural networks.

4.1 Exploring the data

The dataset consists of 505480 submissions from 35 different subreddits. The subreddits differ a lot on total submissions, Figure 4.1 show this in a vertical bar graph. The reason for such a difference in the total amount of submissions is based on how many subreddit subscribers and the type of content that a subreddit has. It also is based on the popularity of a subreddit at the time the dataset was captured. For example, take the subreddit PewdiepieSubmissions, which is a subreddit dedicated to posting memes, at the time the dataset was captured, a YouTube competition between Pewdiepie and another youtube channel gathered much attention and had a vast increase the number of submissions posted.

Because of the significant differences between subreddits when it comes to content and number of submissions, it was decided only to select six subreddits henceforth, and use them independently in machine learning as it is compelled to have different results. The subreddits that were selected are:

- worldnews
- The_Donald
- politics
- legaladvice
- AskReddit
- gaming

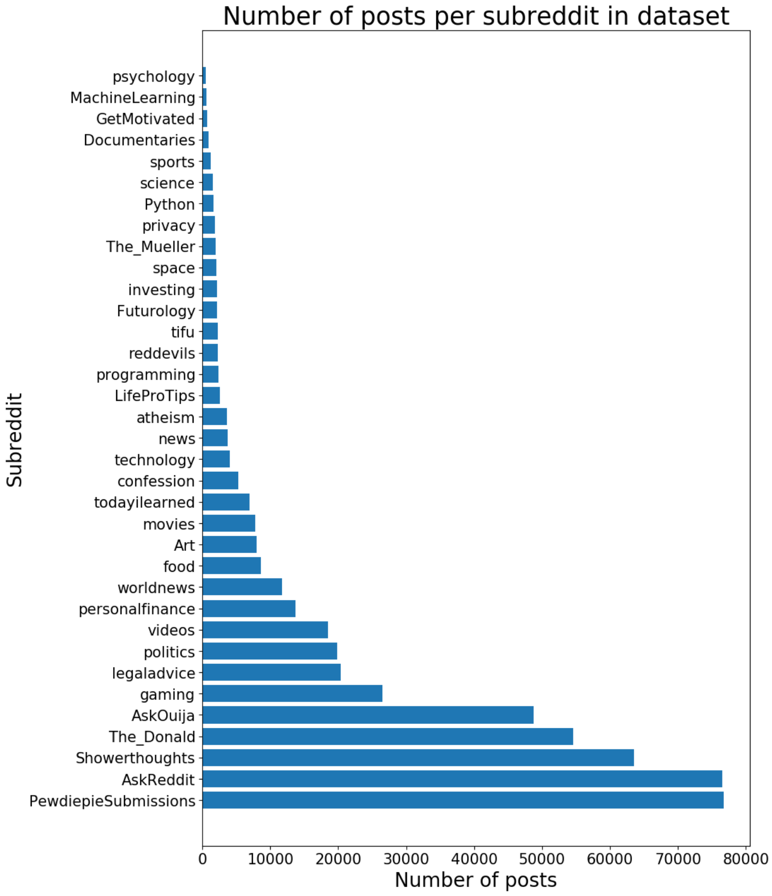


Figure 4.1: Total number of submissions for each subreddit.

In figure 4.2 all numerical features found in the dataset have been plotted in a 50 bin histogram. Looking through all the different plots, we can get a good understanding of how the data looks.

A lot of the features have most of their values around zero. The reason for this is a combination of too many submissions posted at the same time and thus being buried by other submissions, or it might be at the wrong time when for example most users are asleep, and no one is there to upvote it. An excellent example of overflowing submissions was when pictures of the black hole were presented. Many subreddits got flooded with submissions of this picture. Figure 4.3 show this.

What is most interesting might be the Hour and Weekday that clearly show a trend when most users are posting submissions. We see this also in the whole dataset by plotting the number of posts each hour with the average amount of upvotes in figure 4.4, it clearly show the time of day users are submitting and active, but also when the right time to submit a post right before the increase in submissions. Most of Reddit users are from America, and thus Reddit follows their sleep cycle. But not all subreddits follow US time, some subreddits do not follow any particular posting pattern, and so it does not

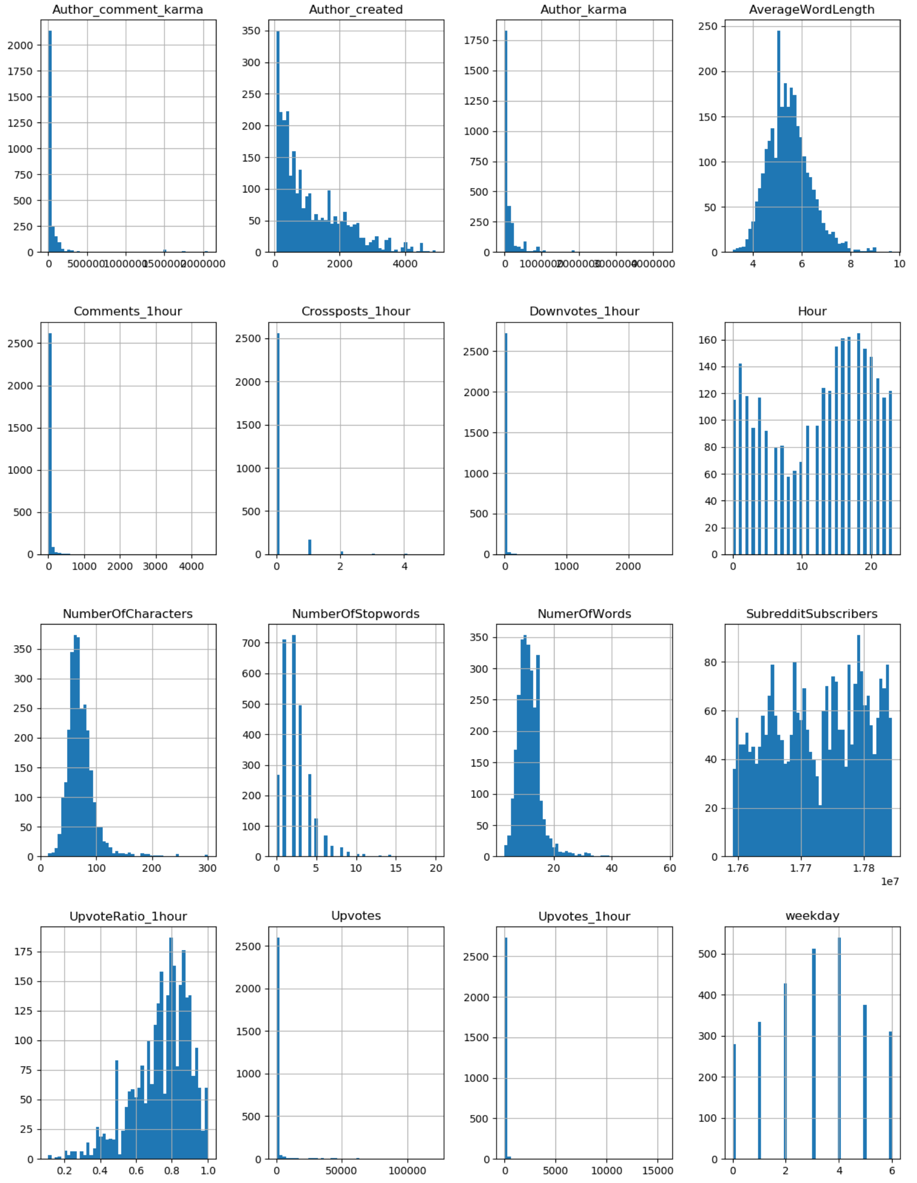
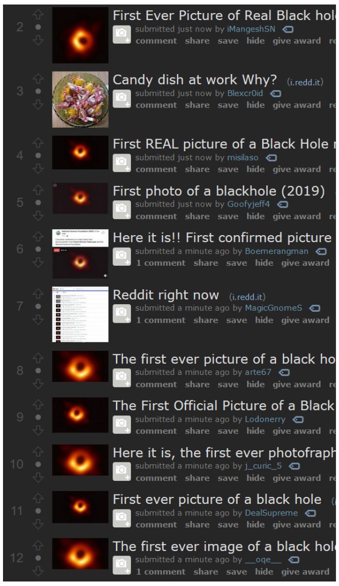


