Social Bots detection

Praveen Malhan
Prerit Chandok
Vishnu Nair
Venkatesh Gopal

Introduction:

We make use of the below API's for the detection of bots on Social Media Platforms. As part of this project, we focus on detecting bots on Twitter.

1. Tweepy - A library provided in Python to access the twitter API. This library helps us to avail the data pertaining to accounts, users, tweets and other user related parameters.
2. Google's Safe Browsing API - An API provided by Google which maintains a list of URL's which are classified as containing Malware, Phishing, Spam, Social Engineering links.
3. Open Source Truthy Project - An API provided by the Indiana University, Bloomington for classifying tweets on the basis of emotional content, user classification, content classification, other temporal and network factors.

We are supplementing the above methods with the prescribed Zerofox functionalities because of the lack of such features for the Twitter platform on Zerofox.

Methodology :

We use several factors for deciphering whether a user is a human or a bot. Once these factors are availed, we make decisions by setting thresholds. The thresholds have been set by manually checking the characteristics of bots on Social Media. In the future, we plan to feed these data sets to machine learning algorithms using Waikato Environment for Knowledge Analysis (WEKA), a Java ML library, to train our tool to set the mark of differentiation between bots and humans. The 6 factors which we define are :

- **Number of Tweets made by a User account with time difference less than 2 seconds**

  This factor defines the total number of tweets made by a user where the time difference between each tweet is less than 2 seconds. A human user cannot practically make multiple tweets within 2 seconds. Therefore, we calculate the time difference between successive tweets and have a counter to indicate the total number of tweets satisfying the aforementioned criteria.

- **Skew Rate**

Skew rate is defined as the ratio of the number of people following a user and the number of people followed by the user.

Skew Rate = Number of people following a user / Number of people followed by the user

For accounts like Donald Trump, Obama , the skew rate would be very high as the number of people following them is large. For normal user accounts this would be in the range 0.1 - 0.9. However, for social bots this would be lesser than 0.00x where x could range from 1 to 9. The reason behind this is that social bots would be following many users, but the number of users following them would be comparatively very small.

- **Classification Scores**

  We make use to the API provided from the Truth project to categorize content based on the below factors. Based on these factors, we defined an aggregate score which would help us to ascertain the final decision.

  1) content_classification
  2) temporal_classification
  3) network_classification
  4) friend_classification
  5) sentiment_classification
  6) user_classification

- **Content in the Tweet**

  The content in the tweet would be analysed for malware, phishing links, Spam links, Social Engineering links. We extract the urls contained in the tweets and provide the same to the Safe Browsing API from Google and classify the content into one of the categories mentioned above.

- **Retweet Count**

  The number of retweets of a particular tweet is taken into consideration for bot detection. A bot would typically retweet multiple tweets it finds on its timeline (or) search in particular for specific accounts and retweet them. We analyse a given tweet and extract the number of times it has been retweeted. Consequently, among the users who have retweeted a particular tweet, we check the total retweet count of the user and consider it as a contributing factor towards bot detection.

- **GeoLocation**

  Tweepy does not provided us with the IP addresses a tweet would originate from. However, it does provide us with the capability to extract the Geo Coordinates from which the tweet originated. Using these values, we zero into the exact location of the

tweet and the corresponding user. This factor is better compared to IP addresses as values of Geo Location cannot be spoofed unlike IP addresses.

- **Verified Account**
  We include this factor to check if this account is verified by Twitter. Usually a blue tick is seen on Twitter if Twitter considers a user's account to be verified.

BotFinder Execution:

