Cyber Security

Applied Data Science

Youssef Taib 14058448

Mendel lion sjin tjoe 16121538

Boyoung Han 17149746

Jorge E Ortiz Durán 17170311

Petri Johannes Pellinen, 18116256

Jaekyu Lee 18132782

Table of Contents

1. Introduction

Social networks have been a powerful platform to achieve globalization in recent years, millions of data are generated every day, it is easy to learn about the things that are happening in the world. It is also easy to share information about food, traditions, the culture of other countries, etc. Therefore, you can learn about anything from social networks. Twitter is one of the most popular forms of social networks. Users can tweet their thoughts, retweet others, favourite other tweets and use hashtags in order to find tweets that are similar to their tweets. This helps certain hashtags to go viral.

In this huge social network, there is a problem called 'hacktivism'. Hacktivism is the use of technology to promote political agenda or social change. The reason for hacktivism is to draw public attention. The hacktivists believe that the issue is important, they address issues like freedom of information or human rights. Hacktivists also use it to oppose something in which case they use images or messages to websites of organizations they believe something wrong. Hacktivism can have several meanings. In some cases, it is about breaking into the security barriers of a computer as well as it can mean hacktivists gain unlawful access into a network.

'Cyber Security' is a project group that aims to create two working classifiers for twitter. The main subject of this project is identifying a hacktivist based on language used in tweets. This is achieved by creating a keywords list, this keywords list is in turn used by 2 classifiers that use machine learning in order to identify the number of times a word has been used. With two classifiers a greater accuracy can be achieved in distinguishing a hacktivist from a non-hacktivist. Machine learning will be used to show patterns and make predictions on the twitter data.

This paper focuses on Naive Bayes and SMV by using data which is extracted from tweets in order to distinguish a hacktivist from a non-hacktivist.

These two classifiers can reach the frequency of some words that there are identified as hacktivism words because we labelled one of the datasets that we are using to be sure that there are some tweets that in fact encourage to the hacktivism, and how do we know that? Because we read almost 10,000 tweets and classified in hacktivist and non-hacktivist tweets, that can give us a ground truth of some context that in this case, a computer can´t give us, because we are able to understand the context of some of the tweets.

The rest of the paper is organized as follows. Chapter 2 provides terminology that is used in the project. Chapter 3 illustrates the approach to achieve the goal. Chapter 4-5 detail the data gathering, labelling, the classifiers used in this project and the results. Chapter 6 describes future work. Chapter 7 concludes this project.

2. Main question and sub questions of the research

2.1 Main question

Can we distinguish hacktivists from normal users by language?

2.2 Sub questions

What kind of behaviour do hacktivists display?

Which of the hacktivists are active, and which are passive?

What are the terms active hacktivists use and what are the terms passive hacktivists use?

What are the trends by hacktivists?

3. Terminology

In this chapter, hacktivism and Tweepy will be explained.
Hacktivism is an online form of activism. Used by hackers to target people as well as organisations.
Hacktivism targets people and organisations in order to make a change in the world they live in by denying service of an app or hacking into for example a Twitter page of a dangerous group, To send out a message that what they're doing is unacceptable. This can be done on a variety of platforms ranging from a twitter page (Which is the subject of this research plan) to banks, government websites, etc.

3.1 Tweepy

Tweepy gives access to the Twitter API, using tweepy is it possible to attain any object. Using Twitter developer page, it is possible to access tweet information. The main purpose of using Twitter is monitoring tweets an important part of tweepy is a stream listener object that monitors tweets in real time and can be turned to csv.
Tweepy is the open-source library that provides access to the Twitter API for python. It relies on Twitter API and that has excellent documentation. Which means it can very much be used as a reliable library. Other libraries like python-twitter provide many functions, but tweepy has a large community and the most code that's added.

3.2 Anonymous

The most famous group of hacktivists is Anonymous. Anonymous is a group of hackers that create missions and tasks in order to make a change in the world they live in. This includes taking down paedophiles and sites that host child pornography, and this along with other missions that serve to better society. They have amassed an enormous following on social media.
Anonymous is also highly active on social media including Twitter. Besides Anonymous having their own social media accounts, it's members also have accounts. Using their account to not only express views and alert about upcoming events, but they also have accounts that they use in their daily lives.
These accounts will be aimed to identify hacktivists.

4. Approach

Our approach was to start from mining the data from Twitter. It is the first step because, without the data, we cannot do anything else. Data mining was done with an API from GitHub, which we modified to fill our purposes better. To get the data extraction working we had to apply and receive a developer code from Twitter. After that, we could start to extract and store tweets in a text file.

Text pre-processing was used to clean the data and pass on only relevant information. In text pre-processing, there are a few different steps of manipulating the data. The goal of pre-processing was to make all the tweets to be in the same format for easier use later. In this step, all the tweets are first in their original form. Then a few functions are applied to the data frame of tweets to make them unified. Punctuation was removed first because it makes further applications of functions easier. While removing punctuation, also emoticons are removed. After that, all the tweets were fully converted to be only lower-case characters. In addition, also English stop words were removed. Stop words include words like "a, an, the, in, on, and" and so on.

Now we have tweets that are stripped from unnecessary words, there is no punctuation, and everything is in lower case for easier manipulation. Text pre-processing is vital for the next step of counting the frequency of words used in the tweets of followed users.

Terms frequency counting was used for comparing the most used words for a keywords list provided to us by the product owner. In addition, new keywords were added to the existing list from the most frequent words used by hacktivists. On addition of counting the frequent words, different counter functions were used to count only hashtags or usernames used in the tweets.

The next step is analysing the data. There was some manual work involved in the analyzation process. We labelled a part of tweets by being related to hacking or not. The tweets labelled were from users that we know are hacktivists. This was a tool used in our classifier to determine the users' involvement in the hacktivist scene.

Data was then visualized to get a better understanding of the results. Visualization helps for example in a situation when one needs to access word frequency of specific dates or the popularity of specific words on different dates.

With visualization, it is also easier to spot false positives and negatives from Naive Bayes classifier. Naive Bayes classifier was chosen because it is optimal for text categorization and offers a solution to the problems of the project. Supervised learning with previously mentioned manually labelled tweets are used to teach the classifier. Lastly, sentiment analysis is used for classifying the tweets in three categories; positive, neutral and negative tweets. With sentiment analysis it is possible to detect a change in the emotion of the tweets, possibly alerting from a future defacement of a website and more.

4.1 Related Work

In this chapter, relevant work will be discussed. Similar research and the result will be displayed in this chapter. For this research, a study concerning social sentiment and a study concerning Twitter will be used for reference. Cyber-attacks are on the rise because of the increasing globalisation. These attacks form a great risk for the following areas; Denial of service, data leaking and application compromising among others.

A variety of anti-threat measures are in effect in order to combat attacks like DDoS. Predictive analysis can be particularly beneficiary for Twitter because of the fact that certain Twitter functions that include retweets, favourites and replies can be characterized and this along with the polarity of the text can improve predicting events like political elections and the release of new products. The predictive power of social networks can be used by investigating published data and statistical modelling and that can assist in identifying statistical similarities between social users on social media. Therefore, it can be concluded that sentiment analysis is a useful way to analyse tweets.