Figure 4.2: Histogram of all features.



2 First Ever Picture of Real Black hol
submitted just now by iMangeshSN
comment share save hide give award re

3 Candy dish at work Why? (i.redd.it)
submitted just now by Blexcr0id
comment share save hide give award re

4 First REAL picture of a Black Hole r
submitted just now by misilaso
comment share save hide give award re

5 First photo of a blackhole (2019)
submitted just now by Goofyjeff4
comment share save hide give award re

6 Here it is!! First confirmed picture
submitted a minute ago by Boemerangman
1 comment share save hide give award

7 Reddit right now (i.redd.it)
submitted a minute ago by MagicGnomeS
1 comment share save hide give award

8 The first ever picture of a black ho
submitted a minute ago by arte67
comment share save hide give award re

9 The First Official Picture of a Black
submitted a minute ago by Lodonerry
comment share save hide give award re

10 Here it is, the first ever photofraph
submitted a minute ago by j_curic_5
comment share save hide give award re

11 First ever picture of a black hole (
submitted a minute ago by DealSupreme
comment share save hide give award re

12 The first ever image of a black hol
submitted a minute ago by __oqe__
1 comment share save hide give award

Figure 4.3: Multiple submissions of pictures of the black hole.

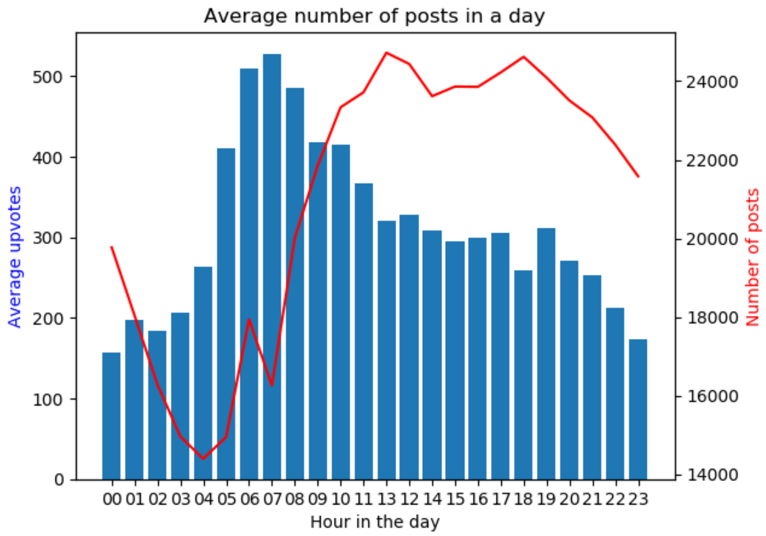


Figure 4.4: Average number of upvotes compared to number of submissions in a day.

matter at what time the user submits a post. In figure 4.5 this shown clearly for the subreddit AskReddit where there is no particular pattern to when users post.

4.1.1 Selecting Features

When selecting the right features to use, it is good to know what makes a Reddit post successful. Six main elements play a role in the success of a popular post.

The first and most apparent is that the content has to be of particularly good quality. As mentioned in related works, quality does not always equal popularity, but for the most part, a popular post will be of good quality.

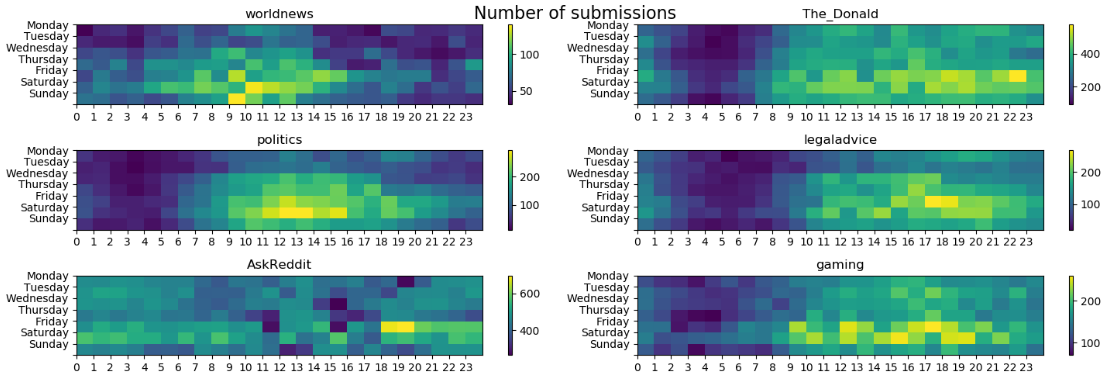


Figure 4.5: Heatmap for each subreddit, showing total submissions in a week

Next timing is important. Submitting a post at a time when there are many users online, but also a time when there are few submissions such that there is less competition.

Another essential element to a successful post is the title. The title combined with the thumbnail is what most people see first and can attract more views towards the post. The thumbnail is not as easy to manipulate as it often is based on the content of the post.

Last and most crucial element that will make or break a post is luck. There is no guarantee that a good post will make it if it does not get lucky with other users noticing the post and upvoting. Most often as stated in related works, a popular submission has often been posted before, so if a post does not succeed, it might be useful to try to submit again. Figure 4.6 is an excellent example of two identical posts receiving different amounts of upvotes.

When going through the features in the dataset, it was decided to only use the submissions from the subreddit news. This way, the features were more consistent and less prone to noise from other subreddits. There is no default way of selecting what features to choose for the models, so trying different approaches and looking objectively at all of them is a good start. First visualizing the features with different types of plots, then using different types of tools and libraries to estimate the best features.

