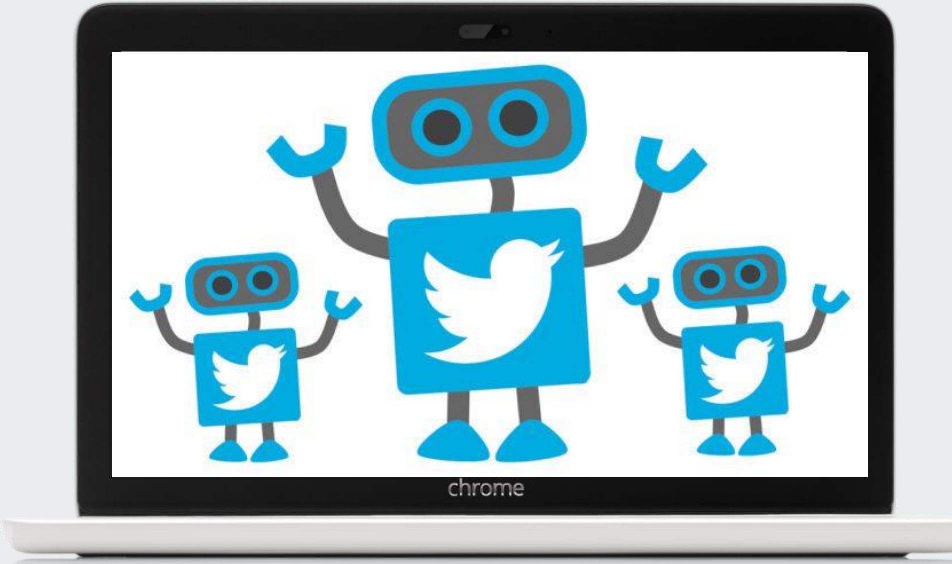
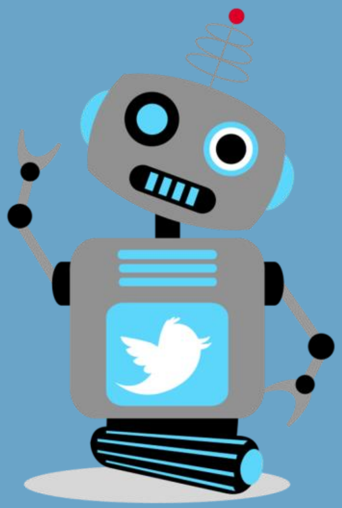
Malicious Twitter Bot Type Detection



Tram Ngoc Le (Jennie)

