NATIONAL INSTITUTE OF TECHNOLOGY, KARNATAKA

Analyzing the Behavior and Bot Detection among Twitter Users

Presented by

Shaikh Sahil Ahmed (192474)
Kushal Mondal (192596)
Sampat Kr Ghosh (192655)

Under the guidance of
Dr. Sowmya Kamath S.

CONTENTS

1. Introduction
2. Literature Review
3. Methodology
   a. Data Collection and preprocessing
   b. Data processing
   c. Data Visualization
   d. Analysis
4. Results
   a. LCS Analysis and Detection of Groups
5. Bot Detection
6. Results
   a. Bot Detection
7. Conclusion

INTRODUCTION

- Social media are powerful tools connecting millions of people across the globe
- Social bots are accounts controlled by software, algorithmically generating content and establishing interactions
- There is a expanding record of malicious applications of social bots
- In this paper, we have analysed the behavior of Twitter bots and compared the performance of standard bot detection algorithms

LITERATURE REVIEW

- Bild [3] studied over aggregate user behavior and the retweet graph. They go through all the points such as Power Law Participation Momentum, Heavy-Tailed Rate Distribution, Heavy-tailed Interevent Distribution, Small-World, Assortative and Clustered Retweet Graph.
- Hana Anber [4] analysed over different information analysis techniques such as different hashtags, identification of influence, twitter’s network-topology, event spread over the network, and analysis of sentiment.
- Chithra R G [5] described to avoid the viral or unwanted or scary tweets from the media and this will helps to give the best recommendation towards the positive users through their experience in the social site.

DATA COLLECTION AND PREPROCESSING

- Dataset was collected from http://mib.projects.iit.cnr.it/dataset.html
- One group of 8.5 million tweets from more than 3000 genuine users
- Three groups of more than 5 million tweets from more than 3000 bot users
- Dataset was observed to find necessary attributes which includes - user id, retweet status, in reply to, hashtag number, url number, mention number
- First category of dataset contains tweet types and second category contains content types
- Required attributes were taken to form a new dataset.
- The new dataset is sanitized to remove records with null values and wrong values.
- The clean dataset then was used to perform necessary modification

