A Comparative Study of Information Spread in Twitter and Reddit

Ravi Sharma*
ravis@mail.usf.edu
University of South Florida
Tampa, Florida

Figure 1: Online Social Networks [9]

ABSTRACT

Challenges arise when comparing data between multiple different platforms. The two main issues are to compare characteristics and resolving digital bias. Different platforms have different characteristics for example popularity, influence, etc. Based on this for instance, Twitter allows for posts to be shared internally in the form of a retweet, however Reddit posts are shared externally via the URL. In this project we try to bring Reddit and twitter into one ground and extract the twitter tweets and Reddit posts for a list of domains and try to find out which social network is more proactive in sharing information. We also, collected list of users who shared first on either platform for our future work to see the correlation between the users of different platform. We observed that for our three sets of extracted data, Reddit users are more proactive in sharing information first. It may not be generalised to all data available but instills curiosity for our future work towards larger Dataset.

CCS CONCEPTS

• **Information systems** → *Social networks.*

KEYWORDS

Twitter, Reddit, Information Spread, Network homophily

1 INTRODUCTION

Online social networks play a major role in the spread of information at very large scale. Their benefits include allowing people to exchange their opinions with others and sharing information with others on the other end of the line [4]. Essentially, this can be used to promote dialogue and a solution without in order to achieve the goal of the initiative. Online social networks allow hundreds of millions of Internet users worldwide to produce and consume content and make connections. When exposed to existing technological applications, they quickly discover that they can create new kinds of relationships. The promise of social networking services is great: individuals can create communities and operate networks for the benefit of others. However, the history of social networking has been riddled with missteps. A well-run network can also cause a lot of harm, such as the improper use of personal information and automated bots that reinforce the sense of community and create a toxic environment for friendship and partnership. Many netizens

*The project is part of course offered - Social Media Mining in Fall 2019

communities, for example Tumblr, thrive on fun and fun-loving culture, with playful memes, funny rants, and hyperbolic social commentary. Their sassy, innocent side comes off as cute to viewers, and the game-like nature of these communities makes it easy to see how the wrong person could end up wasting their time on such things.

While the online social platform(OSNs) has provided us the tools to communicate and spread information, the same OSNs have crippled us with miss information, rumours, fake news, identity thefts, impersonation and many more challenges. In recent years, OSNs user-base has grown at an precedented rate due to technological advancements. Events, issues, interests, etc. happen and evolve very quickly in social networks and their capture, understanding, visualization, and prediction are becoming critical expectations from both end users and researchers [4]. Information spread through Social Media Platforms like Twitter and Reddit, with more than 330 Million users a piece, have grown significantly in recent years. They are used for election polls, news, marketing, socializing, recruitment, etc. [6]. Industries, government, companies, sports and news agencies all take the advantage these Social Media Giants to reach the respective users. Structure of both the platforms is different. Where Twitter is an microblogging and social networking service on which users post and interact with messages known as "tweets" and Reddit is social news aggregation, web content rating, and discussion website containing discussion boards called sub-reddits and users as redditors. [16] The focus of the paper is as follows:

(1) To identify which social platform is more active in sharing the information first from an information source.
(2) To identify the user characteristics across both platform sharing information and finding correlation between them.
(3) To identify the sentiment score of tweets/posts of the users.

The rest of the paper is organized as follows. Section 2 talks about related work. Section 3 first talks about the basics of social media platforms Twitter and Reddit and its structure. It then talks about the terminologies used. Section 4 talks about the method, data-set description and extraction procedure. Section 5 presents the results, finding, limitation and technical challenges with observations. Section 6 describes the concluding remarks and future directions.

2 RELATED WORK

Information spread in social media is widely studied. Most recently [7] et al. collected 600K tweets and 190 posts from reddit for #meToo movement and showed that twitter is helpful in sparking an emotion or information and reddit is more into spreading the information with detailed comments. Information spread in social networks is majorly been studied in detection of false information and rumours. [15] in his thesis developed a model using user dynamics, linguistic style and network propagation to detection of rumour with 75% accuracy. Akhtar et al. [1] used rumour data set provided by [3] and perform stance classification and rumour veracity prediction using hierarchical LSTM approach. Vega et al. [14] investigated the curiosity level of users and proposed a model to suggest the difference between information spread from curious vs non curious users. Doing so they were able to relate diffusion process and dynamical heterogeneous spread in online social platforms like google+ and Facebook.

