
Indian Institute of Technology Delhi

# Subreddit Post Virality Prediction
# ELL880 Social Network Analysis Final Report

Animesh Singh Parihar (2021JCS2235)
Tooba Khan (2021JCS2245)
Ritik Jain (2021JCS2260)

Supervisors:
Sougata Mukherjea
Amit A Nanavati

November 28, 2021

# Abstract

The virality of material in social networks is influenced by user connectedness. In our work, we will be predicting the virality of a post on Reddit which is a content-driven network with no clearly defined user relationships. We will build a model for virality prediction of a post within a subreddit using the temporal(time and date), linguistic(the content of the title), and structural(user-user network) features.

We have used comments and posts of 1 month for a chosen subreddit and constructed a network using this data. Further, we have selected 5 features which includes temporal, linguistic and network features to train a classifier. This classifier predicts whether a given post will be viral in the particular subreddit or not.

# Contents

**1 Introduction** **4**

**2 Literature Review** **5**

**3 Methodology** **6**
3.1 Data Collection . . . . . 6
3.2 Construction of Network . . . . . 7
3.3 Feature Vector . . . . . 8
3.4 Prediction using binary classification . . . . . 8
3.5 Prediction using neural network . . . . . 9
3.6 Difficulties Faced . . . . . 10

**4 Conclusion and Results** **11**
4.1 Conclusion . . . . . 11
4.2 Results and Future Work . . . . . 11

# Chapter 1

# Introduction

A lot of research has been conducted on content virality on user-driven social networks such as Facebook and Twitter in which content is spread through sharing by poster's friends or followers. Comparatively less research has been conducted into how content becomes viral on Reddit, a content-driven social network.

Unlike other social networks, Reddit is unique in that a post's exposure to a widespread userbase is governed by a combination of post timing, quality of a post's content and title, the affinity between content and subreddit, and several other factors. There is a voting system which is the sole mechanism by which the success of a post is defined. Any user can upvote or downvote a post or comment and the post/comment score is defined as the difference between the upvote and the downvote which is used to select the post which is to be displayed on the first page of the subreddit.

In this project, we will explore the virality prediction of a post within a subreddit by training a machine learning model using the temporal, linguistic, and structural features which we've extracted from a subreddit. We will begin with creating a graph from the data of reddit submissions and comments. For this graph, we will calculate the structural properties of nodes. And for each post, we will extract it's temporal, linguistic and the structural features of the node it belongs to and create a feature vector using all these properties. We will then train a classifier for predicting the post virality using the same feature vector.

# Chapter 2

# Literature Review

Cheng et al. [1] use Facebook photo data to predict cascades of reshares. The authors define a family of prediction problems: given that we have already observed k reshares of a photo, can we predict if it will be reshared 2k times at some point in the future? Towards this task, Cheng et al. [1] employed machine learning models for prediction. Many features including temporal, structural, and content features were incorporated into the model.
Cheng et al. [1] provided a useful framework to build a model including structural features of the user-user network for virality prediction.

In their paper "What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media", Lakkaraju et al. [2] build a complex model to predict the virality or ultimate success of different versions (or "reposts") of the same image submitted to Reddit. Their model takes into account the content of the submission, the title of the submission, the community in which the network is posted, and the time when it is posted. The authors demonstrate a set of features that are useful in virality prediction.
We believe that many of the features described by Lakkaraju et al. [2] will also prove useful in our task.

Schonlau et al. [3] in their paper 'The random forest algorithm for statistical learning' have discussed an algorithm for classification problem for credit card holders to predict whether he/she will default or not.
The same model is been used in our task to predict that the post will become viral or not.

In the paper "Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks", Nguyen Nghiep et al. [4] compared the performance of multiple regression and the Neural network model for their task.
This provided the idea behind using two different models to predict the virality using two different models.First using the regression model to predict that the post will become viral or not in a binary form and then using the neural network to predict the actual scores of the post.