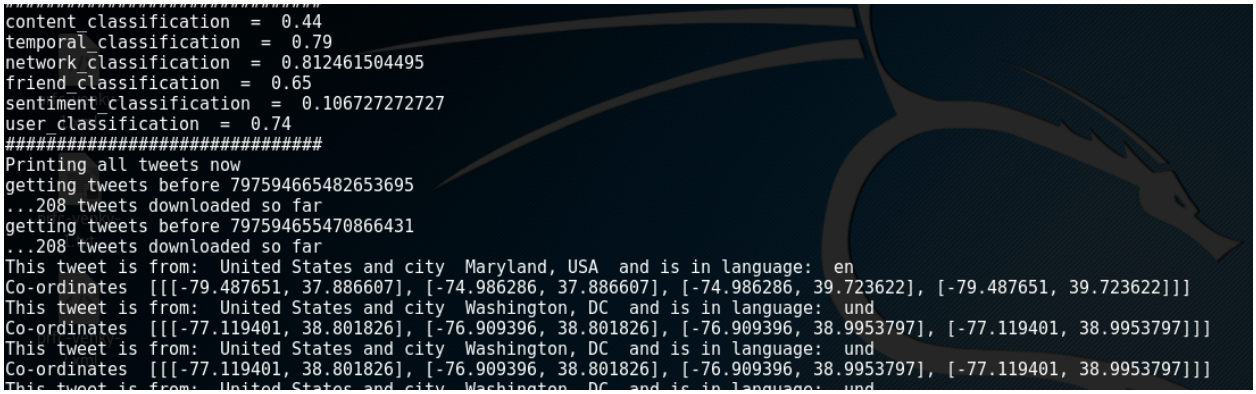
The tool we have functions on the methodology defined above. Since we use the Tweepy API, the person making use of our tool should avail the below information from Twitter. These values are alphanumeric values specific to users.

- Consumer_key
- Consumer_secret
- Access_token
- Access_token_secret

Once these values are availed, we create the object to access Twitter's API.

Example of Monitoring a Bot "Troll Belt" which post lot of Spam Links to twitter :

1.Tool performing classification of user and availing all the tweets.



2. Tool returning the GeoLocation of the source of the tweet.

```
This tweet is from:  United States and city  Maryland, USA  and is in language:  en
Co-ordinates  [[[-79.487651, 37.886607], [-74.986286, 37.886607], [-74.986286, 39.723622], [-79.487651, 39.723622]]]
This tweet is from:  United States and city  Washington, DC  and is in language:  und
Co-ordinates  [[[-77.119401, 38.801826], [-76.909396, 38.801826], [-76.909396, 38.9953797], [-77.119401, 38.9953797]]]
This tweet is from:  United States and city  Washington, DC  and is in language:  und
Co-ordinates  [[[-77.119401, 38.801826], [-76.909396, 38.801826], [-76.909396, 38.9953797], [-77.119401, 38.9953797]]]
This tweet is from:  United States and city  Washington, DC  and is in language:  und
Co-ordinates  [[[-77.119401, 38.801826], [-76.909396, 38.801826], [-76.909396, 38.9953797], [-77.119401, 38.9953797]]]
This tweet is from:  United States and city  Washington, DC  and is in language:  und
Co-ordinates  [[[-77.119401, 38.801826], [-76.909396, 38.801826], [-76.909396, 38.9953797], [-77.119401, 38.9953797]]]
This tweet is from:  United States and city  Washington, DC  and is in language:  und
Co-ordinates  [[[-77.119401, 38.801826], [-76.909396, 38.801826], [-76.909396, 38.9953797], [-77.119401, 38.9953797]]]
This tweet is from:  United States and city  Washington, DC  and is in language:  und
Co-ordinates  [[[-77.119401, 38.801826], [-76.909396, 38.801826], [-76.909396, 38.9953797], [-77.119401, 38.9953797]]]
This tweet is from:  United States and city  Washington, DC  and is in language:  und
Co-ordinates  [[[-77.119401, 38.801826], [-76.909396, 38.801826], [-76.909396, 38.9953797], [-77.119401, 38.9953797]]]
This tweet is from:  United States and city  Washington, DC  and is in language:  und
Co-ordinates  [[[-77.119401, 38.801826], [-76.909396, 38.801826], [-76.909396, 38.9953797], [-77.119401, 38.9953797]]]
This tweet is from:  United States and city  Washington, DC  and is in language:  und
Co-ordinates  [[[-77.119401, 38.801826], [-76.909396, 38.801826], [-76.909396, 38.9953797], [-77.119401, 38.9953797]]]
This tweet is from:  United States and city  Washington, DC  and is in language:  und
Co-ordinates  [[[-77.119401, 38.801826], [-76.909396, 38.801826], [-76.909396, 38.9953797], [-77.119401, 38.9953797]]]
This tweet is from:  United States and city  Maryland, USA  and is in language:  en
```

3. Tool returning the classification of the URL type. Below we see that the URL is susceptible to Social Engineering attacks

