VERACITY OF INFORMATION IN TWITTER DATA
REPORT

TARUN BAKSHI

OBJECTIVE

Twitter is a powerful real time blogging service which provides information instantly to millions of people around the world. Over the past few years some people have used twitter service to spread false information. The objective of this project is to determine whether veracity of information in twitter data is high in unverified users or verified users. I will be comparing the tweets on mainly three factors

1. Diffusion Index
2. Geographic Spread Index
3. Spam Index

DATA

The data I was using for this project is twitter tweets. I have used utilities.scala file to verifying my twitter credentials. I am using spark streaming to extract the tweets from twitter. I have filtered the tweets as I got from the streaming data. Firstly, I have filtered the tweets so that I can get only the tweets that are in English. Secondly, I got the tweets which contain the keyword "sports" in the tweets. Thirdly, I checked if the user of that tweet is verified or unverified. Lastly I have extracted ScreenName, text, location, verified and time zone. Total number of tweets after filtering is 621(verified and unverified).

Size of the data

Unverified: 234 KB Verified: 5 KB

Period of collection:

Collection of tweets was done over irregular intervals of time which amount to 23 hours.

METHODOLOGY

I will be reading two files which will contain unverified and verified processed user/tweet information. The methodology for processing information on unverified and verified data is the same. I will be calculating the following from the given data.

1. Number of unique users
2. Number of unique location
3. Distribution of tweets by users

And

$$Diffusion\ Index = \frac{\#\ \text{Unique users}}{\text{Total tweets}}$$

---------[1][2]

According to [1][2], Diffusion index will help us decide how fast the information has spread through twitter. If diffusion index is high, then the chances of containing false information is greater.

$$Geograpic\ Spread\ Index = \frac{\#\ Unique\ Location}{Total\ tweets}$$

---------[1]

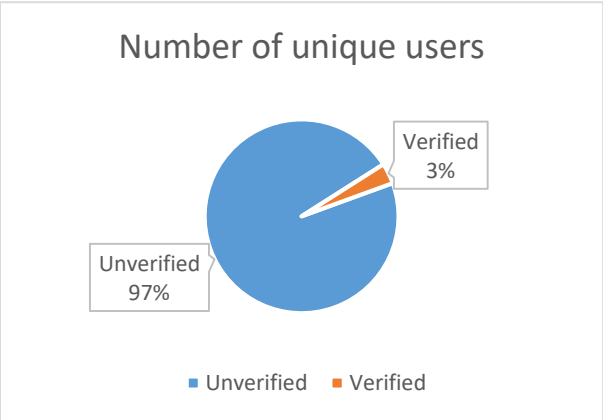
According to [1], Geographic spread index will help us in knowing how far the information has spread in terms of geographic breadth.

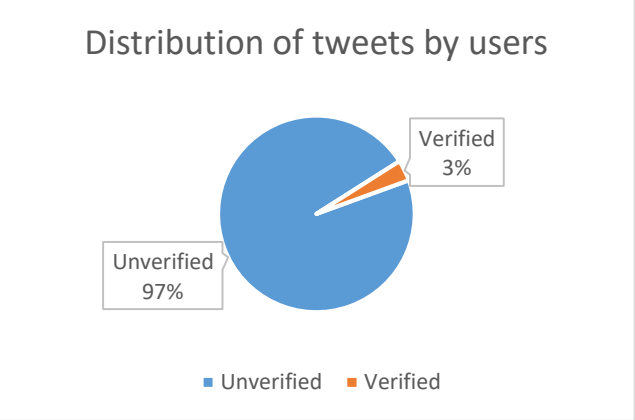
$$Spam\ Index = \frac{\Sigma_{over\ unique\ users} \frac{1}{unique\ user\ tweet\ count}}{Total\ tweets}$$

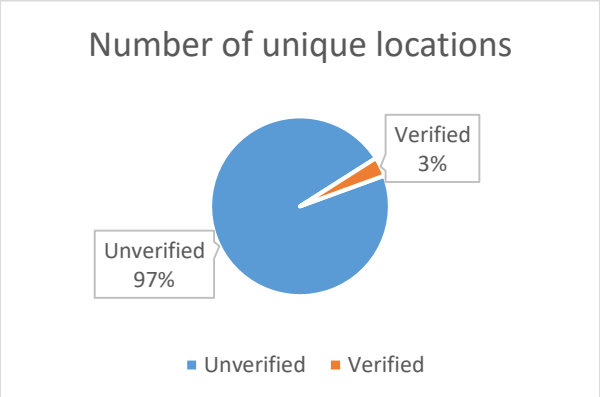
----------[1]

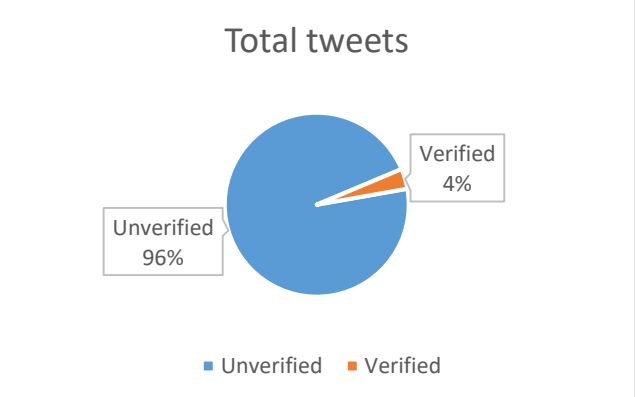
According to [1], Spam index measures the impact of repeated tweets by the same user. This is similar to spamming where the user repeated tweets the same tweet to spread false information.

