

SPAM DETECTION ON TWITTER

Project as part of CS 63001.004: Analysing & Securing Social Media

Aviral Verma (axv190007)
Jitendra Justin (jxj172430)
Kaustubh Deshpande (kxd180005)

1. INTRODUCTION

Twitter gives users an opportunity to share their messages concerning everything including news, occasions, celebrities, political issues, and so on. Compared to various other social media platforms, the connection between users is bi-directional rather than unidirectional connection. This means that a particular user may not be following one of his followers. Twitter was overwhelmed by a lot of malignant tweets that were sent by a huge number of spammed users account; this occurred around April of 2014. The authentic users mistake the spam information for as important one. Spam messages are difficult to regulate. Email administrations such as Gmail, Microsoft etc. are still finding better ways to prevent spam.

Twitter utilises both manual and mechanised administrations to fight spammers. The manual method involves Twitter giving users a chance to report spammers via the spammers' profile pages. These manual methodologies are stressful and may not be sufficient to distinguish between all spammers because of billions of users.

The possession of a huge number of spam protests documented against the account, the following/unfollowing of a huge number of accounts within a brief span, the publishing of a copy of the messages in a single account, the publishing of malignant connections are the ways in which Twitter aims to handle the spam accounts and tweets.

The users are more concerned with spam accounts and tweets on Twitter because there is a high increase in spam account in the Twitter platform since its inception. Average number of spam reports has been around 25,000 a day in 2017 to 17,000 a day in 2018.

2. SUGGESTED MODEL

Twitter API tweepy is used to get tweets from the website. In spam accounts, features like followers_count, favourites_count, profile_image, friends_count, geo_enabled, time_zone, location, will be very low. At the same time, features like total_hashtags and total_links will be very high for spam accounts and spam tweets. `timeline = api.user_timeline(screen_name`

= user.screen_name, count = 100, include_rts = True) is the code to get information from Twitter. Here, API used is the tweepy API.

Twitter considers an account as spam “if a user posts multiple unrelated updates to a topic using the #symbol.” We count the number of hashtags in the 100 most recent tweets of the user, and this feature is used to classify as spam or real.

Data pre-processing in our model is achieved by the processes of Snowballing and TF-IDF Vectorization.

MULTINOMIAL NAÏVE BAYES CLASSIFIER

For the purpose of classification, we use the Multinomial Naïve Bayes classifier. Multinomial Naive Bayes classifier is a specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features.