# Chapter 3

# Methodology

## 3.1 Data Collection

- For our project, we need data from Reddit. So, for this, we collected data from the PushShift dataset using API [5]. The PushShift dataset contains systemized information about authors, comments, subreddit, submissions, etc.
- Each submission entry contains information related to the author, submission id, score, URL, the number of comments, title, etc. Each comments entry contains the author, created time, an id for that post that the comments correspond to.
- For this purpose, we have used the subreddit-Gaming. It is among one of the popular subreddits in the world. Gaming subreddit contains information on (almost) anything related to games - video games, board games, card games, etc. It currently contains 30.7 Million Members to the current date(18 Oct 2021).
- In our project, we have used submissions and comments data of the Gaming subreddit. We have used the data of month Jan 2020.
- For submission entry, we have extracted the title, URL, author, submission id, score, created time, number of comments, Permanent link. For comments entry, we have extracted the id, link id, author, created time.

| Filename | Type | Size (bytes) | Date Modified |
|---|---|---|---|
| 69M_reddit_accounts.csv.gz | 69M_REDDIT_ACCOUNTS.CSV.GZ File | 1,051,903,601 | Sep 23 2018 1:33 PM |
| RA_2018-09.gz | RA_2018-09.GZ File | 1,104,136,829 | Oct 20 2018 4:11 PM |
| RA_2020-06-28.ndjson.zst | RA_2020-06-28.NDJSON.ZST File | 1,933,724,618 | Jun 28 2020 1:30 PM |
| RS_2019-09-01.gz | RS_2019-09-01.GZ File | 256,537,945 | May 14 2020 4:33 PM |
| authors | <Directory> | <Directory> | Apr 23 2019 5:54 AM |
| authors.dat.zst | AUTHORS.DAT.ZST File | 1,444,191,053 | Jul 15 2020 8:19 PM |
| comments | <Directory> | <Directory> | Aug 30 2021 3:17 AM |
| daily | <Directory> | <Directory> | May 14 2020 5:11 PM |
| moderators | <Directory> | <Directory> | Aug 5 2021 7:26 PM |
| poem_for_your_sprog.txt | POEM_FOR_YOUR_SPROG.TXT File | 1,453,290 | Oct 12 2019 1:16 AM |
| requests | <Directory> | <Directory> | Aug 5 2021 7:30 PM |
| rs_2020-11.ndjson.zst | RS_2020-11.NDJSON.ZST File | 10,494,096,522 | Feb 6 2021 6:55 AM |
| staging | <Directory> | <Directory> | Aug 5 2021 8:18 PM |
| submissions | <Directory> | <Directory> | Aug 7 2021 7:41 AM |
| subreddits | <Directory> | <Directory> | Aug 6 2021 1:54 AM |

Figure 1: The pushift data source

## 3.2 Construction of Network

Our project has been implemented in python. We have used the cloud platform, Google collaboratory with 16 GB RAM. The implementation of our code had the following steps:

1. Data collection: Described in the earlier section
2. Creation of Graph:
   (a) For the representation of the graph in python, we have chosen the Networkx library of python. Since our graph has directed edges, we have created a DiGraph object of Networkx.
   (b) Our graph is a directed graph where nodes represent authors. By authors, we mean that people who have posted in the specific subreddit chosen by us are known as authors of their respective posts. Also, Reddit allows people to comment on comments and thus sometimes the comments can also act as posts and become more viral than the original post.
   (c) Every post has some attributes associated with it and PostId and author are two of those attributes. The attributed author is used to create new nodes i.e Nodes of the graph are all the people who posted in the given time frame.
   (d) To get these nodes, we have used the CSV file we created in the previous section using submissions data crawled for a given time frame.
   (e) Next, to create edges, we define the relationship between two nodes. An edge from author A to author B represents that, author A commented on any one of the several posts posted by author B in the given time frame.
   (f) To create edges, we have used the CSV file we created using the comments data crawled for a given time frame.
   (g) For every comment posted, we have extracted the author of the comment and the linkId which is the ID of the post on which the author commented. For every such relation, an edge is created.
3. Saving the graph for future use: We have saved our graph consisting of 212842 nodes and 373026 edges as a gexf file. '.gexf' is a file extension that stands for Graph Exchange XML Format which is used to save complex graph structures.

