Vol.2 Issue.1 June 2023 ISSN 2583 4355

Detection of Spam Tweets in Twitter

P.S.S.N Chandra

PGDM Student

International School of Management Excellence, Bengaluru

Chandrap.isme2123@gmail.com

Stuti Agarwal

Assistant Professor

International School of Management Excellence, Bengaluru

Stuti@isme.in

Abstract

Social media has become a ubiquitous communication platform, with billions of people sharing their thoughts, opinions, and experiences daily. However, with this massive volume of user-generated content, there is also an increasing amount of spam, which can be both annoying and potentially harmful to users. Spam messages can include fake news, scams, phishing attempts, and other unwanted content that can compromise the safety and security of social media users. To combat this problem, many social media platforms use automated spam detection systems to identify and filter out spam messages. Machine learning is a popular technique used to develop these systems, as it can automatically learn patterns in data and make accurate predictions based on those patterns. In this project, to distinguish between tweets that are spam and those that are not, an attempt has been made to train a machine learning model. This paper has used information of tweets collected from twitter platform using some random keywords. Later these tweets have been manually labeled as spam or non-spam using an open-source application. Some basic preprocessing of the data was done by removing irrelevant columns, handling missing values, cleaning, and standardizing text data. Post which TF-IDF was used for vectorization to text in terms of features. The final model not only considers the tweet but also the meta data of tweets for spam detection. The model performance was tested on various testing datasets to validate the model and suggest future directions for improving spam detection in social media.

Keywords: Machine Learning, Regression Modeling, Model Validation, Statistical modeling predictive analytics, spam detection, manual labelling.

Introduction

An Online Social Network (OSN) is a web based service that enables users to create a public or semi-public profile, which is used to communicate and listen to other users of the service in their connection. Users can access and browse both their own and other users' lists of connections on such platform for better connectivity, reach and networking. Currently some of the popular OSNs are Facebook, WhatsApp, Twitter, and Instagram. However, as a downside, the number of spammers has also expanded in tandem with the expansion of social networks. Spammers are users who utilize platforms to send out unwanted or harmful communications to others. Spamming has become a big challenge for these platforms from the point of both safety and experience.

Twitter is one of the very popular networking platforms that enables users to send 280-character microblogs known as tweets to its connection. (Beveridge, 2022) On an average 200 billion tweets are posted every year. Thanks to this huge growing trend, this Online Social Network has attracted many users alongside spammers. Web Attacks that have appeared on Twitter can Scam, Spam, Phishing etc., of which spam may be a sort of Platform Manipulation which is used not only for bulk messaging but also to harm safety. Platform Manipulation is defined as an activity meant to negatively damage people's Twitter experiences. This includes concerns that are unwelcome or frequently repeated. Spam can include malevolent automation as well as platform exploitation such as fraudulent accounts.

Spamming is the practice of continually sending unwanted messages to many receivers for the purpose of commercial advertising, non-commercial preaching, or any illegal reason. It has been observed that most Spam Tweets contain shortened URLs to deceive users into clicking on them. They also prefer to tweet similar popular subjects to attract a wider audience because resources, such as tweets, may be shared. This type of Web Attacks not only disturbs the user experience but also spread incorrect or inappropriate content across the globe. To handle the consequences, twitter has a one-click solution to report a scam tweet on its platform. Such reports are constantly analysis and used to suspend the spammer’s account. (Kumar, 2022) reports that around 10% tweets are spam in nature, which is huge in volume when there are 500 million tweets daily (Beveridge, 2022) which is a major Problem today. Hence there is a need to find a machine-driven solution – a machine learning approach which can detect a possible scam and report the same.

Review of Literature

A survey based approach with three sections: spam detection, real-time spam detection, and spammer detection has been discussed by (Borse et al., 2022), where a thorough review has been done on usefulness and problems of current research work using various Twitter characteristics, leveraging the power of ensemble learning system.

A study by (Devi & Kumar, 2022) advises utilising statistically based features that are modelled using the boosting approach known as stochastic gradient boosting to assess twitter data sets in English. Simulation results are used to assess the performance of their suggested model.

