Predicting information diffusion on Twitter – Analysis of predictive features

Arbaz Khan
23100243@lums.edu.pk

Muhammad Murtaza Hassan
23100252@lums.edu.pk

1. PROBLEM STATEMENT

In this project, we are attempting to forecast if a post will be forwarded or not. Additionally, we want to forecast how much it will diffuse. We will find the spread of tweets using the number of retweets, and we will do the sentiment analysis of our data to find the sentiments that people have on our selected topic. We will be working with Twitter which we have extracted using the Twitter developer account and a python and a Python script. By using the attributes of a tweet, we will try to predict the number of retweets a tweet may get.

2. BACKGROUND

The use of Twitter has increased, and most discussions occur on this forum. The spread of any specific news is carried by its retweet feature.

Users with more followers tend to have more retweets and favorites on their posts.

To find the spread of information, we will use a machine learning algorithm to predict the number of retweets on a specific post. Posts containing '#CPEC' will only be included in our dataset.

The reason for selecting this topic is to find the rate at which the information will spread in the south Asia region. An algorithm to find the sentiments of the tweets is also used to record the reviews of the people for this project.

3. MOTIVATION

Information propagation on online social networks focuses much attention on almost every domain. It is essential to model information dispersal in these expanding communication channels to comprehend better and manage information spread.

Twitter is one of the most famous social networks. We will be working on the CPEC topic, a hot topic in South Asia. For the past 10-15 years, China and Pakistan have been working on this project. Due to its vast impact on Pakistan and China, it has become a critical project in South Asia. With the results of our classifier, we will be able to find the spread of the information and will be able to predict diffusion in future tweets.

We will also be analyzing the tweets' sentiments using unsupervised learning. This will then help us to find the reviews or sentiments of the people who belong to Pakistan and China. We will be able to find the thoughts of the people affected by the project, and we can plan to compensate them better.

Overall, with our research, we can predict the number of retweets a tweet may get. We will also be able to analyze the sentiments from the tweet, which will give us the response of the people of this region.

4. LITERATURE REVIEW

Several surveys or literature reviews have been conducted regarding the study of information diffusion on social networks (SN) in general and not specifically on Twitter. No previous review

articles performed a complete systematic literature review.

Kakar and Mehrotra conducted a review of 90 filtered papers from six databases: namely, Scopus, Science Direct, ACM Digital Library, Springer, IEEE, and Google Scholar. This work focuses on three research areas under the umbrella of information diffusion in social networks, namely influence modeling, influence maximization, and retweet prediction. However, it did not discuss the metrics and measures used by researchers.

Riquelme and González-Cantergiani conducted a survey on the size of a user's influence on Twitter. This work collected and classified various measures of influence on Twitter. Some were based on simple metrics, and some were based on complex mathematical models. Various criteria were given to determine the most influential users on Twitter. However, an information diffusion model was not discussed in this work.

There are multiple research papers on social network diffusion, which can be replicated and improved by adding to that work.

5. DATASET

At the start, we crawled data using a python script, scraped data without Twitter APIs, and collected around 100,000 tweets. However, we need the data of the followers and their location. Therefore, we again used the streaming API of Twitter to collect data and successfully crawled hundreds of tweets. Twitter offers a sample stream through its streaming API, which returns a 1% sample of all public tweets.

Each tweet includes text and metadata, such as the timestamp, the user's screen name, and the application used to post the tweet. The data contain information about millions of users. Data extracted using the Twitter API contains almost 300 different attributes which can be used to for various research. In our case, we require mainly the number of retweets and if user has enabled location on his device, then we can access his coordinates too.

To get more detail about the tweet we implemented the sentiment analysis to separate the tweets into different classes.

6. METHODOLOGY

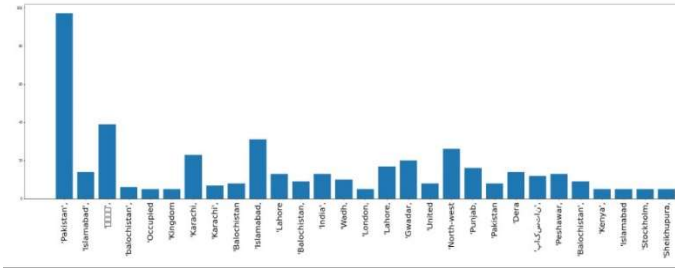
Firstly, after collecting the data we analyzed the tweets and we were successful in extracting various attributes of the tweets for example user data, date, retweet count, like count etc. Then we spent a few days in analyzing that data so that we can decide a suitable approach for our research.

