Threat Detector for Social Media Using Text Analysis

Saidul Haq Sadi, Md Rubel Hossain Pk, Akram M Zeki
Department of Information Systems, Kulliyyah of ICT, International Islamic University Malaysia, Kuala Lumpur, Malaysia.
sadi190896@gmail.com

Abstract— The scam is one of the most important security threats among social media users. It is required to detect scams not only to protect social user's data that is stored online but also to secure the social network. Besides, machine learning techniques are becoming more popular in the text analysis sector. To fraud detection, the most used supervised machine learning techniques are Naïve Bayes (NB) and Support vector machine (SVM). In this project, a machine learning model is developed for detecting threats from Twitter tweets. Accordingly, the Naïve Bayes classifier and flask micro web framework were used to build the model by using the python programming language. The model provided 91% accuracy in detecting tweet scam threats. This finding will benefit the social network users to be aware of threats as well as social media network providers to enhance their security system.

Keywords— threat detector, social media, Twitter, supervised machine learning, Naïve Bayes classifier

I. Introduction

Threats are criminal activities intended to trick someone into having personal information or stealing money. Methods are continually evolving as scammers rummage for new ways of committing fraud and avoiding detection. Consumers can be reached easily by phone, by post, by email, or at their doorstep. Given the long history of scams to focus on possible victims, the internet can be a relatively new place for fraudsters and can easily adapt old tricks to brand new digital platforms [1].

On July 15 2020, within the US, many immoderate-profile Twitter debts of personalities along with Barack Obama, Bill Gates was given hacked. The tweeted message stated that any bitcoin dispatched to the hyperlink in the tweet would be dispatched returned doubled[2]. Fake loan schemes have been another conventional tactic as scammers were given away with a minimum of $41.3 million in general between January and March 2020 [3]. Another example of a scam usually occurred when shortening the URL to share with multiple users or sites, including Twitter. They can also be used to mask a malicious net web page [4].

Many scams occurred in Malaysia in recent years, like A 45-year-old rubber tapper who took only 24 hours to lose RM184,999 of his savings to a Macau scam syndicate [5]. Besides that, a Pahang woman who decided to borrow RM15,000 after reading an Instagram advertisement is now RM11,060 poorer than that [6].

Some known indicators of scams are included below:

- Use an urgency such as, "Win the golden prize!"
- Giant income pledge within a short time frame.
- Overuse of Jargon and Buzzword.
- Demanding data statements or sensitive information.

Some scams rely to some degree on simple human attributes that everybody has. Many of those attributes are not inspiring enough. They include such traits as anxiety, greed, and covetousness. For several years, Con artists have leveraged these characteristics — playing on someone's greed and being able to persuade them up is down and cold is dry. The objective of the study is to build an online threat detector by analysing the text of social media. The study dwell on the fact that Scammers are increasingly using sophisticated methods to focus on customers, and everyone can become a victim. The experts interviewed did not identify any significant variations in the number of victims in terms of old or gender. Nonetheless, if they think it will maximise success, fraudsters in different demographic groups will likely focus on such a system. Trading standards within the UK, for example, estimate that online scams involving drugs are more likely to target young people, while middle-aged women may be targeted by scams involving dietary pills. The Canadian Public Interest Advocacy Centre said younger people appear to be at the forefront of cryptocurrency scams. In vulnerable situations, consumers may find it more difficult to make informed decisions or say no to high-pressure sales, which may increase their risk of fraud. Citizens Advice within the UK, for example, has found that older people, particularly those on social media, maybe more at risk of online fraud. So, this project can cope with the problems listed above.

To build an online threat detector by analysing the text of social Islam encourages human beings to do potential discussion and spreading useful knowledge. Besides that, it additionally encourages to make an environment or social system where people can share beneficial things, and others can be availed in a good way from there. However, one of the most important sins is if all people take advantage of the system that may be harmful to humans.