First, we begin by plotting a heatmap that shows correlations between features. A correlation matrix that shows how the features are related to each other. Correlation can either be negative or positive, based on how correlated the values are. The closer the correlation between two features is to one another, the more correlated they are.

Looking at the heatmap in figure 4.7 one can see that for Upvotes which is what we want to predict, correlates heavily with Upvotes_1hour, Downvotes_1hour, UpvoteRatio_1hour, Comments_1hour, Crossposts_1hour.

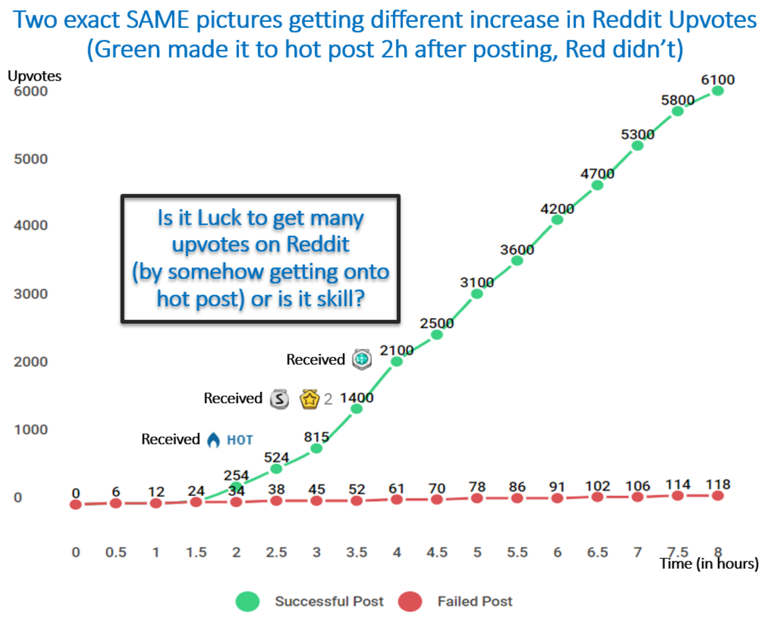


Figure 4.6: Comparing two identical submissions.[13]

The correlation between CreatedUtc and SubredditSubscribers is not anything special, as the time increases, so do the subreddit subscribers. Time in a linear fashion is not that interesting compared to the day of the week or hour in the day. The same with Author_created_utc a better feature would be to transform it into days old.

An adverse correlation does not immediately mean that it is a bad feature; there is still hope by combining them with other features that will help with predicting.

Next plot that will help us select the right features for our model is a scatter-plot. Some of the features are plotted up against each other to see if there are a visual pattern that shows some form of correlation. Looking at Figure 4.8, we see five features scattered with each other. The diagonal line is showing the histogram of the values since using scatter plot with itself would only produce a straight line.

There are a few interesting patterns when looking at different scatter-plots. First, the obvious one is Upvotes_1hour with Upvotes has a small upwards trend and a linear pattern to it.

Another interesting scatter-plot is between Author_karma and Upvotes. Higher Author karma does not necessarily equal a very successful submission, the lower the user karma is the more upvotes it seems the post can achieve.

After looking at the heat mat and scatter-plots, there seems like there are no features that stand out other than the values recorded 1 hour after posting. Luckily there are automatic functions that will rank our features for us and show which features

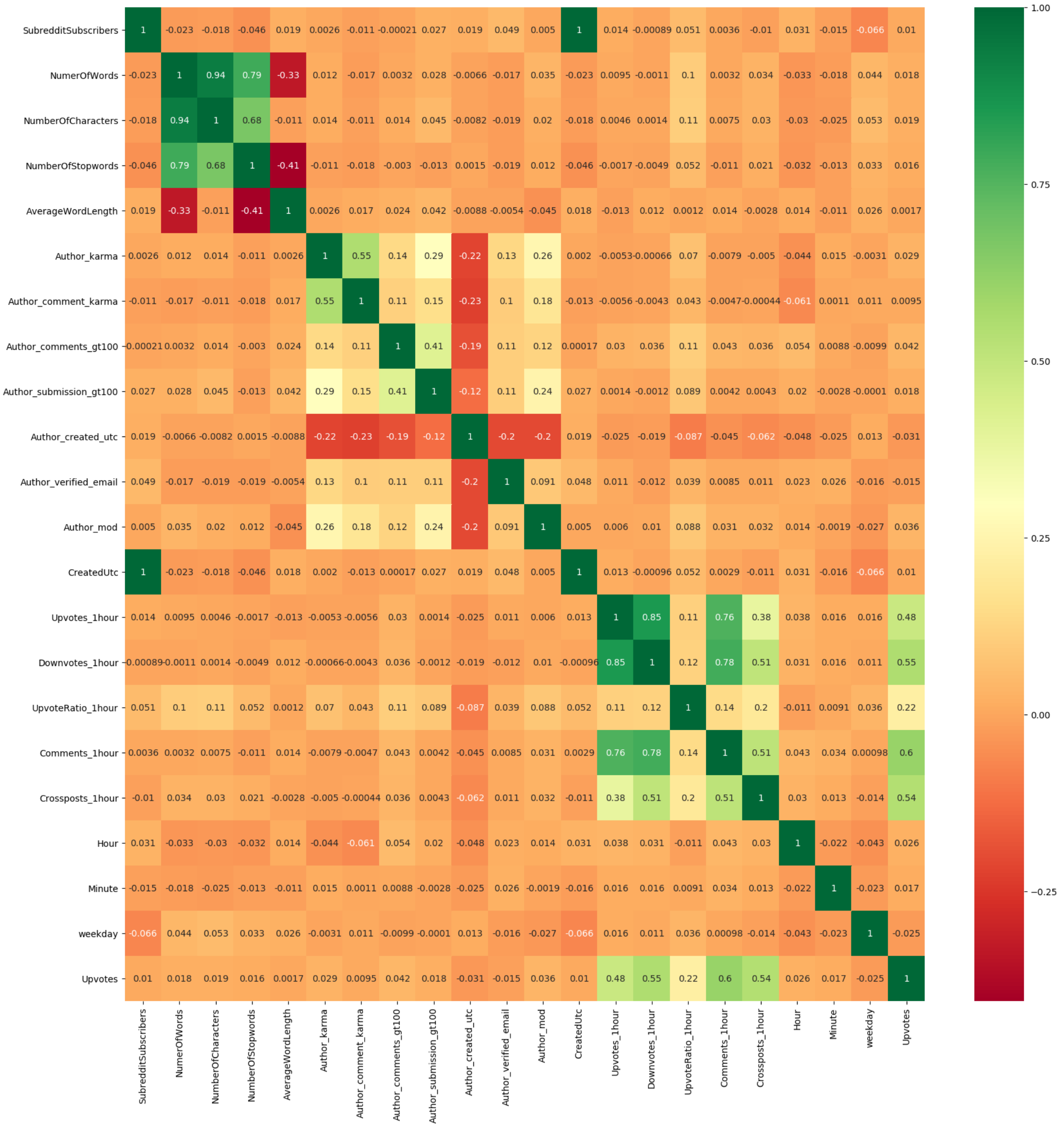


Figure 4.7: Heat-map of all features

are the best for our dataset. The functions we are going to use are SelectKBest and feature_importances_ that we get from RandomForestRegressor. The result from both functions in table 4.1. There is a big difference between the functions but for the most part, similar features in the top ten list. What makes us trust RandomForestRegressor more is that it use regression. It also matches better with the heatmap than what SelectKBest does. Thus the features that will be used for machine learning are the ones from RandomForestRegressor.