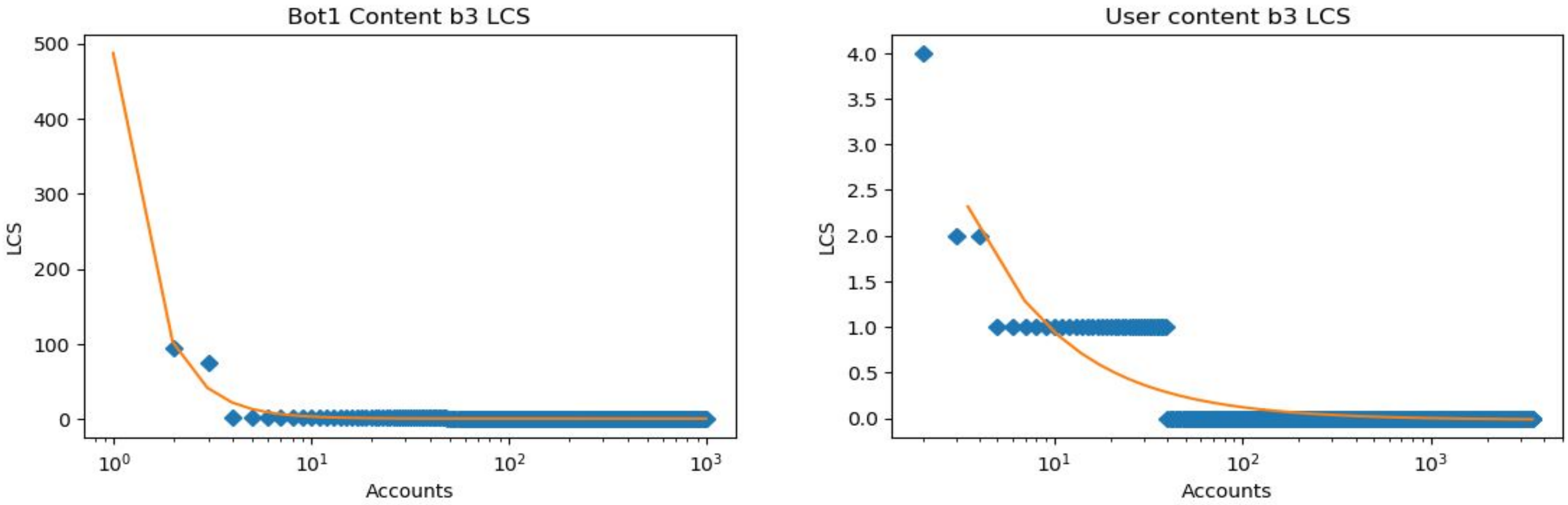
DATA PROCESSING

- First category of dataset was encoded with the following scheme:
  - A = normal tweet
  - C = reply to a tweet
  - T = retweet
- Second category of dataset was encoded in following manner:
  - N = tweet contains no entity
  - E = tweet contains one or more entity of one type
  - X = tweet contains entity of mixed types
- Second category was further encoded in following manner:
  - N = normal text tweet
  - U = tweet that contains one or more URL
  - H = tweet that contains one or more hashtag
  - M = tweet that contains one or more mentions
  - X = tweet that contains one or more mixed types

DATA PROCESSING

- For each user id we created the encoded strings and created another data file for that.
- Thus for each group and each category type we created one data file.
- For each of this file LCS was calculated among the number of strings present in the file.
- These LCS were visualized and analysed.
- We tried to fit the LCS data points with exponential curve and later analysed the area under the curve to get an insight for the behavior of the groups.

DATA VISUALIZATION

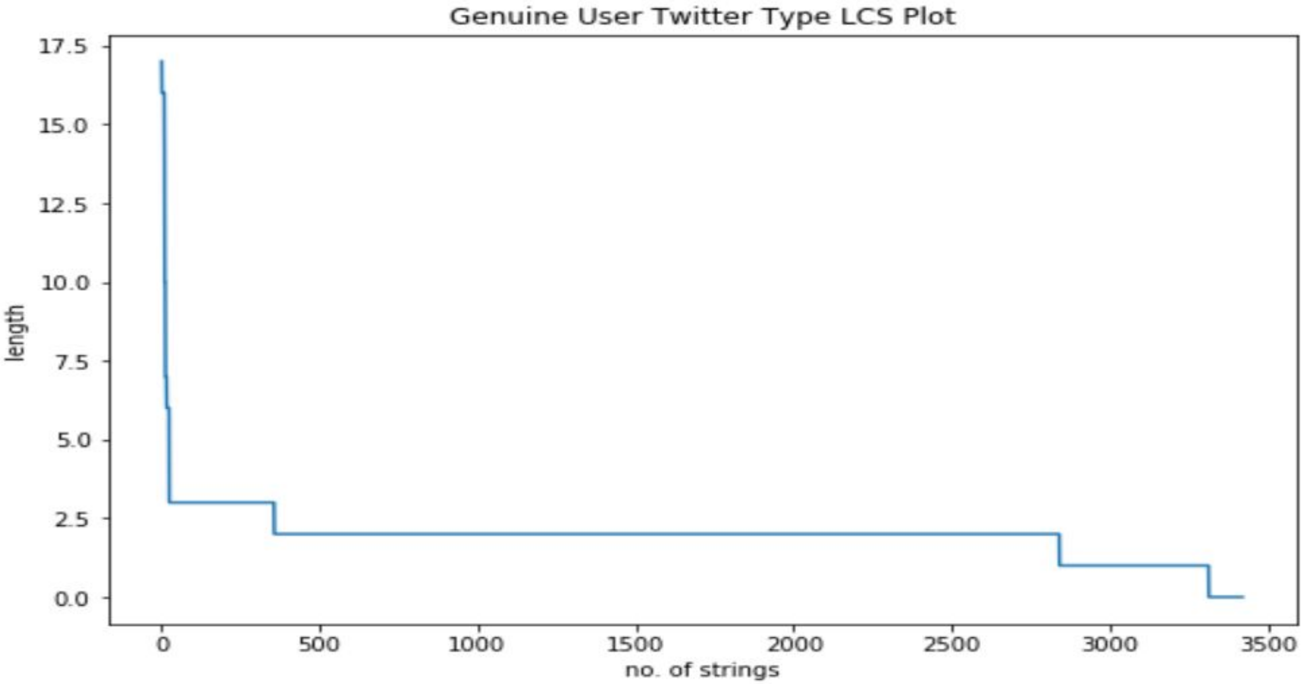


LCS plot and curve fitting example for tweet content category for genuine users group and Bot 1 users group.