According to the holy Quran
يَا أَيُّهَا الَّذِينَ آمَنُوا إِن جَاءَكُمْ فَاسِقٌ بِنَبَإٍ فَتَبَيَّنُوا أَن تُصِيبُوا قَوْمًا بِجَهَالَةٍ فَتُصْبِحُوا
عَلَىٰ مَا فَعَلْتُمْ نَادِمِينَ

"O you who have true faith! If one who publicly and openly commits sins brings you any news (concerning another person), then ascertain its truthfulness carefully (before you spread it) lest you harm people through (your own) ignorance (through accepting and following false reports) and then regret what you have done" [49:6] [7].

According to hadith
عَنْ عَبْدِ اللَّهِ بْنِ مَسْعُودٍ قَالَ قَالَ رَسُولُ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ أَلَا أُنَبِّئُكُمْ مَا الْعَضْهُ
هِيَ النَّمِيمَةُ الْقَالَةُ بَيْنَ النَّاسِ

"Abdullah ibn Mas'ud reported: The Messenger of Allah, peace, and blessings be upon him, said, "Shall I not tell you about calumny? It is gossip that is spread among people." [Ṣaḥīḥ Muslim 2606]

Islam has inspired Muslims no longer to simply accept any records that involve them from anywhere which they don't know any longer understand the real deliver even as Muslims are handiest supposed to simply accept the records from a completely actual supply or the folks who fear Allah[8].

II. Related Works

Person-to-person communication sites are turning out to be famous step by step. As much as Social Networking Sites (SNS) facilitate the life of people, it likewise offers to emerge to different issues looked by clients utilising these systems administration destinations. There are various organising destinations like Twitter, Facebook, and Linked In, and so forth. The severe issue looked at by clients utilising these different organising locales is spam. Spam is any inappropriate or disallowed behaviour that legitimately or roundabout violates any program administration site's other standards. These papers concentrate mainly on spam identification using multi-layer perceptron learning on Twitter. Twitter has faced confrontations to date.

A research was conducted to provide an automatic detection system for identifying and removing spam from social networking sites by Singh et al. in 2019. Researchers used a Twitter dataset and developed a prediction model using Artificial Bee Colony (ABC), Artificial Neural network (ANN), naïve Bayes, and Support vector machine (SVM.). The accuracy of the proposed system to detect spam in the Twitter site of about 99.14% has been achieved. (ABC with ANN) has performed well compared to (Naïve Bayes and SVM) [9].

In 2018, Mokhsin et al. also researched to investigate what possibly be the cause of online shopping scams to occur in the social media environment. They had a total of 201 respondents with selected features. They used Descriptive Analysis (DA), Reliability Analysis (RA), and Log Regression Analysis (LRA). The researchers compared the accuracy level of Descriptive Analysis, Reliability Analysis, and Log Regression Analysis. The research result was 0.696, which is closer to 1. Hence, the respondent has a high chance of getting scammed [10].

In 2015, Dhingra et al. attempted to compare the various networking sites of spam that directly or indirectly violate certain rules of any networking site. They have collected tweets live through the API created. According to the researcher, more than 50.0% of data were missing, and they filled it with average values using forward-filling and back-filling. They chose MLP, Special character removal, word separation, tokenisation, stop word removal, and stemming. The accuracy has turned out 82 percent maximum [11].

Another research was conducted to build technology to detect and summarise an overall sentiment by Agarwal et al. in 2011. Researchers employed Support Vector Machines (SVM) and report averaged 5-fold cross-validation test results. The proposed system Model Avg. Accuracy & Std. Dev. (%). Unigram 71.35, 1.95. Senti-features 71.27, 0.65. Kernel 73.93, 1.50. Unigram + Senti-features 75.39, 1.29. Kernel + Senti-features 74.61, 1.43 accordingly. Also, they will explore even richer linguistic analysis, for example, parsing, semantic analysis, and topic modelling [12].

In 2009, Markines et al. researched the motivations of social spam automatic detection of spammers in a social tagging system. Researchers used a spam dataset. The authors used the SVM, SMO, and AdaBoost prediction models. They achieved over 98% accuracy in detecting social spammers with 2% false positives [13].

III. Methodology

Scammers have been around us since the beginning of electronic communication adapted through the development of technology. The scam is a serious issue, and it has been studied for a long time to detect scams. Scammers are turning into the fast-growing in these platforms. Therefore, in that section, the process of detecting scams on social media and proposed a system on online supervised classification method.