3 BASICS OF SOCIAL MEDIA - TWITTER AND REDDIT

There exists a vast number of online social networks (OSNs) offering different features to facilitate their users. They have different overlying structure but still follow the simple relationship structure. Informally, an online social network consists of a platform that facilitates its users to open an online personalized account which allows then to create their profile page and post messages. OSNs allows the user to connect with other users and share their posts. This creates social ties or relationships among the various users. Formally, OSNs can be seen as a graph with nodes representing the users and edges representing the relationship among them. Like any graph, the OSNs can be directed or undirected graphs depicting the type of relationship among users. For example: 1) A is a friend of B implies B implies B is a friend of A. This is an example of undirected graph. 2) A follows B but B does not follows A. This is an example of directed graph. The former can be seen as bilateral relationship and latter as unilateral relationship. The information shared among users can be images, text, links, documents, videos, etc. The type of information can range from personal messages with friends, families, etc., opinions about a topics such as sharing political views, products reviews, etc. Along with the center information, the additional data about the users is also shared which includes time of creation of post/message, author name/author screen name and other similar characteristics. The subsequent subsections talks about the structure of Twitter and Reddit.

3.1 Twitter

Twitter is an microblogging online platform with more than 330 million users. It allows user to post messages called 'tweets'. The ubiquity, accessibility, speed and ease-of-use of Twitter have made it an invaluable communication tool. People turn to Twitter for a variety of purposes, from everyday chatter to reading about breaking news [5]. Twitter allows users to tweet, retweet, follow other users and reply to other users tweets. The following terms are generally more common than others in twitter. [13]:

- Tweet: A Tweet is a message (up to 280 characters) which may contain photos, GIFs, videos, and text.
- username:A username (or handle) is how you're identified on Twitter, and is always preceded immediately by the .
- Timeline: A timeline is a real-time stream of Tweets. Any user timeline will show the users and their friends tweets who they follow.
- Follower: A follower is another Twitter account that has followed you to receive your Tweets in their Home timeline.
- Retweet: A Tweet that you forward to your followers is known as a Retweet.
- Like: A heart icon on a tweet indicates the appreciation. By tapping the like icon a user is informed that their tweet is been appreciated.
- Hash Tag: A hashtag is any word or phrase immediately preceded by the # symbol. This is generally used for spreading the information with a specific word or phrase.

- Direct Messages: Direct Messages are private messages sent from one Twitter account to another account(s). They can be used for both group as well as single user.

3.2 Reddit

Reddit is a social news aggregation, web content rating, and discussion website. Reddit Users called 'Redditors' are submit content to the site such as links, text posts, and images, which are then voted up or down by other members. Posts are organized by subject into user-created boards called "subreddits", which cover a variety of topics including news, science, movies, video games, music, books, fitness, food, and image-sharing [16]. Reddit has more than 330 Million users worldwide and provides rich source of information via discussions, link sharing and other user generated content. The following terms are generally more common than others in reddit:

- Redditor: Reddit user is known as redditor. This includes users and moderators. Any user is denoted as $/u/$<username>. Moderator is the user who manages the subreddit.
- : Post: A Reddit post is a submitted content which can be either text, messages, links, urls, images or videos. The post can be a reply to an already posted content on a thread(called subreddit) or a single post to begin with.
- Subreddit: Every Reddit post by a redditor is a part of discussion. This discussion board is called as subreddit which is categorised via different topics such as location, news, etc. A subreddit is denoted by $/r/$<subreddit-name>
- Upvote and downvote: Reddit users can show their support or appreciation using upvote or disliking or oppose using downvoting. The cumulative score of upvotes and downvotes decides the popularity of the post.

4 MODEL DEFINITION AND DATASET COLLECTION

To identify which social platform is more active in sharing the information first from an information source, we modelled our method as follows:

(1) First we selected list of top 10 domains from the top 500 sites on ranking website Alexa based on country and news category each.
(2) We then take this lists and search for these domains in twitter and Reddit for specified dates.
(3) We extracted the matching URls from both the platforms and then compare the timestamp of corresponding tweet and Reddit post.
(4) We count the number of Urls having timestamp smaller than the other for each platform.

4.1 Alexa

Alexa traffic rank also known as Alexa rank [2] is a ranking system of websites based on the popularity of the website. The popularity of the website is analysed by the amount of internet traffic on the website. This includes time spent on the website, number of queries, number of unique IP addresses visiting the website, number of page views of the website, etc. The rank is a result of the analysis done for all available domains browsed by the alexa servers over a period of 3 months. The Alexa ranking has many criteria for ranking. Along with the global scores, it provides ranking based on countries as well as categories such as news and sports. We selected two sets of domains from Alexa. The following table lists the domains used for data extraction.