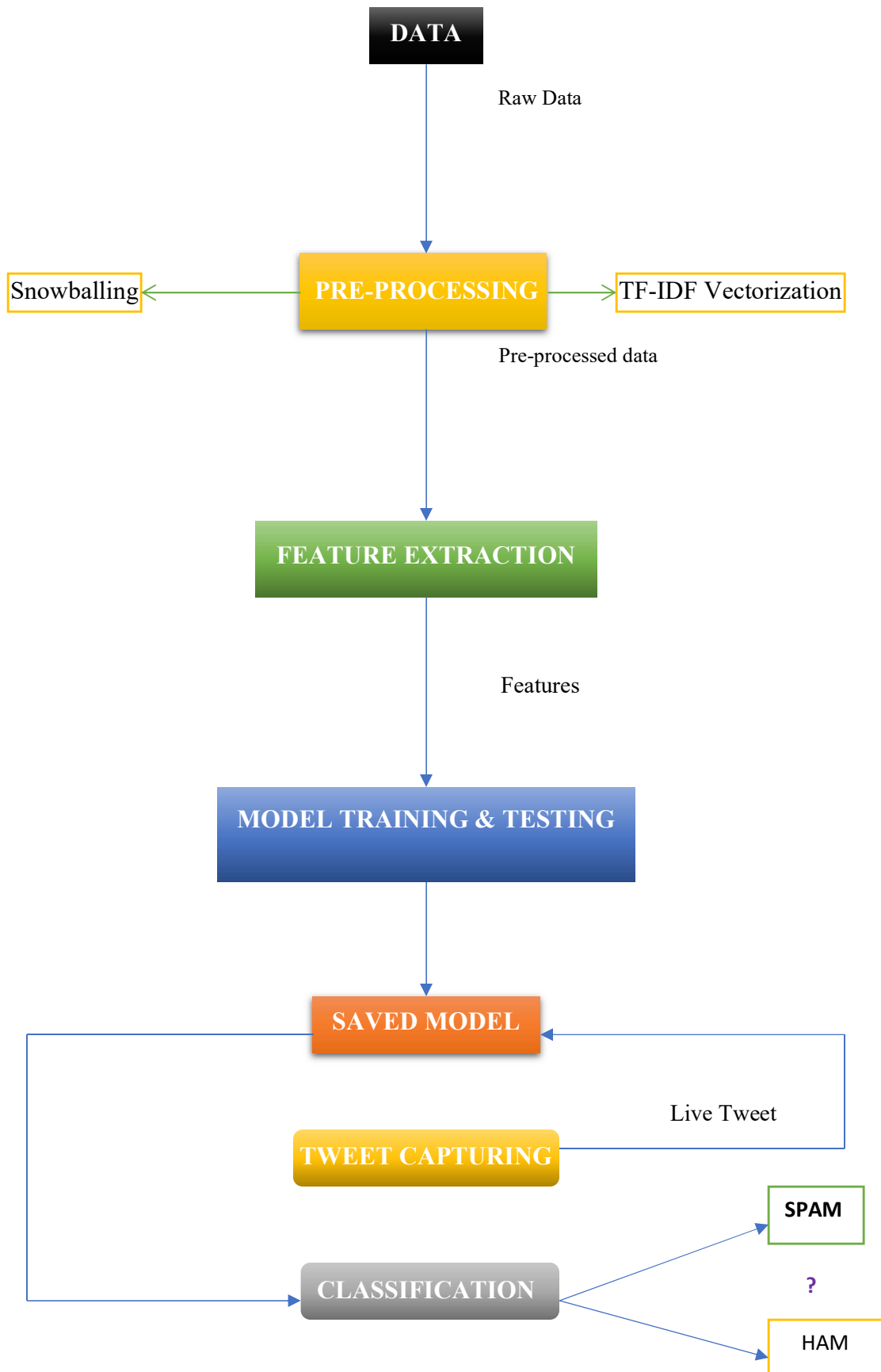
This works well for data which can easily be turned into counts, such as word counts in text. This is especially suitable for our data. This is a type of classifier which works on tokens, with spam or ham, in the tweets. It then uses Bayes theorem to calculate the probability that a particular tweet is spam or not. The technique can classify almost any sort of data.

The Dataset used in our model has been created through the Twitter API. The API was used to retrieve information regarding the Tweets and their posting accounts. Since images cannot be processed, we instead use the links these images direct to, contained in the HTTP header of the image. The Tweet texts were concatenated with these links for more efficient spam detection mechanism.

CHALLENGES:

- Difficulty in gathering labelled dataset in a limited time.
- Learning the required NLP (Natural Language Processing) concepts used in the project.
- Understanding Machine Learning concepts and Naive Bayes Algorithm from scratch, since most of us had not taken any related courses till now.

ARCHITECTURE DIAGRAM



3. WORKING

Step 1: Data Collection and Categorization

Accounts that have verified attribute as TRUE are real accounts. Tweets which do not have profile picture, background image, followers but use many links in the tweets and follow many accounts are considered spam accounts. Tweets which use many links, external links, shortened URLs, defamatory abusive language can be classified as spam. We also performed manual checking of the rest of the tweets and accounts in order to classify as spam or real from the Twitter data.

Step 2: Data Pre-processing

In order to increase the efficiency of our mechanism, data pre-processing is a necessary function. Data pre-processing involves the processes of Snowballing and TF-IDF Vectorization.