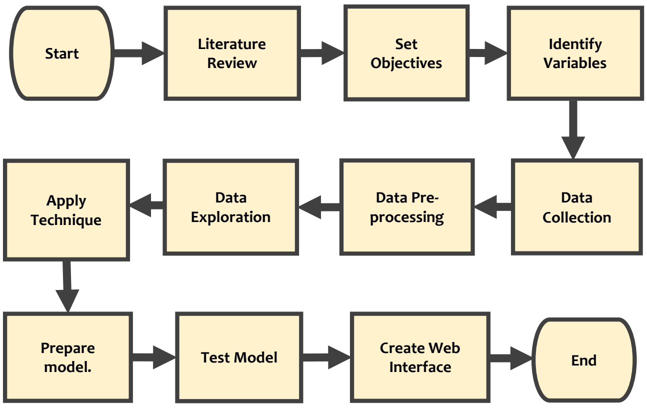


Fig. 1 Diagram of workflow

This diagram shows that the present project paper's workflow and the diagram details have been discussed.

A. *Data Collection*

For threat detection, the dataset named "Social Honeypot dataset" was collected from http://infolab.tamu.edu/data/. They have been collected these datasets for seven months. There are 41,499 users and 5,613,166 tweets in the Social Honeypot dataset. This dataset contains 22,223 spammers and 19,276 legitimate users. They provide a Twitter social honeypot dataset collected between December 30, 2009, and August 2, 2010. The data was used in the ICWSM 2011 paper, Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. Twenty-two thousand two hundred twenty-three content polluters, their number of followers over time, 2,353,473 tweets, 19,276 legitimate users, their number of followers over time, and 3,259,693 tweets are included in them the dataset [14].

B. *Proposed Model*

Naïve Bayes is a classifier of excessive bias, low-variance, and it can create an excellent version despite a limited statistical collection. It is easy to use and cheaper in terms of computing. Typical use cases involve the categorisation of text, consisting of unsolicited identification of mail, the examination of sentiment, and systems of recommendation.

Naïve Bayes' basic principle is that the price of any attribute is unbiased from the cost of any other attribute, given the price of the mark (the elegance). "This assumption, strictly speaking, is not often accurate (it's "Naïve"!), but it indicates that the Naïve Bayes classifier often works nicely. The presumption of freedom significantly simplifies the calculations to create the opportunity model for Naïve Bayes.

To complete the version of the possibility, it is very important to make a few assumptions about the conditional distributions of probability for the individual attributes despite the beauty. In order to model the attribute information, this operator utilises Gaussian opportunity densities.

The Twitter data set contains a large volume of data corresponding to two different types of Twitter data: Polluted and legitimate.

In these projects, a predictive model for detecting threats using the Twitter dataset. Based on the characteristics of Twitter data, one is legitimate tweets for the normal message and another one is polluted tweets which are a scam. The output of this model is displayed in three steps:

1) The whole Twitter dataset will be displayed.
2) A subset of the Twitter dataset, along with the Naïve Bayes-based projections, will be shown.
3) It will present a confusion matrix indicating that the forecasts are strongly correlated with the dataset.

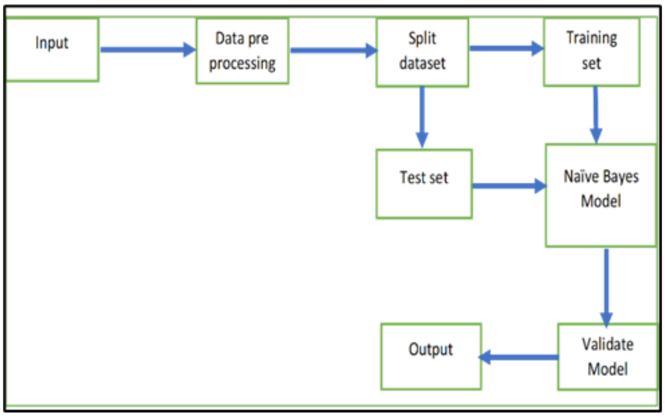


Fig. 2 Model

The original data set is split into two parts by the Operator Split Data: one is used to train Naïve Bayes, and the other to test the model. The result shows that a good fit for the Iris data set can be created by this simple model.