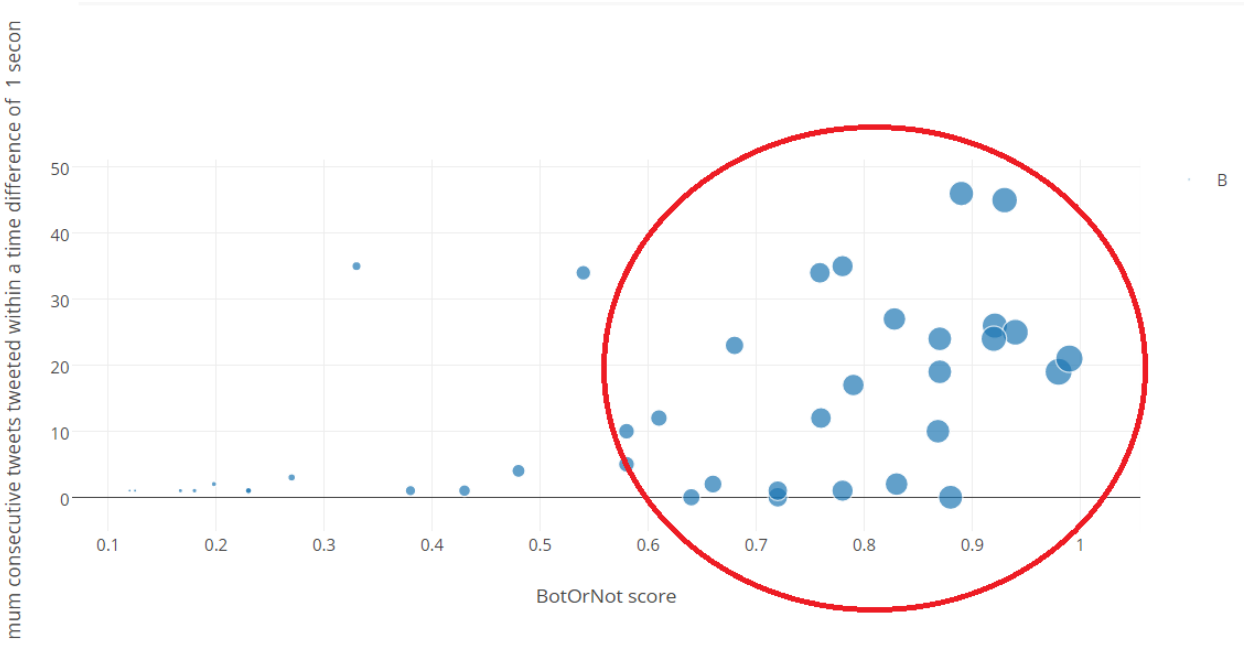

```
Response:{
  "matches": [
    {
      "threatType": "SOCIAL_ENGINEERING",
      "platformType": "WINDOWS",
      "threat": {
        "url": "websitebuildersinfo.in"
      },
      "cacheDuration": "300s",
      "threatEntryType": "URL"
    },
    {
      "threatType": "SOCIAL_ENGINEERING",
      "platformType": "LINUX",
      "threat": {
        "url": "websitebuildersinfo.in"
      },
      "cacheDuration": "300s",
      "threatEntryType": "URL"
    },
    {
      "threatType": "SOCIAL_ENGINEERING",
      "platformType": "OSX",
      "threat": {
        "url": "websitebuildersinfo.in"
      },
      "cacheDuration": "300s",
      "threatEntryType": "URL"
    },
```

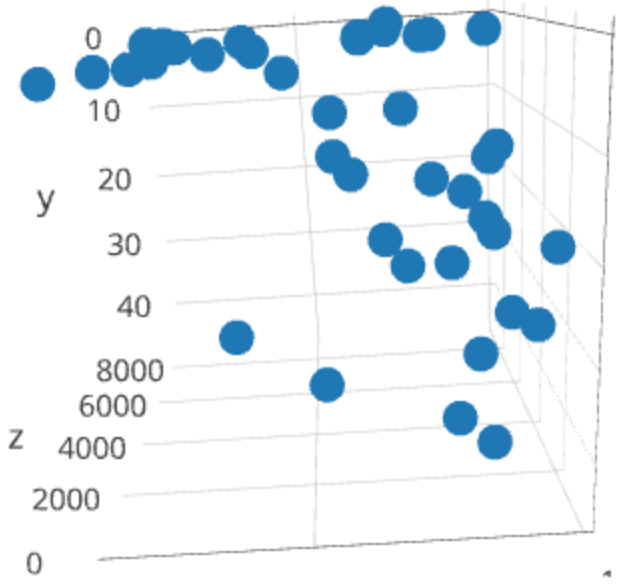

4. Tool making the final decision based on the factors defined under Methodology.

```
Number of successive tweets with a time difference less than 1 second: 177
The number of followers for TrollBelt  are  26  and he is following  250
The skew rate is:  False
TrollBelt  has  0  retweets
TrollBelt  is a bot
```

5. Graphical Detection & Statistics : All Dots in the highlighted region below correspond to Twitter Accounts being detected by our tool as Bots. The X axis corresponds to the BotOrNot

score and the Y axis corresponds. As the graph clearly shows there is a clear correlation between high scores (0.00 - 1.00) and high volume of messages tweeted in short spans of time.