Snowballing consists of the following actions,

- Separate the sentence into individual words
- Convert all letters to lowercase
- Remove stop words
- Don't care for non-English words

TF-IDF stands for term frequency-inverse document frequency.

The TF-IDF weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus.

The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

TF: Term Frequency, which measures how frequently a term occurs in a document.

$TF(t) = (\text{Number of times term } t \text{ appears in a document}) / (\text{Total number of terms in the document})$

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as

"is", "of", and "that", may appear a lot of times but have little importance.

$IDF(t) = \log(\text{Total number of documents} / \text{Number of documents with term } t \text{ in it})$

Visualization: Using the occurrence weights obtained by TF_IDF Vectorization, we present a Word Cloud Visualization for spam and ham data. A word cloud is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or colour.

This format is useful for quickly perceiving the most prominent terms to determine its relative prominence. The word clouds generated for spam and ham data are provided in the Appendix A, Fig A.1 and A.2.

Step 3: Feature Selection

Favorite_Count, retweet_count, friends_count, total_mentions, total_hashtags, total_links and sample_tweet. The features that can individually categorise a tweet spam or ham are retweet_count, total_mentions, total_links and total_hashtags.

Step 4: Model Training and Testing

We scramble dataset and separate data for training and testing purposes. We use 80% of the data as Training data and remaining 20% as Testing data. Scrambling is random. We perform initial fitting of parameters on the training set and transform to a particular set of examples.

We extract the top features under the categories features_train, features_test, labels_train, labels_test. The top features extracted are used for improving the performance of the model.

We tested the trained model with the test data to obtain accuracy and F1 scores using Multinomial Naive Bayes. The trained model is saved to be further used in real time spam detection in a tweet.

Step 5: Real Time Tweet Catching and Classification

In order to catch a tweet in real time, the Tweepy API is used. A user must provide their Twitter Developer account credentials in the *credentials.py* file to capture a live tweet.

Using the model we generated by our MNB classifier, we classify the live tweet as spam or ham. The tweet captured is fed to the model and based on its features, it is classified by the model as Spam or Not Spam.

4. RESULTS

Running the *spam_classifier.py* file with valid credentials in *credentials.py* provides us with results on a live tweet captured by *tweet_catch.py*.

We were able to successfully classify a number of tweets as spam or ham.

After numerous iterations we found our model to display the following statistics,

| CLASSIFIER | ACCURACY | PRECISION | F1 SCORE |
|-------------------------|----------|-----------|----------|
| Bernoulli Naïve Bayes | 88.07 | 71.22 | 80.37 |
| Multinomial Naïve Bayes | 94.51 | 85.78 | 89.11 |