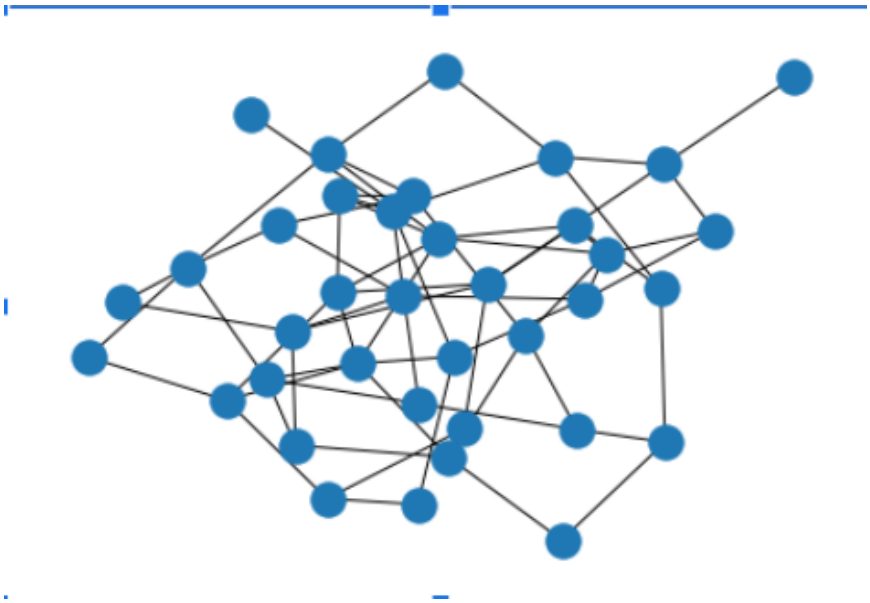

Figure 2: The subgraph obtained using erdos renyi graph with degree more than 5

## 3.3 Feature Vector

- For capturing all the details of the post, we have created a feature vector.
- We first ran Page Rank and Degree Centrality algorithms for structural features.
- For linguistic features, we have performed sentimental analysis on the post title.
- We have stored sentiment score as an attribute for each title.

| | PostID | Title | Date | NumberOfComm | SentimentScore | PageRank | DegreeCentrality | Score |
|---|---|---|---|---|---|---|---|---|
| 2 | ei7nwx | How i celebrate I | 2019-12-31 19:3 | 0 | 1 | 1.26E-06 | 0 | 1 |
| 3 | ei7o1k | The Elder Scrolls | 2019-12-31 19:3 | 4 | 0 | 6.12E-06 | 2.82E-05 | 1 |
| 4 | ei7oio | THIS SHOULD E | 2019-12-31 19:3 | 0 | 0 | 1.26E-06 | 0.0149193703 | 0 |
| 5 | ei7rxa | They say, "The w | 2019-12-31 19:4 | 2 | 1 | 1.26E-06 | 0.0149193703 | 6 |
| 6 | ei7vkp | This game is ruin | 2019-12-31 19:5 | 0 | -1 | 1.26E-06 | 0.0149193703 | 1 |
| 7 | ei829v | this game just en | 2019-12-31 20:0 | 1 | 0 | 1.26E-06 | 0.0149193703 | 1 |
| 8 | ej1jkw | Wife took the kid | 2020-01-02 17:5 | 0 | -1 | 1.26E-06 | 0.0149193703 | 1 |
| 9 | ej1ktq | Can we all show | 2020-01-02 17:5 | 6 | 1 | 1.26E-06 | 0.0149193703 | 29 |
| 10 | ej1n7p | Just a bandit cav | 2020-01-02 18:0 | 2 | 0 | 1.26E-06 | 0.0149193703 | 3 |
| 11 | ej1twe | 2019 was the wo | 2020-01-02 18:1 | 445 | -1 | 1.26E-06 | 0.0149193703 | 538 |
| 12 | ej1uli | Sometimes you j | 2020-01-02 18:1 | 0 | 0 | 1.26E-06 | 0.0149193703 | 1 |
| 13 | ej20wd | If I get a ps4 plus | 2020-01-02 18:2 | 8 | 1 | 1.26E-06 | 0.0149193703 | 0 |
| 14 | ej21hf | Finally my Strear | 2020-01-02 18:3 | 2 | 1 | 1.26E-06 | 0.0149193703 | 1 |
| 15 | ej24t0 | Scorpion in Mord | 2020-01-02 18:3 | 4 | 0 | 1.26E-06 | 0.0149193703 | 1 |
| 16 | ej28aq | New Year Same | 2020-01-02 18:4 | 0 | 0 | 1.26E-06 | 0.0149193703 | 1 |
| 17 | ej2dno | low effort meme | 2020-01-02 18:5 | 0 | -1 | 1.26E-06 | 0.0149193703 | 1 |
| 18 | ej2g56 | Brother dead, wif | 2020-01-02 18:5 | 0 | -1 | 1.26E-06 | 0.0149193703 | 1 |
| 19 | ej2gah | Fresh dose of sir | 2020-01-02 18:5 | 8 | 1 | 1.26E-06 | 0.0149193703 | 19 |
| 20 | ej2hf6 | How would you r | 2020-01-02 19:0 | 2 | 0 | 1.26E-06 | 0.0149193703 | 1 |
| 21 | ej2l5d | The Character M | 2020-01-02 19:0 | 32 | -1 | 1.26E-06 | 0.0149193703 | 86 |
| 22 | ej2mrp | My Skyrim cospla | 2020-01-02 19:1 | 0 | 0 | 1.26E-06 | 0.0149193703 | 1 |
| 23 | ej2nxj | How would you g | 2020-01-02 19:1 | 0 | 0 | 1.26E-06 | 0.0149193703 | 1 |