The role of user profile in spam detection has been discussed by (Chowdhury et al., 2022), highlighting both profile data and content-based spam detection. The study makes three important contributions - First and foremost, extensive use of natural language processing (NLP) approaches. Second, the training of a unique cutting-edge hybrid machine learning model that was constructed entirely utilising a blend of machine learning (ML) and deep learning (DL) frameworks. Lastly, the classical logistic regression-based approach was used to derive at the probabilities of authenticity for a user.

Vijayalakshmi et al. (2022), has trained a Multinomial Naïve-Bayes Classifier for detection of spam using the tweet embedding as feature set. This social network spam has piqued the interest of many experts, who have proposed several ideas for spam categorization and identification. The major purpose of this endeavour is to develop a high precision Twitter spam detection system for accurately identify spammer, as well as to improve Twitter user security.

The concept of drift in context of twitter spam is has very well discussed by (R. Priyanka & Dr. Bhuvana, 2022). To address the difficulty in the drift, a system with a semi-supervised learning technique was proposed. The two step approach helped in understanding the domain structure to deal with the drift.

A Multi-Objective Genetic Algorithm and CNN-Based Deep Learning Architectural Scheme based approach was taken by (Rosita & Jenifer, 2022) for developing an effective and efficient spam detection engine. The application of such complex approach is unique for Twitter spam detection. This research discussing a rolling test train split by varying training ratio. The developed model was assessed in terms of multiple classification accuracy criteria like precision, recall and their Harmonic Mean popularly known as F1-Score.

One of the key challenges while working with such classification problem is unbalanced dataset. (Kumar, 2022) suggested that approximately 10% of the entire tweet base are spam leading to a problem of rare event. Although, Detection of Social Network Spam has been discussed extensively by many researchers. (Zhang Zhijie et al., 2020) proposed an Improved Extreme Machine Learning approach for spam detection based on features extracted not only by user profile and their activity but also the content and relationships. This unique and niche classification algorithm called the Improved Incremental Fuzzy-kernel-regularized Extreme Learning Machine (I2FELM) is designed based on the concept of regularised extreme learning machine. One of the key advantage of using this algorithm was its ability to handle and identify both balanced and unbalanced datasets, which is very effective and efficient for these type of problem.

Ensemble models are designed to handle the issue of unbalanced data. (Ainapure et al., 2022), has proposed a Deep Ensemble Model for twitter spam classification problem. This method dealt with the embedding of the tweet and the extracted the sentiment of it. The researcher suggested the usage of sentiment score to extraction sentiment and further use it to investigate the differences that exists between positives (Non Spam) and the negatives (Spam) in the data. An Optimised Deep Ensemble approach based on Neural Networks, Support Vector Machine, Random Forest and Convolutional Neural Networks is developed to accomplish this

The social honeypots concept was used in the framework of Recommendation Systems by (Elmendili et al., 2020) to detect spam tweets. It suggested identifying of spams using a unique security technique based on social honeypots. The research offered an approach based on content filtering to discover tweets those that are similar to spam tweets using the honeypot approach. The proposed method has substantially enhanced the classification Quality in terms of accuracy, performance and design. The algorithm is fast and easy to

Deploy in production which is a big plus from the point of implementation. The experimental findings suggest that the approach is robust as it has an optimal level of bias and variance.

Spam has been always thought of a binary problem. (L.Velammal, 2021), used the concept to detect spams in multiple phases. A combination of domain based classification, content based classes and tweet hash based classification has been done. The embedding for this frame work has been extracted leveraging the power of natural language process. The embedding acted as the feature set for the suite of classification models like Random Forest, Naïve Bayes and Logistic Regression. The popular bagging method of voting is used to assign labels to tweets obtained after the classification phase.