In research (5) Twitter has researched using sentiment analysis, using natural language processing and Machine learning techniques to interpret sentimental tendencies related to user opinions and make predictions about user opinions and make predictions about real events.

In the paper, a Social Sentiment Sensor in Twitter has been used in order to collect historical tweets. This in order to classify negative, positive and security-oriented tweets. When 3 different classification algorithms were used to evaluate results, maximum entropy provided the most effective results. Naive Bayes and Support vector machine followed and were responsible for less accurate results. In this research, accuracy of 80% has been achieved of identifying the tweets related to hacking.

Which proves that using sentimental analysis for Twitter can be used as a reliable base to analyse tweets.

This research (9) looks into movie reviews in order to attain the sentiment of the reviewers. Where it starts by stating that research towards this topic has come from a shift of interest from the topics of discussion. This can be for example sports or politics. The researchers have concluded that this shift has changed into the interest for the sentiment of the users rather than just the discussion topics. This can be about a product review, focusing on whether the review would be positive or negative. On Rotten Tomatoes (a movie critic website) the reviews are labelled with a rating system in addition to the review. The problem that is being examined is a sentiment classification problem. The similarities between this related research and the research that is being presented right now becomes apparent, while it’s possible to extract the keywords (because keywords hold a clear value because its only one word) with sentiment comes a variety of possibilities to express emotions. For example; Without a rating, a sentence like: Well this was great, Contains zero words with a negative value. However, when you also take in account a rating system 1-10 and it has been graded by the reviewer with a low score, it becomes apparent that the reviewer used sarcasm in order to get their point across.

Three machine learning methods were used in order to see what classification method Naive Bayes classification, maximum entropy classification, and support vector machines. Two of which were also used in this research. The accuracy for Naive Bayes and support vector machine both topped 80%. Where the accuracy lacked was in the sentences that had an abundance of positive words. For an example a sentence where the effort put into performance was graded in the review rather than the performance itself.

In the relevant works that have been examined, there has been a clear pattern is revealed.

Also, it confirms that it is difficult to differentiate the positive or negative sentiment for either a tweet or a movie review. And researchers confirm that using Supporter Vector machine and Naive Bayes is completely legitimate and reliable ways of using machine learning in order to attain the sentiment of a sentence.

5. Methodology

5.1 Theory

5.1.1 Baseline

As a baseline, we use project owner'r list of keywords. This list consists of 100 negative words related with hacktivists. For each tweet, we count the number of negative keywords that appear. This classifier returns the polarity with the higher count.

5.1.2 Naive Bayes

Naive Bayes is a simple model which works well on text categorization [5]. We use a multinomial Naive Bayes model. Class c∗ is assigned to tweet d , where

$$c^* = \text{argmac}, PNB(c|d)$$

$$PNB(c|d) := \frac{(P(c)\sum_{i=1}^{m} P(f|d)^{ni(d)}}{P(d)}$$

In this formula, f represents a feature and n i(d) represents the count of feature f i found in tweet d. There are a total of m features. Parameters P(c) and P(f|c) are obtained through maximum likelihood estimates, and add-1 smooth-ing is utilized for unseen features.

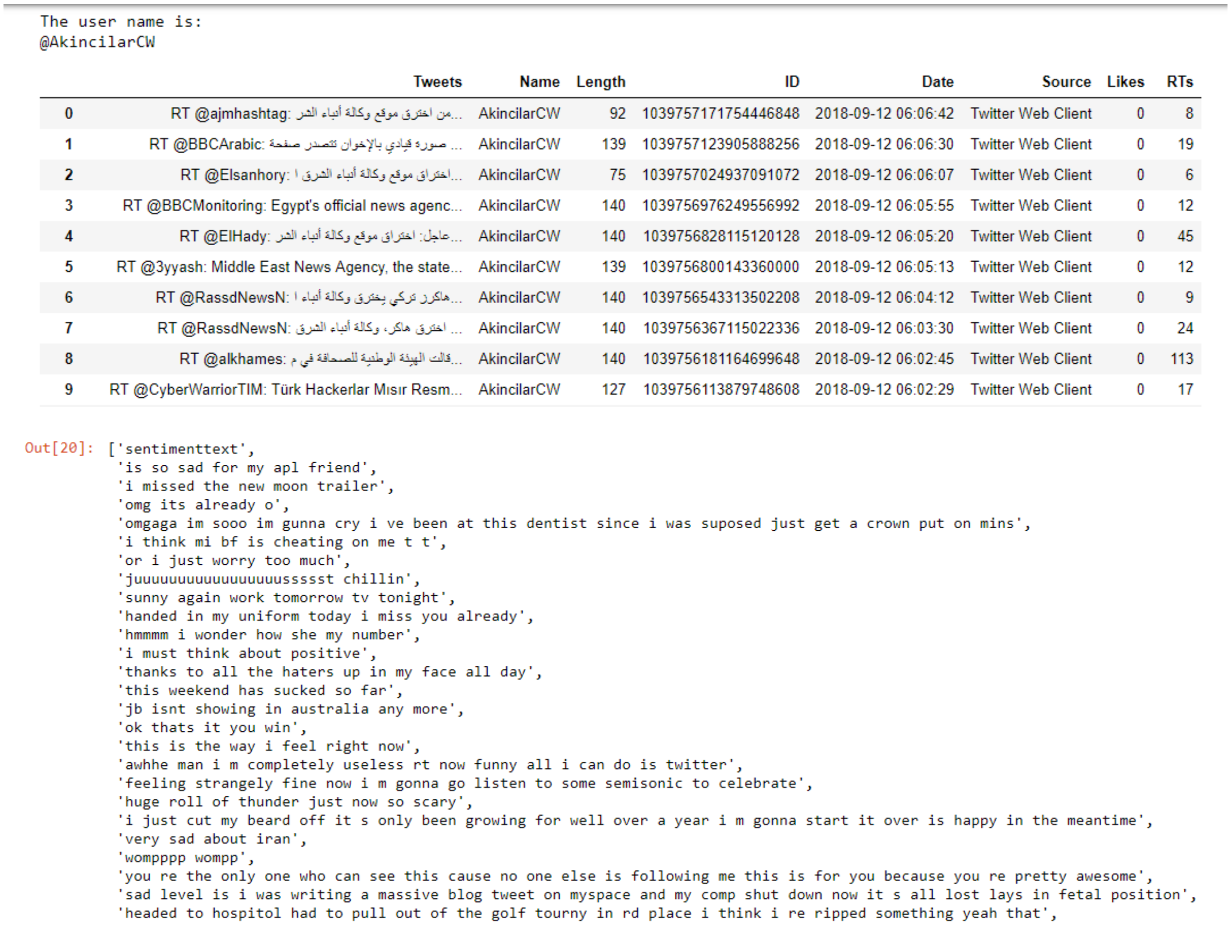
5.1.3 Support Vector Machines

Support Vector Machines is another popular classification technique. We use the SVM light software with a linear kernel. Our input data are two sets of vectors of size m. Each entry in the vector corresponds to the presence a feature. For example, with a unigram feature extractor, each feature is a single word found in a tweet. If the feature is present, the value is 1, but if the feature is absent, then the value is 0. We use feature presence, as opposed to acount, so that we do not have to scale the input data, which speeds up overall processing.

