Insights from Reddit Submissions dataset Using Machine Learning Algorithms

Sree bhargavi balija, Mounika padala

Abstract -

Since the advent of social media, data transmission has improved tremendously. People could share data(text, video, images, audio, etc;) in the form of posts. The most common platform where people post about diverse topics is Reddit. With the diverse communities on Reddit, people could access more relevant information from these subreddits. From the Reddit Submissions dataset, we have tried to predict the success of the post or submission, based on different features such as the submission time, posted community, title length, etc; We have used several Regression, TF-IDF, transformer models to evaluate the co-dependencies among different features in the Dataset.

***Keywords*- TF-IDF, Word2vec, Bigrams, Most combined unigram and bigrams, Siamese BERT network, sentence, Transformers, Semantic search, Natural language processing, false negatives**

I. Introduction

Reddit is a social news aggregation, web content rating, and discussion website. The site quickly gained popularity after its creation in 2005 by two University of Virginia students. The goal of Reddit is to submit online content in the form of links, text posts, and images, which can further be rated up or down by other users. The posts are categorized into "subreddits," where users can share specific topics and interests related to the topic check at hand. In its early years, Reddit began to rise in popularity, with NSFW. Programming and Science are the top trending subreddits of the time. As of 2019, Reddit is ranked the 18th top site globally, according to Alexa Internet.

In this work, we analyzed all the Reddit posts from three years of data (over 510 million posts!). The initial goal was to find the most correlated factors of a post (such as title length, posted time, upvotes, downvotes, score, and general reactions to the post. Posts may vary in topics, arguments, time posted, and many more variables, but we felt as if the popularity really depended on the post's title length and the time it was posted. We could determine which length is too short to gain attention and which is long enough to bore an audience. We also looked at the most popular subreddit posts and time of day to see any upvote relation. We hope to give enough information and analysis to provide clarity, understanding, and a newfound interest to readers that are unfamiliar with the social foreground. And hopefully, fellow Reddit users will gain some insight into how to optimize their posts to gain the most traction.

II. Dataset

This dataset is a collection of 132,308 reddit.com submissions. Each submission is of an image that has been submitted to Reddit multiple times. For each submission, we collected features such as the number of ratings(positive/negative), the submission title, and the number of comments it received. 16,736 of which are unique. In other words, each image we obtained has been submitted 7.9 times on average. This data consists of roughly 5 million comments, and 250 million ratings (56% upvotes, 44% downvotes), from 63 thousand users to 867 communities ('subreddits').

Attribute	Description
Image_id	id of the image, submissions with the same id are of the same image
Unixtime	time of the submission (Unix time)
raw time	the raw text of the time
title	submission title
total_votes	number of upvotes + number of downvotes
reddit_id	id of the submission on Reddit
number_of_ upvotes	number of upvotes
Subreddit	subreddit
number_of_ downvotes	number of downvotes
local time	time of the submission (unix time)
score	number of upvotes - number of downvotes
number_of_ comments	number of comments the submission received
username	name of the user who submitted the image

Table 1: Reddit Submissions Dataset Attributes

The dataset contains over 1.7 billion posts from the month of May 2015. Features available in the original dataset include subreddit labels, the text of the post, as well as metadata about the post, including its timestamp and the number of up-votes and down-votes. The original dataset was subsetted to span five subreddit categories. These five categories were chosen to make the task more tractable by selecting five distinct categories where domain-specific vocabularies would be less likely to overlap. Of this subset, only posts with positive scores were kept, with the assumption that posts with positive scores (i.e. posts that

received more up-votes than down-votes) were less likely to exhibit significant label noise that would undermine the interpretability of the results. The final dataset used in the analysis includes 1,004,560 samples from five subreddit categories.