Objective

The objective is to develop a model using both tweet metadata and tweet content to identify probable spam tweets and validate the model on independent out of time samples to check for the robustness of the model

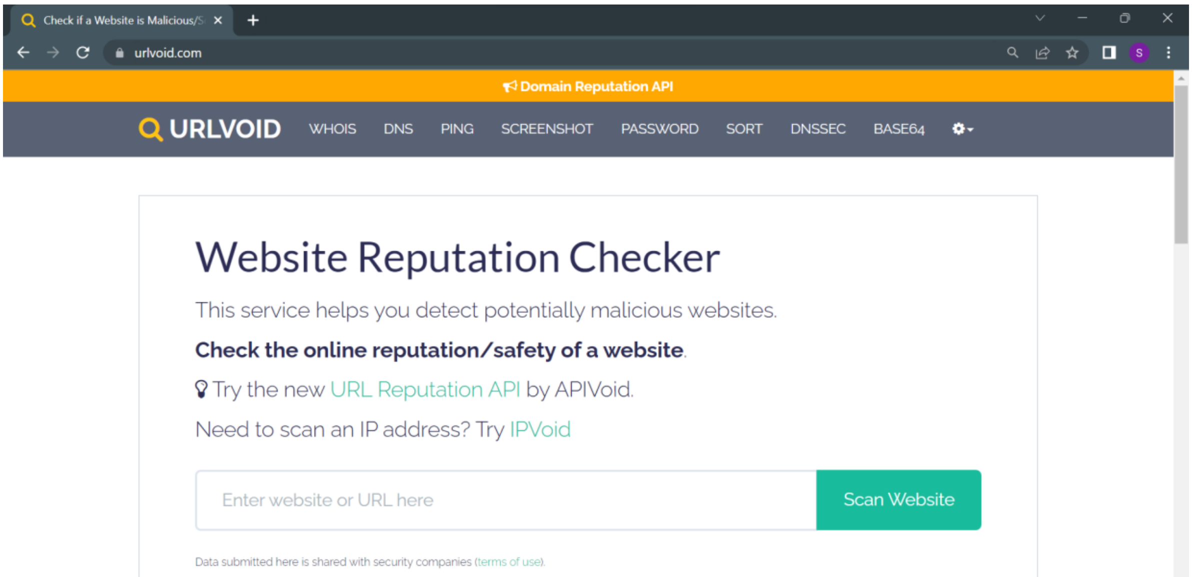
Methodology

As the objective is to use semi structured data of the tweets i.e. meta data of tweet along with the tweet content, the information first will be converted into a structured format leveraging the power of natural language processing and then fit a model on tweet embedding and tweet metadata. Once the model is developed, it will be validated on out of time sample for its robustness.

Data Analysis & Interpretation

- Data Collection: A sample of 1,000 tweets along with the metadata was extracted using the official API of twitter. To ensure a balanced data the tweets were collected using the following keywords and hashtags related to spam and bots: "spam", "bot", "scam", "fraud", "phishing", "clickbait", "malware", and "virus". The attributes considered for spam classification is not only tweet text but also other tweet information like Uniform Resource Locators (URLs) and retweet status, retweet count day, time etc. To develop a robust model a subset of 10% of the sample was manually classified as spam or ham using the website "urlvoid.com". The website runs on a pre-trained algorithm to detect spam based on URLS.

Figure I:



Website Reputation Checker

This service helps you detect potentially malicious websites.

Check the online reputation/safety of a website.

Try the new URL Reputation API by APIVoid.

Need to scan an IP address? Try IPVoid

Data submitted here is shared with security companies (terms of use).

Source: https://www.urlvoid.com/

- Data Preprocessing: As a first step, a basic exploratory data analysis was done based on which the irrelevant and sparse columns/ features were removed from the data. The tweets content being text data required necessary and standard pre-processing and cleaning for creation of embedding. The text was preprocessed by removing special characters, digits, and stop words. Stop words are commonly used words that are filtered out from text data during the preprocessing stage because they don't carry much meaning and may hinder the analysis. The natural language toolkit (NLTK) package in python was used to remove stop words from the text data. The stop words module from NLTK imports a pre-defined list of stop words, which was then used to remove stop words from the text data. The method used was the stop words. words ('English') method, which removes all English stop words from the text data. Removing stop words can improve the performance of natural language processing models by reducing noise and increasing the signal-to-noise ratio of the text data. Further, stemming was done to standardize the words in the text. Also, the emoji were converted into textual data.