Table 1: Alexa Top sites

Country(US)	News-category
google.com	nytimes.com
youtube.com	news.google.com
amazon.com	cnn.com
facebook.com	theguardian.com
yahoo.com	shutterstock.com
reddit.com	indiatimes.com
wikipedia.org	washingtonpost.com
ebay.com	news.yahoo.com
bing.com	forbes.com
netflix.com	foxnews.com
office.com	weather.com

4.2 API Structure and Data Extraction Process

The tweets and posts were extracted for a set of 3 specific time frames.

- November 4^{th} to November 14^{th} - set 1 domains
- November 29^{th} to December 2^{nd} - set 2 domains
- December 3^{rd} to December 9^{th} - set 2 domains

4.2.1 ***Twitter.*** In order to extract the data from twitter, we performed the following steps:

(1) To extract realtime tweets, we make use of filter API [12] provided by twitter developers platform.
   - We take the list of domains from set 1 i.e. the top 10 sites in US and begin tracking the domains all at once. This means whenever any tweet containing the domain listed in set 1 appears in realtime, we will be extracting that tweet. We used the endpoint 'track' to filter the tweets in stream. In order to cover url and sub-url of the domain listed, we used the query perimeter as "domain com" instead of "domain.com". This results in excessive redundant tweets which we filter later to keep only relevant tweets with valid url.
   - similar approach was carried out for list of domains from set 2 i.e. top 10 sites from news domain. Some of the domain were omitted from the list in order to remove redundancy. In place of the omitted domains the next popular domains were inserted in the list.
(2) We used the Tweepy [11] which is a python wrapper for twitter API to extract the tweets. Streaming of tweets with filter API consist of following steps:
   - First create a class which will inherit the source class StreamListner.
   - Using the class created above, we created a Stream object that will listen to the stream of tweets. To keep the Urls

from truncating we extracted the tweets in extended mode. $stream = Stream(auth, listen_r, tweet_mode =' extended')$
- Using the stream object we connect to twitter API and filter the tweets based on the our track. Only the English language tweets were extracted for the sake of uniformity. $stream.filter(track = queries, languages = ['en'], stall\_warnings = True)$

4.2.2 ***Reddit***. In order to extract the data from Reddit, we performed the following steps:

(1) We used the PRAW [8] which is a python wrapper for Reddit API to extract the posts, submissions and other data from reddit. Streaming of posts consists of following steps:
   - We extracted all the submissions from Reddit and check weather there exist a domain present in the set1. This can be checked by extracting submission.url and checking it is not equal to the peramlink of the post. $submission\ in\ reddit.subreddit('all').stream.submissions(skip\_existing = True)$
(2) We also calculated the posts for the set 1 using DomainListing model as follows:
   - reddit.domain('domain.com').new()
(3) similar approach was followed for set 2 of domains as well to extract the posts.

4.3 Data Distribution - Twitter

For set 1, tweets were collected from 4th November till 14th of November. A total of arount 7.6 Million tweets were parsed containing the keywords "<domain> <com/org>". For set 2, two set of datasets were extracted. One for date range November 29 - December 2 and other for date range December 3 - December 9. A total of 1.6 Million and 2.9 million tweets were parsed respectively for the given dates. This is summarized in table 2

Table 2: Tweets Extracted

set	Date-range	Tweets-before-Filtering	After-filtering
1	11-04 to 11-14	7.6 mi	837
2	11-29 to 12-04	1.6 mi	13008
2	12-03 to 12-09	2.9 mi	20880

4.4 Data Distribution - Reddit

Similar set range and dates were applied for extraction of posts from reddit. For set 1 a total of 7.9 million posts were extracted. For set 2 and 3 a total of 20612 and 42796 posts were extracted. This is summarized in table 3

Table 3: Posts Extracted

set	Date-range	Posts-before-Filtering	After-filtering
1	11-04 to 11-14	7.9 mi	9021
2	11-29 to 12-04	20612	11821
2	12-03 to 12-09	42796	18987

4.5 Data Filtering

Filtering process was based on following parameters:

- No retweets: Tweets beginning with 'RT @' were not considered. This was done in order to avoid the urls shared again and again. This is necessary since we are only interested in the tweets with urls first shared.
- Contains Urls: Only tweets which contains a url were considered. This was done using the regular expression to find a url in a tweet text. [10]
- No redirection to any twitter/reddit status: Tweets/posts which contains Urls as twitter/reddit links which redirects to another twitter status or Reddit post which in turn contains domains listed in set 1 or 2 were ignore to keep things simple and less computationally expensive. Additionally, all tweets/posts with only twitter/reddit links were removed from parsing.
- No repeated Urls: No repeated urls were kept. If two urls matched, one which has small timestamp(came first) was kept.