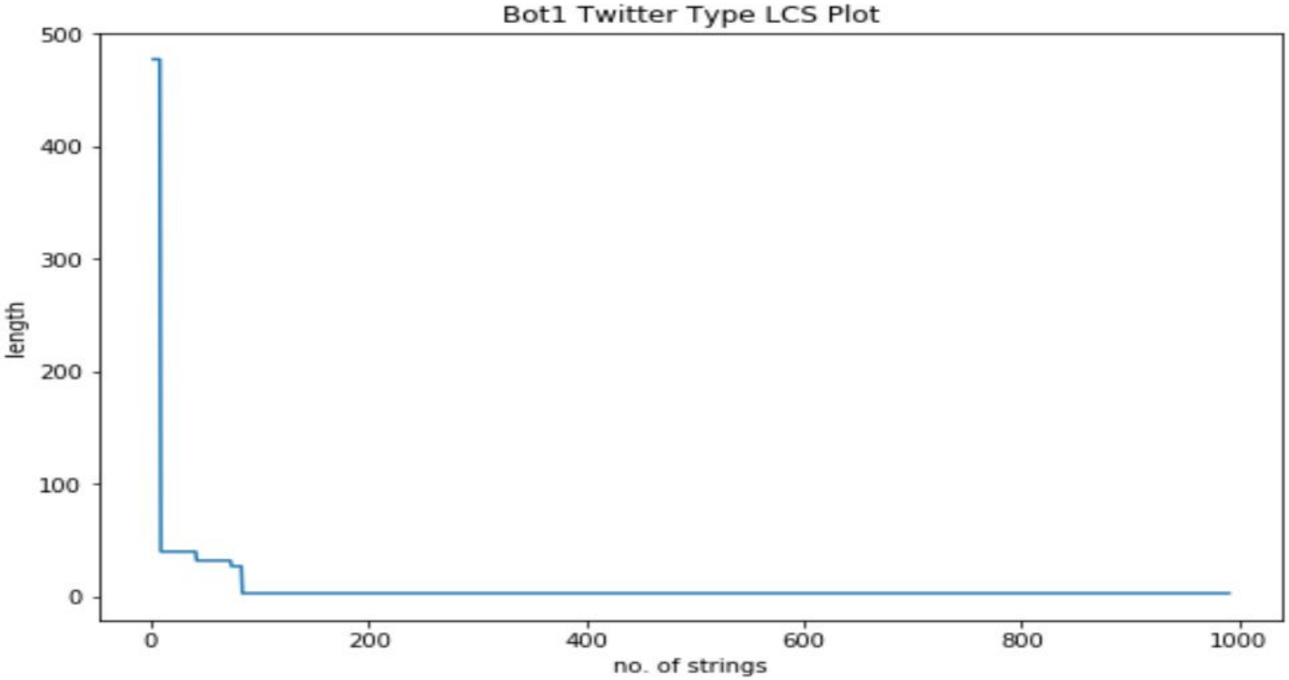
ANALYSIS

- LCS has a time complexity of O(n+m)
- When this approach is extended to k number of strings then the time complexity becomes O((n+m)*k)
- We have tried to find the length of longest of common substrings from the k value of 2 to n number of strings
- We faced limitation of computing ability of our machines
- We observed that probability of finding the LCS is very high in longer strings than the shorter ones since each string had very small number unique characters
- We also observed that if L is the LCS of k strings then L is a superstring of the LCS of k+1 strings

ANALYSIS


Genuine User Twitter Type LCS Plot
length
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
500
1000
1500
2000
2500
3000
3500
no. of strings