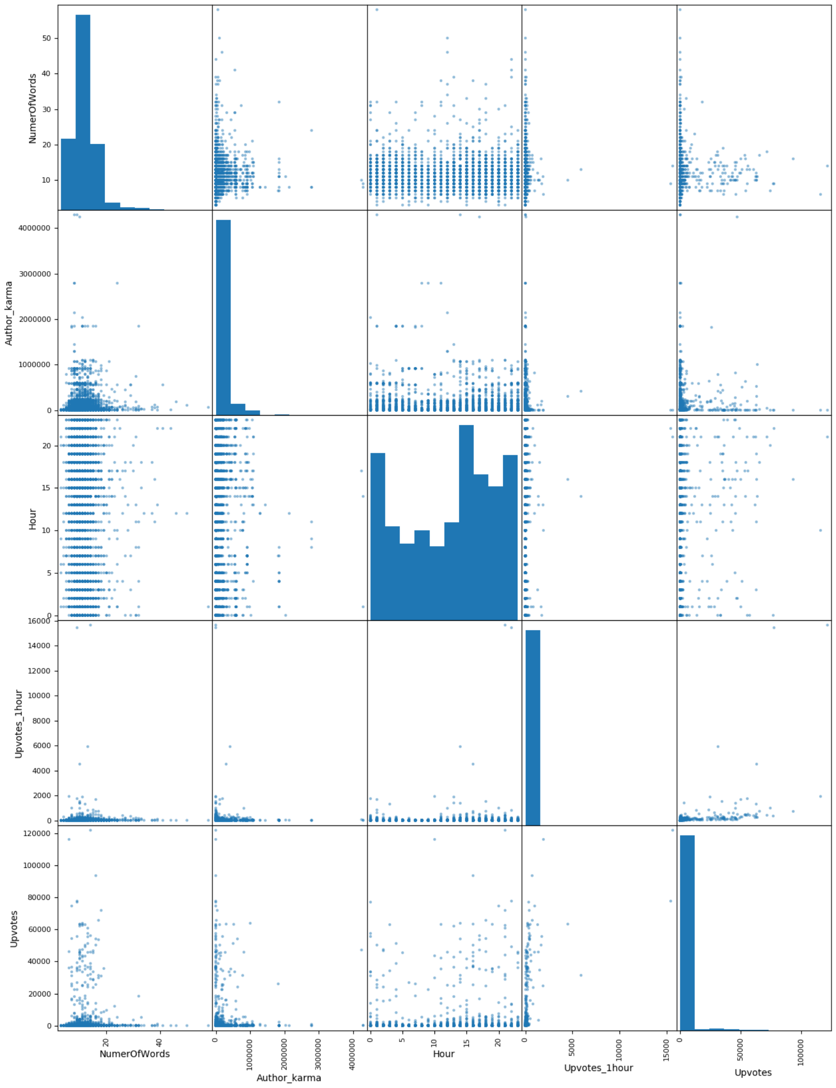


Figure 4.8: Scatter plot of different features.

Table 4.1: Feature importance Results

SelectKBest Results		RandomForestRegressor	
Rank	Features	Rank	Features
1	Author_created_utc	1	Upvotes_1hour
2	Author_karma	2	Downvotes_1hour
3	Author_comment_karma	3	Author_karma
4	Upvotes_1hour	4	Crossposts_1hour
5	Downvotes_1hour	5	UpvoteRatio_1hour
6	Comments_1hour	6	Author_comment_karma
7	Author_Created	7	Hour
8	CreatedUtc	8	Comments_1hour
9	SubredditSubscribers	9	SubredditSubscribers
10	Minute	10	NumberOfCharacters

4.1.2 Features

The Title of every submission was not taken into account when selecting the right features because we knew it is an important feature. The Title of each post will be used with TF-IDF and scored separately and together with other features.

The features that will be used in machine learning models are:

- Title
- Upvotes_1hour
- Downvotes_1hour
- UpvoteRatio_1hour
- Comments_1hour
- Crossposts_1hour
- Author_karma
- Author_comment_karma
- Hour
- NumberOfCharacters

4.2 Machine Learning

All data scientists know of the saying "There ain't no such thing as a free lunch". The statement no free lunch tells us that nothing comes to us for free, and we have to make assumptions about the data, or else there is no reason to prefer one model over any other. No model will work with all kinds of data, and thus, the only way to know which one is best is to evaluate them all. Unfortunately, this is not practical, and we need to make some assumptions about the data and evaluate the most reasonable models available. The models are Linear regression, Naive Bayes, SVM, and Random Forest.

To be able to compare the models with each other, a few performance metrics functions need to be involved. Mean Squared Error, Root Mean Squared Error, and R Squared. Mean Squared Error measures the average squared error between the model and the closer the value is to zero, the better is the model.

$$Mean\ Squared\ Error:\ \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.1}$$

Root Mean Squared Error is the same as Mean Squared Error; only the result is squared back to original values more comparable to the actual values that the upvotes are.

$$Root\ Mean\ Squared\ Error:\ \sqrt{\frac{1}{n}\sum_{t=1}^{n}e_t^2} \tag{4.2}$$

R Squared is a bit different in that it is calculated by dividing the sum of the first errors by the sum of the second errors and subtracting the derivation from 1. R squared tells us how well each prediction fit on the model, the value will most often be between 0 and 100, but it can also be negative if the model is not predicting well.

$$R^2:\ 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.3}$$

Where: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2$

Because of memory error and time restrictions, TF-IDF has to be dropped as a feature for now as this created problems under testing. So the features will be the already mentioned above, but without the Title.

We will begin with default parameters on the different models and later try to improve them with RandomizedSearchCV. Cross-validation will also be used as a default and use

	worldnews	The_Donald	politics
Min:	-3.31	0.32	-0.14
Max:	0.45	0.58	0.49
Mean:	-0.51	0.52	0.33
SD:	1.42	0.1	0.23
MSE:	51474279.81	459733.5	10733333.79
RMSE:	7174.56	678.04	3276.18
R^2:	-0.57	0.58	0.44

	legaladvice	AskReddit	gaming
Min:	-0.53	-0.15	0.26
Max:	0.45	0.37	0.52
Mean:	0.07	0.11	0.42
SD:	0.33	0.17	0.09
MSE:	505679.99	472446.45	27096948.03
RMSE:	711.11	687.35	5205.47
R^2:	0.2	0.58	0.44

Table 4.2: Linear regression Result.

a five-fold split, and show the results for the worst score, best score, take the mean of the results and the standard deviation of the results.