After filtering process tweet set 1 was reduced to 837 tweets, set 2 for 1st range of dates reduced to 13008 tweets and for range two reduced to 20880 respectively. Similarly, post set 1 was reduced to 9021 posts, set 2 for 1st range of dates reduced to 11821 posts and for range two reduced to 18987 posts respectively. These tweets and posts contains unique links in their respective files and needs to be checked with each other.

5 RESULTS AND OBSERVATIONS

After the tweets and posts were saved in respective csv files, a column wise comparison is performed. Every url of tweets file is compared with every url of post file. After this comparison 123 links were common for set 1, for data ranges in set 2 a total of 766 and 1375 links were common respectively.

5.1 Url Comparison

Table 4: URLs Compared

Total Links	Total Overlapped	T-1st	R-1st
9858	126	23	103
24829	766	176	590
39867	1375	206	1169

As shown in the table 4 and figure 2, Reddit users seems to outnumber twitter users for different set of domains. Though, comparison fails to follow any pattern with the % of twitter users first at 22% then increases close to 30% and then again decreases to 17.6%. One significant difference between the URLs obtained for set 1 and set 2 is that even though a fraction of tweets and posts were there for set2, number of URLs obtained after filtering is much higher than set1. This may be because of particularly the following reasons:

(1) The tweets/posts gets filtered for set 1 due to most of the links redirecting to twitter and Reddit for domains listed in set 1.

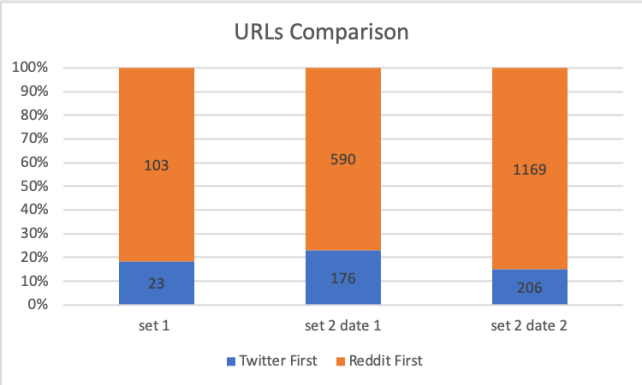

Figure 2: Comparison of Matched URLs Across Reddit and twitter

(2) The tweets gets filtered because of the URLs being shared in retweets.
(3) The tweets/posts extracted were mostly repetitive where a large amount of data gets filtered because of the URLs being shared again and again.
(4) People may not have shared URLs with domains in our set and data contains simply '<domain>' or 'com/org'.

Overlapping URLs are just 1%, 2% and 3% of 3 date ranges. This signifies that amount of URLs shared across the two platforms consist of very other different URLs from same domain range. When it comes to news domains, both twitter and reddit are more proactive in sharing the information or links first. This may be because of the large user-base of both the platforms and news agencies have their own channels across both the platforms. The information shared across such a large OSNs may result in both popularity of their network as well as the spread of information faster. Also, when news articles are shared across these platforms, people tend to share the information based on their opinions.

5.2 Limitations

The URL comparison though gives us Reddit a clear winner in terms of link sharing, the methodology is seriously crippled by a series of limitations described as follows:

- First of all, the standard API provided by both twitter and reddit have different bottlenecks. While twitter streaming API allows us to extract all the tweets which are less than 1% of all tweets(roughly 60 tweets per second at any moment on average, though it may vary), reddit on the other hand allows us to extract in the range of 25-100 posts per API call. This may result in a large number of missed posts and tweets.
- The extraction query only considers the URLs which are having domains from our sets. This poorly covers the entirety of the URLs being shared.
- Though our set 2 results in more overlapping URLs it does not generalises the fact that news related information spreads faster in both Twitter and Reddit. We have omitted the links which redirects to reddit and twitter status and also embedded tweets. This resulted in a massive reduction of tweets obtained. Without all the tweets we cannot justify the users preference of news links vs other resources such as memes, videos or e-commerce links.
- We also omitted the retweets from our final data set. This also is a huge loss of data. Though it avoids getting repetitive links, it also make us lose the URLs which are embedded in retweets and retweets which have comments. These comments may have additional URLs.