5.2 Data and Methods

5.2.1 Overview

This work involves the integration of many techniques. This section attempts to give an overview of all of these parts so that the reader can understand and place in context all of the methods covered in the subsections that follow. Figure 1 shows the transformation and handling of the data from source to result.



```
The user name is:
@AkincilarCW
```

	Tweets	Name	Length	ID	Date	Source	Likes	RTs
0	RT @ajmhashtag: من اخترق موقع وكالة أنباء الشر...	AkincilarCW	92	1039757171754446848	2018-09-12 06:06:42	Twitter Web Client	0	8
1	RT @BBCArabic: صوره قيادي بالإخوان تتصدر صفحة ...	AkincilarCW	139	1039757123905888256	2018-09-12 06:06:30	Twitter Web Client	0	19
2	RT @Elsanhory: اختراق موقع وكالة أنباء الشرق ا...	AkincilarCW	75	1039757024937091072	2018-09-12 06:06:07	Twitter Web Client	0	6
3	RT @BBCMonitoring: Egypt's official news agenc...	AkincilarCW	140	1039756976249556992	2018-09-12 06:05:55	Twitter Web Client	0	12
4	RT @ElHady: عاجل: اختراق موقع وكالة أنباء الشر...	AkincilarCW	140	1039756828115120128	2018-09-12 06:05:20	Twitter Web Client	0	45
5	RT @3yyash: Middle East News Agency, the state...	AkincilarCW	139	1039756800143360000	2018-09-12 06:05:13	Twitter Web Client	0	12
6	RT @RassdNewsN: هاكرز تركي يخترق وكالة أنباء ا...	AkincilarCW	140	1039756543313502208	2018-09-12 06:04:12	Twitter Web Client	0	9
7	RT @RassdNewsN: اخترق هاكر، وكالة أنباء الشرق ...	AkincilarCW	140	1039756367115022336	2018-09-12 06:03:30	Twitter Web Client	0	24
8	RT @alkhames: قالت الهيئة الوطنية للصحافة في م...	AkincilarCW	140	1039756181164699648	2018-09-12 06:02:45	Twitter Web Client	0	113
9	RT @CyberWarriorTIM: Türk Hackerlar Mısır Resm...	AkincilarCW	127	1039756113879748608	2018-09-12 06:02:29	Twitter Web Client	0	17

```
Out[20]: ['sentimenttext',
          'is so sad for my apl friend',
          'i missed the new moon trailer',
          'omg its already o',
          'omgaga im sooo im gunna cry i ve been at this dentist since i was suposed just get a crown put on mins',
          'i think mi bf is cheating on me t t',
          'or i just worry too much',
          'juuuuuuuuuuuuuuuusssst chillin',
          'sunny again work tomorrow tv tonight',
          'handed in my uniform today i miss you already',
          'hmmmm i wonder how she my number',
          'i must think about positive',
          'thanks to all the haters up in my face all day',
          'this weekend has sucked so far',
          'jb isnt showing in australia any more',
          'ok thats it you win',
          'this is the way i feel right now',
          'awhhe man i m completely useless rt now funny all i can do is twitter',
          'feeling strangely fine now i m gonna go listen to some semisonic to celebrate',
          'huge roll of thunder just now so scary',
          'i just cut my beard off it s only been growing for well over a year i m gonna start it over is happy in the meantime',
          'very sad about iran',
          'wompppp wompp',
          'you re the only one who can see this cause no one else is following me this is for you because you re pretty awesome',
          'sad level is i was writing a massive blog tweet on myspace and my comp shut down now it s all lost lays in fetal position',
          'headed to hospitol had to pull out of the golf tourny in rd place i think i re ripped something yeah that',
```

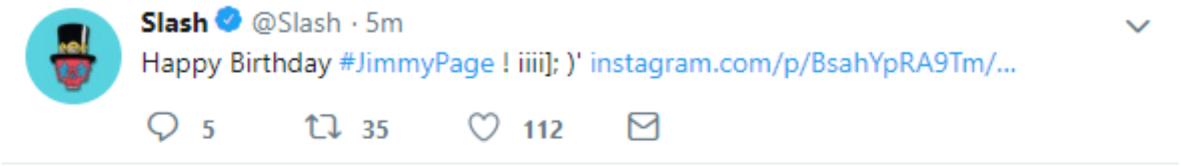
Figure 1Overview of data collection, pre-processing, cleaning and final analysis processes

Our framework for the study of predicting hacktivist with Twitter data proceeds in four stages. The first one deals about collecting Twitter data, group together by user and then put the data in a csv file (cf. Sections 3.2-3.4). Once we obtain the tweets from the users, we have to clean that data and pre-processing in order to make it clear to read. Then we proceed to label those tweets by hand looking for some tweets related with hacktivism or with hacktivists to have a ground true. After that we split the data in training and test datasets in order to apply the statistical tests to determine the relation between the tweets that we already check and the new tweets. We made this using the predictive models to predict by his tweets if a user is a hacktivist or if he is not.

5.2.2 Twitter Data

Twitter is an online microblogging platform that allows its users to build social net-works. It scores functionality is to share messages with the messages with the members of one´s own net-work, known as followers.

In the Twitter domain, these messages are known as tweets and are limited to a maximum of 140 characters by design. Figure 2 shows an example of a tweet; it has some parts with nothing:



Slash @Slash · 5m

Happy Birthday #JimmyPage ! iiii];)' instagram.com/p/BsahYpRA9Tm/...

5 35 112

Figure 2 Example of a tweet

— Other people can be mentioned or replied to by using the @ symbol followed by their user name. User names are alphanumeric strings of up to 15 characters. Underscores are allowed as well.

— Tweets beginning with the expression RT @[\w_]{1,15}are called retweets and are a handy way to share information with the people in your network.

— Words within a message preceded by the # symbol are known as hashtags and are mostly used to assign messages to topics or to mark keywords.

Retrieving Tweets

Even though working with Twitter data is becoming very common, one major problem is the lack of standard datasets. In April 2010 Twitter updated the API terms of service introducing a rule that does not allow third parties to redistribute Twitter Content without the company's prior written approval [Twitter 2010b; Twitter 2010a]. Therefore, attempts to release Twitter corpora, like the Edinburgh Corpus presented in [Petrovic et al. 2010], have failed. These restrictions make it very hard to reproduce previous results; thus, we had no choice but to collect our own data using the APIs permitted by Twitter, which work under restricted terms of use imposing bounds on the number of requests per hour that can be made, volume of tweets that can be retrieved, and other limitations. After evaluating all these APIs we chose to work with the Twitter Streaming API and Tweepy, which are intended for developers with data intensive needs and it works by establishing a single HTTP long-lived connection that is kept alive indefinitely and over which new tweets are sent as they are being posted; it also has a filtering method which is very convenient for the task we are trying to accomplish. The retrieval of tweets (or listening) began on 22 March 2011.