- Feature Extraction: In the Feature Extraction step, tweet text was pre-processed and converted into a numeric matrix having a feature set using Term Frequency – Inverse Document Frequency (TF-IDF) technique. This technique calculates a numerical value for Each word in the text that reflects its importance in the tweet relative to the entire corpus. The

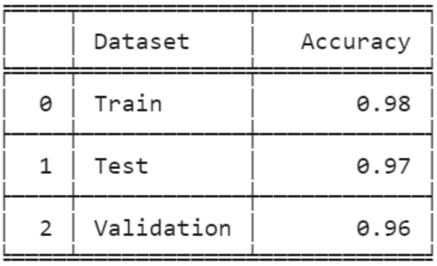
Resulting feature matrix had a vocabulary size of 5000, which means that only the top 5000 most important words in the corpus were included as features in the matrix. TF-IDF is a popular approach for feature extraction in text classification applications. The number of times a term appears in a tweet is referred to as Term Frequency (TF), whereas Inverse Document Frequency (IDF) is a measure of how frequent or unusual a word is over the whole corpus. Each word's TF-IDF value is derived as the product of its TF and IDF values. This strategy is useful for giving extra weight to terms that are relevant in a specific context but are uncommon in the overall corpus.

- Model Selection: Further the dataset was split into training and testing sets 80% being the training data and remaining 20% for testing. After multiple iteration a logistic regression model for spam classification was used. The logistic regression model was chosen for spam classification because it has been shown to perform well on text classification tasks especially in cases where the classes are linearly separable. Logistic regression is a binary classification approach that predicts the likelihood of an occurrence falling into one of many classes. It models the relationship between the input features and the target variable by estimating the coefficients of the linear equation using the maximum likelihood estimation method. The logistic function is then applied to the linear equation to produce a probability value between 0 and 1, which is then thresholder to predict the class label. In addition to its effectiveness in text classification tasks, logistic regression has several advantages that make it a suitable choice for this data analysis. Firstly, it is a simple and interpretable model that can be easily understood and explained to stakeholders. Secondly, it performs well on datasets with many features, as it is not affected by multicollinearity between the input features. Finally, it can deal with non-linear correlations between input characteristics and target variables by employing polynomial or interaction terms. Overall, the logistic regression model was a suitable choice for this data analysis due to its effectiveness in this text classification tasks.

- Model Evaluation: Based on the accuracy values presented in the **Table** , it can be observes that the model has performed well on all three datasets. The training dataset has the highest accuracy of 0.98, indicating that the model has learned the patterns in the training data very well. The test dataset has an accuracy of 0.97. This suggests that the model is generalizing Well to unseen data. The validation dataset has an accuracy in the same range. Overall, the

Model appears to be performing well and is likely to be useful for making predictions on new, unseen data.