Figure 3: Sample Feature vector for model

## 3.4 Prediction using binary classification

- We have implemented Random Forest Classifier which is a binary classifier.
- We have split our data into 80% training data and 20% testing data.
- By obtaining mean of the score, we have fixed a threshold of 5. If any post have score more than 5, we have assumed to be viral or output=1 and vice-versa.
- We have achieved training accuracy of 99.94% and testing accuracy of 88.12%.

| Predicted | 0 | 1 | All |
|---|---|---|---|
| Actual | | | |
| 0 | 3781 | 497 | 4278 |
| 1 | 258 | 1711 | 1969 |
| All | 4039 | 2208 | 6247 |

Figure 4: Confusion Matrix obtained using RandomForestClassifier

- From Confusion matrix, we can observe that there are lots of True Positives and True Negatives. There are few False positive and False negatives. Overall,Our Classifier works fine.

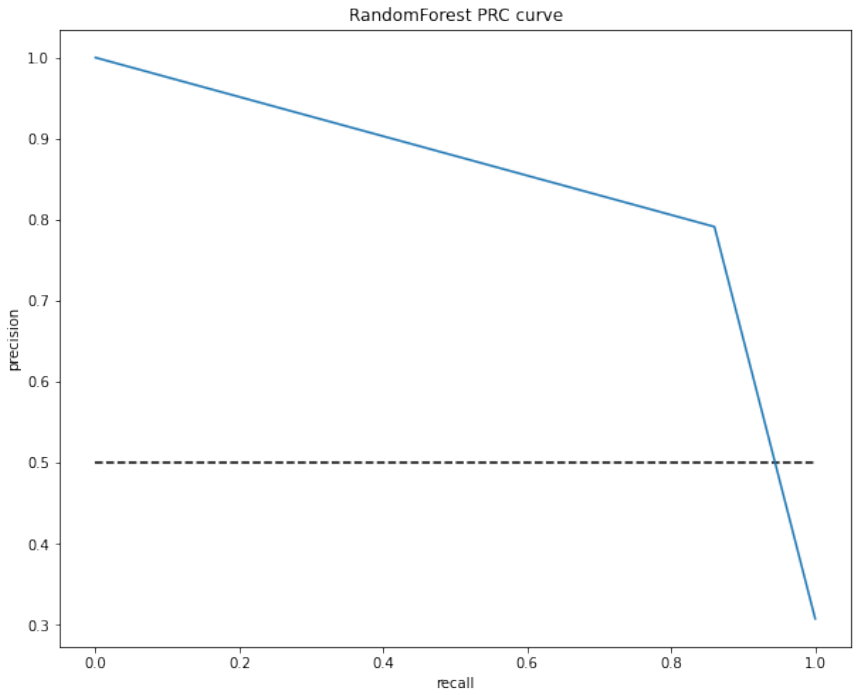


Figure 5: RandomForest PRC curve

## 3.5 Prediction using neural network

In our prediction using neural network, we have tried to predict the actual score rather than virality. We have extended and included this also in our project.

- We have implemented our neural network using Keras.
- We have splitted our data into 80 % Training and 20% Testing data.
- We have achieved training accuracy of 94.48% and testing accuracy of 94.67%.This training and testing accuracy shows that overfitting occurs.
- We could have reduced this problem using larger dataset of more than 1 month.
- We observed that few instances of high scores our model doesn't work well in predicting that score.