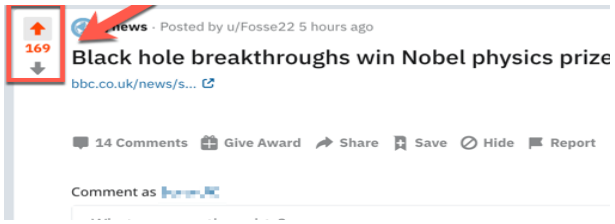
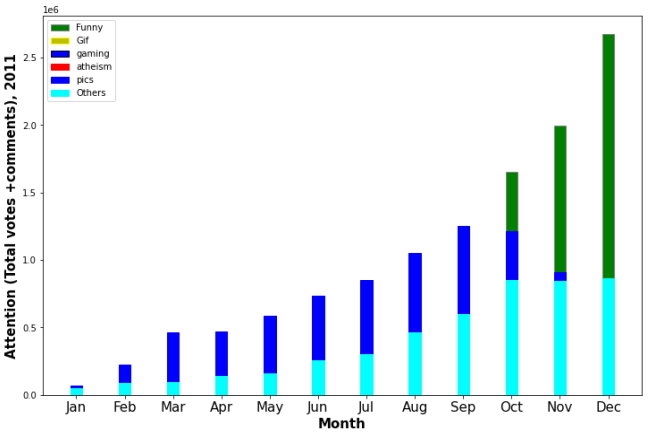


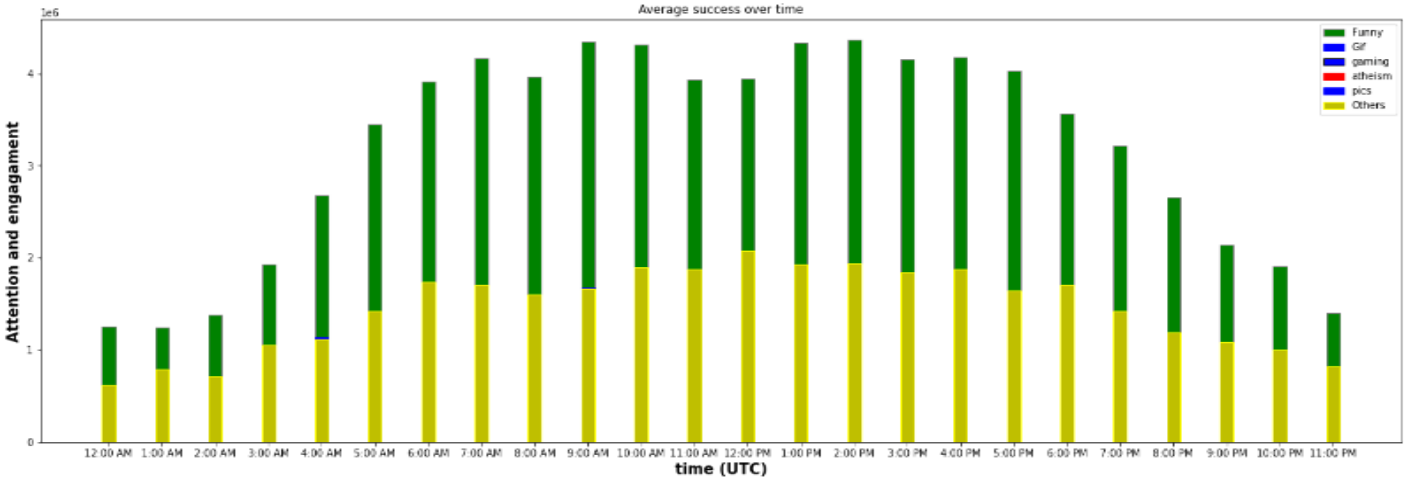
Fig 1: Upvotes and Downvotes on Reddit



Graph 2: Month-wise analysis of Attention and Comments (engagement) for the year 2011

Exploratory Data Analysis:

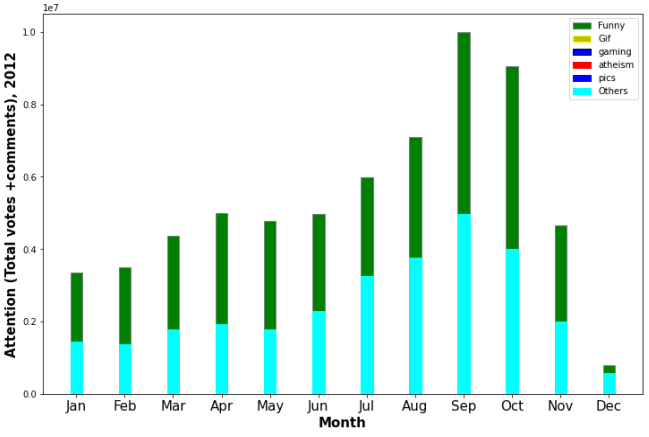
After analyzing the dataset, we can regard a few of the features with realistic parameters: the rating (upvotes - downvotes) that the image receives, the attention (upvotes + downvotes) and the engagement (total number of comments).