6 CONCLUSION, DISCUSSION AND FUTURE WORK

We analysed the URLs shared across the two platforms. With two different types of domains resulting in two different types of datasets spanning across 3 date ranges. The results obtained are preliminary and does not generalise the fact that information is spread or shared quickly if its a news article neither does it shows that reddit is more responsive in sharing the URLs. Our work does not correlates with lydia et al. [7] where twitter was first to share the information across platform and reddit was more into spreading the information. However, we have not verified the fact that weather the information spread is more common for what type of data or category (e.g. news vs jokes). Also, since our data extraction spans across 20 domains and does not rely on a specific url to track or follow, our research is more diverging as compared to all the related work listed.
Our future works consist of 3 fold research:

(1) We have conducted a date range based extraction of tweets for a comparatively shorter time span. We plan to work on larger range of date ranges in future. This will allow us to conduct a more detailed analysis of the data. We will try to extract, out of limit data not covered by the streaming API using Rest APIs and removing the redundant tweets in process. As proposed by Vega et al. [14] we want to carry extraction of data for both near an event, product launch or sports game as well as holiday season to see the impact of curiosity on both platform users while sharing the information.
(2) We have extracted the username of the users who were involved in sharing the information first on Reddit and Twitter for the date range December 3_{rd} - December 9_{th}. We want to find the correlation between these users across platform. Doing so, we want to analyse whether they form a Network homophily with their cross platform counterpart.
(3) Finally, we want to analyse the behaviour of these users in their respective platform to find their influence and popularity. We want to analyse and predict the impact of these users in propagation of any information across their network.

7 APPENDICES

GitHub-link : https://github.com/ravis3011/social-Media-Mining

ACKNOWLEDGMENTS

To Dr. Giovanni for providing excellent resources for the course Social Media Mining and the project.

REFERENCES

[1] M. S. Akhtar, A. Ekbal, S. Narayan, and V. Singh. 2018. No, That Never Happened!! Investigating Rumors on Twitter. *IEEE Intelligent Systems* 33, 5 (Sep. 2018), 8–15. https://doi.org/10.1109/MIS.2018.2877279

[2] Alexa: . 2019. The top 500 sites on the web. https://www.alexa.com/topsites [Online; accessed 10-December-2019].

[3] Genevieve Gorrell, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. RumourEval 2019: Determining Rumour Veracity and Support for Rumours. *CoRR* abs/1809.06683 (2018). arXiv:1809.06683 http://arxiv.org/abs/1809.06683

[4] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A. Zighed. 2013. Information Diffusion in Online Social Networks: A Survey. *SIGMOD Rec.* 42, 2 (July 2013), 17–28. https://doi.org/10.1145/2503792.2503797

[5] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07)*. ACM, New York, NY, USA, 56–65. https://doi.org/10.1145/1348549.1348556

[6] Anders Olof Larsson. 2017. Top Users and Long Tails: Twitter and Instagram Use During the 2015 Norwegian Elections. *Social Media + Society* 3, 2 (2017), 2056305117713776. https://doi.org/10.1177/2056305117713776 arXiv:https://doi.org/10.1177/2056305117713776,

[7] Lydia Manikonda, Ghazaleh Beigi, Huan Liu, and Subbarao Kambhampati. 2018. Twitter for Sparking a Movement, Reddit for Sharing the Moment: #metoo through the Lens of Social Media. *CoRR* abs/1803.08022 (2018). arXiv:1803.08022 http://arxiv.org/abs/1803.08022

[8] PRAW Developers . 2019. PRAW: The Python Reddit API Wrapper. https://praw.readthedocs.io/en/latest [Online; accessed 10-December-2019].

[9] Software Engineer Training . 2019. Social Logo. http://software-engineer-training.com/social-media-wars-2014-facebook-vs-twitter-vs-google-vs-pinterest-vs-instagram-vs-tumblr-vs-reddit/ [Online; accessed 10-December-2019].

[10] Stackoverflow: Allan. 2019. Regex matching a set of characters except when one of them is last in python. https://regex101.com/r/03VgN5/5/ [Online; accessed 10-December-2019].

[11] Tweepy Developers . 2019. Tweepy: The Python twitter API Wrapper. http://docs.tweepy.org/en/latest/streaming_how_to.html [Online; accessed 10-December-2019].

[12] Twitter Developers . 2019. Twitter Filter API. https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter [Online; accessed 10-December-2019].

[13] Twitter Developers. 2019. Twitter Glossary. https://help.twitter.com/en/glossary [Online; accessed 10-December-2019].

[14] Didier A. Vega-Oliveros, Lilian Berton, Federico Vazquez, and Francisco A. Rodrigues. 2017. The Impact of Social Curiosity on Information Spreading on Networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (ASONAM '17)*. ACM, New York, NY, USA, 459–466. https://doi.org/10.1145/3110025.3110039

[15] Soroush Vosoughi. 2015. Automatic detection and verification of rumors on Twitter. (01 2015).

[16] Wikipedia contributors. 2019. Reddit — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Reddit&oldid=928191923 [Online; accessed 10-December-2019].