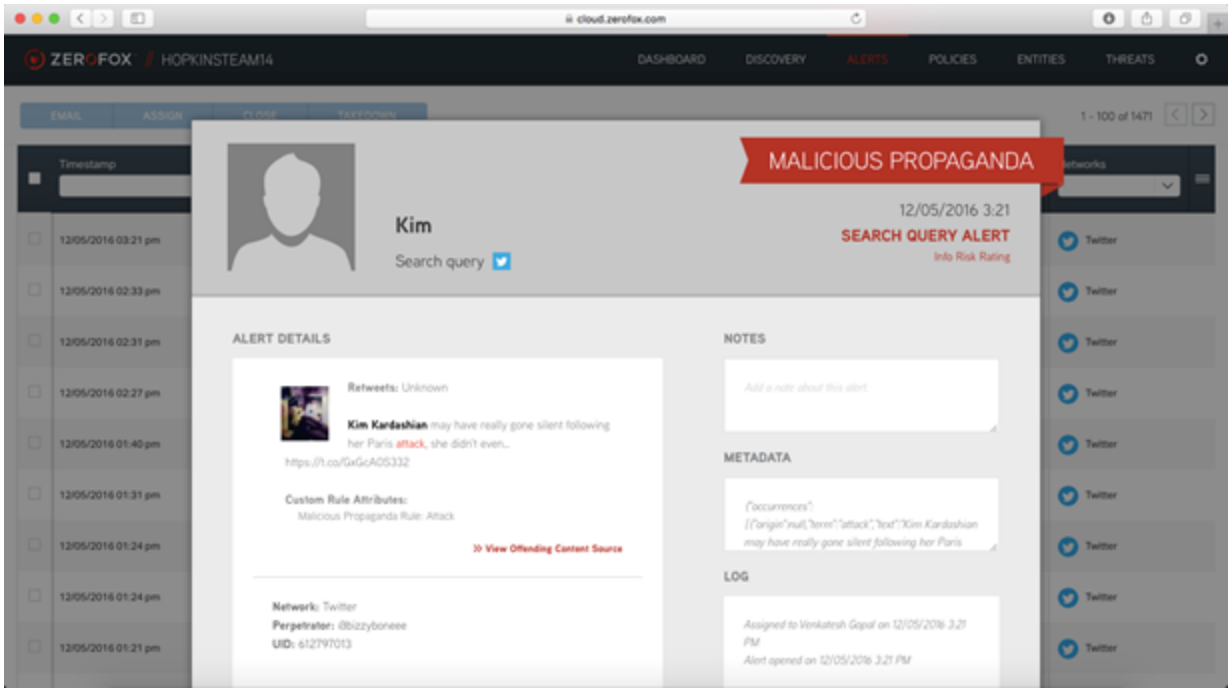




3D view (X -botornot score, Y- Maximum consecutive messages, Z - skew rate)

Zerofox :

We only use the Zero Fox platform to identify malicious links/posts tweeted or retweeted by bots. The above is a screenshot of an alert generated by the rule identifying an account posting about the Protected Entity (Kim Kardashian). If there are any words used in the post, which seem to be malicious, an alert is generated.



Challenge & Future Work with our Tool:

1. The GeoLocation of the user is available only when the user tweets with the location service turned on. User would seldom do this. Availability of IP address of the source would be useful. Though IP packets are susceptible to spoofing attacks, this evaluation is usually handed by the underlying routers.
2. The Heuristics we used could also capture lot of false positives if an normal user behaviour resembles a bot.
    a. A user who has just created his account but never used it.
    b. If the user simultaneously tweets from more than 4 to 5 machines. Though, we have put a hard limit of 2 seconds and checking for consecutive tweets, a user could potentially perform an operation of posting more than 6-7 tweets within 2 seconds from multiple machines.