APPENDIX A

| | spam | ham |
|------|------|---|
| 6980 | ham | Just witnessed the 2nd most intense lightning of my life. 2nd only to the plasma ball i saw when i was 13. I love florida |
| 6981 | ham | For now, hookah break! |
| 6982 | ham | The fire wings are coming along so nicely that i decided to figure out how to make em more portable to add to the challenge. For now, ho ... |
| 6983 | ham | Decorated my hookah vase. It says PLURtastic on it and has stars, glowsticks, fire, and two burningman logos on it |
| 6984 | ham | @jas2cool4u2005 its coming be patient |
| 6985 | ham | Lol i just read a bumper sticker that said "if you dont like the way i drive go punch yourself in the face |
| 6986 | ham | I just read an article that said 2007s burningman was prematurely burned and ther rebuilt it in two days with a phoenix cut out of it |
| 6987 | ham | The person in the apartment under me is moving out today! That means no noise issues for the weekend! Get as loud as possible! |
| 6988 | ham | Thats interesting. They found, caught, and killed a new kind of animal in texas that some are believing is the chupacabra |
| 6989 | ham | Check it out! New hookah video on our hookah science myspace! |
| 6990 | ham | Finally modified the prototype double firestaffs SO now they are hundred percent ready to play with! |
| 6991 | ham | I'm wide awake and its 2am. Dunno what to do. |
| 6992 | ham | Beautiful thunderstorm while smoking hookah on my porch and planning for burningman 2010 |
| 6993 | spam | OMG you have got to check this out now - it's unbelievable! >>> http://bit.ly/InternetTimeMachine |
| 6994 | ham | @jas2cool4u2005 i would assume so yes |
| 6995 | ham | Thankfully two drill holes and a cut aint a hard thing to redo |
| 6996 | ham | Quicksteel plus plastic sportsbottle equals FAIL! I need to get another and redrill it cuz the whole bottle smells like epoxy |
| 6997 | ham | @keithsuperk aww i miss MoS. ITS A FUCKING TRAIN STATION! |
| 6998 | ham | If you choose to do illegal substances please know your limit! RIP DJ AM |
| 6999 | ham | @VITCI oh yeah they always take forever but theyre the only ones that sell kilos |
| 7000 | spam | This is the game changer - only \$5 >>> http://bit.ly/InternetTimeMachine |
| 7001 | spam | RE: @ZainR Signed the better world petition. Thank you for your support. http://tinyurl.com/vj5w9fw |
| 7002 | ham | @VITCI that sucks what site? |
| 7003 | ham | Letting jasper play around on the porch while i smoke hookah and watch the sunset |
| 7004 | ham | @jas2cool4u2005 hooray! i wanna move back! |
| 7005 | spam | A 2400 Dollar SEO project: I need 10 coder from 10 different zone. Each will do same job. What i need is given bel... http://bit.ly/bm8ubY |
| 7006 | ham | @VITCI yes, yes you are |
| 7007 | ham | @VITCI Then you are a hooker too |
| 7008 | ham | @VITCI your moms a hooker |
| 7009 | | |