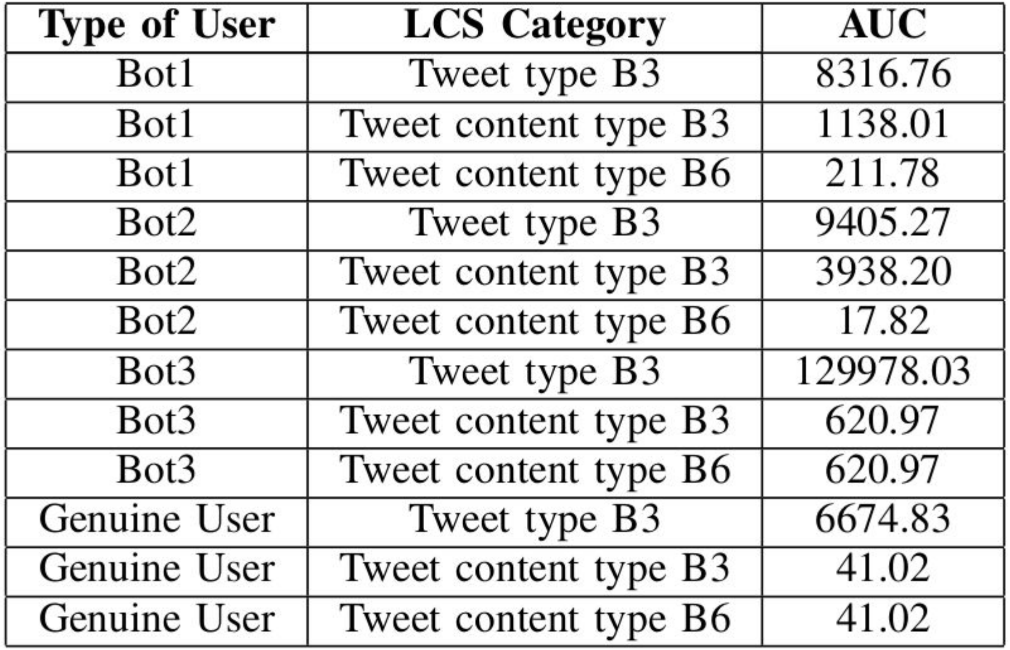
ANALYSIS


Bot1 Twitter Type LCS Plot
length
500
400
300
200
100
0
0
200
400
600
800
1000
no. of strings

RESULTS

a) LCS Analysis and Detection of Groups

i) Calculate AUC of the LCS curves using definite integrals on the curves

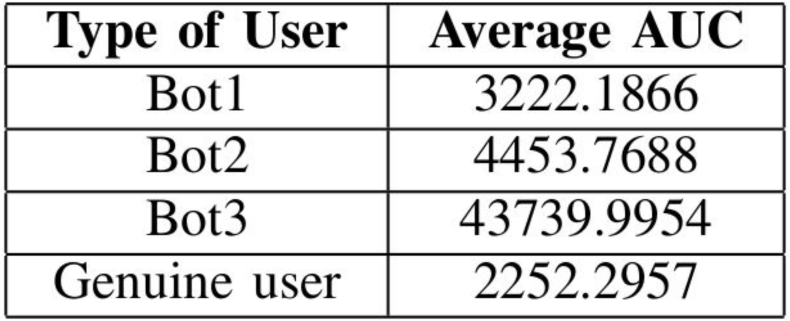


Type of User	LCS Category	AUC
Bot1	Tweet type B3	8316.76
Bot1	Tweet content type B3	1138.01
Bot1	Tweet content type B6	211.78
Bot2	Tweet type B3	9405.27
Bot2	Tweet content type B3	3938.20
Bot2	Tweet content type B6	17.82
Bot3	Tweet type B3	129978.03
Bot3	Tweet content type B3	620.97
Bot3	Tweet content type B6	620.97
Genuine User	Tweet type B3	6674.83
Genuine User	Tweet content type B3	41.02
Genuine User	Tweet content type B6	41.02

RESULTS

ii) Calculate the average AUC for each specific account groups

iii) Genuine users has least LCS hence shows heterogeneity whereas bots have higher AUC which indicates behavioral similarity



Type of User	Average AUC
Bot1	3222.1866
Bot2	4453.7688
Bot3	43739.9954
Genuine user	2252.2957

BOT DETECTION

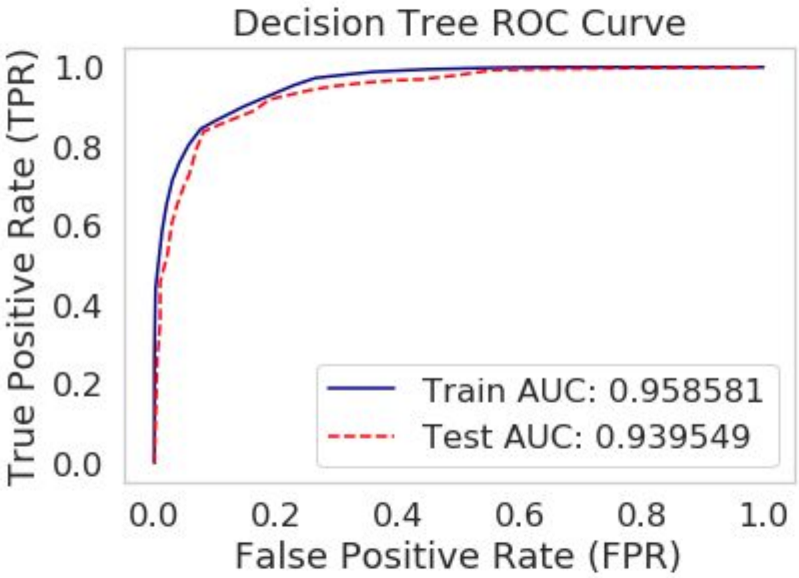
- We used the dataset as referenced in [1].
- The training dataset is chosen as “training_data_2_csv_UTF.csv” .
- Algorithms Used:

1. Decision Tree
2. Multinomial Naive Bayes
3. Gaussian Naive Bayes
4. Bernoulli Naive Bayes
5. Random Forest
6. AdaBoost
7. Logistic Regression
8. K Neighbors Classifier

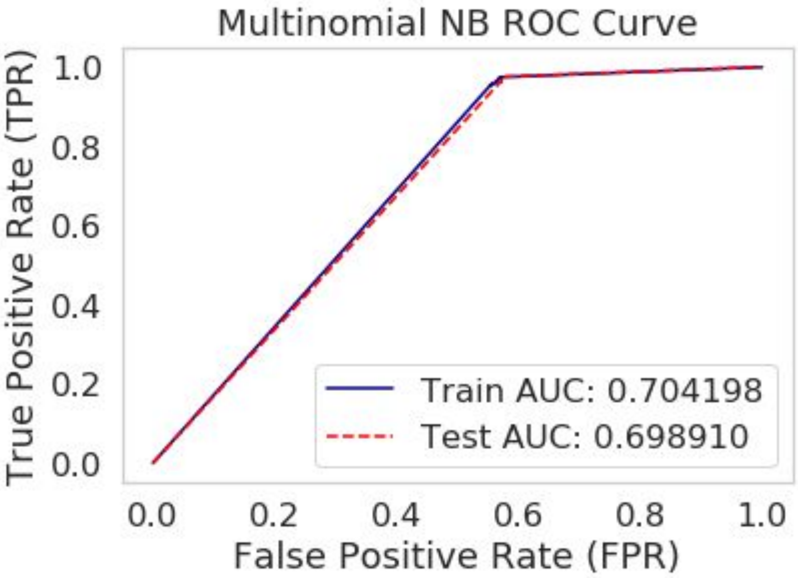
RESULTS

a) Bot Detection

1. Decision Tree

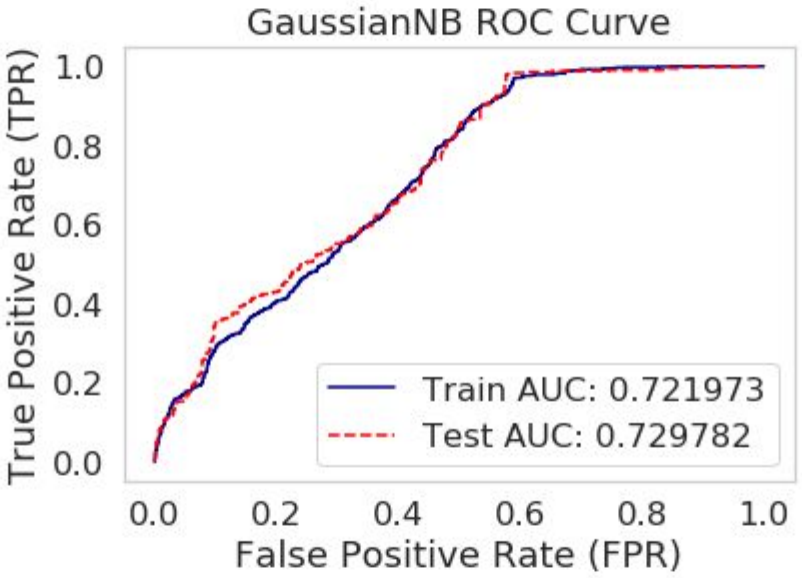


2. Multinomial Naive Bayes

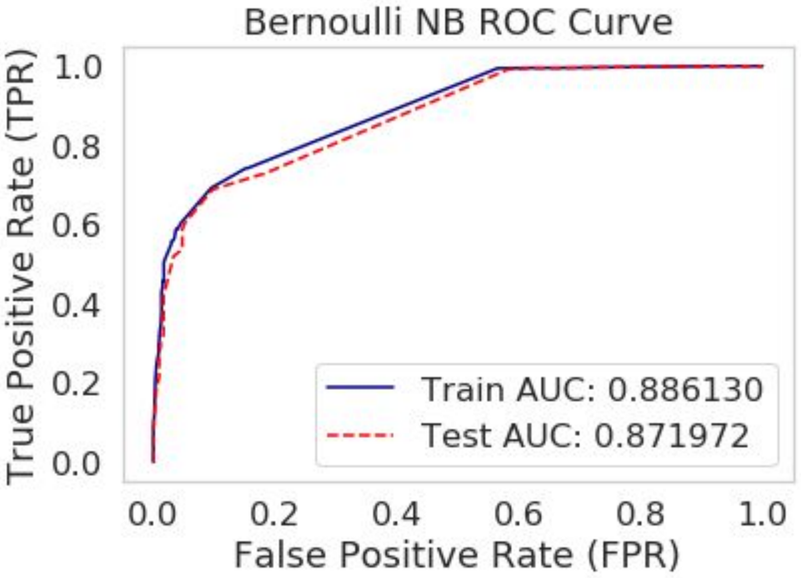


RESULTS

3. Gaussian Naive Bayes

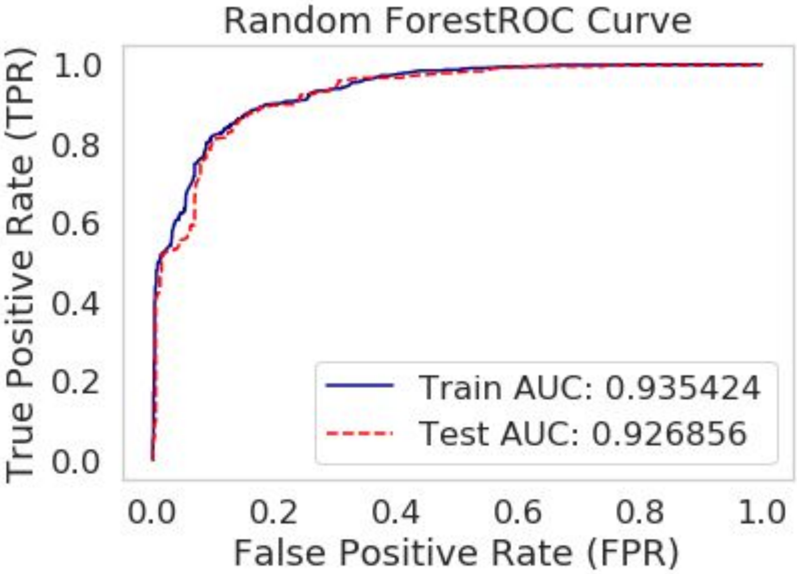


4. Bernoulli Naive Bayes

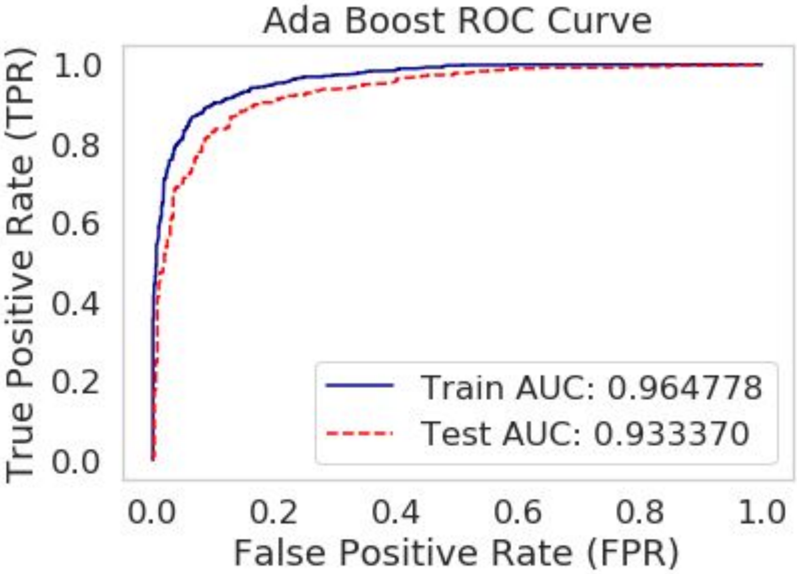


RESULTS

5. Random Forest

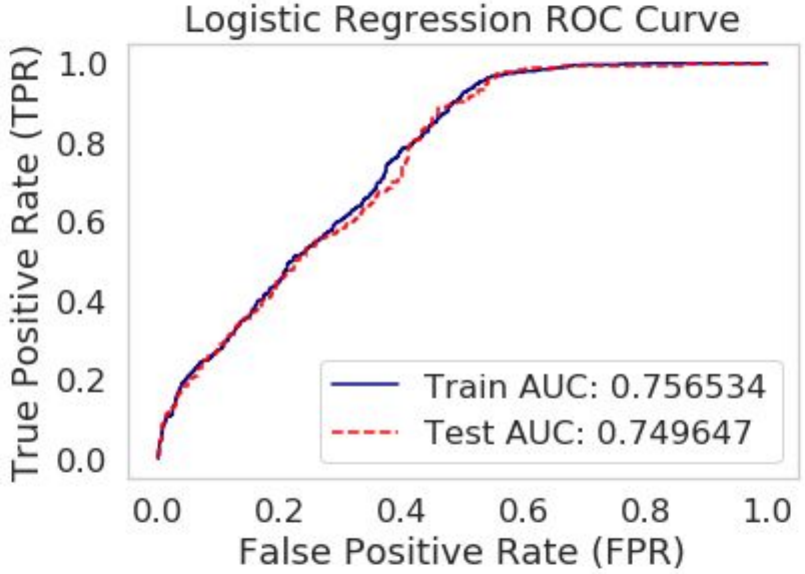


6. AdaBoost

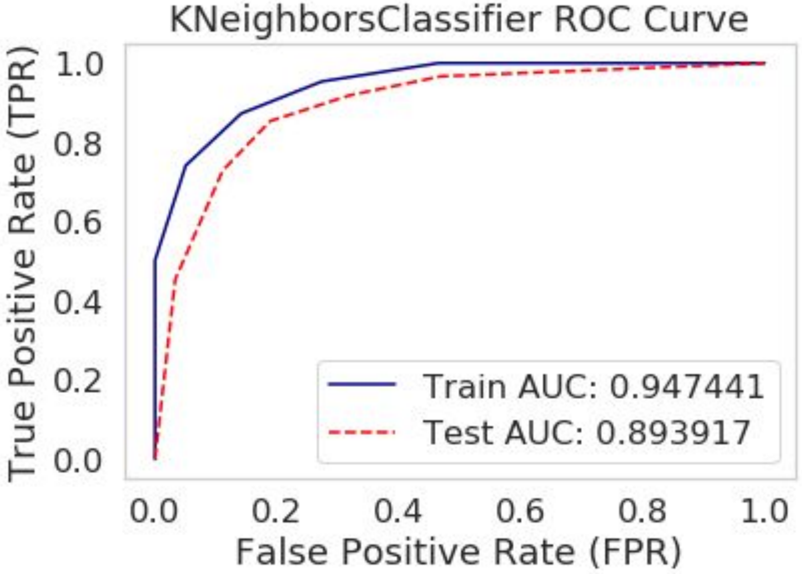


RESULTS

7. Logistic Regression

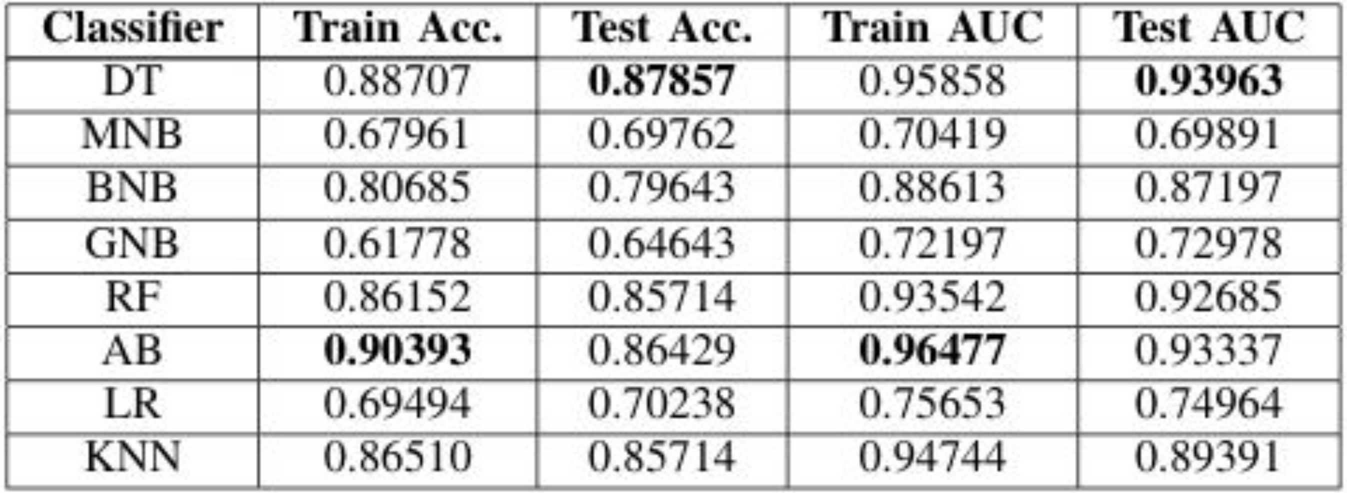


8. K Neighbors Classifier



RESULTS

