RESEARCH ARTICLE

A Sentiment Analysis-Based Model for Spam Filtering in Social Networks

Mohammad Mirahmadi, Mahboubeh Shamsi[*] and Abdolreza Rasouli Kenari

[1]Faculty of Electrical and Computer Engineering, Department of Computer Science and Engineering, Qom University of Technology, Qom, Iran

[*]**Corresponding Author:** Mahboubeh Shamsi, Faculty of Electrical and Computer Engineering, Department of Computer Science and Engineering, Qom University of Technology, Qom, Iran, E-mail: Shamsi@qut.ac.ir

Citation: Mohammad Mirahmadi, Mahboubeh Shamsi, Abdolreza Rasouli Kenari (2024) A Sentiment Analysis-Based Model for Spam Filtering in Social Networks, J Electron and Telecommun Eng 2: 1-19

Abstract

Spam, is a persistent issue across various digital platforms. This paper presents a unique and innovative approach to spam filtering on social networks, harnessing the power of sentiment analysis. By integrating emotional features into the detection process, our model strives to elevate the precision of spam identification.

Often, we use defined features and appropriate machine learning algorithms to identify spam at the tweet level. However, relying solely on user account features for detection can have drawbacks. Therefore, the proposed method also incorporates emotional features, employing a combined approach to synergize the results of multiple methods and achieve more accurate outcomes. This study combined a feature-based model with a sentiment analysis-based model and machine learning algorithms. The proposed method's performance has improved due to the use of emotional features in combination with user-level features. The proposed selection of emotional features outperformed the classic user-based or tweet-based features studied in the baseline paper. The cited research reports a maximum accuracy of 93%. But, this research's artificial neural network estimated the accuracy of the proposed method to reach up to 98%. This high accuracy holds promise for significantly reducing the amount of spam on social networks and enhancing the user experience and trust.

Keywords: Spam; Sentiment Analysis; Tweet; Machine Learning; Artificial Neural Network

Introduction

Social networks are defined as people interacting with each other and sharing information, needs, activities, and thoughts. Today, online social networks are a popular tool for collaboration and communication and a vital part of our digital lives, attracting millions of Internet users. Online social networks such as Facebook, LinkedIn, Twitter, etc., are among the most popular of these applications. These networks, especially those with ordinary and non-commercial applications, are virtual places where people briefly introduce themselves and provide the opportunity for communication between themselves and like-minded people in various fields of interest. Social networks on the Internet will become even more critical in the future. These networks are becoming more and more popular every day. With social networks, people are no longer alone in finding like-minded people in various matters. Therefore, our research on sentiment analysis-based spam filtering in social networks is highly relevant and significant in this context.

As mentioned, the use of social networks has increased significantly these days. People with different educations, ages, genders, and occupations are members of these networks, and users publish millions of different content items daily. However, living in such an environment has its etiquette that must be paid special attention to avoid serious problems. Millions of people around the world use social networks to connect with friends, meet new people, network with colleagues, and more [1].

Currently, based on the received statistics, Twitter has over 600 million users, who send over 400 million tweets per day and perform over 16 million searches. One of the features of Twitter is the limitation to typing only 140 characters. In addition to adding text, sending videos, photos, and audio on Twitter is possible. By the end of August 2011, Twitter had been translated into 11 living languages of the world, but to increase the languages of this social network, a team of 200,000 people is currently translating it into other popular languages.

The Spam Problem in Social Networks

As mentioned, spam is now causing increasing problems for social networks. For example, research has shown that about 40% of Facebook user accounts and 8% of posts on this network are spam. With the increasing influx of spam on social networks, the success of real-time search and exploration tools depends on the ability to distinguish valuable posts from spam. Facebook and Twitter have provided users with several ways to report spam [1].

Spam Filtering Mechanisms

The security system used in these networks has two advantages in dealing with attackers: user feedback and global knowledge. User feedback itself includes two explicit and implicit methods. Explicit feedback includes marking the found item as spam or reporting the user in question, and implicit feedback includes deleting a post or rejecting a user's request. Both of these feedbacks are valuable and important in the defense issue.