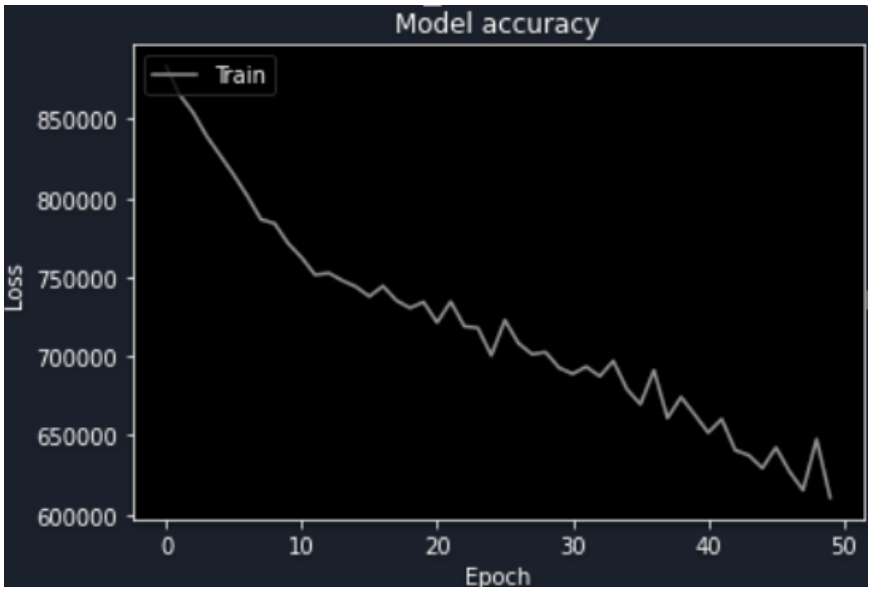


Figure 6: Loss as a function of epochs

## 3.6 Difficulties Faced

It can be noted, that while implementation, we have faced the following difficulties:

- The enormous amount of data available in the submissions and comments of Reddit requires much more resources than google colab can provide us because of which we had to limit ourselves to a shorter time span for collecting data.
- A shorter time span means that many authors who have commented on the posts might not have posted in that period of time and their nodes will have indegree 0. If we could have taken more data, the results would have been better and we could have got a large graph for analysis.

# Chapter 4

# Conclusion and Results

## 4.1 Conclusion

In this project, we have made a classifier to predict whether the given post will become viral in the chosen subreddit or not. We selected a subreddit and extracted it's authors and comments data. We converted this data into a Network such that there will be a directed edge from author A to B if the former commented on a post of the later. After this we extracted network features like Degree centrality and Page Rank. Using network features like degree centrality and page rank, linguistic features like sentiment score and temporal features like timestamp of posting the post and number of comments on it, we constructed a feature vector. Using this feature vector, we trained a random forest classifier.

## 4.2 Results and Future Work

We further extended our project by training a neural network which could not only predict if the post will become viral or not but it will also be able to predict an exact score that the post can achieve. The neural network model did not perform well because of limited data set.

- The classifier gave us 99.94% train accuracy and 88.12 % test accuracy.
- Loss of the neural network decreased in subsequent epochs but it did not come to a minimum.
- We suggest that if the neural network is trained on a larger data set of approximately more than a year, we can get a better result from it.
- More structural features can be worked with to compare what network features are best for the classifier and neural network.

# Bibliography

[1] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" *Proceedings of the 23rd international conference on World wide web - WWW '14*, 2014. [Online]. Available: http://dx.doi.org/10.1145/2566486.2567997

[2] H. Lakkaraju, J. McAuley, and J. Leskovec, "What's in a name? understanding the interplay between titles, content, and communities in social media," in *ICWSM*, 2013. [Online]. Available: https://cseweb.ucsd.edu/~jmcauley/pdfs/icwsm13.pdf

[3] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," in *The Stata Journal*, 2020. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/1536867X20909688

[4] N. Nghiep and C. Al, "Predicting housing value: A comparison of multiple regression analysis and artificial neural networks," in *Journal of Real Estate Research*, 2001. [Online]. Available: https://core.ac.uk/download/pdf/7162826.pdf

[5] B. K. M. S. J. B. Jason Baumgartner, Savvas Zannettou, "The pushshift reddit dataset," in *ICWSM*, 2020. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/download/7347/7201/10577