4.2.1 Linear regression

Linear regression tries to fit a linear model that match the structure of the data. If the data have a linear structure, most likely the best bet is to use a linear model. Table 4.2 show the results, and it does not seem that this dataset is very linear.

There is a significant difference between subreddits. It might have to do with how often posts on a subreddit reaches the top of Reddit and achieves an extreme amount of upvotes. Even though some of the subreddits are achieving ok results, the model does not seem to fit the data very well. If we look through the predictions, there are positive, but also negative values, this should not be possible as one can not have negative upvotes. The linear model does not consider that. A negative R Squared score is a good indicator that this is not a good model. It might work on some subreddits, but a model that works on all subreddits would be better.

	worldnews	The_Donald	politics
Min:	0.01	0.02	0.14
Max:	0.02	0.03	0.27
Mean:	0.01	0.02	0.22
SD:	0.0	0.0	0.05
MSE:	34854947.19	1069119.12	16241473.95
RMSE:	5903.81	1033.98	4030.07
R^2:	-0.06	0.03	0.16

	legaladvice	AskReddit	gaming
Min:	0.3	0.35	0.1
Max:	0.37	0.39	0.17
Mean:	0.33	0.37	0.13
SD:	0.03	0.01	0.03
MSE:	626879.23	1115971.03	41375147.22
RMSE:	791.76	1056.4	6432.35
R^2:	0.01	0.0	0.14

Table 4.3: Naive Bayes Result.

4.2.2 Naive Bayes

To be able to use the Naive Bayes for regression, a Gaussian distribution is needed. This way, it is possible to predict values based on the Gaussian spectrum. Naive Bayes looks at all the features individually and does not make any assumptions amongst them. Looking a the results in table 4.3 this did not work well with the data. A reason for this might be that a large number of features are centered around zero. So Naive Bayes have trouble moving away from those.

4.2.3 SVM

SVM stands for Support Vector Machine, and to use it with regression the model SVR is used which stands for Support Vector Regression. It can be compared to Linear regression in that it will fit a straight line, but SVR can also create a non-linear line that might fit the data better. The difference between SVM classifier and SVR is instead of trying to divide the data with a line or a street as it often is called, SVR tries to get as many values inside of the street while limiting margin violations. This model did not fit at all, and the result in table 4.4 show all having negative R Squared.

	worldnews	The_Donald	politics
Min:	-0.02	-0.1	-0.06
Max:	-0.01	-0.08	-0.05
Mean:	-0.02	-0.09	-0.06
SD:	0.0	0.01	0.0
MSE:	33311670.29	1207894.77	20283909.33
RMSE:	5771.63	1099.04	4503.77
R^2:	-0.01	-0.1	-0.05

	legaladvice	AskReddit	gaming
Min:	-0.01	-0.0	-0.02
Max:	-0.0	-0.0	-0.02
Mean:	-0.0	-0.0	-0.02
SD:	0.0	0.0	0.0
MSE:	633594.31	1118492.24	49047888.47
RMSE:	795.99	1057.59	7003.42
R^2:	-0.0	-0.0	-0.02

Table 4.4: SVM Result.

4.2.4 Random Forest

Random Forest is a model that handles tabular data, and data with numerical features well. All features in our data are numerical, so we can assume it will work. The data did not work well with a linear model, making us believe the data is non-linear. Random Forest can work with data that is non-linear and can capture interactions between the features and target. Looking at the results in table 4.5, this went a lot better than previous attempted models. Here all R Squared values are, and all subreddits seem to do somewhat good on this model.

4.2.5 Improving parameters

After having tried four different models, there is only one that did somewhat well, and that is Random Forest. Therefore this is the model that will be improved. The way one improves a model is to change its parameters. The parameters are tweaked by testing out multiple different values and combinations and tested to see which parameters performed the best. To do this, we used the function RandomizedSearchCV. By giving it a range of values, it will randomly go through and test them with each other to see which one is the best. The parameters are n_estimators, max_features, max_depth, and min_samples_split. The default values is listed in Table 4.6.

Running RandomizedSearchCV, and testing multiple parameters. New and better parameters were found shown in Table 4.7.

	worldnews	The_Donald	politics
Min:	0.23	0.72	0.66
Max:	0.69	0.74	0.73
Mean:	0.5	0.73	0.69
SD:	0.16	0.01	0.03
MSE:	13161352.77	294732.98	5330157.61
RMSE:	3627.86	542.89	2308.71
R^2:	0.6	0.73	0.72

	legaladvice	AskReddit	gaming
Min:	-0.09	0.1	0.56
Max:	0.56	0.66	0.72
Mean:	0.39	0.44	0.66
SD:	0.24	0.2	0.06
MSE:	373075.62	435258.48	17835447.43
RMSE:	610.8	659.74	4223.2
R^2:	0.41	0.61	0.63

Table 4.5: Random Forest Result.

n_estimators	10
max_features	auto
max_depth	None
min_samples_split	1

Table 4.6: The default parameters for Random Forest.

	worldnews	The_Donald	politics
n_estimators	20	200	110
max_features	auto	sqrt	auto
max_depth	23	45	23
min_samples_split	5	10	5

	legaladvice	AskReddit	gaming
n_estimators	20	200	155
max_features	sqrt	auto	sqrt
max_depth	45	45	45
min_samples_split	10	5	10

Table 4.7: Improved parameters for Random Forest.

	worldnews	The_Donald	politics
Min:	0.56	0.56	0.56
Max:	0.72	0.72	0.72
Mean:	0.66	0.66	0.66
SD:	0.06	0.06	0.06
MSE:	9767387.62	263422.78	4815416.39
RMSE:	3125.28	513.25	2194.41
R^2:	0.7	0.76	0.75

	legaladvice	AskReddit	gaming
Min:	0.56	0.56	0.56
Max:	0.72	0.72	0.72
Mean:	0.66	0.66	0.66
SD:	0.06	0.06	0.06
MSE:	356079.61	381957.17	15823511.61
RMSE:	596.72	618.03	3977.88
R^2:	0.44	0.66	0.67

Table 4.8: Results for Random Forest with improved parameters.

Noticing that each subreddit have different parameters is a good indication that subreddits are different, and creating a model for each subreddit is a good idea. Also great to see that all R Squared values increased. The results are in Table 4.8.

4.2.6 Sentimental

Sentimental learning use NLTK.SentimentIntensityAnalyzer to determine if a title is either negative, neutral, or positive. Running the sentiment analyzer on all titles and plotting them in figure 4.9 to see how the different subreddits vary in negative, neutral, and positive submissions. What might surprise us is how different all of the subreddits are.