C. *Model Selection*

There are four types of machine learning algorithms: supervised, semi-supervised, unsupervised, and reinforcement. In supervised machine learning has three techniques which are Classification, Regression and Forecasting[15].

The purpose of this model is to detect or identify the threat from the text. Since the classification is generally used for fraud detection [15], therefore, Naïve Bayes (NB) under classification technique is used for this study.

Under the tree of classification, the most used techniques for threat detection are:

1) Support Vector Machine (SVM):

Support Vector Machine algorithms are supervised learning models for classification and regression analysis that analyse data. They essentially filter data into categories by providing a set of training examples, each of which is labelled as belonging to one of the two categories. The algorithm then goes to work creating a model that assigns new values to one of the two categories[15].

2) Naïve Bayes (NB):

The Naïve Bayes classifier uses Bayes' theorem to classify all values as independent of each other. It allows us to use probability to predict a class/category based on a set of features. Despite its simplicity, the classifier performs admirably and is frequently used because it outperforms more complex classification methods[15].

Based on the above descriptions, Naïve Bayes (NB) and Support Vector Machine (SVM) both are appropriate for this study. However, past studies reported the highest accuracy rate for Naïve Bayes (NB) compare to other techniques [10],

[11]. Therefore, this study has employed Naïve Bayes (NB) technique to detect threats from Twitter.

IV. Project Implementations

To build the whole idea, the threat detector is a system that uses Python. The programming language is used to build the back end of the framework in Python. The web programming languages Flask, HTML, CSS, and JavaScript have been used for front-end or GUI design. Besides, app.py is used to transfer the front-end data to the system model.

A. *System Analysis*

The Multinomial Naïve Bayes model was used in the system. Initially, 20,000 data went to experiment, and after pre-processing, it was split into a train and test where 16,000 were trained, and 4,000 were tested data.

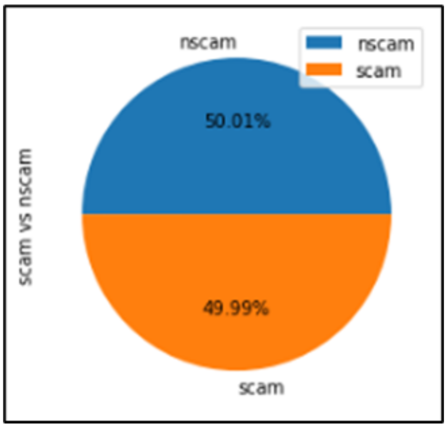
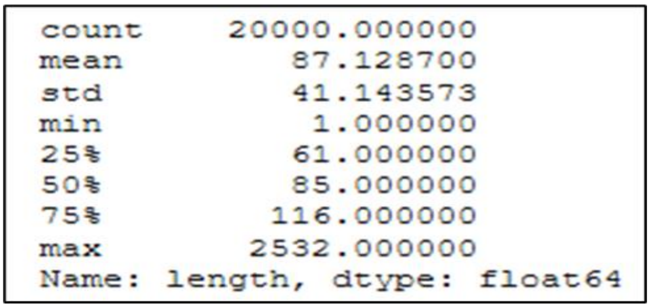


Fig. 3 Data visualisation

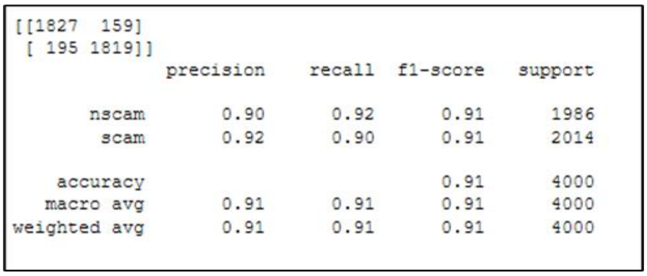
The legal data was 50.01% called "nscam" while the polluted data was 49.99% called "scam".



```
count     20000.000000
mean         87.128700
std          41.143573
min           1.000000
25%          61.000000
50%          85.000000
75%         116.000000
max        2532.000000
Name: length, dtype: float64
```