For making the data suitable for the sentiment analysis, we started off with the preprocessing of the data, first we changed the alphabets to lowercase letters. Then, we replaced all the break statements to single space, we also removed the stopwords from the tweets. Stopwords are the words that do not have any meaning, so they do not affect the sentiment of the tweet. We also removed numbers and punctuation marks from the tweet as it does not have any effect while carrying out sentiment analysis. At the end we saved it into a new csv file so that we can use it later.

To perform sentiment analysis, we used the python's scikit learn library. We have used the KNN classifier to find the sentiments of the tweets.

	Unnamed: 0	content	retweetCount	sentiments
0	0	key lesson china pakistan economic corrido...	0	0
1	1	dgpr khi meeting col mukhtar butt manzar naq...	1	0
2	2	ccp xijinping curse humanity anyone eve...	0	0
3	3	adipkr ccp xijinping curse humanity any...	1	0
4	4	happynewyear coastwaybuilders gwadar cpec ...	1	0
5	5	war within pakistan establishment powerful fr...	1	-1
6	6	pak china failed convert cpec actionable inst...	0	0

To predict the retweets, we modified our data so that it can easily be used in our classifier. We used random forest classifier to predict the retweet count for the future tweets. We also made a graph for number of tweets versus location of the user.



When we calculated accuracy, it occurred to be 46.15% which means our 46.15 per cent of predicted values exactly matched with the true values. As we are predicting a number, if our predicted value fluctuates by 1 or 2 (which is common in the case of predicting numbers) with the true value, it counts as a false prediction. So, to cope with this challenge, we calculated the accuracy for ranges. If our predicted value lies within a range of +- 10, then we are saying our prediction is correct. By doing this, our accuracy reaches 73.78%, which is a good accuracy score. We have also attached the result of our random forest classifier.

	text	predicted_retweet
1382	RT @shahsabg: .@GermanyinPAK trip #Gwadar, bea...	4
1062	RT @JiRongMFA: The #China-#Pakistan Economic C...	21
438	Rural #Pakistan: New #Infrastructure Driving #...	0
868	RT @iqbal_iips: China has agreed to provide a ...	1
1219	RT @ChinaUrdu_: گوادر میں انٹرنیشنل ایئرپورٹ کے...	83
408	RT @ChanakyaForum: Chinese Port In Myanmar – T...	27
1335	RT @uniofgwd: Students of the @uniofgwd partic...	8
856	RT @CPEC_Official: BRI's biggest beneficiary i...	1
549	RT @TheDailyCPEC: The 5th Rashakai Special Eco...	32
1102	RT @MakranUpdates: سی پیک کو وقتی طور پر ایک ...	4

7. LIMITATIONS

Due to hardware constraints, we were not able to work with the complete data as it would take hours to train our model. If we had access to GPUs we could have used a lot more data than what we are using now.
We got access of the twitter developer account too late, so we had to start from scratch to work with the new data.

With the change in the machine learning classifier, it could have improved our results. Due to non-availability of ground truth data, we were unable to test the data and find the accuracy of our results.

8. CONCLUSION

In this paper, we describe a machine learning approach for predicting retweets on a tweet and analyzing its sentiment.

We collected data using Twitter API and we implemented different data cleaning techniques to make it suitable for our machine learning models. Due to time and data constraints, our proposed model could not perform strongly enough.

Nevertheless, evaluation showed that our approach can predict the number of retweets and in future we can also use the data of people who are retweeting to find complete chain of the spread of tweet.

Overall, our machine learning approach can be improved and enable more accurate detection of abusive third-party applications.

9. REFERENCES

Kakar, S.; Mehrotra, M. A review of critical research areas under information diffusion in social networks. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 383–396.

Riquelme, F.; González-Cantergiani, P. Measuring user influence on Twitter: A survey. *Inf. Processing Manag.* **2016**, *52*, 949–975.

Felfli, Z.; George, R.; Shujaee, K.; Kerwat, M. Potential-driven model for influence maximization in social networks. *IEEE Access* **2020**, *8*, 189786–189795.

Arora, A.; Bansal, S.; Kandpal, C.; Aswani, R.; Dwivedi, Y. Measuring social media influencer index-insights from facebook, Twitter and Instagram. *J. Retail. Consum. Serv.* **2019**, *49*, 86–101.