When taking a closer look at the subreddit worldnews, there is a large number of submissions that are negative compared to positive. One would assume negative submissions are more popular or get more upvotes. However, if we take a look at figure 4.10 here we take the average number of upvotes for all posts in each category, and we see positive and negative titles receive almost the same amount of upvotes. Also taking a look at the subreddit leagaladvice, neutral titles are higher than both negative and positive but have the lowest average upvotes. The reason for this might be that the particular subreddit likes submissions more if they are either negative or positive, but users do not realize this, so most of the submissions are neutral.

Most of the subreddits do not seem affected by the sentiment and therefore give no advantages to predict the number of upvotes a post receives. Even though they might

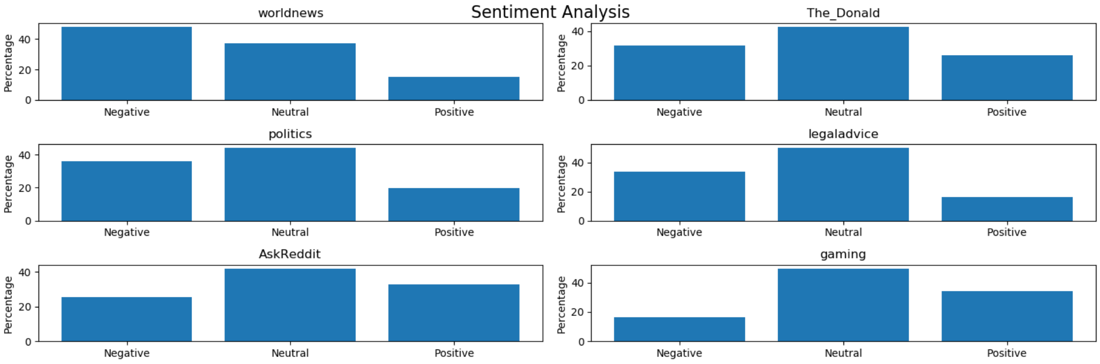


Figure 4.9: Sentiment analysis

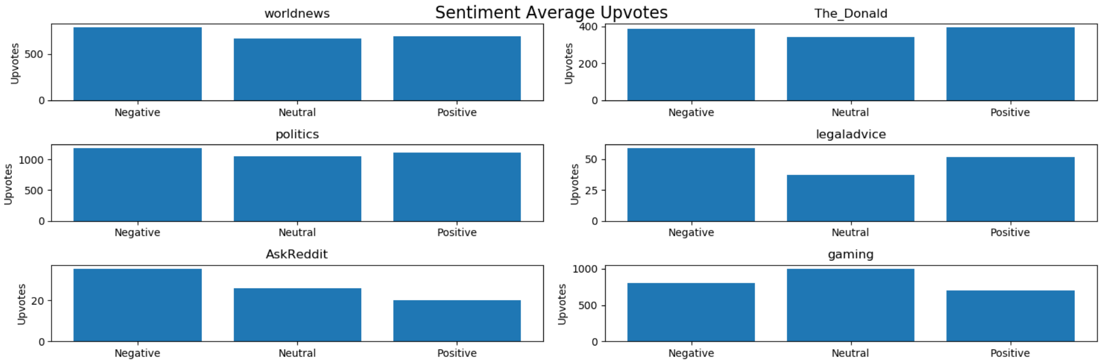


Figure 4.10: Sentiment analysis average upvotes

have more submissions for a particular sentiment, does not show any differences in the average number of upvotes.

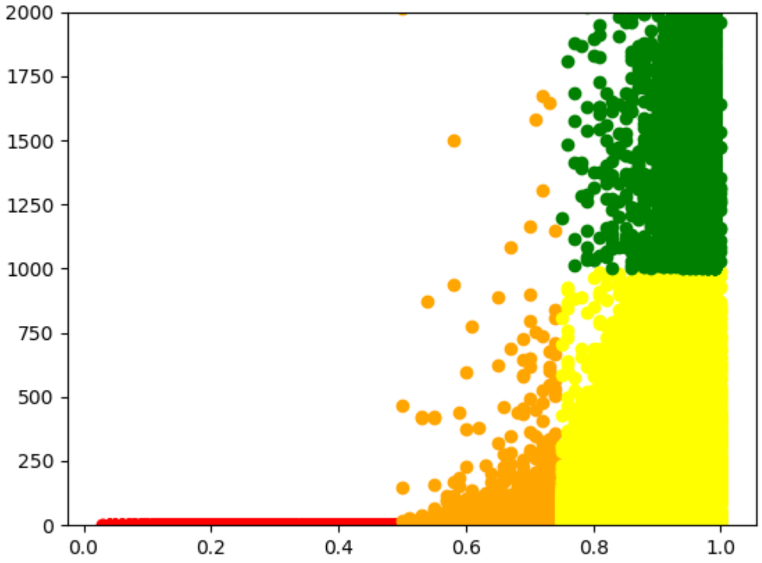


Figure 4.11: Scatterplot of Upvote Ratio to Upvotes where the colors are based on the different classes.

4.3 Advanced Machine Learning

4.3.1 Classification

If instead of predicting upvotes, but categorize the submissions into classes would maybe make it easier to predict. To do this the posts are classified into four different types, and those are *Bad*, *Controversial*, *Ok*, and *Great*.

The Reddit frontpage can deem a submission *Controversial* by looking at the number of upvotes to downvotes, and if it is between 0.5 to 0.75 in upvote ratio, it is deemed *Controversial*.

If the submission is below 0.5, it is considered Bad, and everything over 0.75 is good.

The *Good* posts are split into two, to be able to separate submissions that did *Ok*, and submissions that did *Great*. This limit is at a 1000 upvotes. We deem a post successful if it receives more than a 1000 upvotes. By doing this it is possible to predict if a submission will only be *Ok* or *Great*. Looking at figure 4.11 we have *Bad* posts in red, *Controversial* in orange, *Ok* in yellow, and *Great* posts in green. The plot stops at 2000 upvotes to show the distribution better. In figure 4.12, all points are visible, and the yellow is almost gone.

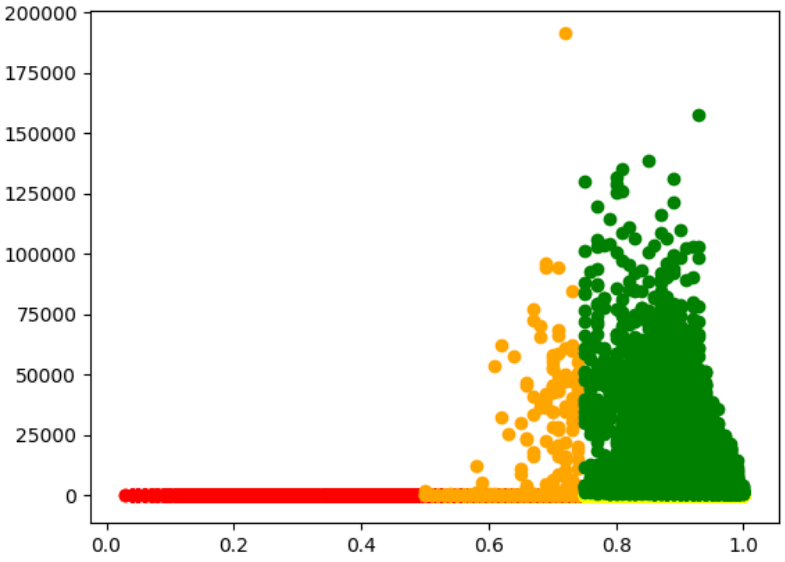


Figure 4.12: Full Scatterplot of Upvote Ratio to Upvotes where the colors are based on the different classes.