Outline



Problem Statement

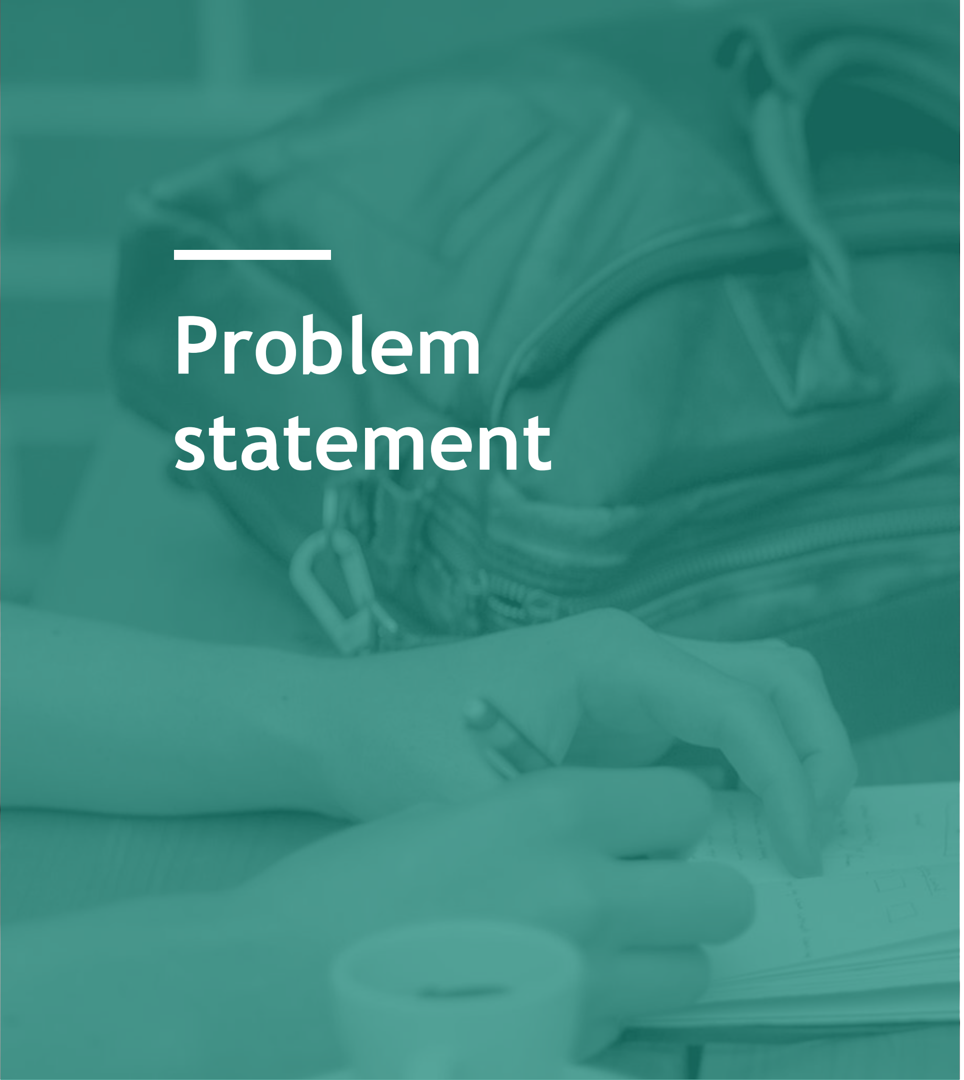
Solution Proposal

Data Processing

Methodology

Result Analysis

Conclusion



Problem statement

Although Twitter is able to identify most of the bot accounts, the company could have done better in Malicious Bots Classification.

How can Twitter classify different types of malicious bots with better accuracy ?

Solution Proposal

Create a Better Method that Helps Twitter Detect and Classify Malicious Bot Types

Our Mindset and Workflow

- Classify major types of bad bots
- Crawl features from bot accounts
- Build models to test bad bots' Behavior and User Profile features
- Leverage Natural Language to find keywords from tweets
- Build keyword dictionaries for each type
- Add a new indicator (TFIDF) and test if the dictionaries work

Data Processing 01

Bot Repository: Get **900** IDs of bots in each group, and **2700** IDs in total.
https://botometer.iuni.iu.edu/bot-repository/datasets.html

1. cresci-2017
Description: A dataset of (i) genuine, (ii) traditional, and (iii) social spambot Twitter accounts, annotated by CrowdFlower contributors. Released in CSV format.

2. pronbots-2019
Description: Pronbots shared by Andy Patel (github.com/r0zetta/pronbot2).

Malicious Bot	Description
Fake Follower	Robot or inactive accounts that inflate number of followers of another account.
Scam Bot	Accounts that advertise scam sites.
Spam Bot	Accounts that spam different kinds of information by sending messages with the same content multiple times.

Data Processing 02

Figure out what features we want to analyze in our model.

Category	Features Will Be Used
Tweet Syntax	The average number of retweets of tweets for each account
	Percentage of tweets containing URL or hyperlink for each account
Tweet Semantics	Keyword TFIDF
Temporal Behavior Features	Average number of tweets per day
User Profile Features	Number of followers of one account
	Number of friends of one account
	Number of tweets that one account has
	Using default profile
	Using default profile image
	Using geography or location enabled

Data Processing 03

Use Python (tweepy, nltk, pandas, numpy, csv, etc) to crawl all the information we want and create three dictionaries of keywords with more than 0.05% term frequency rate for each group of malicious bots.

Sample Code:

```python
for i in range(0, len(Tweet)): # read each tweet
    tweet = nltk.FreqDist(Tweet.iloc[i, Tweet.columns.get_loc('text')].replace('b\'RT', '').replace('b"RT', '').rep

    for term in tweet.keys():
        if term in np.array(dict_fake['keyword']).tolist() or term.lower().startswith('http'):
            count1 += tweet[term]

        if term in np.array(dict_scam['keyword']).tolist() or term.lower().startswith('http'):
            count2 += tweet[term]

        if term in np.array(dict_spam['keyword']).tolist() or term.lower().startswith('http'):
            count3 += tweet[term]

    if count1 > 0:
        df_fake.append(1)
    else:
        df_fake.append(0)
    TF_fake.append(count1/len(tweet))

    if count2 > 0:
        df_scam.append(1)
    else:
        df_scam.append(0)
    TF_scam.append(count2/len(tweet))

    if count3 > 0:
        df_spam.append(1)
    else:
        df_spam.append(0)
```