Fig. 4 Data describe

The cumulative data was 20,000, and the minimum length was 1, with an overall length of 2532. on the other hand, 87.128 was the mean or average, and 41.143 was the standard deviation.



```
[[1827  159]
 [ 195 1819]]
              precision    recall  f1-score   support

       nscam       0.90      0.92      0.91      1986
        scam       0.92      0.90      0.91      2014

    accuracy                           0.91      4000
   macro avg       0.91      0.91      0.91      4000
weighted avg       0.91      0.91      0.91      4000
```

Fig. 5 Precision, recall, and f1 score

TF-IDF and counter vectorisation for analysis has been used for feature extraction. The outcome was similar to the counter vectoriser, so I used that in this part.

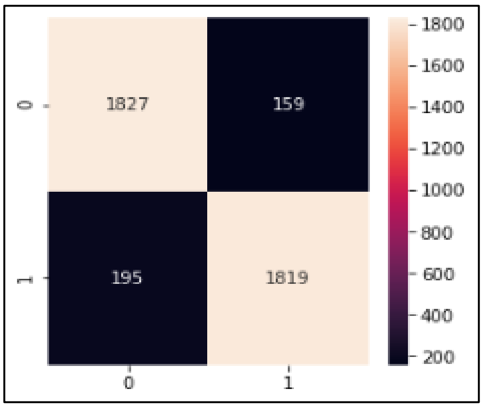


Fig. 6 Confusion matrix

It can be observed from fig. 6; this model has been detected 1827 non-scam tweets and 1819 scam tweets accurately, whereas 354 tweets are identified wrongly from 4000 tweets. The accuracy of the model is 91%.

B. *Web Interface*

This section will briefly discuss all testing procedures done throughout the system development. The findings from the user acceptance test are discussed, as are the possible improvements that can be made from the findings.

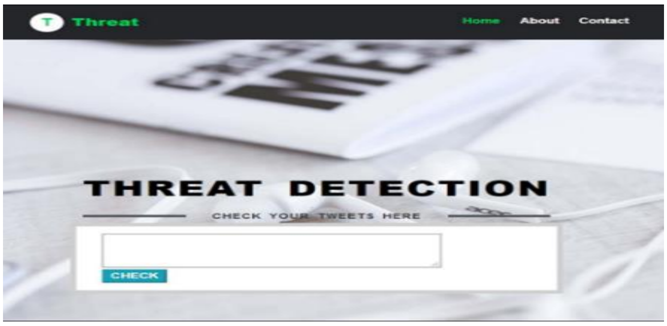
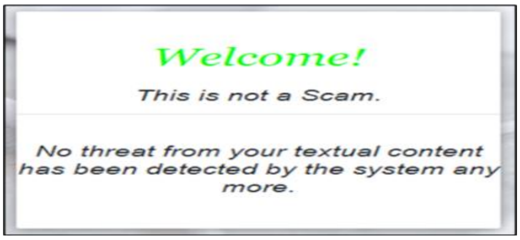


Fig. 7 Home page

This is the home page of the threat detector system. Users will see the text box to put their tweets to check the threat of that.

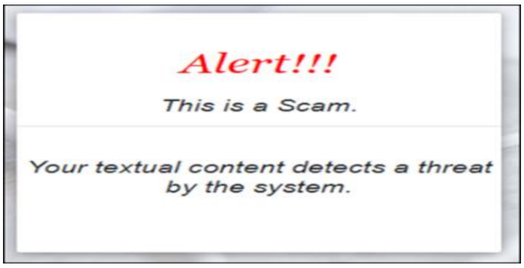


Welcome!

This is not a Scam.

No threat from your textual content has been detected by the system any more.

Fig. 8 Result of threat free text/ tweet



Alert!!!

This is a Scam.

Your textual content detects a threat by the system.

Fig. 9 Result of threat text/ tweet

In the output part, if the user's tweet is regular or normal, it will show the "This is not a Scam" refer to the fig. 8, while if the tweet is detected threat, it will show the text "Alert!!! This is Scam" output of fig. 9.