Fig A.1 Dataset in .csv format



Fig A.2 HAM WORD CLOUD

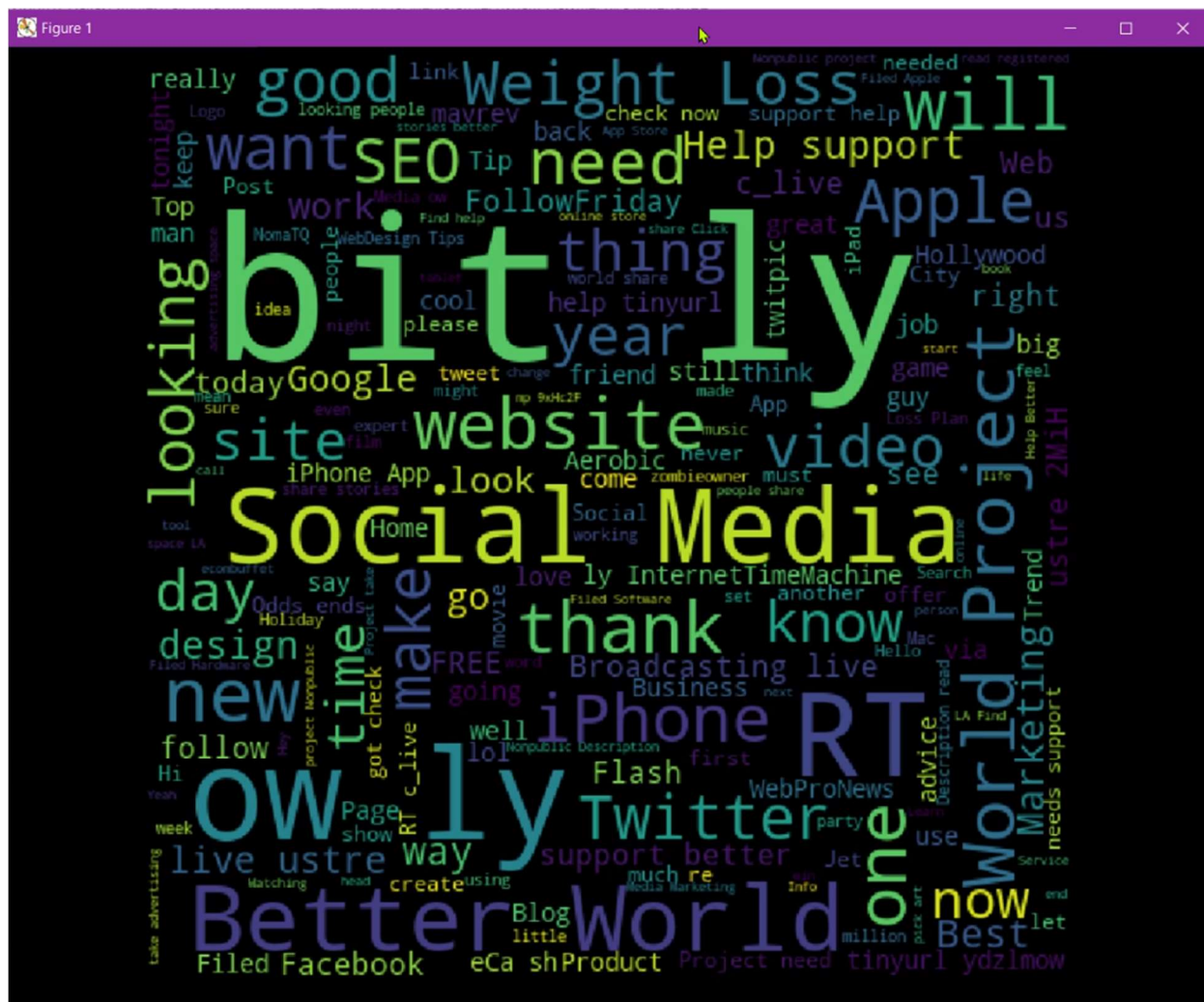


Fig A.3 SPAM WORD CLOUD

```
Python 2.7.15 Shell
File Edit Shell Debug Options Window Help

***** HAM WORD CLOUD *****

-----

***** SHOWCASE OF TRAINING DATA *****
v1 v2 length
0 ham Apple Defends iPhone App Approval Process - th... 84
1 ham Preparando una entrada bien cargadita de fotos... 60
2 ham @rrey estoy probando el tweetie ahora que no l... 91
3 ham @LuY ya le probé pero no me terminaba de conve... 81
4 ham ...until something better smelling with a pre... 78
5 ham Dear God the smell is fantastic and YUM- Yanke... 130
6 ham @dstagg is a mixmaster with vodka and applejui... 111
7 ham Quickie @coffeegroundz with @jrcohen before he... 137
8 ham RT @qcait: Lots of love to @twiterpated for th... 116
9 ham @twodeuces that's great news; hope recovery is... 79
10 ham @danielwcarlson try http://kickyoutube.com 42
11 ham @realdawnsummers um... i may have been twitter... 67
12 ham @ericaogrady I admit I'm jealous of your journ... 70
13 ham @peacecorn oh honey I am so sorry. hugs to you... 78
14 ham @realdawnsummers DIRTY SANCHEZ you too! 39

***** TRAINING DATA GROUPED BY LABEL *****
ham 3400
spam 2014
Name: v1, dtype: int64

-----

MULTINOMIAL NAIVE BAYES PREDICTION: 85.78%
BERNOULLI NAIVE BAYES PREDICTION: 71.22%

MNB Classifier accuracy 94.51%
BNB Classifier accuracy 88.07%

MNB F1 score is: 87.11
BNB F1 score is: 80.37
>>> |
```

Fig A.4 CLASSIFICATION RESULTS

Examples:

Analysed Tweet:

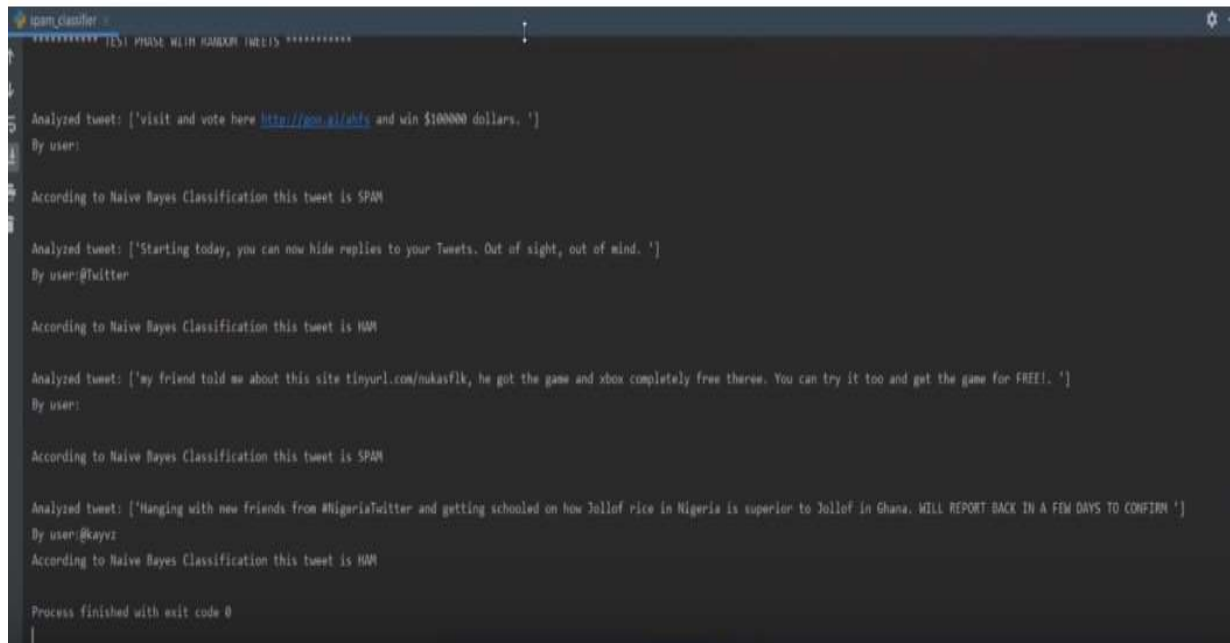
“Starting today, you can now hide replies to your Tweets. Out of sight, out of mind “

According to Naive Bayes Classifier, this tweet is: HAM

Analysed Tweet:

“my friend told me about this site tinyurl.com/nukasflk, he got the fame and xbox completely free from there You can try It to and get the game for FREE! “

According to Naive Bayes Classifier, this tweet is: SPAM

A terminal window titled 'spam_classifier' with a dark background and light blue text. It shows the results of a Naive Bayes classifier testing four tweets. The first tweet is classified as SPAM, the second as HAM, the third as SPAM, and the fourth as HAM. The process ends with 'Process finished with exit code 0'.

```
spam_classifier
***** TEST PHASE WITH RANDOM TWEETS *****

Analyzed tweet: ['visit and vote here http://gon.g/ahfs and win $100000 dollars. ']
By user:

According to Naive Bayes Classification this tweet is SPAM

Analyzed tweet: ['Starting today, you can now hide replies to your Tweets. Out of sight, out of mind. ']
By user: @Twitter

According to Naive Bayes Classification this tweet is HAM

Analyzed tweet: ['my friend told me about this site tinyurl.com/nukasflk, he got the game and xbox completely free there. You can try it too and get the game for FREE!. ']
By user:

According to Naive Bayes Classification this tweet is SPAM

Analyzed tweet: ['Hanging with new friends from #NigeriaTwitter and getting schooled on how Jollof rice in Nigeria is superior to Jollof in Ghana. WILL REPORT BACK IN A FEW DAYS TO CONFIRM. ']
By user: @kayyz

According to Naive Bayes Classification this tweet is HAM

Process finished with exit code 0
```

Fig A.5 Real Time Tweet Classification