Data Size:

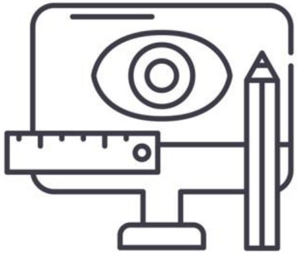
Data Description	Data Size
Number of valid ID	**2042**(some have been suspended, and some don't have tweets)
Number of tweets	**216173**
Total number of words	**138042** for Fake Follower, **160245** for Scam Bot, **161956** for Spam Bot
Number of keywords in a dictionary	**120** for Fake Follower, **126** for Scam Bot, **170** for Spam Bot

Data Processing 04

Use Python and Excel to calculate the normalized TFIDF of each tweet and the average normalized TFIDF of each ID.

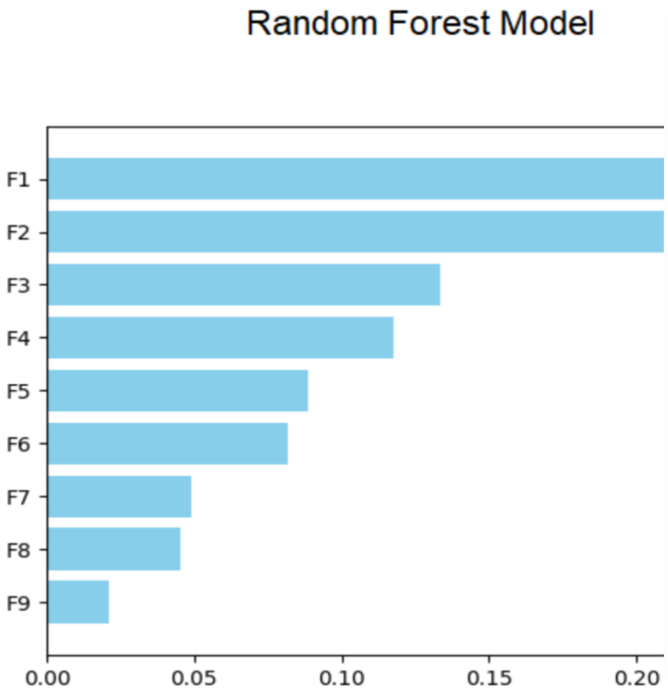
$$\textbf{Normalized TFIDF}=\frac{Keyword\ frequency\ in\ a\ tweet}{Number\ of\ terms\ in\ a\ tweet}\times\left(\log\frac{Number\ of\ tweets\ for\ each\ account+1}{Number\ of\ tweets\ include\ keyword+1}+1\right)$$