— Finally, a tweet can also contain URLs. While this is not Twitter-specific, it is important to note that links are very common and are to be expected.

5.2.3 Pre-Processing Twitter Data

We applied standard data cleaning and pre-processing techniques for preparing the Twitter data to build the sentiment index. These techniques include lower-case conversion, stop-word removal, duplicate removal (mostly corresponding to retweets), and language detection.

Language detection

Twitter's user interface is currently translated to 22 languages [Twitter 2012] and the default language is English. Therefore, it does make sense to do some sort of language detection. We use the Guess Language 2 library, to associate a language to each tweet at the time of retrieval. Internally, this tool applies some heuristics and looks at the frequencies of trigrams for each of the considered languages. The brevity of the tweets means that this method may fail to detect the correct language much more often than what it would for more extensive texts. This problem worsens if we take into account that people tend to use some English words in their native languages.

Handling negation

Negation can play an important role in the task of sentiment classification. Consider the sentences we think it was Good and we think it was not Good. While they only differ in one word and would score highly in most of similarity measures, their sentiment polarities are completely opposite. We have attempted a very simple form of negation handling for English texts by tagging words between common polarity shifters such as not, don't or haven't. For instance, the sentence from the previous example would become I think it was not NOT good. With this transformation, a word and its negated counterpart are considered to be different words, significantly increasing the size of the vocabulary. This translates to a larger set of features thus a larger dataset is preferable. That is why we have to label the data by hand and read all the tweets, to understand the context of the sentence and then determinate if that was a bad sentiment or a hacktivism tweet or i fit has not. Moreover, this technique only covers a small subset of negations where there is a valence shifter involved.

Relevance filter

One of the problems found by skimming through the collected data is that there is a considerable number of tweets that, while containing one or more of the filtering keywords, are not relevant to the task we are trying to accomplish. This is usually caused by homonyms, polysemous words, proper names or words that are part of an idiom.

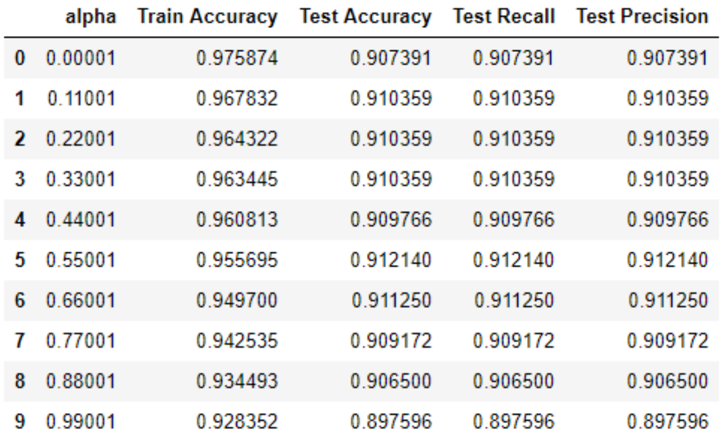
We have created two different datasets with which we will train two different LDA-based filtering models. The first dataset consists of 8000 tweets containing all the users' tweets that they were given to us. The second dataset contains 1,000,000 random tweets from Stanford link's sentiment 140 public project.

To evaluate the performance of the two trained models, we have used the labelled data set as our training model and then split it with one test set without labelled tweets to proof the efficiency of our models. These include accounts from hacktivist, among others. It should be noted, though, that these labels have been set by hand but maybe some errors may be expected.

5.2.4 Sentiment Classifier

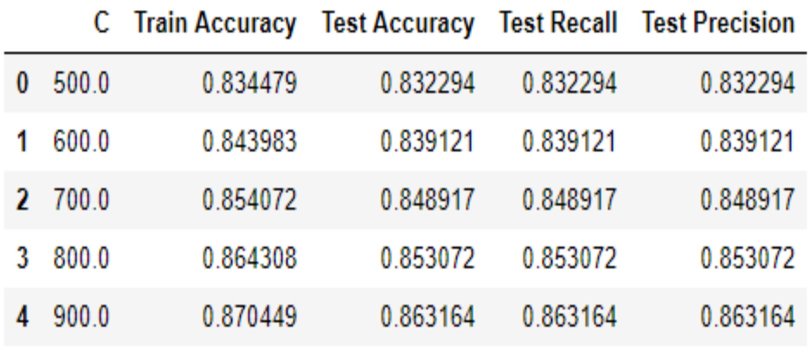
The goal of a textual sentiment classifier is to determine whether a text contains positive or negative impressions on a given subject, in order to determine if that user is a hacktivist or if is not. After a pre-processing step, where at this stage the noisy terms from the tweets are removed (cf. Section 3.3), the next problem that must be addressed is the development of a corpus from which to train a sentiment classifier. As a first step for constructing such a corpus, we implemented in this project a recent labelling idea which consists on label tweet by hand to create a ground true. Several attempts at multi-class sentiment analysis have been made. For instance, in [Ahkter and Soria 2010] Facebook messages are classified into Happy, Unhappy, Sceptical and Playful. Nevertheless, we only focus on the binary classification problem. Multiple datasets have been extracted from our tweet collection for training the sentiment classifier. Each dataset has a balanced number of positive and negative instances and thus the total amount of tweets is limited by the number of negatives in the collection. Retweets have not been included in these datasets.

We have trained several sentiment classifiers using Multinomial Naive Bayes and Supported Vectors Machine with varying pre-processing steps and multiple sets of feature words to represent the documents for each of datasets. Three feature lists have been considered: alpha value, C value, Train Accuracy, Test Accuracy, Test Recall and Test Precision. To evaluate the classifiers generated, we have collected two additional independent datasets (English and multi-language), containing 50000 and 200000 tweets respectively. Only the best scoring classifiers are listed in Figure 3 and Figure 4 that are considered throughout the rest of this paper.



	alpha	Train Accuracy	Test Accuracy	Test Recall	Test Precision
0	0.00001	0.975874	0.907391	0.907391	0.907391
1	0.11001	0.967832	0.910359	0.910359	0.910359
2	0.22001	0.964322	0.910359	0.910359	0.910359
3	0.33001	0.963445	0.910359	0.910359	0.910359
4	0.44001	0.960813	0.909766	0.909766	0.909766
5	0.55001	0.955695	0.912140	0.912140	0.912140
6	0.66001	0.949700	0.911250	0.911250	0.911250
7	0.77001	0.942535	0.909172	0.909172	0.909172
8	0.88001	0.934493	0.906500	0.906500	0.906500
9	0.99001	0.928352	0.897596	0.897596	0.897596

Figure 3 Table of the result in Naive Bayes Classifier



	C	Train Accuracy	Test Accuracy	Test Recall	Test Precision
0	500.0	0.834479	0.832294	0.832294	0.832294
1	600.0	0.843983	0.839121	0.839121	0.839121
2	700.0	0.854072	0.848917	0.848917	0.848917
3	800.0	0.864308	0.853072	0.853072	0.853072
4	900.0	0.870449	0.863164	0.863164	0.863164

Figure 4 Table of the results in SVM classifier

The sentiment classifier that we use has been built using standard methodology, and our results are comparable to most state-of-the-art methods, some publicly available through APIs.