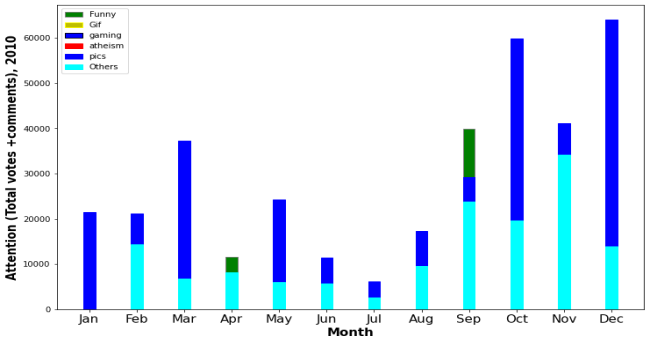
Graph 1: Attention and engagement of different users during the day for different top Reddit classes

From the data, we specifically analyzed the components (attention and engagement) over time. (Time sorted in ascending order). The insights are graphically represented in terms of the day, week, and month.

The above plot is plotted between attention and engagement to the Utc time. It indicates that users' total attention and engagement are more at 9 am and 1 pm. As we all know, people tend to use online platforms during eating hours; from the above plot, we can observe that attention and engagement are more during lunch and breakfast times.



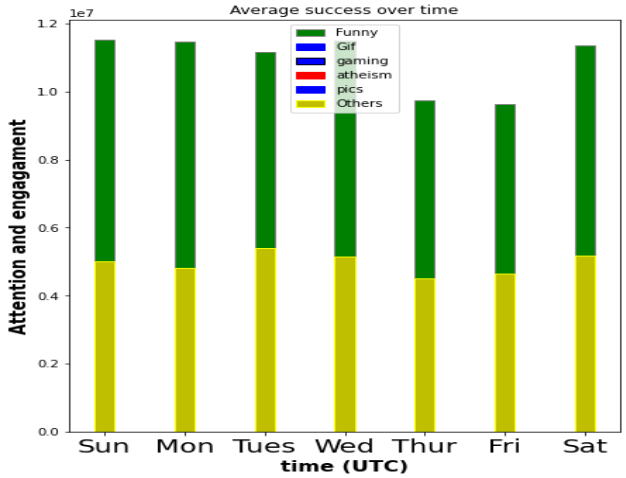
Graph 3: Month-wise analysis of Attention and Comments (engagement) for the year 2012



Graph 4: Month-wise analysis of Attention and Comments (engagement) for the year 2010

An interesting truth: People tend to give more attention and engagement in the latter half of the year. The graph shows that people tend to use Reddit more at the end of the year. Maybe it's because of the cold weather. When people have fewer choices of activities, visiting the Reddit web-site and watching those images for fun can be an attractive way to relax. And also, there might be cases where people have more holidays during December, so people spend most of their free time on online social platforms.

From the above graph, it is observed that the total attention of users increased from Jan to September and decreased from then.

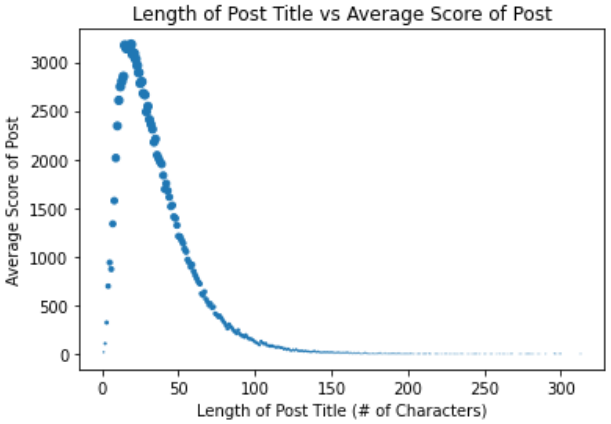


Graph 5: Week-wise analysis of Attention and Comments (engagement) to the success of the submission

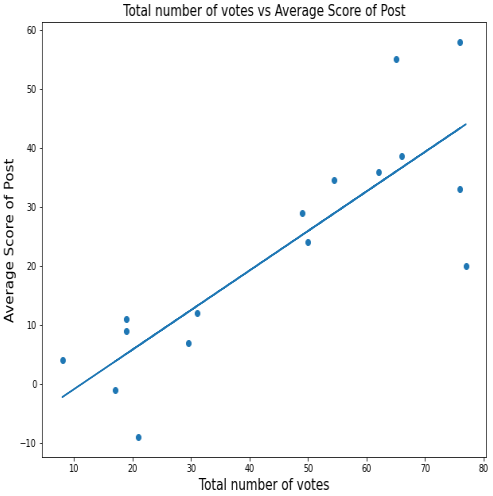
From the above plot, we can identify that the Success of posts is more on weekends, that is, Saturday and Sunday, over weekdays, success is shown with respect to the total number of upvotes.