Table I



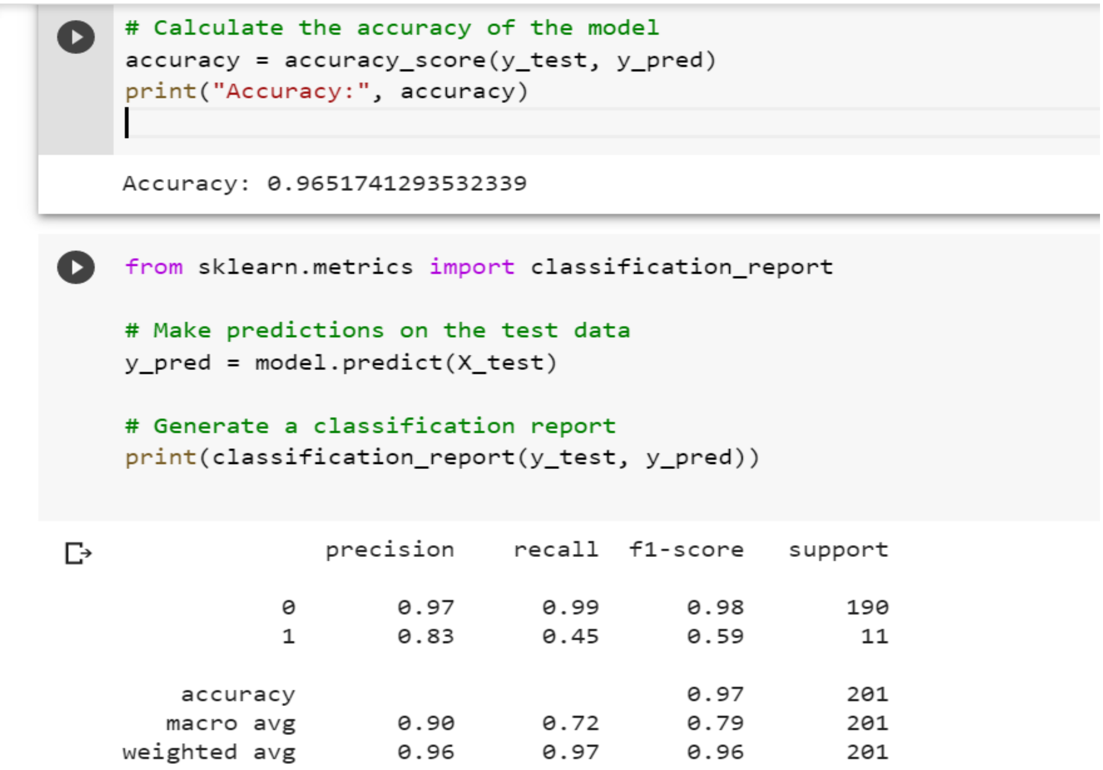
```
╒════╤════════════╤════════════╕
│    │ Dataset    │   Accuracy │
╞════╪════════════╪════════════╡
│  0 │ Train      │       0.98 │
├────┼────────────┼────────────┤
│  1 │ Test       │       0.97 │
├────┼────────────┼────────────┤
│  2 │ Validation │       0.96 │
╘════╧════════════╧════════════╛
```

Source: Primary Data

Measures like accuracy, precision, recall, and their harmonic mean, F1-score is used to assess the model's performance on the testing set. As shown in the

Figure I model has a 96.5% accuracy rate. The confusion matrix for the spam classification helps in computing the model's precision, recall, and f1-score for each class (0 – Ham and 1-Spam), as well as the overall accuracy and macro/micro-averaged precision, recall, and f1-score for each class. The accuracy for ham is 0.97, which implies that 97% of tweets predicted as ham were truly the ham cases. The recall is 0.99, indicating that 99% of the actual spam tweets were properly predicted as such. The overall F1 Score is 0.98. The accuracy for spam tweets is 0.83, which implies that 83% of tweets identified as spam were truly spam tweets. The model's overall accuracy is 0.97 implying that 97% of the tweets where correctly classified as spam or ham. The macro-averaged precision, recall, and f1-score are computed as the weighted average of the precision, recall, and f1-score for each class. The accuracy is 0.90, the recall is 0.72, and the f1-score is 0.79 on a macro-averaged basis. The weighted average precision, recall, and f1-score are determined by taking the average of the precision, recall, and f1-score for each class and multiplying it by the number of samples in each class. The average weighted precision is 0.96, the average weighted recall is 0.97, and the average weighted f1-score is 0.96. The classification report showed that the model had a high precision and recall for the ham class (0), but a lower precision and recall for the spam class (1). This suggests that the model may need further tuning to improve its performance on the spam class.

Figure I



```
# Calculate the accuracy of the model
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

Accuracy: 0.9651741293532339
```

```
from sklearn.metrics import classification_report

# Make predictions on the test data
y_pred = model.predict(X_test)

# Generate a classification report
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       190
           1       0.83      0.45      0.59        11

    accuracy                           0.97       201
   macro avg       0.90      0.72      0.79       201
weighted avg       0.96      0.97      0.96       201
```