In addition to user feedback, the system knows general patterns of normal and abnormal behavior based on identifying, clustering, and collecting anomaly features. The system generally uses these methods to detect and respond to spam posts.

Types of Posts in Social Networks

As mentioned, one of the most essential activities of users on social networks is sending and viewing posts. These posts can generally include text, URLs, photos, or videos. Unfortunately, spammers can also achieve their goals by sending posts, such as accessing users' personal information, spreading viruses, etc. So, posts can be classified into two categories: spam and non-spam.

With the growing popularity of social networks, spammers are also targeting this platform to spread their content. Twitter is one of the most popular social networks where users discuss various topics and interact with each other. Most spam filtering methods on Twitter focus on identifying spammers (people who publish spam) and blocking them. However, spammers can create new accounts and send new spam messages again. In 2013, Twitter was announced as one of the top ten websites on the list of most popular websites, and it was also given the title of Internet SMS (Top Sites, Alexa Internet, 2019).

Since 2018, Twitter has had over 321 million monthly active users (USA Today, 2013). This massive user base is an attractive place for spammers to prey on their victims. Although producing and distributing spam is very costly for spammers, prevention methods are much more expensive for hosting companies. According to the U.S. legislature, the cost of spam in the U.S. was an estimated

$13 billion in 2007, including decreased productivity, wasted equipment, and workforce (Spamlaws, 2013). The direct financial impacts of spam include overload on computer systems and network resources and waste of time and human resources. Additionally, spam has costs from several dimensions. This cost is of even greater importance in the case of a company like Twitter with millions of users. Therefore, there is a need for solid spam detection techniques at the tweet level. These techniques can prevent spam immediately.

Account-based Spam Detection Methods

Account-based spam detection methods are based on the features (or combinations of them) of the account. This method is standard in other social networks and distinguishes spam from non-spam accounts. The focus of this method is on user account information. For example, the number of followers and followees in regular accounts is much higher than in spam accounts. As another example, the lifespan of a spam account is significantly shorter than that of a standard account. Another essential feature is Reputation, which is different for spam and non-spam accounts. The Reputation feature in a spam-generating account is 100% or very low, while it is around 30% to 09% in a standard account. This factor is very effective in distinguishing spam accounts from non-spam accounts. Although this method has high detection power, there are also spam- generating accounts with many followers in exceptional cases; in this case, the algorithm makes mistakes. Usually, these methods are used in conjunction with other methods.

Chen et al. (2015) investigated six machine learning algorithms and achieved the best F-measure with Random Forest. They used features such as number of followers, number of followees, and account lifespan feature. One of the weaknesses of this method is that when a spam-generating account is closed, it creates a new account again. Also, spam generators can eventually deceive the detection methods by circumventing these features [2].

Lee et al. (2010) proposed a honeypot-based approach to detect spam in social networks. The features they consider for spam detection are Twitter account lifespan, average tweets per day, the ratio of number of followers to number of followees, percentage of mutual friends, ratio of number of URLs in 20 recently sent tweets, ratio of number of unique URLs in 20 recently sent tweets, the ratio of the number of usernames in 20 recently sent tweets, and the ratio of the number of unique usernames in 20 sent tweets [2].

Tweet-based Spam Detection Methods

Tweet-based spam detection methods are based on tweet features (or combinations). All account-based and graph-based methods have a significant problem. After the algorithm blocks the user account, the spam generator creates a new account and continues its activity. Therefore, recent research has focused on the content of the tweet text itself. In this method, after identifying a spam tweet, it is prevented from being published without considering the sender of the spam. Since spam uses similar destructive words and topics, tweets containing these words and topics can be spam. The detection techniques in this method are based on natural language processing.

N-gram-based features are also divided into three categories: Uni-gram, Bi- gram, and Tri-gram. Five classifications are applied to these features: Naive Bayes, KNN, SVM, Decision Tree, and Random Forest algorithms. According to this research, the results on both datasets give the best output with Random Forest and SVM algorithms.