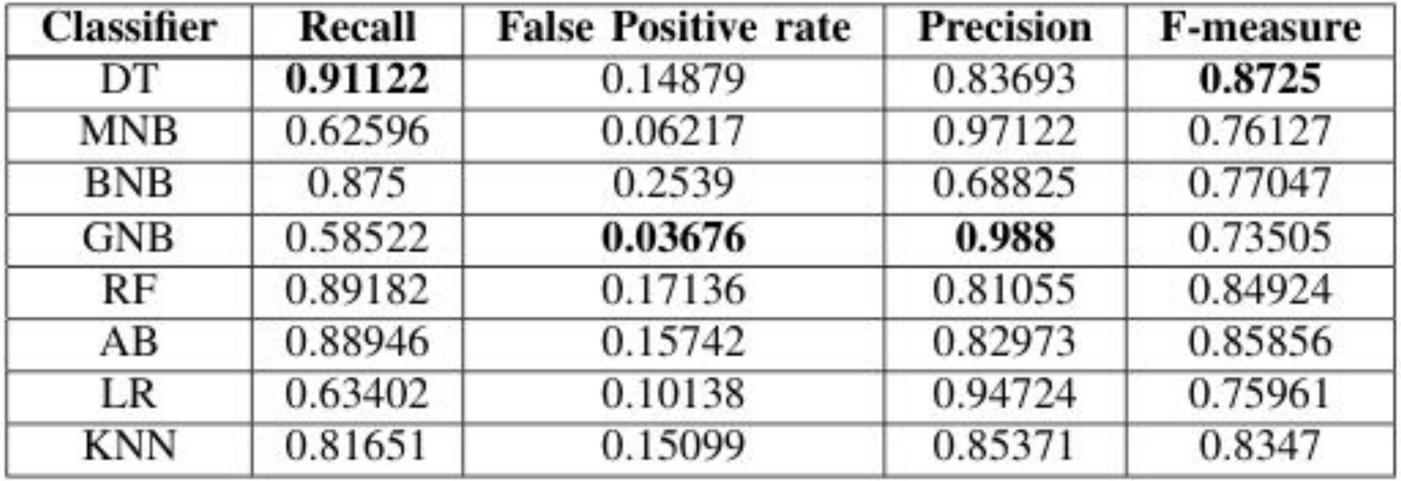
Train and Test AUC (Area Under the Curve)



Classifier	Train Acc.	Test Acc.	Train AUC	Test AUC
DT	0.88707	**0.87857**	0.95858	**0.93963**
MNB	0.67961	0.69762	0.70419	0.69891
BNB	0.80685	0.79643	0.88613	0.87197
GNB	0.61778	0.64643	0.72197	0.72978
RF	0.86152	0.85714	0.93542	0.92685
AB	**0.90393**	0.86429	**0.96477**	0.93337
LR	0.69494	0.70238	0.75653	0.74964
KNN	0.86510	0.85714	0.94744	0.89391

RESULTS

Performance Metrics



Classifier	Recall	False Positive rate	Precision	F-measure
DT	**0.91122**	0.14879	0.83693	**0.8725**
MNB	0.62596	0.06217	0.97122	0.76127
BNB	0.875	0.2539	0.68825	0.77047
GNB	0.58522	**0.03676**	**0.988**	0.73505
RF	0.89182	0.17136	0.81055	0.84924
AB	0.88946	0.15742	0.82973	0.85856
LR	0.63402	0.10138	0.94724	0.75961