Source: Primary Data

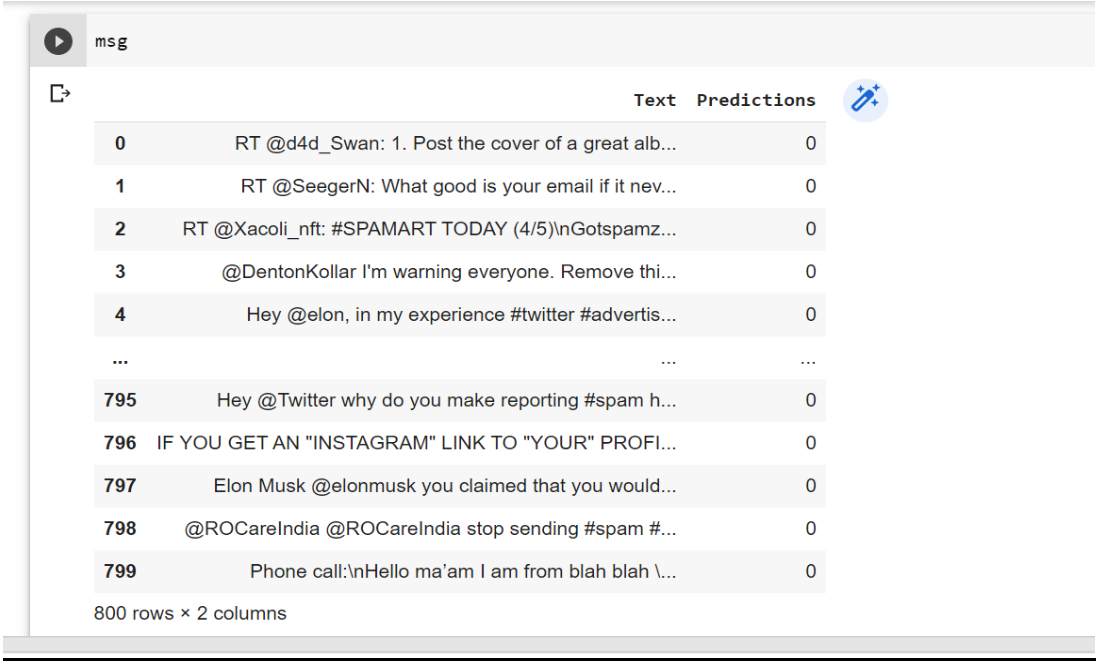
Out of sample Data Analysis: A new dataset of 800 tweets was further created using the Twitter API to validate the model on a real dataset. The same pre-processing was done and the model was deployed on the new dataset to predict if a tweet is spam or ham.

Figure II shows the output of the model:

Number of spam messages: 13

Number of ham messages: 787

Figure II



```
msg
```

	Text	Predictions
0	RT @d4d_Swan: 1. Post the cover of a great alb...	0
1	RT @SeegerN: What good is your email if it nev...	0
2	RT @Xacoli_nft: #SPAMART TODAY (4/5)\nGotspamz...	0
3	@DentonKollar I'm warning everyone. Remove thi...	0
4	Hey @elon, in my experience #twitter #advertis...	0
...	...	...
795	Hey @Twitter why do you make reporting #spam h...	0
796	IF YOU GET AN "INSTAGRAM" LINK TO "YOUR" PROFI...	0
797	Elon Musk @elonmusk you claimed that you would...	0
798	@ROCareIndia @ROCareIndia stop sending #spam #...	0
799	Phone call:\nHello ma'am I am from blah blah \...	0

800 rows × 2 columns

Source: Primary Data

Findings and Conclusion

Based on the out of sample dataset, it can be concluded that logistic regression model using TF-IDF feature extraction technique performed well for spam classification of tweets. The model had an accuracy of 96.5% on the test set, indicating that it can effectively distinguish between spam and ham messages. However, the model had a lower precision and recall for the spam class, suggesting that it may need further tuning to improve its performance on identifying spam messages. It may be necessary to explore other feature extraction techniques or try different models to improve the spam classification accuracy.

The analysis also revealed that the new dataset of 800 tweets had a relatively low proportion of spam messages, with only 13 out of 800 tweets being classified as spam. This suggests that spam activity on Twitter may have decreased, or that spammers are using more sophisticated techniques that are not easily detected by the current methodology. Overall, the project demonstrates the importance of using machine learning techniques for spam detection along with usefulness of multi-level verification of tweets and the need for continuous adaptation and improvement to stay ahead of spammers' evolving tactics.