Predictive classifiers

Using the top-scoring sentiment classifiers we obtain a collection of predictions about which users were hacktivist and if we were expecting that they were hacktivist and users who were not hacktivist and we were not expecting them to be hacktivist that will represent the evolution of the general mood towards a specific item, expressed in terms of one or more filtered words. Each predictive classifiers is represented as a time series where every value corresponds to the percentage of labelled tweets as hacktivist tweets over the total number of messages that were posted on a given data.

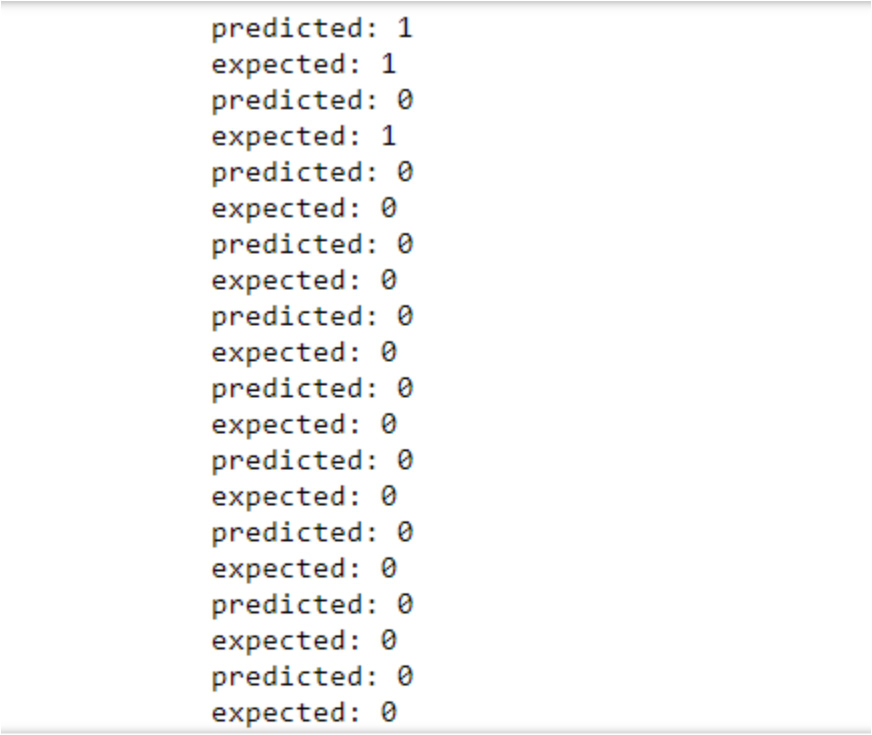
It should be noted that, in contrast to the training of the sentiment classifier, retweets are removed during the predictive index generation.

5.2.5 Model Adequacy

We attempt to be mathematically rigorous and, hence, in order to have some certainty of the adequacy of Twitter as part of a predictive model for our data sets, we run some widely accepted tests to assess, first, for a relationship among the target labelled tweets and neutral tweets build from Twitter data, which can either be a predictive index (as described in Section 3.4) or a simple count of tweets (i.e., volume); We briefly comment in this section on the tests we use for the assessment of our models.

5.2.6 Predictive Models and Evaluation

As has been stated in the Introduction our primary goal is to study the effect of using Twitter data when doing predictions in different users to see if they are hacktivist or not with various machine models popular among the Machine Learning community. The predicted values for the tweets are obtained by training a machine learning model and providing them with past observations (lags) of both the labelled tweets and the normal ones. This is a binary classification task, so the models considered in the experiments have been chosen accordingly. The models that we have considered are among the most popular and effective in Machine Learning. These are: Support Vector Machines and Naive Bayes. Detailed descriptions of these models can be found in standard textbooks such as [Mitchell 1997; Hastie et al.2003]. However, since the amount of available data in our collection is rather limited in terms of the number of daily observations, we have taken a prequential approach [Gama et al. 2009; Bifet et al. 2010] for evaluating the experiments. This means that, for each prediction, a model is fitted with all the available past data and also with new data that we collect by ourselves and labelled. Once the actual value is known, it is included in the training set so it can be used for the next prediction. After repeating this process for all the available observations, we get a result of hits and misses like the one below in the Figure 5.



```
predicted: 1
expected: 1
predicted: 0
expected: 1
predicted: 0
expected: 0
predicted: 0
expected: 0
predicted: 0
expected: 0
predicted: 0
expected: 0
predicted: 0
expected: 0
predicted: 0
expected: 0
predicted: 0
expected: 0
predicted: 0
expected: 0
```

Figure 5 The results of hits and misses

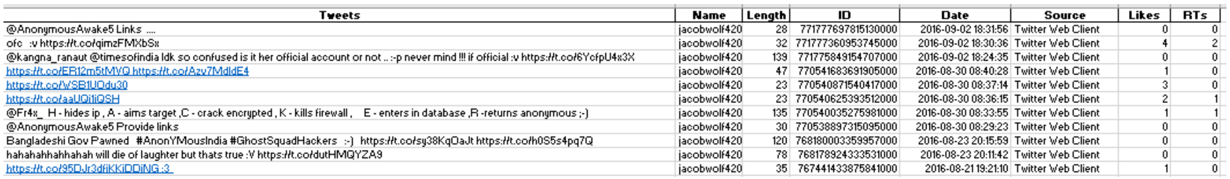
6. Evaluation

There are not any large public data sets of Twitter messages with sentiment, so we collect our own data. There are six steps to estimate possible malicious users by analysing collected tweets. After pre-processing, overall users' sentiment levels were scored and selected possible malicious users. Dataset was consisted based on the result of topic modelling. Machine learning algorithms have been used to classify and cluster possible hacktivist. Finally, selected users who are in red group of possibility were adapted to the condition of noncompliance of information security and compared whether he/she corresponds to the condition of noncompliance. Similarly, W. Park, Y. You, and K. Lee [3] researched correlation between tweets and actual behaviour in the real world. 4,000 Tweets of 2016

U.S. presidential election nominees (Donald Trump and Hillary Clinton) were crawled. After sentiment analysis process, daily events were correlated with the negative tweets based on daily average sentiment score. Through the machine learning process, this research proved that individual's social media reflects the real world's situation.

6.1 Pre-processing

Our dataset consisted of .csv file. It has User ID, date, length, likes, source, RTs, tweets and sentiment of tweet in the file. Some tweets have mathematical character "=". When this character is at the front of the sentence, excel program perceives it as a function. So, we had to remove "=" at the front of sentence. Also, we removed unessential words including the parts of sentence such as pronouns, prepositions, and articles. Through this process, we could get the groups of words and were erased by the topic modelling. It is shown in Figure 6.