Methodology
+
Result Analysis



Phase One: Detect Bad Bot-Like Behaviors

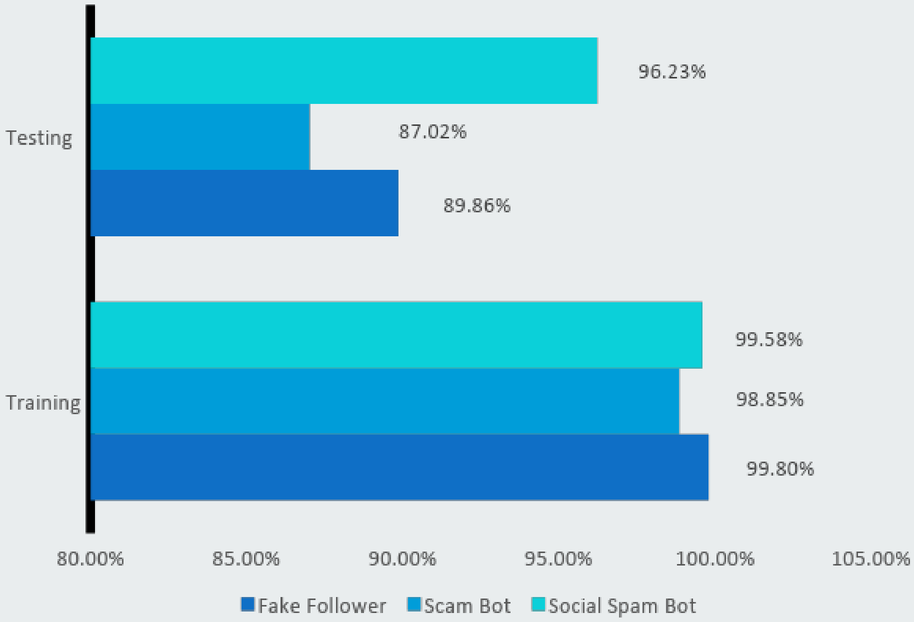
RANDOM FORESTS METHODOLOGY



Features names for short

Original field name	Field name on gra
Average of retweet_count	F1
num_of_followers	F2
average_tweet_per_day	F3
status_num	F4
num_of_friends	F5
percentage_of_url_tweet	F6
default_profile	F7
Average of favorite_count	F8
geo_enabled	F9

How Accurate is Random Forests ?



Results for output field bot_group

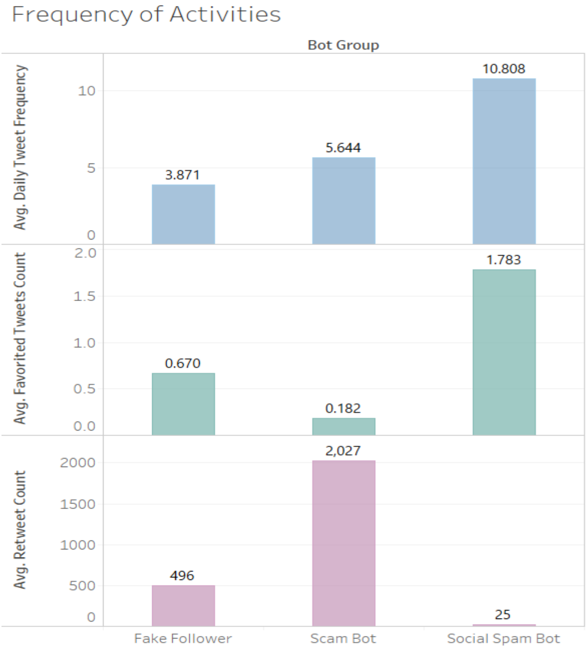
Comparing $R-bot_group with bot_group

'Partition'	1_Training		2_Testing	
Correct	1,397	99.43%	580	91.05%
Wrong	8	0.57%	57	8.95%
Total	1,405		637	

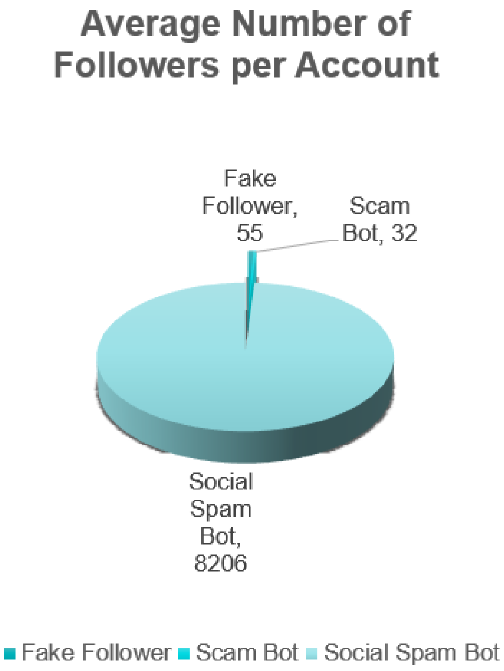
Coincidence Matrix for $R-bot_group (rows show actuals)