We tried to explore if there is any relationship between the title's length to its success. The graph below is the correlation plot. The success of the posts is inversely proportional to the size of the post.

From this, the takeaway is that the authors, who would like to gain more popularity on their posts can tailor the length of their Titles, thereby it would increase the chances of submission success.



Graph 6: Analysis of the Length of the Title post to the success of the submission



Graph 7: Interpolation of a line graph between Total Votes and Score (Success of Post)

Word cloud generation with Reddit posts titles

III. Predictive Task

1. As part of the First Predictive Task, based on our data and associated correlations, we would like to predict the scores, given the number of upvotes, imageId, and UserID.

f (userId, imageId, upvotes) =scores (Success of Submission).

g(userId, imageId, total votes)=comments (Engagement)

We have evaluated this model using RMSE and R-squared error.

2. Text Classification Using Sentence Embedding | Sentence Transformers:

Built sentence transformer model for classifying the Reddit titles with the subreddit classes. We have used **all-mpnet-base-v2 and multi-qa-mpnet-base-dot-v1** semantic search models to map title words with 473-dimensional dense subreddit vector spaces. Further, we have evaluated the models using accuracy and precision, recall metrics. more details about the model were mentioned in model section

3. Text Classification using feature vectors, Word2Vec| Text Mining:
We have trained Uni gram, bi gram, and combination (Unigrams and bigrams) Bow models with feature representations of the text to capture the frequencies of word occurrences in a text corpus and further built a model with attention and engagement metric (Total votes and comments)

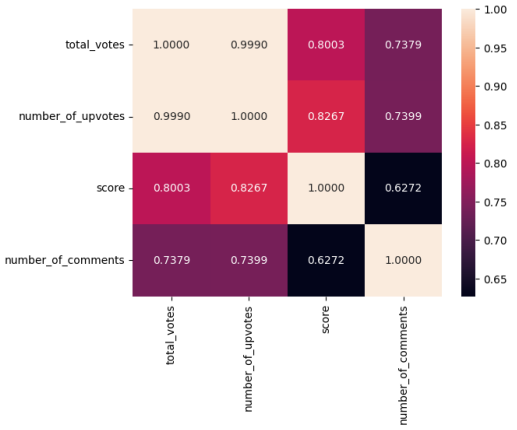
f(title words (unigrams))= Votes, Comments
g(title words (bigrams))= Votes, Comments
h(title words (most common unigrams and bigrams))= Votes, Comments

These models are evaluated by calculating their MSEs.

IV. Models

Model 1:

From the dataset, we have considered scores, total votes, Upvotes, downvotes, timestamps, and the number of comments. Considering each of these features individually, we have plotted a correlation matrix.



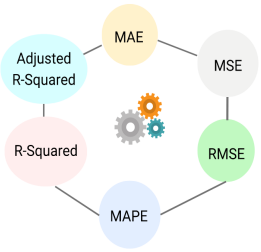
Graph 7: Correlation matrix among Dataset Features

Besides the multiple posts of the same image, The Reddit submissions dataset contains total votes(attention), score(rating/quality of content), and the number of comments(engagement) as some of the major features. To get better insights, we built a correlation matrix to exploit the dependencies. From the correlation matrix, we could observe strong relation between the following features:

a. Upvotes and score
b. Total Votes and Comments
c. Total Votes and Score
d. Upvotes and Comments

For each of these features, we have explored the Rmse, by fitting different models. The obtained Rmse and mae values in case of each of the cases are described in the tables below.