Tweets	Name	Length	ID	Date	Source	Likes	RTs
@AnonymousAwake5 Links	jacobwolf420	28	771777697815130000	2016-09-02 18:31:56	Twitter Web Client	0	0
ofc :v https://t.co/qimzFMXbSx	jacobwolf420	32	771777360953745000	2016-09-02 18:30:36	Twitter Web Client	4	2
@kangna_ranaut @timesofindia Idk so confused is it her official account or not .. :-p never mind !!! if official :v https://t.co/6YcfpU4x3X	jacobwolf420	139	771775849154707000	2016-09-02 18:24:35	Twitter Web Client	0	0
https://t.co/ER12m5tMVQ https://t.co/Azv7MdldE4	jacobwolf420	47	770541683691905000	2016-08-30 08:40:28	Twitter Web Client	1	0
https://t.co/WSB1UOdu30	jacobwolf420	23	770540871540417000	2016-08-30 08:37:14	Twitter Web Client	3	0
https://t.co/aaUQi1iQSH	jacobwolf420	23	770540625393512000	2016-08-30 08:36:15	Twitter Web Client	2	1
@Fr4x_ H - hides ip , A - aims target ,C - crack encrypted , K - kills firewall , E - enters in database ,R -returns anonymous ;-)	jacobwolf420	135	770540035275981000	2016-08-30 08:33:55	Twitter Web Client	1	1
@AnonymousAwake5 Provide links	jacobwolf420	30	770538897315095000	2016-08-30 08:29:23	Twitter Web Client	0	0
Bangladeshi Gov Pawned #AnonYMousIndia #GhostSquadHackers :-) https://t.co/sy38KqOaJt https://t.co/h0S5s4pq7Q	jacobwolf420	120	768180003359957000	2016-08-23 20:15:59	Twitter Web Client	0	0
hahahahhahhahah will die of laughter but thats true :V https://t.co/dutHMQYZA9	jacobwolf420	78	768178924333531000	2016-08-23 20:11:42	Twitter Web Client	0	0
https://t.co/95DJr3dfiKKiDDiNG :3	jacobwolf420	35	767441433875841000	2016-08-21 19:21:10	Twitter Web Client	1	0

Figure 6 The tweets and some of the parameters that we collected

6.2 Analysing hacktivist users

At first, sentiment level was scored through the sentiment analysis of the whole data. After that, the average of the sentiment score and the ratio of negative tweets are calculated. Figures 7 shows the dates of the tweets and figure 8 show the number of likes and RTs that the users had. In figure 9 we can se the most used sources for create the tweets in this dataset, we classified 70 hacktivist users.

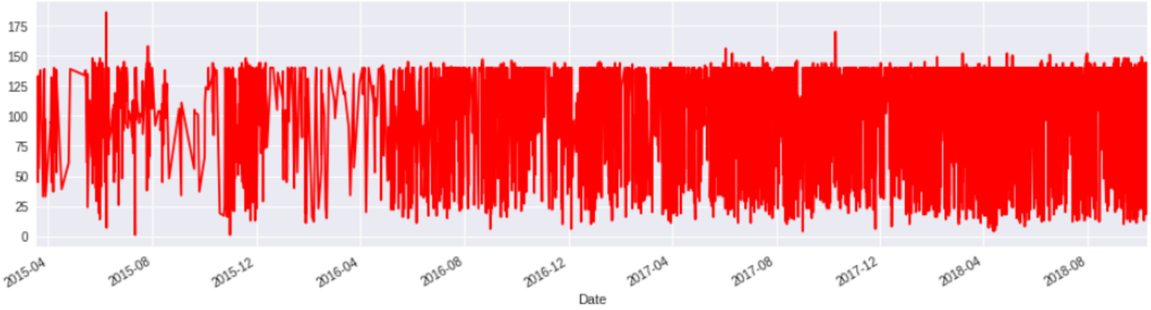


Figure 7 The most common dates for tweeting

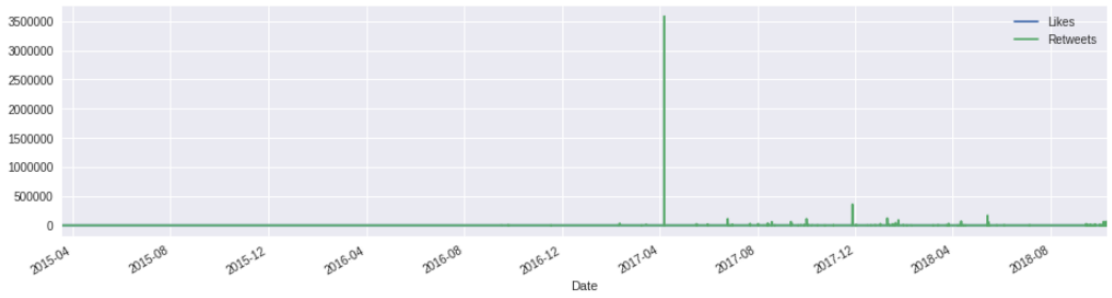


Figure 8 The most liked and re-tweeted tweets from our users

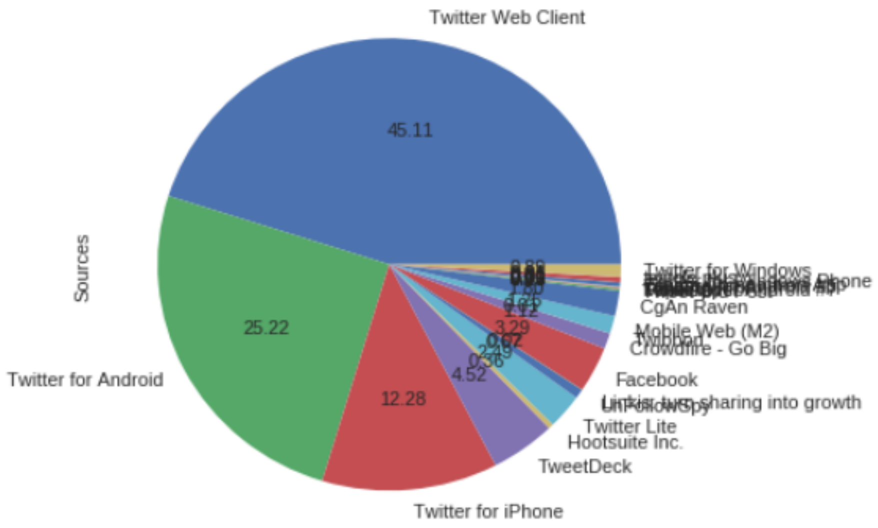
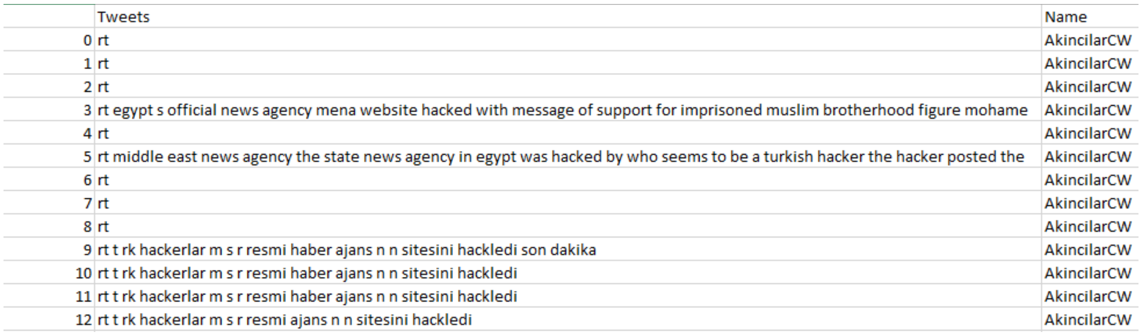


Figure 9 The most used sources to publish the tweets

6.3 Cleaning the Data

We cleaned the data, we take away those special characters that does not have relevance for us, for example "@" in the beginning of every user´s name, all the links to another pages, all the images, the tweets that were not in English, we throwed away, we only leave the information that we considered it relevant for the project. There is an example in figure 10.