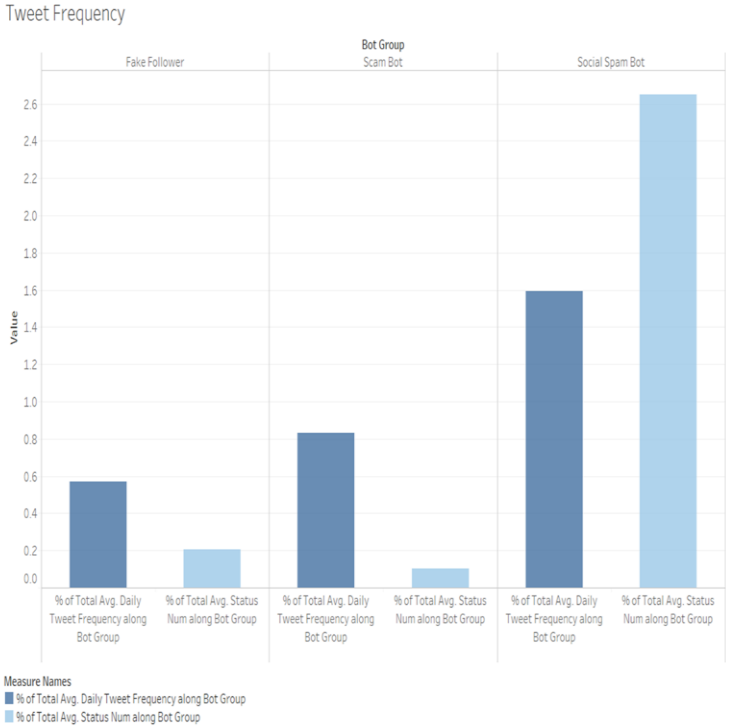
'Partition' = 1_Training	Fake Follower	Scam Bot	Social Spam Bot
Fake Follower	491	1	0
Scam Bot	5	428	0
Social Spam Bot	2	0	478
'Partition' = 2_Testing	Fake Follower	Scam Bot	Social Spam Bot
Fake Follower	195	14	8
Scam Bot	24	181	3
Social Spam Bot	6	2	204

Phase One: How Do You Behave, Bad Bots?



Average of Daily Tweet Frequency, average of Favorited Tweets Count and average of Retweet Count for each Bot Group. For pane Average of Daily Tweet Frequency: The marks are labeled by average of Daily Tweet Frequency. For pane Average of Retweet Count: The marks are labeled by average of Retweet Count. The view is filtered on average of Daily Tweet Frequency, which keeps all values.