Evaluation metrics of models:



	model	r2	rmse	mae
0	Baseline	-0.001189	466.859371	315.321152
1	Linear Regression	0.685596	261.620672	153.315660
2	Lasso Regression	0.685574	261.629810	153.411395
3	Ridge Regression	0.685596	261.620672	153.315660
4	Elastic Net Regression	0.685571	261.630986	153.423316
5	KNN Regression	0.842251	185.315373	63.650633
6	Decision Tree	0.860707	174.137962	59.419740
7	Random Forest	0.857572	176.086370	60.111870
8	Gradient Boosting	0.855463	177.385550	58.744104

Fig. 2. RMSE, Mae for different algorithms while fitting the model for **Upvotes and score**

	model	r2	rmse	mae
0	Baseline	-0.000766	129.203539	55.861549
1	Linear Regression	0.587951	82.905376	23.101804
2	Lasso Regression	0.587981	82.902322	23.113404
3	Ridge Regression	0.642439	278.999378	166.399260
4	Elastic Net Regression	0.587982	82.902269	23.113610
5	KNN Regression	0.458336	95.054630	24.346398
6	Decision Tree	0.566607	85.025438	22.288011
7	Random Forest	0.554388	86.215744	22.513170
8	Gradient Boosting	0.567842	84.904201	22.032161

Fig. 3. RMSE, Mae for different algorithms while fitting the model for **Total Votes and Comments**

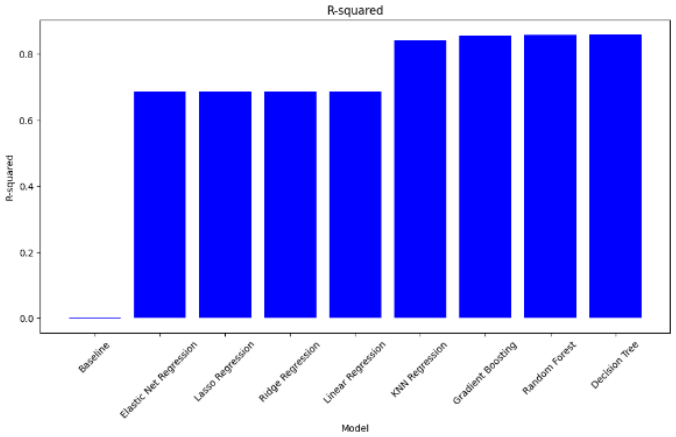
	model	r2	rmse	mae
0	Baseline	-0.001189	466.859371	315.321152
1	Linear Regression	0.642439	278.999378	166.399260
2	Lasso Regression	0.642417	279.007889	166.487069
3	Ridge Regression	0.642439	278.999378	166.399260
4	Elastic Net Regression	0.642416	279.008452	166.492711
5	KNN Regression	0.819164	198.413447	72.503323
6	Decision Tree	0.842570	185.128018	67.388149
7	Random Forest	0.836363	188.741887	68.379805
8	Gradient Boosting	0.842926	184.918528	66.426678

Fig. 4. RMSE, Mae for different algorithms while fitting the model for **Total Votes and Score**

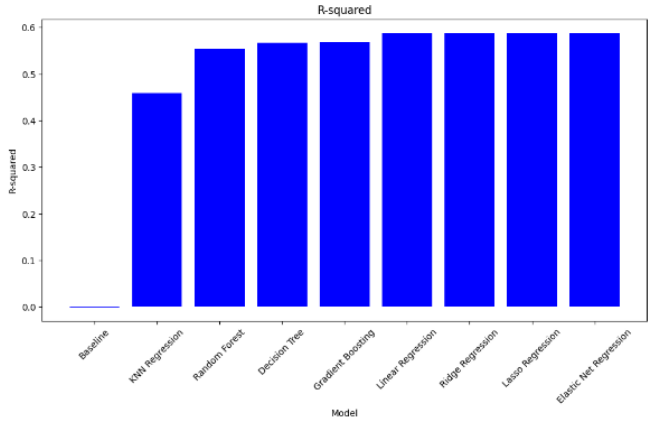
	model	r2	rmse	mae
0	Baseline	-0.000766	129.203539	55.861549
1	Linear Regression	0.592557	82.440674	22.332473
2	Lasso Regression	0.592587	82.437637	22.341944
3	Ridge Regression	0.592557	82.440674	22.332473
4	Elastic Net Regression	0.592588	82.437538	22.342262
5	KNN Regression	0.458113	95.074158	24.745733
6	Decision Tree	0.569603	84.731068	22.517741
7	Random Forest	0.556644	85.997256	22.678510
8	Gradient Boosting	0.523931	89.113431	22.162099

Fig. 5. RMSE, Mae for different algorithms while fitting the model for **Upvotes and Comments**

The corresponding graph is plotted according to the sorted values of R-Square Values.
The plots of Total Votes and Comments, Upvotes, and Scores are shown below.