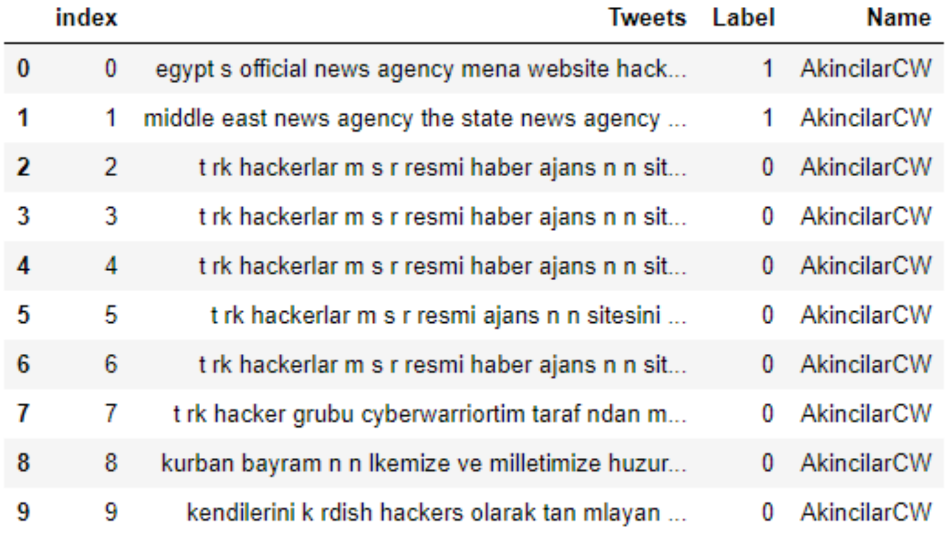


	Tweets	Name
0	rt	AkincilarCW
1	rt	AkincilarCW
2	rt	AkincilarCW
3	rt egypt s official news agency mena website hacked with message of support for imprisoned muslim brotherhood figure mohame	AkincilarCW
4	rt	AkincilarCW
5	rt middle east news agency the state news agency in egypt was hacked by who seems to be a turkish hacker the hacker posted the	AkincilarCW
6	rt	AkincilarCW
7	rt	AkincilarCW
8	rt	AkincilarCW
9	rt t rk hackerlar m s r resmi haber ajans n n sitesini hackledi son dakika	AkincilarCW
10	rt t rk hackerlar m s r resmi haber ajans n n sitesini hackledi	AkincilarCW
11	rt t rk hackerlar m s r resmi haber ajans n n sitesini hackledi	AkincilarCW
12	rt t rk hackerlar m s r resmi ajans n n sitesini hackledi	AkincilarCW

Figure 10 Cleaned data

6.4 Label the tweets

As a way to create a ground true we labelled 10,000 of tweets by hand, read it each tweet carefully to determinate if the tweets were neutral or hacktivist context tweets, we split the dataset and together we create a dataset with which al the sentiment in it was verified as true, in other words, we were sure about the labelled, so we can use it to train the models in the future. This labelled process is in figure 11.



	index	Tweets	Label	Name
0	0	egypt s official news agency mena website hack...	1	AkincilarCW
1	1	middle east news agency the state news agency ...	1	AkincilarCW
2	2	t rk hackerlar m s r resmi haber ajans n n sit...	0	AkincilarCW
3	3	t rk hackerlar m s r resmi haber ajans n n sit...	0	AkincilarCW
4	4	t rk hackerlar m s r resmi haber ajans n n sit...	0	AkincilarCW
5	5	t rk hackerlar m s r resmi ajans n n sitesini ...	0	AkincilarCW
6	6	t rk hackerlar m s r resmi haber ajans n n sit...	0	AkincilarCW
7	7	t rk hacker grubu cyberwarriortim taraf ndan m...	0	AkincilarCW
8	8	kurban bayram n n lkemize ve milletimize huzur...	0	AkincilarCW
9	9	kendilerini k rdish hackers olarak tan mlayan ...	0	AkincilarCW

Figure 11 An Example of labelled tweets

6.5 Machine Learning

As a way to detect hacktivist users, Machine Learning methodology was used. Supervised learning algorithms were adapted in this research and the result of learning. This process has a meaning as a way to identify possible hacktivist users and improve the accuracy. Topic Modelling was used to rank according to how often the words in tweets were used. And then, we listed negative words. Figure 12 shows the example of dataset with that negative words identified. We focused negative words because we assumed that people's negative thinking related to the threat would be presented by the words. To make a dataset, negative words were selected as features. Each of the features (word) and the number of how many times they were used were included in the dataset, then we labelled the tweets in order to classify that specific tweet like a tweet that has possible hacktivism content.

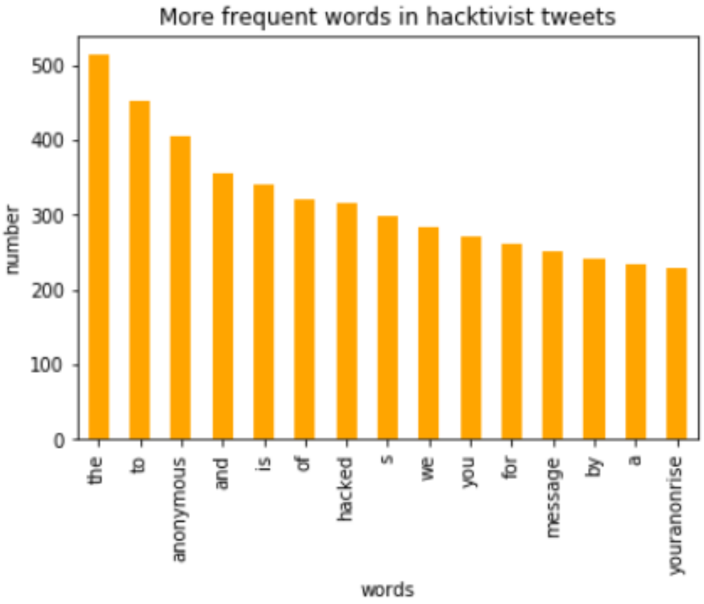


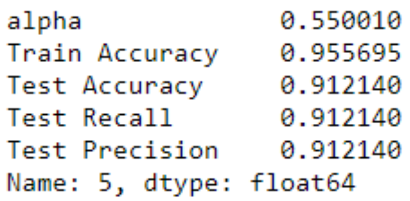
Figure 12 Frequent words used by hacktivists

6.5.1 Feature Selection

There are so many words which consist all of tweets. Among them, we chose negative words by frequency. After that, we composed a dataset and checked if the word is in the sentence.