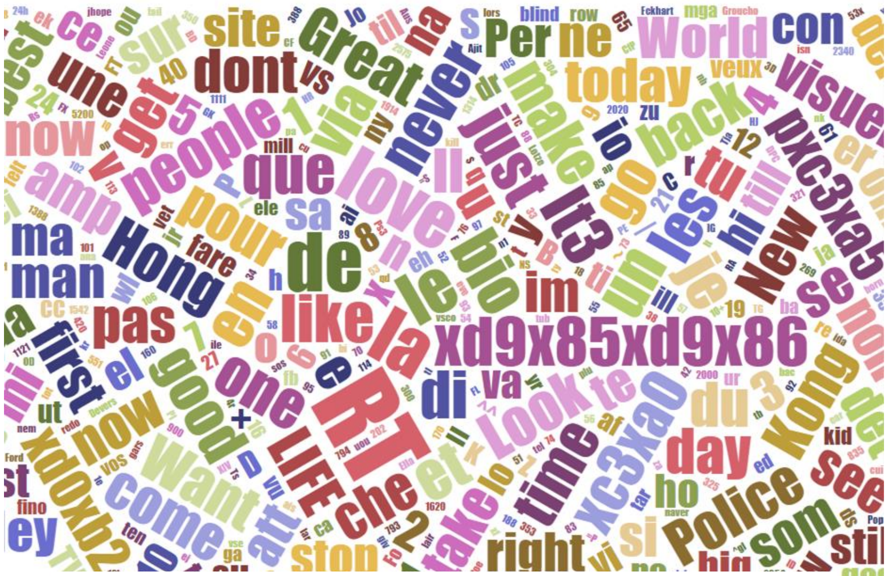




PHASE TWO: DETECT BAD BOTS WITH BEHAVIORS AND TWEET SEMANTICS

WORDCLOUD

Type 3



Type 2

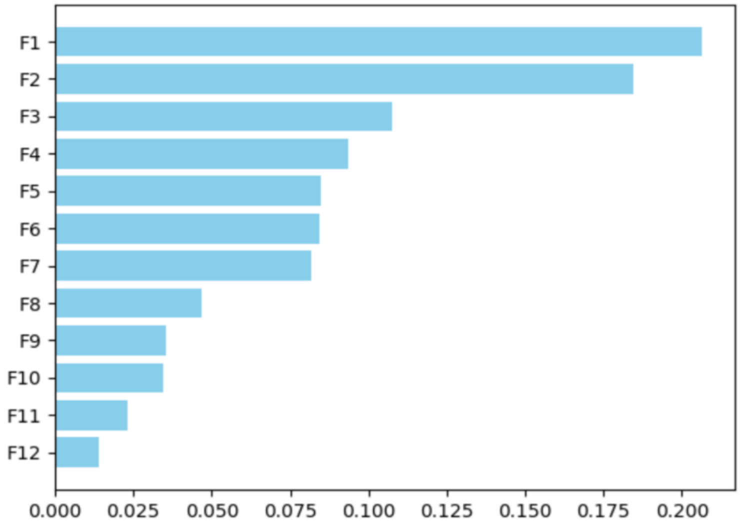
PHASE TWO: DETECT BAD BOTS WITH BEHAVIORS AND TWEET SEMANTICS

SEMANTIC ANALYSIS & RANDOM FORESTS METHODOLOGY

Features names for short

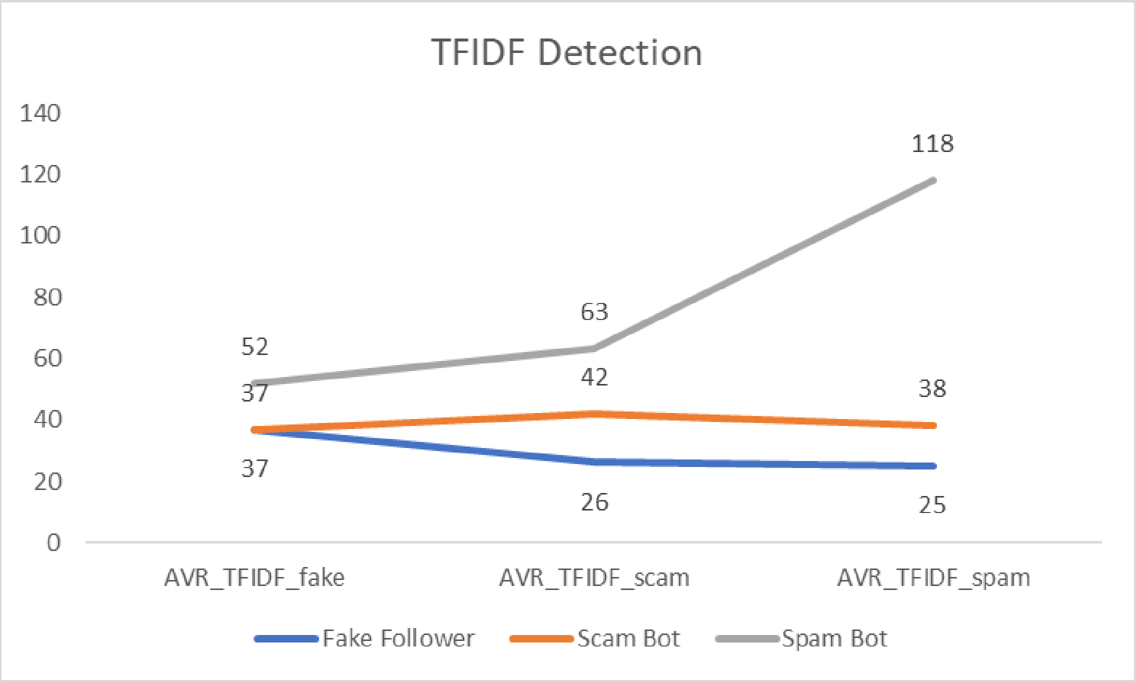
Original field name	Field name on graphic
Average of retweet_count	F1
num_of_followers	F2
AVR_TFIDF_spam	F3
percentage_of_url_tweet	F4
num_of_friends	F5
average_tweet_per_day	F6
status_num	F7
AVR_TFIDF_scam	F8
AVR_TFIDF_fake	F9
Average of favorite_count	F10
default_profile	F11
geo_enabled	F12