Graph 8. Upvotes and Scores



Graph 9. Total Votes and Comments

Based on the analysis of the different features, The Elastic Net Regression, and Decision Tree are found to be effective when compared to the rest of the regression methods.

Unsuccessful Attempts:

In the given Reddit Submissions Dataset, the user and image instances are not repeated multiple times. This data cannot yield enough distinct pairs of users and images to draw a correlation.

We have tried to predict the future interaction using Jaccard, Cosine Similarity, and Latent Factor Models. However, the models returned an accuracy of 0.0039 which is very low. It further indicates that such a correlation drawn from the given dataset is not effective.

The Item2vec model built between user and image interactions yielded a very high mse error, the reason is that there are very few data records available for user image posts.

Strengths and Weaknesses:

Since our focus is on highly correlated components in the dataset, we could achieve higher accuracy with an efficient algorithm.

However, one of the shortcomings is that people consider different features to build their models. And on top of that, each model may not be interpreted in the same way.

One more drawback to consider is that the data considered from social platforms is highly volatile. And even when all the predictions have high accuracy, a user might act differently and could stop posting. Such unusual activities are highly unpredictive in the case of social media users.

Model 2: Text Classification Model Using Sentence Embedding | Sentence Transformers

The transformer models were trained on all available training data for word feature vectors. Initially input title words are passed through the sentence transformer model for obtaining the contextualized word embeddings.

The all-mpnet-base-v2 model provides the best quality, while all-MiniLM-L6-v2 is 5 times faster and still offers good quality. so we have used these models. Later logistic regression model is used to classify the feature representations with the subreddit classes; we obtained an accuracy of 48% and 55% for these two models. The precision, recall and accuracy values of these two models remained the same which means that false negatives count is the same as false positives count.

Model	Precision	recall	accuracy
all-mpnet-base-v2	0.48	0.48	0.48
multi-qa-mpnet-base-dot-v1	0.55	0.55	0.55

Minor issue faced with the transformer model:-
The transformer models cannot be trained for the posts with unique labels, So for this dataset, we have ignored the unique subreddit classes, in total we have removed 100 records before training the model.

Model 3: Text classification using Word2Vec

Most common unigrams	*Most common Bigrams*	*Most combined unigrams and bigrams*
girl by a eaten little ass at is looking	'how i' 'i feel' 'when i' 'this is' 'in the' 'of the' 'xpost from' 'on the' 'i see' 'feel when'	i the a this to my of in is when

Table 4. Unigrams and Bigrams

Text feature extraction, Unigram and bi grams:

	Model	Mse
Ridge regression model	Uni-gram	0.3425515714
	Bi- gram	0.3433493487
	Combined Unigrams and bigrams	0.342131700
Linear regression model	Uni-gram	0.342
	Bi- gram	0.344
	Combined Unigrams and bigrams	0.34

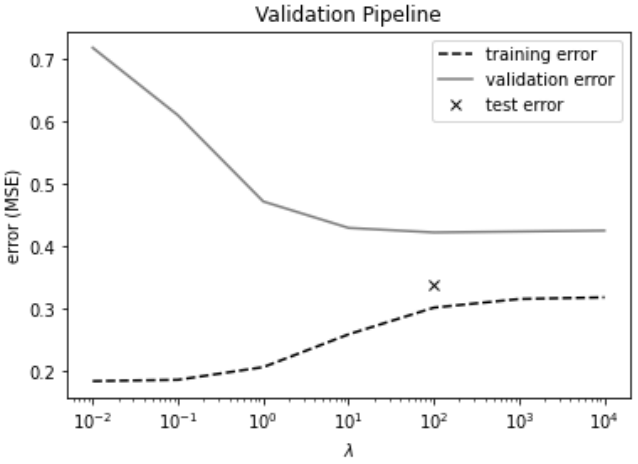
Table 5. Different models with corresponding MSE and Accuracy

Ridge regression:-

We have trained Uni gram, bi gram, and combination (Unigrams and bigrams) Bow model with feature representations of the text for capturing the frequencies of word occurrences in a text corpus and correlating them with attention and success of posts, Generally attention and success are called as total votes and the number of comments added to the posts. The accuracy obtained with the two models is 0.34, 0.36, and 0.38. The table summarizes the accuracies and mse for the linear regression model with varying maximum n-gram sizes of the feature vectors (i.e. the bigram model includes first and second-order n-grams).