Xiao and Liang [4] used a hybrid approach using machine learning algorithms to identify spam in the comments of YouTube videos [4]. The research by Zhang [5] evaluates machine learning-based methods for spam detection at the tweet level. Wu et al. (2018) reviewed spam detection methods for Twitter with a comparative analysis (Wu et al., 2018).

Twitter hashtag-centric spam data was created by Sedhai and Sun [6]. The authors collected 14 million tweets and named the data as 14HSpam. Sedhai and Sun [6] obtained a spam detection framework in their research. They used four lightweight identifiers to identify spam at the tweet level [6].

A deep learning-based method for spam detection is presented in the research by Alom [7]. This research uses two methods based on complex neural networks simultaneously. One complex neural network is responsible for classifying tweet text, and one uses metadata classification [7]. Also, Bazzaz et al. [8] used classification based on various feature selection analyses, content analysis, user analysis, tweet analysis, network analysis, and combined analysis to identify spam on Twitter [8].

In the research by Le and Mikolov [9], the tweet vector was constructed by combining the tweet document vector (obtained by modeling the paragraph vector). These combined vectors act as input features for machine learning algorithms (random forests and neural networks) [9].

The study by Madisetty and Desarkar [10] also used two n-gram features [10]. With Uni-grams and Bi-grams features, they compared the results of their proposed research model with the research by Wang [11]. Despite the higher execution time, the proposed method significantly improved spam detection results [11].

Blacklist-based methods are a subset of tweet-based methods that are very slow to protect users because there is a delay before malicious URLs are entered into the database. Like account-based features, tweet-based features are light enough for real-time spam detection, requiring immediate analysis.

According to the research by Grier, about 09% of clicks on spam URLs occur in the first two days, while it takes an average of about four days for a new URL to be blacklisted, which is a significant delay in blacklisting updates. During this time, spam spreads rapidly, which is a significant weakness of this method.

Much research has been done in this area. For example, Patil (2018) used decision trees and statistical features to identify malicious URLs. Some of their features include the length of the URL and the presence of an IP address in the Hostname [12].

Graph-based spam detection methods use graph data structures to model Twitter features as nodes and edges. Graph data models are a suitable solution for representing data, where information about the connection of data or topology is at least as important as the data itself. Therefore, graphs are commonly used by social networks like Facebook and Twitter, mostly built on users, topics, and two-way interactions.

This method extracts features based on the social graphs of Twitter users based on the relationships between followers and followees. There has been much research in this area of social networks. In graph-based methods, which are somewhat similar to account-based methods, each user account is considered a node, and the input degree of the node indicates the number of followers and the output degree indicates the number of followees. Neighborhood-based features also fall into this category. These features are used in machine learning classifiers.

Song et al. (2011) extract the distance and relationship between the sender and the tweet notes. While distance defines the shortest path length between the tweet sender and the mentioned items, connection specifies the strength of the relationship between users.

Unlike account-based and tweet-based features, manipulating graph-based features is difficult. Extracting these features requires deep analysis of the vast and complex Twitter graph, which takes time and resources. Therefore, graph-based features are not light enough for real-time spam detection.

Combined Spam Detection Methods

Combined spam detection methods combine the methods explained in the previous sections to provide more robust spam detection that can more comprehensively assess the possibility of spam.

Wang et al. propose a spam detection method based on account, tweet, natural language processing (NLP) [1] and sentiment features. Some of the unique features they use when detecting spam are profile name length, automatic or manual emotional vocabularies, number of exclamation marks, number of question marks, maximum word length, average word length, number of capital words, number of spaces, and part of speech tags (POS) [2] in each tweet [11].

Lee et al. present a social honeypot that can collect spam profiles from social networking communities. Each time an attacker tries to connect to the honeypot, an automated robot retrieves some observable features from malicious users, such as the number of friends. Then, this set is analyzed to create a spam profile and train the corresponding classifiers [3].