Subreddit:	Result:
worldnews	0.72
The_Donald	0.93
politics	0.89
legaladvice	0.80
AskReddit	0.91
gaming	0.84

Table 4.9: Random Forest Classifier Result.

The type of classifier used is the Random Forest Classifier. Since it did well on the data for regression, it seems like a good idea to use it for classification.

After using the classifier on the data, the results are auspicious. The result in table 4.9 is how many correct it predicted divided by total predictions. 0.72 means it got 72% correct predictions.

With some subreddits reaching 90% correct the model did, and this dividing up the submissions helped a lot with being able to predict. The reason for such good results might have to do with all the extremely successful posts that can hurt the results of regression in a big way. So by grouping them, helped eliminate problems which that creates.

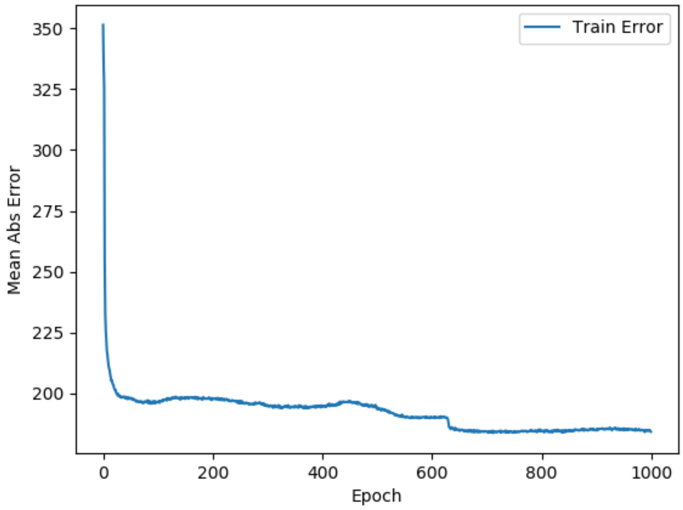


Figure 4.13: Neaural network run for 1000 ephocs.

4.3.2 Neural Network

is what one calls a deep feedforward network.

Neural networks use a loss function that tells them how far away they are from the correct output and then they use backpropagation and use a decent gradient algorithm that moves the values for each node such to minimize the loss-function. To calculate gradient decent RMSProp (Root Mean Square Propagation) is selected, and the loss function use Mean Squared Error.

All nodes use an activation function that alters the output of a node to a specific form. Sigmoid function, for example, changes the value to be between 0 and 1. What is used in this neural network is the ReLu activation function, which takes anything negative and set it to zero.

Each time the model backpropagates and alters values for all nodes are called an epoch. We run the model for a certain amount of epochs so that the loss function can try to converge, but hopefully not too long to overfit the data.

Running the model for a 1000 epochs did suffice. It converges at a loss function of 189 and has a root mean squared error of 529. Figure 4.13 show this. These are similar scores to Ransom Forest.

In figure 4.14, the predictions, and test values are plotted against each other in a scatter-plot and show a great pattern that is close to a linear. The model has both submissions that are predicted to high or predicted to low. The plot has a reasonable distribution

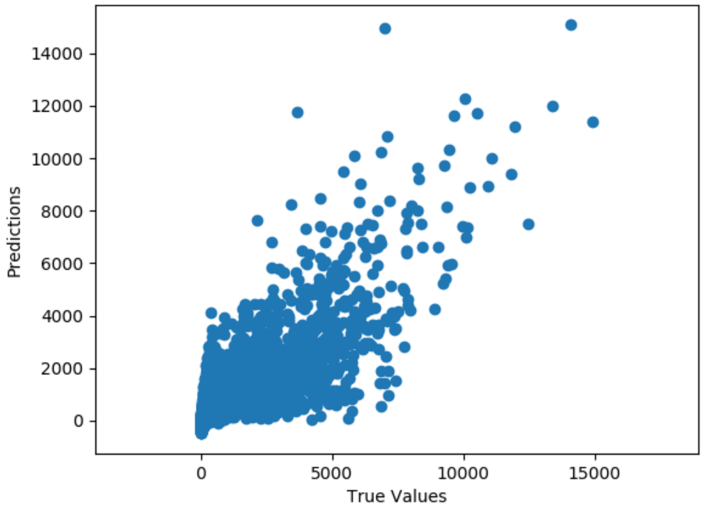


Figure 4.14: Scatter plot of Neural networks predicions against the truth.

of both and might mean the features are good, but be missing a feature that can help further down the road.

Chapter 5

Discussion

Looking back on the different models that were tested, Random Forest was the model with the best results. The data does not seem to be linear and therefore, did not work well with Linear Regression, Naive Bayes, or SVM, all of which are models that work best on linear data.

Looking at the results from Random Forest, all subreddits did well, and when improving the parameters, we can see that the subreddits prefer different values. Showing us that the subreddits all have a particular structure to them, and should be taken into consideration when making models for different subreddits.

During our testing of sentiment analysis, there did not seem to be a huge difference for how many upvotes a submission receives based on its sentiment. To find out if a title was either negative, neutral, or positive, the function NLTK.SentimentIntensityAnalyzer was used. This function is based on its a corpus of predetermined words. A better way to do this would have to create our own corpus of negative and positive words from our data. If for example, taking the title of controversial posts and defining them as negative, and successful posts titles used for positive. Then maybe the results would have been better as it is based on the dataset and are more established in what is popular and calculates the sentiment based on popularity instead of positive and negative words. Unfortunately, we did not have time to try this.