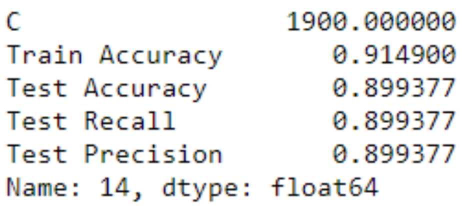
6.5.2 Supervised Learning

Two supervised learning algorithms were adjusted to analyse the accuracy of learning. Figure 13 and 14 shows the result of supervised learning who had the highest accuracy of SVM, and Naive Bayes. A kind of probabilistic model Naive Bayes assumes that the population follows Gaussian or polynomial distributions. The SVM can produce high accuracy when it classifies two categories by creating a non-probabilistic binary linear classification model, and the linear algorithm has similar characteristics.



```
alpha              0.550010
Train Accuracy     0.955695
Test Accuracy      0.912140
Test Recall        0.912140
Test Precision     0.912140
Name: 5, dtype: float64
```

Figure 13 Naive Bayes best model

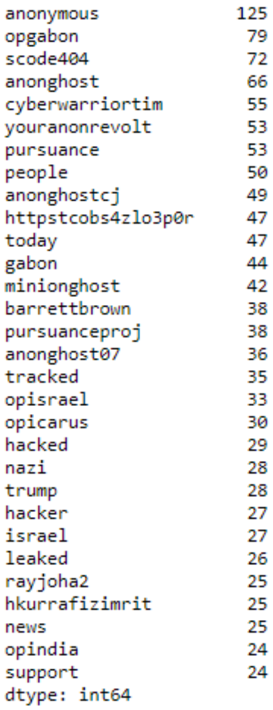


```
C                  1900.000000
Train Accuracy        0.914900
Test Accuracy         0.899377
Test Recall           0.899377
Test Precision        0.899377
Name: 14, dtype: float64
```

Figure 14 SVM best model

6.6 Get Frequently Used words

Topic modelling was used and ranked into the frequency of how often the words were used. We focused on negative words because we assumed that people's negative thinking related to the threat would be presented in the words. There are a lot of words and each score was evaluated by several people. Figure 15 shows an example of the most negative words that the hacktivist usually uses



```
anonymous            125
opgabon               79
scode404              72
anonghost             66
cyberwarriortim       55
youranonrevolt        53
pursuance             53
people                50
anonghostcj           49
httpstcobs4zlo3p0r    47
today                 47
gabon                 44
minionghost           42
barrettbrown          38
pursuanceproj         38
anonghost07           36
tracked               35
opisrael              33
opicarus              30
hacked                29
nazi                  28
trump                 28
hacker                27
israel                27
leaked                26
rayjoha2              25
hkurrafizimrit        25
news                  25
opindia               24
support               24
dtype: int64
```

Figure 15 Most negative user's words

7. Conclusions

This paper presented a methodology to distinguish hacktivists by using our own classifiers and sentiment analysis in Twitter. Our methodology collects tweets from 70 hacktivists and classifies them as negative, positive by hand-written sentiment classifier. Specifically, we have shown that the proposed methodology can detect whether the user is hacktivist. And we notice that Machine learning algorithms can do high accuracy for distinguishing users when using our methodology.

Social media can help people around the world communicate freely. we can use the social media as useful analysis media to detect the potential threats. In this paper, it has been shown that it is efficient to analyse individual tendencies to detect possible hacktivist users. To do this, sentiment analysis, and word count frequency was conducted on the collected tweets, and we classified the users by the level of threat according to our criteria. And then, the classified possible hacktivist users were verified by performing information security compliance matching process. Machine learning algorithms were applied to detect possible hacktivist users. In this way, a methodology has been proposed to prevent damage to the organization's information systems. This paper contributes to the analysis of data on social media to show that the criteria for detecting hacktivism threats are based on the sentiment level, word count frequency and the ratio of negative emotions, and it can be verified based on the concept of information security compliance. Above all, to improve the level of information protection of the organizations, it is necessary that not only the information protection person but also the management's active interest and efforts are put together.

Our work is actually not limited to detect hacktivists. Our future goal is making complete machine learning algorithm for predicting cyberattacks from hacktivists.

7.1 Future Work

In this work we have explained a robust methodology for the study whether the addition of Twitter data helps in forecasting. Our work could be improved in many ways, we proceed by explaining a few of the ideas we have not implemented yet. For example, we are currently using libraries to recognize sentiment in the sentiment analysis. In order to train a sentiment classifier, supervised learning usually re-quires hand-labelled training data. With the large range of topics discussed on Twitter, it would be very difficult to manually collect enough data to train a sentiment classifier for tweets. Because uses of context is very difficult to explain through the vocabulary that people use online. like shortening words, emoticons, words with numbers and words with punctuation marks. Also analysing URL from a tweet was very difficult to accomplish. Machine learning techniques perform well for the classifier and classifying sentiment in tweets. We believe that the accuracy could still be improved.

8. References

1. AHKTER, J.AND SORIA, S. 2010. Sentiment analysis: Facebook status messages.nlp.stanford.edu, 1-19.
2. Arias, M., Arratia, A., & Xuriguera, R. (2013). Forecasting with twitter data. ACM TIST, 5, 8:1-8:24.
3. Bahia, J. (2018). The online battleground : the use of online platforms by extremist groups and hacktivists to form networks and collective identities (T). University of British Columbia. Retrieved from https://open.library.ubc.ca/collections/ubctheses/24/items/1.0369285
4. Go, A. (2009). Sentiment Classification using Distant Supervision.
5. Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Perez-Meana, H., Olivares-Mercado, J., & Sanchez, V. (2018). Social Sentiment Sensor in Twitter for Predicting Cyber-Attacks Using ℓ_1 Regularization. Sensors (Basel, Switzerland), 18(5), 1380. doi:10.3390/s18051380
6. Munkhdorj, B., & Sekiya, Y. (2017). Cyber attack prediction using social data analysis. J. High Speed Networks, 23, 109-135.
7. Park, W., You, Y., & Lee, K. (2018). Detecting Potential Insider Threat: Analyzing Insiders' Sentiment Exposed in Social Media. Security and Communication Networks, 2018, 7243296:1-7243296:8.
8. W. Park, Y. You, and K. Lee, "Twitter sentiment analysis using machine learning," Research Briefs on Information & Communication Technology Evolution ,http://rbisyou.wixsite.com/rebicte/volume-3-2017, 2017.
9. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf
10. Rahman, A. (2018, 11 april). Hacking and Hacktivism: What it is and How Can it Affect Us. C https://medium.com/@rahman.alif1/hacking-and-hacktivism-what-it-is-and-how-can-it-affect-us-611e3e341967
11. Novalić, AHMET Introduction to tweepy, Twitter for Python - Python Central. van https://www.pythoncentral.io/introduction-to-tweepy-twitter-for-python/