Table 1: Summarizes the research done on this topic

Approach	Features	Algorithm	Researchers
Account-based spam detection	Number of followers, number of following, account age	Various machine learning algorithms	Chen et al., 2015
Account-based spam detection	Account age,average daily tweets, follower- to-following ratio, percentage of mutual friends,etc.	Honeypot-based approach	Lee et al., 2010
Account-based spam detection	Follower-to- following ratio, number of tweets to account age, average time between posts, changes in posting time, maximumidle hours	Artificial neural network	Gee & Hakson, 2010
Tweet-based spam detection	Number of followers and followings, length of user profile name, length of profile description, user account age inhours	Hybrid methods based on user, tweet content, andN-gram	Wang et al., 2015
Tweet-based spam detection	Four lightweight identifiers for identifying spamat the tweet level	Genetic algorithm	Zhang et al, 2014; Wu et al, 2018
Tweet-based spam detection	Number offollowers and followings, length of user profilename	Convolutional neural network	Alom et al., 2020
Tweet-based spam detection	Combining tweet vector with tweet document vector (obtained by paragraph vectormodeling)	Artificial neural network	Le & Mikolov, 2014
Tweet-based spam detection	Using word embedding features of each tweet	Convolutional neural network	Madisetty & Desarkar, 2018
URL-based spam detection	URL length, presence of IP address in Hostname, etc.	Decision tree and statistical features for identifying malicious URLs	Patil & Patil, 2018

Graph-based spam detection	Graph density and average shortest path	Graph-based approach	Yang et al., 2013
Graph-based spam detection	Extracting distance and relationship between sender and tweet notes	Graph data structure	Song et al., 2011
Hybrid spam detection	Ratio of friend requests, total number of user tweets, similarity of tweets sent by the user, number of tweets sent by the user, number of user's friends	An approach based on account and tweet-based features	Stringhini et al, 2010
Hybrid spam detection	Number of mutual links, ratio of bidirectional links, centrality, clustering coefficient along with tweet and account-centric features like number of followers	A Twitter spam detection method based on a combination of graph, tweet, and account-based features	Yang et al., 2011
Hybrid spam detection	Profile name length, emotional lexicons automatically or manually, number of exclamation marks, number of question marks, maximum word length, average word length, number of capital words, number of part of speech tags whitespaces and	A spam detection method based on account, tweet, natural language processing	Wang et al, 2015; Tolosana et al, 2020; Majeed et al, 2020
Hybrid spam detection	Number of followers and followings, length of user profile name	A spam detection method based on account, tweet, natural language processing	Lee et al., 2010
Hybrid spam detection	Classification based on various feature selection analyses, content analysis, user analysis, tweet analysis, network analysis, and hybrid analysis	A hybrid approach of artificial neural network and support vector machine methods	Bazzaz et al, 2023
Hybrid spam detection	Sentiment Analysis Textual features of YouTube comments	Hybrid approach of random forest, artificial neural network, support vector machine, and decision tree methods	Xiao & Liang, 2024

This research will use several valid databases that were reviewed in many articles from 2015 to 2020. The following datasets are also used:

6.5 million tweets on MTV Lady Gaga 2017 and ICC Championship Trophy 2017 [13] and Creaci-2017 data set has been used for experiments. The main data set includes nearly 7 million tweets, and the process of collecting the data set was done in two months and finally Real world dataset.

As in previous researches, various features have been used to detect spam at the tweet level. Including the words used in tweets as a feature, the specific features of each user and user information as well as content-based features will be included in the system. But many spam producers use features related to the user's emotions in order to encourage the user to click and follow the links in the spam. In other words, they try to emotionally encourage the user to click on the spam links in the tweet. In this research, in addition to some features used in previous works, the emotional features used in the research of Alum et al. (2020) are also effectively added to create a stronger composite model.

A convolutional neural network model will be used to detect spam. Two other methods based on machine learning were also used. Logistic regression and simple Gaussian Bayesian are these two methods

For each published tweet, the indicators related to the content of the tweets will be analyzed. Some of these indicators include the number of retweets, likes, mentions, time interval between tweets, etc. NLTK library in Python is used for sentiment analysis.