Limitations

i. The Limited labeled data: Manually labeling the data during URL Void was time consuming, hence a sample dataset was used to train the model. Obtaining a large amount of accurately labeled data for training a machine learning model can lead to more accurate model.

ii. Evolving spam techniques: Spammers constantly change their methods to avoid detection, making it difficult to keep up with new spamming techniques.

iii. Language barriers: Spam messages can be written in different languages, and machine learning models may not perform well when dealing with languages they were not trained on.

iv. URL obfuscation: Spammers often use obfuscated or complicate URLs to bypass spam filters, and identifying these URLs can be a challenge

Scope for Future Study

The dataset used in this project was limited to tweets related to spam and bots. Future work could explore different types of text data, such as emails or text messages, to see how well the model performs on those types of data. Also the model could be integrated into a real-time spam detection system to identify and filter out spam messages as they are received. This could be useful for social media platforms or email providers that want to protect their users from unwanted messages. Future analysis will concentrate on the link between accounts and their proclivity to post spam tweets. While labelling a tweet as spam, probability of being spammer can also be considered. Like account with very high number following as compared to followers can be an indication of being a spammer. Since, the majority of spam tweets are neutral in terms of sentiment one can also learn how to determine the emotions of spam tweets. To conclude with there are multiple areas of learning and improvement in spam detection to ensure a spam free and secure web experience.

Reference

Ainapure, Bharati, &. B., Mythili, &. K., & Chandra, &. J. (2022). Deep Ensemble Model for Spam Classification in Twitter via Sentiment Extraction: Bio-Inspiration-Based Classification Model. International Journal of Image and Graphics.

Beveridge, C. (2022, March 22). Hootsuite. Retrieved from blog.hootsuite: https://blog.hootsuite.com/twitter-statistics/

Borse, Dipalee, &. B., & Swati. (2022). State of the Art on Twitter Spam Detection. In Applied Computational Technologies (pp. 486-496).

Chowdhury, Ratul, &. D., Kumar, &. S., Banani, &. B., & Samir. (2020). A Method Based on NLP for Twitter Spam Detection.

Devi, K., & Kumar, G. (2022, January). Stochastic Gradient Boosting Model for Twitter Spam Detection. Computer Systems Science and Engineering, pp. 849-859.

Elmendili, Fatna, &. E., & Younes. (2020, October). A Framework for Spam Detection in Twitter Based on Recommendation System. International Journal of Intelligent Engineering and Systems.

Kumar, S. (2022, June 08). Global Data. Retrieved from Business Fundamentals: https://www.globaldata.com/media/business-fundamentals/10-twitters-active-accounts-posting-spam-content-says-globaldata/

L.Velammal, &. N. (2021, July). Improvised Spam Detection in Twitter Data Using Lightweight Detectors and Classifiers. International Journal of Web-Based Learning and Teaching Technologies, pp. 12-32.

Martin, M. (2023, March). 29 Twitter Stats That Matter to Marketers in 2023. Retrieved from Hootsuite: https://blog.hootsuite.com/twitter-statistics/

R. Priyanka, & Dr. Bhuvana. (2022). A Semi-Supervised Learning Approach for Tackling Twitter Spam Drift. International Journal for Research in Applied Science and Engineering Technology, 1337-1340.

Rosita, & Jenifer, &. J. (2022). Multi-Objective Genetic Algorithm and CNN-Based Deep Learning Architectural Scheme for effective spam detection.

Vijayalakshmi, K., K Tanvi, L Rao, & Y Sindhu. (2022, July). spam detection in twitter using multinomial navie bayes classifier.

Zhang Zhijie, &. H., Rui, &. Y., & Jin. (2020). Detection of Social Network Spam Based on Improved Extreme Learning Machine.

Acknowledgement

'This article was presented at 11th International Conference on Contemporary Issues in Management organized by International School of Management Excellence, Bangalore held on 24th and 25th February 2023’