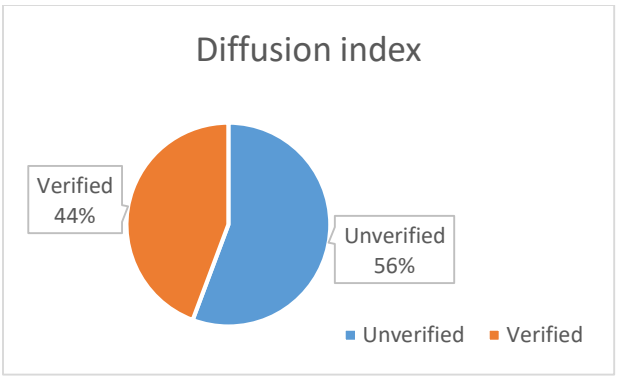
If diffusion index, geographic spread index and spam index is greater, then our assumption is those tweets are carrying false information. After the calculation of the above we get the following results (all results are converted into graphical format for representation)

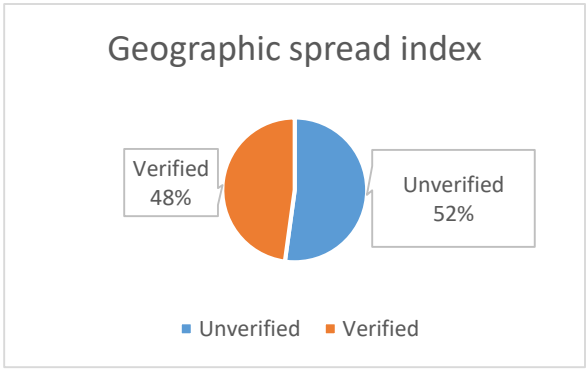












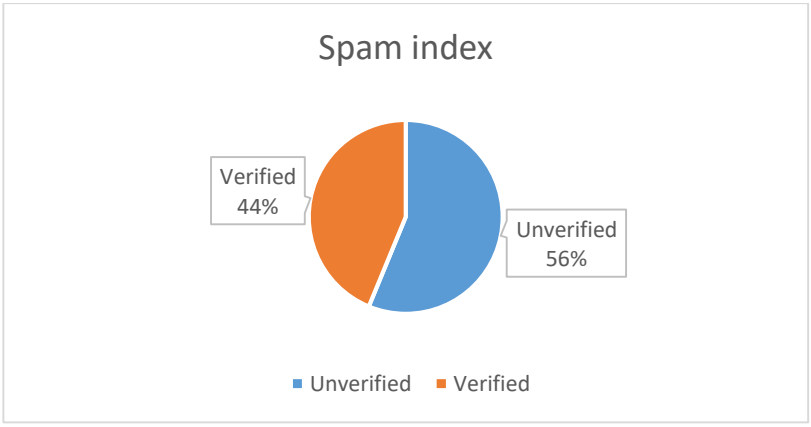


Table 1: Veracity results for "sports" keyword

Content	Unverified	Verified
Number of unique users	582	17
Distribution of tweets by users	504	18
Number of unique locations	505	17
Total tweets	599	22
Diffusion index	0.9716193656093489	0.7727272727272727
Geographic spread index	0.8430717863105175	0.7727272727272727
Spam index	0.9632721202003339	0.75

From the above table we can assume that veracity of information is more in verified dataset

For further understanding of data, I have congregated the tweets using keywords "golf"," cricket" and "nba" and calculated the veracity for each of the category. I have got the following results

Table 2: Veracity results for “golf” keyword

Content	Unverified	Verified
Number of unique users	2	1
Distribution of tweets by users	2	1
Number of unique locations	2	1
Total tweets	3	1
Diffusion index	0.6666666666666666	1.0
Geographic spread index	0.6666666666666666	1.0
Spam index	0.6666666666666666	1.0

From the above table we can assume that veracity of information is more in unverified dataset.

Table 3: Veracity results for “nba” keyword

Content	Unverified	Verified
Number of unique users	17	0
Distribution of tweets by users	12	0
Number of unique locations	17	0
Total tweets	17	0
Diffusion index	1.0	NA
Geographic spread index	1.0	NA
Spam index	1.0	NA

The veracity of tweets cannot be compared as there are no tweets for verified data. But we can assume that unverified tweets may contain false information as the 3 factors are equal to one.

Table 4: Veracity results for “cricket” keyword

Content	Unverified	Verified
Number of unique users	3	1
Distribution of tweets by users	4	1
Number of unique locations	3	1
Total tweets	4	2
Diffusion index	0.75	0.5
Geographic spread index	0.75	0.5
Spam index	0.625	0.5

From the above table we can assume that veracity of information is more in unverified dataset

CONCLUSION

After above analysis of twitter data, I have observed that percentage of veracity is more in verified data than unverified data of twitter users. Our results can be used in understanding the veracity of tweets. Further applications of this project can be used to know where and how the false information is spreading based on topic and location and predict if the user might produce false information.

REFERENCES

[1]Ashwin Kumar TK, Prashanth Kammarpally, KM George. Veracity of Information in Twitter Data: A Case Study

[2] Tapia, A. H., Moore, K. A., and Johnson N. J. Beyond the Trustworthy Tweet: A Deeper Understanding of Microblogged Data Use by Disaster Response and Humanitarian Relief Organizations. Proceedings of the 10th International ISCRAM Conference – Baden - Baden, Germany, May 2013, pp: 770-779.

[3] http://spark.apache.org/docs/latest/streaming-programming-guide.html

[4] https://www.tutorialspoint.com/scala/