Validation pipeline:-

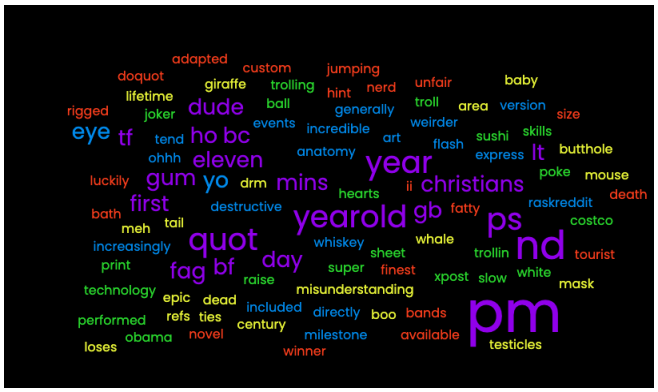
We explicitly measured the performance of multiple models in terms of a formal statistical framework. So far, we have compared (regression) models in terms of their Mean Squared Errors though as we explored in Section 2.2.2.



Training, validation, and test error on a real pipeline.

Word clusters based on word2vec similarities:-

All words from Reddit post titles are grouped into four clusters based on word2vec similarities



Word2vec similarities:-

Word	***Word2vec similarities***
Table	'mall', 'desk', 'bedroom', 'elementary', 'law', 'fridays', 'security', 'lamp', 'theater'
father	'grown' 'mood' 'business' 'windy' 'sleeping' 'hat' 'looking' 'mother' 'place' 'soon'
family	*'accident'* *'old'* *'cousin'* *'halloween'* *'letter'* *'brothers'* *'wallpaper'* *'photograph'* *'wasted'* *'costumes'*

Table 3. Word2vec Similarities

V. Literature Survey

Analysis of several factors contributing to the success of social media content has been done by incorporating many techniques. One approach is to use the Community model and Language model (Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec 2013). Where they have accounted for the effects of choosing a particular subreddit community on the success of the content. In their Language model, they have analyzed the impact a title has on the submission's success. These two models, along with the time stamps could give a better understanding of the vital factors that lead to the Reddit Submission's success.

Some works have focused more on user loyalty towards particular sub-communities. (William L. Hamilton , Justine Zhang, Cristian Danescu-Niculescu-Mizil 2017) have exploited the general patterns to predict the future rates of loyalty. From the Reddit user networks dataset, they could analyze why few users consistently prefer few communities over others.

The clip transformers are the current state-of-art methods to analyze the text and image data. Since our Reddit data constitute these relevant data. The transformers technique could yield better insights.

Although several methods have attempted to predict popularity before an item is submitted (Tsagkias, Weerkamp, and de Rijke 2009; Bandari, Asur, and Huberman 2012), one method uses metrics of an item's early popularity, such as view counts on youtube and digg.com, to predict its future success (Lee, Moon, and Salamatian 2010; Tatar et al. 2011). Others have even utilized these platforms to forecast the results of uncontrollable events, such as predicting box office receipts using tweets (Asur and Huberman 2010).

Others have analyzed the user interaction by the actions performed by the user. One of the key factors for the prediction task in Chen and Pirolli's (2012) study on participation in the OccupyWallStreet movement was how many tweets a movement follower had posted. The number of posts played a role in the participation prediction test conducted by Sadeque et al. (2015) as well. Additionally, Liu et al. (2016) considered a user's number of posts in their churn prediction work. Milosevic et al. (2017) used the game of session counts and click counts within the game to quantify the level of activity. In each of these situations, the level of activity was positively correlated with future involvement, meaning that a higher level of activity increased the likelihood of future engagement.

VI. Results

In this project, we performed exploratory data analysis of the overall data set and tried to make clear what features can be used to predict the popularity (upvotes, comments, scores, and so on) of one image based on the correlation plots and analysis. We further applied text classification techniques like bert models and text mining techniques like Word2Vec models to classify the post titles with different subreddit content labels based on feature representations. We obtained the same precision, recall and accuracy values which means that false positives count is equal to the false negatives count. The titles, count of votes, and comments are passed to different models as input parameters and user characteristics were retrieved in turn.

We created a linear and non-linear model for prediction with a feature vector. Bert models have resulted in good accuracy, precision, and recall values. We ultimately achieved a satisfactory outcome by gradually enhancing features and the modeling itself. Feature representation with the Item2Vec model has not worked well as the interaction data between user and image posts is very low.