Random Forest Model

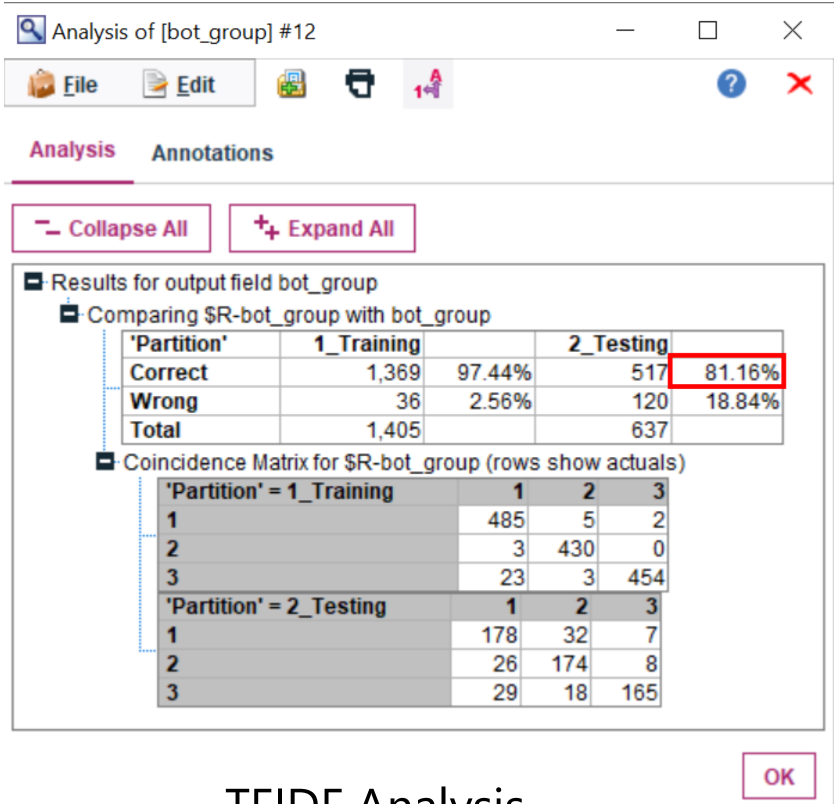


PHASE TWO: DETECT BAD BOTS WITH BEHAVIORS AND TWEET SEMANTICS

TFIDF SCORE VISUALIZATION



Phase Two: Detect Bad Bots' Tweet Keywords


Analysis of [bot_group] #12

File Edit

Analysis Annotations

Collapse All Expand All

Results for output field bot_group

Comparing $R-bot_group with bot_group

'Partition'	1_Training		2_Testing	
Correct	1,369	97.44%	517	81.16%
Wrong	36	2.56%	120	18.84%
Total	1,405		637	

Coincidence Matrix for $R-bot_group (rows show actuals)

'Partition' = 1_Training	1	2	3
1	485	5	2
2	3	430	0
3	23	3	454

'Partition' = 2_Testing	1	2	3
1	178	32	7
2	26	174	8
3	29	18	165

OK

TFIDF Analysis

Accuracy Comparison Between Phase One and Phase Two

PHASE ONE: BOT-LIKE BEHAVIOR

Results for output field bot_group

Comparing $R-bot_group with bot_group

'Partition'	1_Training		2_Testing	
Correct	1,397	99.43%	580	91.05%
Wrong	8	0.57%	57	8.95%
Total	1,405		637	

Coincidence Matrix for $R-bot_group (rows show actuals)

'Partition' = 1_Training	Fake Follower	Scam Bot	Social Spam Bot
Fake Follower	491	1	0
Scam Bot	5	428	0
Social Spam Bot	2	0	478
'Partition' = 2_Testing	Fake Follower	Scam Bot	Social Spam Bot
Fake Follower	195	14	8
Scam Bot	24	181	3
Social Spam Bot	6	2	204

PHASE TWO: BOT-LIKE BEHAVIOR AND SEMANTIC ANALYSIS

Results for output field bot_group

Comparing $R-bot_group with bot_group

'Partition'	1_Training		2_Testing	
Correct	1,403	99.86%	584	91.68%
Wrong	2	0.14%	53	8.32%
Total	1,405		637	

Coincidence Matrix for $R-bot_group (rows show actuals)

'Partition' = 1_Training	Fake Follower	Scam Bot	Social Spam Spot
Fake Follower	492	0	0
Scam Bot	2	431	0
Social Spam Spot	0	0	480
'Partition' = 2_Testing	Fake Follower	Scam Bot	Social Spam Spot
Fake Follower	195	14	8
Scam Bot	20	185	3
Social Spam Spot	7	1	204

CONCLUSION
YOU'RE UNDER ARREST BAD BOTS !!!

Fake Follower

Low Activities Frequency

Rarely Retweet

Low Engagement in Connecting with Others

Average TFIDF_fake ranks highest among keywords of other type

Scam Bot

Low Engagement in Connecting with Others

High Engagement in Retweet

Average TFIDF_scam ranks highest among keywords of other type

Spam Bot

High Activities Frequency

High Number of Followers and Friends

High Daily Tweet Frequency

Average TFIDF_spam ranks highest among keywords of other type