The information processing process in this research was done with a computer system with 16 GB of RAM and 7-core CPU. For each of the processing steps, between twenty and thirty minutes were spent, and based on that, the results were reported in this article.

Sentiment Analysis

One of the new areas in social network text-based research is sentiment analysis. Sentiment analysis, also known as opinion mining or sentiment mining, is one of the most important subfields of natural language processing (NLP) and is widely used in data mining, web mining, and text mining. Sentiment analysis systems are used in almost every business and social field because opinions play a vital role in all human activities and are one of our most influential behaviors. Our beliefs, perceptions of reality, and choices are primarily conditioned by how others see and evaluate the world. Therefore, when we need to make decisions, we often look for the opinions of others. This is true not only for individuals but also for organizations.

Russell's research developed a two-dimensional circular model of descriptive effect, pleasantness/unpleasantness, and arousal (the degree of reactivity to a stimulus). The polar dimensions refer to the degree of positive or negative feeling, while the arousal dimension is related to the degree of calmness or excitement. The range of both dimensions is from 1 (entirely negative or calm) to 9 (wholly positive or excited). As a result, most research in sentiment analysis has focused on the pleasantness/unpleasantness factor.

Many studies have addressed the relationship between different emotions, and they are recommended for research on emotions in particular. Today, there is much research on sentiment analysis of text; therefore, there are many ready-made libraries for sentiment analysis. However, there has been less research on sentiment analysis in the field of spam detection on Twitter.

Adel Majeed and his colleagues (2020) focused on sentiment detection from Roman Urdu text. A comprehensive sentence dataset was collected from domains and annotated with six classes (happy, sad, angry, scared, love, and neutral). Word2Vec was used for feature extraction. For classification, various base algorithms, such as k-nearest neighbors, decision trees, support vector machines, and random forests, were applied to the dataset. After testing and evaluation, according to this study, the support vector machine model achieved better results than other classification algorithms, with an accuracy of 54.69%.

Gwala and Patel [14] presented a general Twitter spam detection framework. This framework, which is almost the same in all research, is shown in Figure 1 with minor changes and using a hybrid approach.