One problem with the features that are measured 1 hour after posting is how the majority of submissions have already stalled and are not receiving any more upvotes. Eighty percent of submissions do not receive more than ten upvotes after the first hour. These features help a lot in predicting the outcome of the submissions, but it might create problems in that it is too good, and 1 hour is too late of a measurement that it already has a good indication of how well the submission is going to do. This ties together

Subreddit:	Result:
worldnews	0.50
The_Donald	0.88
politics	0.55
legaladvice	0.56
AskReddit	0.59
gaming	0.52

Table 5.1: Random Forest Classifier Result.

	worldnews	The_Donald	politics
Min:	0.56	0.56	0.56
Max:	0.72	0.72	0.72
Mean:	0.66	0.66	0.66
SD:	0.06	0.06	0.06
MSE:	9767387.62	263422.78	4815416.39
RMSE:	3125.28	513.25	2194.41
R^2:	0.7	0.76	0.75

	legaladvice	AskReddit	gaming
Min:	0.56	0.56	0.56
Max:	0.72	0.72	0.72
Mean:	0.66	0.66	0.66
SD:	0.06	0.06	0.06
MSE:	356079.61	381957.17	15823511.61
RMSE:	596.72	618.03	3977.88
R^2:	0.44	0.66	0.67

Table 5.2: Results for Random Forest without 1 hour features.

with what we mentioned in related works is how short-lived the media in today's age. Moreover, with Reddit submissions living is dependant on how well it does.

If we remove the features and try the same classification model, we get lower scores on all subreddits seen in table 5.1. With more than 50 percent correct on all subreddits seem somewhat useful.

If we do the same to random tree regressor, we see again a drop in score table 5.2

Removing the features in the neural network and run it for a 1000 epochs does not yield as good results and it gets a root mean absolute error of 1050. The model does not seem to want to predict values over a 1000 upvotes, as shown in figure 5.1.

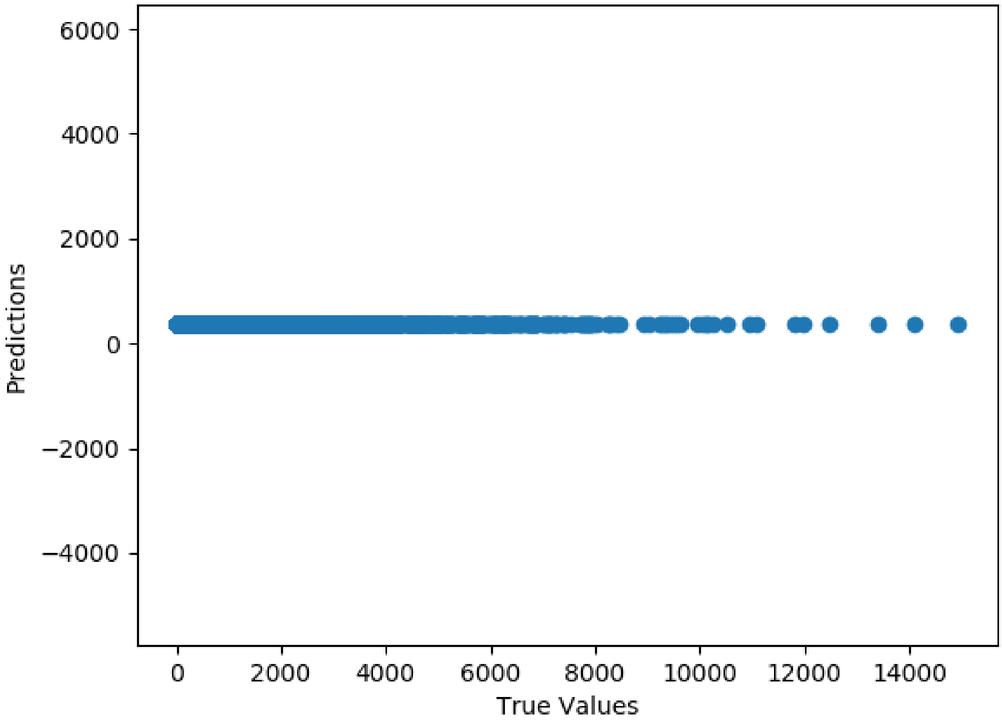


Figure 5.1: Scatter plot of Neural networks without 1 hour features.

Chapter 6

Conclusion and Future Directions

Predicting submissions on Reddit is not straight forward, and there are multiple factors to take into account for what makes a post popular. Using features collected 1 hour after posting seemed like a good idea when gathering the data, but might have been too good of a feature to consider for any useful information. The number of submissions that do not make it and are dead within an hour does not help to predict successful posts as they are in extreme minority. Many submissions would have been popular if it was lucky in receiving upvotes within the first hour or so, and so one can not assume a post is bad, a lot of them are successful later when reposted, which was mentioned in related works. Then if one only looks at the successful, the data is not as plentiful and would have trouble knowing if it is a bad post.

If we could have done a few things differently, we would have spent more time messing around with the Title and sentiment analysis and tried combining models to see if they could predict better together.

This thesis has shown that the success of a Reddit post can be explained by a multitude of factors and are all critical in predicting if a post will be successful. Timing is important but also depends on the subreddit. Most of the submissions go unnoticed and never get any interactions. If precise predictions are not necessary, classification is excellent at predicting defined classes. However, if precision is important, Random Forest Regressor is the model that worked best. The neural network also worked great, but the only problem was when removing the 1-hour features, which made all models worse and had trouble predicting high values.

List of Figures

List of Tables

Bibliography

[1] J Mander. Daily time spent on social networks rises to over 2 hours. *Global Web Index*, 2017.

[2] Katerina Eva Matsa and Elisa Shearer. News use across social media platforms 2018| pew research center. *Pew Research CenterâĂŹs Journalism Project, Pew Research Center*, 10, 2018.

[3] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. Echo chambers on facebook. *Available at SSRN 2795110*, 2016.

[4] Jens Stoltenberg. Nato: good for europe and good for america. *"https://www.nato.int/cps/en/natohq/opinions_165210.htm"*, 2019.

[5] SmarterEveryDay. Who is manipulating facebook? - smarter every day 215 [12:15]. *"https://www.youtube.com/watch?v=FY_NtO7SIrY"*, 2019.

[6] SmarterEveryDay. Who is manipulating twitter? - smarter every day 214 [16:37]. *"https://www.youtube.com/watch?v=V-1RhQ1uuQ4"*, 2019.

[7] NATO StratCom COE. The black market for social media manipulation. *ISBN: 978-9934-564-31-4*, 2018.

[8] Greg Stoddard. Popularity and quality in social news aggregators: A study of reddit and hacker news. *arXiv preprint arXiv:1501.07860*, 2015.

[9] Tim Weninger, Thomas James Johnston, and Maria Glenski. Random voting effects in social-digital spaces: A case study of reddit post submissions. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 293–297. ACM, 2015.

[10] Andrew Tsou. How does the front page of the internet behave? readability, emoticon use, and links on reddit. *First Monday*, 21(11), 2016.

[11] Andrei Terentiev and Alanna Tempest. Predicting reddit post popularity via initial commentary. *nd): n. pag*, 2014.

[12] Eric Gilbert. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 803–808. ACM, 2013.

[13] pineapplezach. Two exact same post getting different upvotes on dataisbeautiful, one was hot post after 2 hours. is it luck or skill that affects whether a post is successful? [oc]. *"https://www.reddit.com/r/dataisbeautiful/comments/b6gt3v/two_exact_same_post_getting_different_upvotes_on/"*, 2019.