VERACITY OF INFORMATION IN TWITTER DATA
REPORT

TARUN BAKSHI

TABLE OF CONTENTS

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 language. 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. In the last I have extracted ScreenName, text, location, verified and time zone.

Size of the data

Unverified: 234 KB Verified: 5 KB

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. Number of unique tweets

And

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

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.

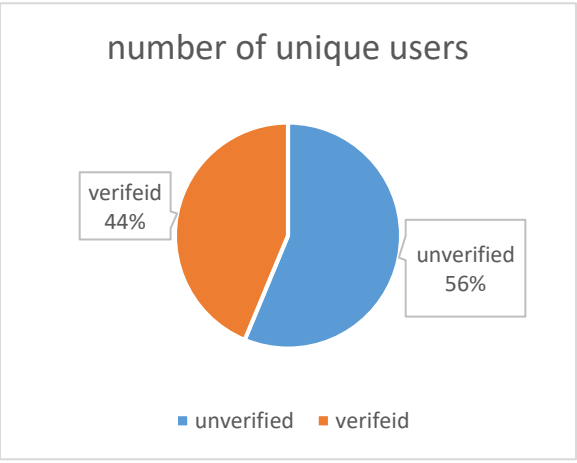
$$Geograpic\ Spread\ Index = \frac{\#\ Unique\ Location}{Total\ tweets}$$

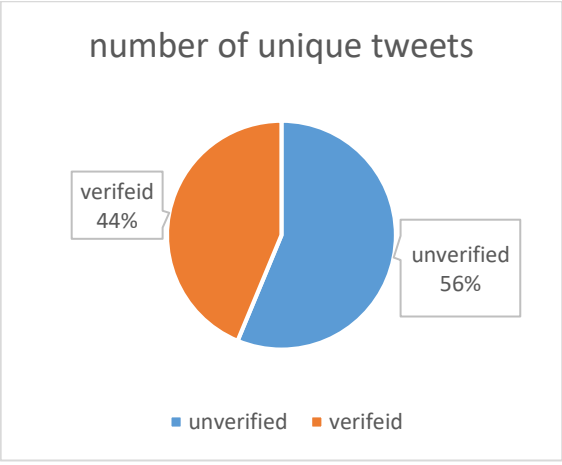
Geographic spread index will help us in knowing how far the information has spread in terms of geographic breadth.

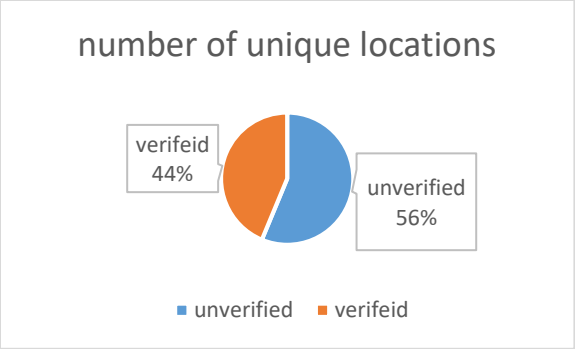
$$Spam\ Index = \frac{\sum_{over\ unique\ users} \frac{1}{unique\ user\ tweet\ count}}{Total\ tweets}$$

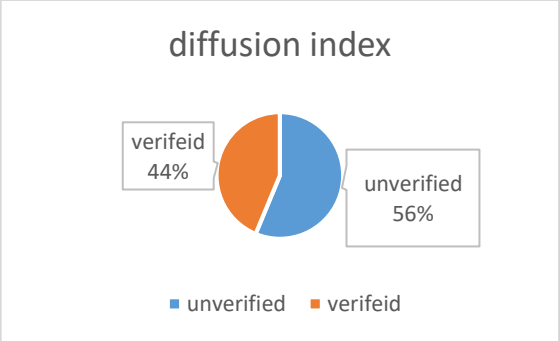
Spam index measures the impact of repeated tweets by the same user. This is similar to spamming where the user repeated tweets the same tweet to spread false information.

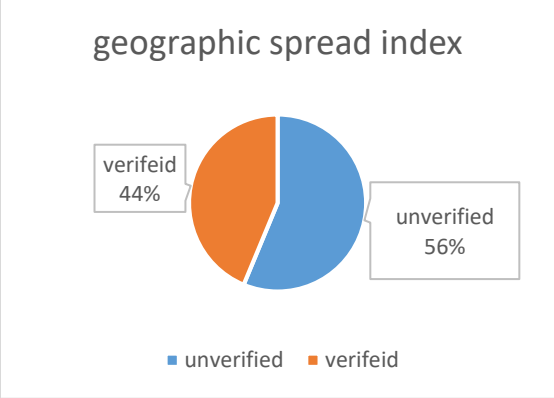
If diffusion index, geographic spread index and spam index is greater, then our assumption is those tweets are carrying false information. After the calculation of the above we get the following results (all results are converted into graphical format for representation)

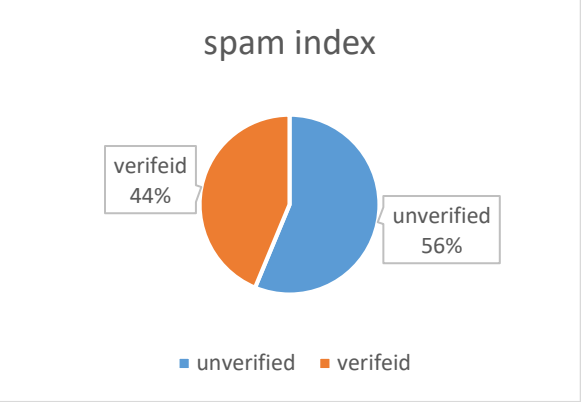












CONCLUSION

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 to introduce a way to compare the tweets qualitatively rather than quantitatively.