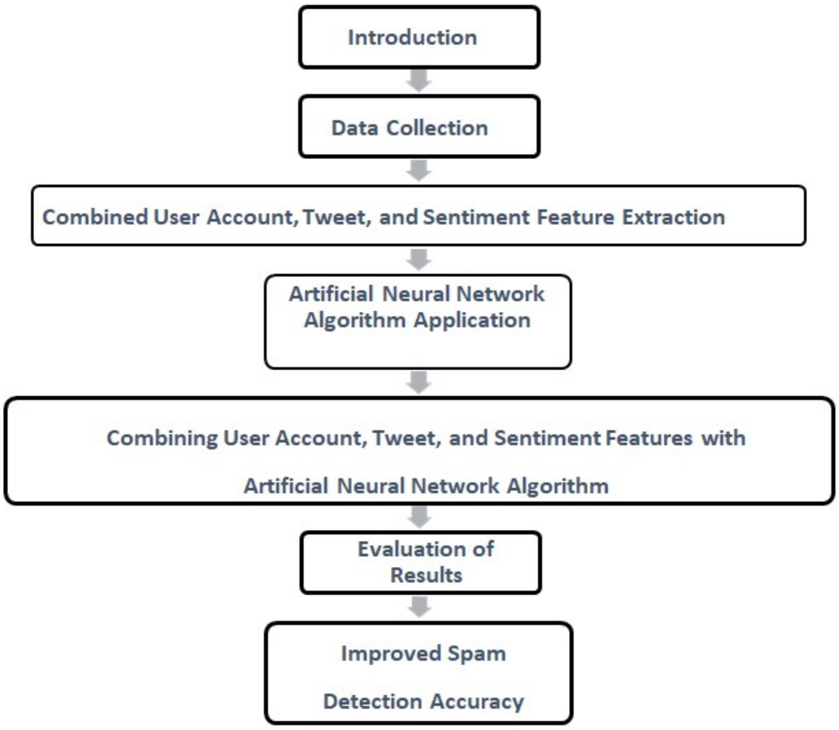

Introduction
Data Collection
Combined User Account, Tweet, and Sentiment Feature Extraction
Artificial Neural Network Algorithm Application
Combining User Account, Tweet, and Sentiment Features with Artificial Neural Network Algorithm
Evaluation of Results
Improved Spam Detection Accuracy

Figure 1: General Framework for Spam Detection using a Combined Machine Learning Approach

Materials and Methods

General Method

As in previous researches, various features have been used to detect spam at the tweet level. Including the words used in tweets as a feature, the specific features of each user and user information as well as content-based features will be included in the system. But many spam producers use features related to the user's emotions in order to encourage the user to click and follow the links in the spam. In other words, they try to emotionally encourage the user to click on spam links in the tweet. In this research, an attempt has been made to effectively add emotional features used in Alum et al.'s research (2020) in addition to some features used in previous works to create a stronger composite model.

A convolutional neural network model was used to detect spam. Two other methods based on machine learning were also used. Logistic regression and simple Gaussian Bayesian are these two methods.

Dataset

This study used the creaci-2017 dataset for experiments. The original dataset consists of 7 million tweets, and the collection process took two months.

This dataset belongs to the Twitter social network and has two main categories: legitimate users and bot users. Additionally, legitimate and spam tweets can be separated into two other parts of this data.

These two categories were combined in this research, and a part of this data was used for training. For this purpose, simple Python commands such as pandas, numpy, tqdm, glob, and matplotlib were used in Python to read the dataset. The dataset used in this research is relatively large regarding the number and volume of data; therefore, its results can be reliable. The volume of the received data in CSV format is 440 megabytes. For example, this dataset includes 3474 legitimate users, about 3 million tweets from these users, and 4912 spammers (probably bots) with about 3.5 million spam tweets in this dataset.

Forty-three raw features (features [2]) are defined for each user (including humans and bots).

This dataset includes the information of these users and tweets, all written in English and compressed in form, including ID, Name, screenname, followers, friends count, language, etc.

After merging all the data, there will be four general categories of data: legitimate users, tweets of legitimate users, spam users, and tweets of spam users. They merge these results to determine the number of legitimate and spam users. In total, 8386 users and 43 features will be the basis for the analysis.

Since the research focuses on using NLP and natural language processing, it must be used wherever text data is available. For example, for user analysis, in addition to examining the time of account creation and the number of followers, it is possible to analyze the description of each user since each user on Twitter has a description. For example, how are bots described? Moreover, how do real human users write their account descriptions? Even the text of the tweets should be analyzed. The popularity of each post (like), the number of responses and conversations related to each post (mention), and the number of hashtags used will likely be influential factors in identifying spammers. Text analysis of each post can also be practical. Combining numerical and text features can improve spam identification (spam posts).

In one part of this research, by examining the length of the username of each user (screen name), which is like a URL or ID, a new feature was created in the dataset to identify non-human users (bots). The name of this feature in the dataset is defined as user_name_length.

Another essential feature is when a user has been on a social media platform. Those new to Twitter are more likely to be identified as bots. For example, those who have created a new Twitter account may tweet less and are likelier to like other users' posts or re-tweet (retweet) a post. Identifying and defining some bot- like behaviors can better define the path to identifying bots and spam posts on this media. For example, having a profile picture for a user account will reduce the chances of being a bot. Therefore, several columns are added as new features in addition to the dataset's 43 raw columns (features). Simple mathematical calculations add columns containing essential features to the dataset as new features. For example, the account creation history in days (which is obtained by subtracting the current time from the account creation date). Some features, as mentioned, were added to the dataset using the review of previous texts and innovatively during the research (for example, the length of the screen_name string or the number of numbers used in the screen_name).