V. Conclusions

In all aspects of online existence, including personal communication, business, and economics, social media plays a major role. Scams on social media will spread rapidly across social networks such as Twitter and Facebook. Social media scams are typically cunningly designed to appear genuine by a scammer, using official brand logos, composed of terms and conditions, and providing a link to enter user information. The scammers build social media profiles and pay for user timelines to get promoted scam messages to the user. Scammers seek to take advantage of social media ads' credibility, which has evolved to be used on social media to see and trust deals from actual advertisers. Moreover, clicking on these links, unknown to the victim, sends personal information to third parties while activating the user's links' sharing function, often with an added status message. During this project, the developed threat detector can help social media users identifying threats' messages. The Twitter data set contains a large volume of data corresponding to two different types of Twitter data: Polluted and legitimate. For training and testing the model, the dataset named "Social Honeypot dataset" was accumulated. UserID, tTweetID, tTweet, and tCreatedAt were significant for threat detection in the feature selection process. Data analytics, and the application of artificial intelligence (AI), offer the ability to learn and enhance processes without being programmed but automatically from experience. This threat detector used the Naïve Bayes algorithm, a machine learning technique, to detect the threat with a 99.0% accuracy rate. Social networking sites need to make it easy to report threat scams and other negative content for users.

References

[1] Consumer International, "Social Media Scams : Understanding the Consumer Experience to Create a Safer Digital World," *Consum. Int.*, no. May, pp. 1–25, 2019.

[2] P. Dutta, "Biggest Twitter Breach: Accounts of US High-profiles Hacked in Bitcoin Scam," *Kratikal Blogs*, 2020. https://www.kratikal.com/blog/biggest-twitter-breach-accounts-of-us-high-profiles-hacked-in-bitcoin-scam/ (accessed October 11, 2020).

[3] C. Wong, "Scam victims lost $41.3 million in Q1; e-commerce and loan scams among most common," *The Straits Times*, Singapore, May 04, 2020.

[4] Joan Goodchild, "URL-Shortening Cons," 2018. .

[5] T.N.Alagesh, "Rubber tapper loses RM185k in 24 hours in Macau scam," *New Straits Times*, Kuala Lumpur, November 17, 2019.

[6] T. Lokman, "Woman loses RM11,060 in online loan scam," *New Straits Times*, Kuala Lumpur, October 01, 2019.

[7] J. Subhani, "The Sin of Making up Rumours," *Al-Islam*, 2003. https://www.al-islam.org/islamic-moral-system-commentary-surah-hujurat-jafar-subhani/sin-making-rumours (accessed November 10, 2020).

[8] Ismail, "THE VERDICT OF ISLAM ON SPREADING RUMORS OR FALSE GOSSIP," *The Siasat Daily*, July 31, 2018.

[9] A. P. Singh and M. Dutta, "Spam Detection in Social Networking Sites using Artificial Intelligence Technique," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 8S3, pp. 2278–3075, 2019.

[10] M. Mokhsin, A. A. Aziz, A. S. Zainol, N. Humaidi, and N. A. A. Zaini, "Probability Model: Malaysian Consumer Online Shopping Behavior towards Online Shopping Scam," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 9, no. 1, pp. 746–755, Feb. 2019, doi: 10.6007/ijarbss/v9-i1/5478.

[11] A. Dhingra and S. Mittal, "Content Based Spam Classification in Twitter using MultiLayer Perceptron Learning," *Int. J. Latest Trends Eng. Technol.*, vol. 5, no. 4, pp. 9–19, 2015.

[12] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment Analysis of Twitter Data," in *Proceedings ofthe Workshop on Language in Social Media (LSM 2011)*, 2011, pp. 30–38, [Online]. Available: http://www.webconfs.com/stop-words.php.

[13] B. Markines, C. Cattuto, and F. Menczer, "Social Spam Detection," *AIRWeb*, vol. 09, pp. 1–8, 2009.

[14] K. Lee, B. D. Eoff, and J. Caverlee, "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2011, pp. 185–192, [Online]. Available: http://bit.ly/a8rMRH.

[15] K. Wakefield, "A guide to the types of machine learning algorithms and their applications," *SAS*, 2021. https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html.