# VERACITY OF INFORMATION IN TWITTER DATA

## REPORT

TARUN BAKSHI

### TABLE OF CONTENTS

OBJECTIVE

DATA

METHODOLOGY

CONCLUSION

#### OBJECTIVE

The objective of this project is to determine whether veracity of information in twitter data is more 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 is twitter tweets. I have used utilities.scala file to verifying my twitter credentials. The method that we are importing the tweets is using spark streaming. 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, get only the tweets which contain the keyword “sports” in the tweets. Thirdly, check 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

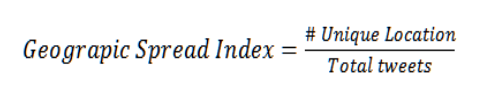
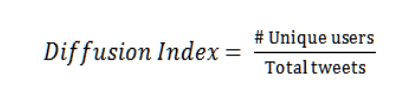
Unverified: 234 KB Verified: 5 KB

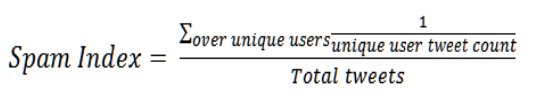
#### METHODOLOGY

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





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 if more in verified when compared to unverified tweets. 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 quantitative.