Using equation (1), the number of years of media activity was added to the data as a new feature.

Equation (1) [account_age]=all_data[account_age(days)]/365

Another feature is the result of dividing follower_ccount by account_age, named followers_growth_rate. This straightforward and exciting feature was used to identify bot users. The number of years a user has been on Twitter and how

many followers they have in those years can be crucial because usually, those who have a short account history but have a very high number of followers are probably either bots or celebrities who have suddenly entered Twitter and have had high account growth.

The following growth rate of an account is another essential feature calculated using the account_age/friends_count relationship.

Another innovative feature is the calculation of the popularity of a user account, which was added to the dataset under the name popularity and was calculated using the following relationship (the result of dividing the number of followers by the number of followers plus the number of friends):

[popularity]=all_data [followers _count]/all_data[followers _count]+ all_data[friends_count]

This factor tells us the level of popularity. For example, a user with ten thousand followers but only follows five people is probably very popular. The closer the defined factor (popularity) is to one, the more popular it is.

Another feature is location, which is checked to see if a user has defined the location on their account. As mentioned before, having a profile picture for a user account, considering it as a human indicator, and entering the account owner's location can also confirm this.

These features are defined at the user level. However, another essential part of identifying spam posts is reviewing the text of tweets. These are based on the characteristics of the account. First, a user is checked from this point of view; for example, their popularity is estimated, and then we will enter the phase of reviewing the tweets. The distinguishing feature of this research is the combination of users, profiles, and tweets, which means that both a user's profile and tweets are reviewed.

The number of unique_mentions, unique_URLs, and other features were used in the analysis. For example, if a person tweets repeatedly, they are likely to be a bot, but the time interval between tweets for humans is slightly longer. The number of unique_URLs used in a tweet is also essential. (Humans are likely to use fewer URLs when tweeting, but bots use more URLs in their tweets.) For example, human users use fewer hashtags in their posts. For example, when human users tweet, they usually get likes, retweets, replies, mentions, etc., but posts by bots are usually not liked, retweeted, or mentioned.

Some of the features used in this study are based on the research by Rodríguez-Ruiz et al. (2020), which are shown in Table (2) and Table (3).

Table 2: Features used in previous research (Rodriguez-Ruiz et al., 2020)

Description	Defined Feature
Ratio of retweets to tweets	retweets
Ratio of replies	replies
The ratio of favorite tweets to tweets	favoriteC
The ratio of hashtags to tweets	hashtag
The ratio of URLs to tweets	url
The ratio of mentions to tweets	mentions
Average seconds between posts	intertime
The ratio of friends to followers	ffration
Number of favorite tweets on this account	favorites
Number of listed tweets on the account	listed
The ratio of unique hashtags to tweets	uniqueHashtags
The ratio of unique mentions to tweets	uniquementions
The ratio of unique URLs to tweets	uniqueURL

Table 3: Features used in the present study (at tweet and user account level)

Description	Defined Feature
Ratio of retweets to tweets	retweets
Ratio of replies	replies

The ratio of favorite tweets to tweets	favoriteC
The ratio of hashtags to tweets	hashtag
The ratio of URLs to tweets	url
The ratio of mentions to tweets	mentions
Average seconds between posts	intertime
Number of tweets	Tweet_count
Length of account description	Description length
Popularity	Popularity
Ratio of followers to account age	Log_friends_growth_rate
Ratio of followers to account age	Log_followers_growth_rate
Account age (days)	Account_age
Length of username string	User name length
Number of digits used in screen name	Number of digits in screen name
Number of characters in the screen name	Screen name length
Number of saved tweets	Listed_count
Number of saved posts	Favorites count
Number of friends	Friends count
Number of followers	Followers count
Number of statuses	Statuses count
Number of favorite tweets on this account	favorites
Number of listed tweets on the account	listed
The ratio of unique hashtags to tweets	unique hashtags
The ratio of unique URLs to tweets	unique URL pen_spark