KNN	0.81651	0.15099	0.85371	0.8347

CONCLUSION

- In group detection AUC of digital DNA approach on the tweet categories produces significant result
- Human groups shares less similarity among their tweets whereas bots share more similarity
- Some bot groups have lower AUC than other bot groups, these groups are harder to detect
- Decision Tree shows better test accuracy as well as the test AUC is greater than other classifiers.
- Decision Tree also gives better recall and F-measure.
- Gaussian NB shows least False Positive Rate and better precision among all.

REFERENCES

[1] https://www.kaggle.com/charvijain27/detecting-twitter-bot-data/data#.

[2] U. Ramakrishnan, R. Shankar, Ganesha Krishna, ”Sentiment analysis of twitter data: Based on user-behavior” in International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 7, (2015), pp. 16291-16301.

[3] Bild, D. R., Liu, Y., Dick, R. P., Mao, Z. M., and Wallach, ”Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph” in ACM Transactions on Internet Technology, Vol. 15, No. 1, Article 4, Publication date: February 2015.

[4] Hana Anber, Akram Salah, A. A. Abd El-Aziz, ”A Literature Review on Twitter Data Analysis”, International Journal of Computer and Electrical Engineering, Volume 8, Number 3, June 2016.

[5] Chithra R G, Harshitha G M, Anuprakash M P, Rakshitha H B, ”Behavioural Analysis of Twitter data : A Classification Approach” in International Journal of Engineering Research Technology (IJERT), ISSN: 2278-0181, 2019.

Thank You