Another crucial indicator used in this research is sentiment analysis. For each tweet published, the researchers analyzed the indicators related to the content of the tweets. These indicators include the number of retweets, likes, mentions, time intervals between tweets, etc. The NLTK library in Python was used to analyze sentiment. The NLTK library [1] is one of Python's most comprehensive and oldest natural language processing libraries. This library is a foundation and standard for text-processing libraries and is excellent for research applications. One of the good features of this library is the ability to connect to different text corpora, which can be very useful in identifying spam posts. The output of this tool is an index called SA [2], which shows the polarity of a tweet, which can be one of the positive numbers, one, zero, or negative. For each user, the average of this index is calculated among all the tweets of that user. A positive one indicates that the tweet is identified as a non-spam tweet, a negative one means the tweet is identified as spam, and zero indicates indecision. The Vader module [3] exists in the NLTK tool, and this module was used in Python in this research based on the research results of Rodríguez-Ruiz et al. (2020).

Relationship between Features

At this stage, behavioral differences between real users and spammers were investigated using statistical tools. The results showed that many defining features at the user level are not very effective for identifying spammers due to the behavioral similarities of both groups of real users and spammers Figures (2) and (3), but some features such as popularity can be very effective in distinguishing bot users from real users Figure (4), (5) and (6).

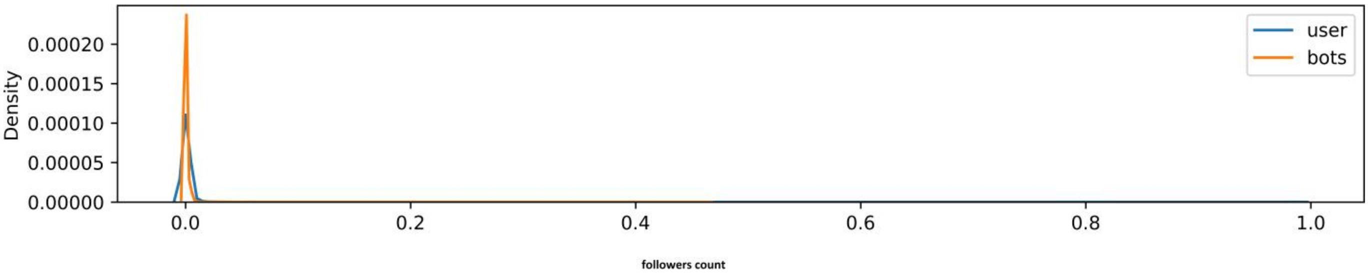


Figure 2: The graph of the amount of followers in two groups of bots (orange) and real users (blue)

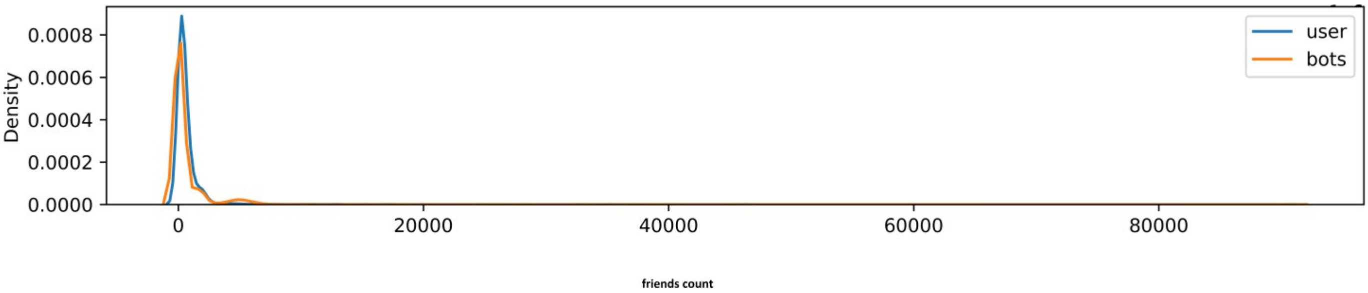


Figure 3: Number of followers

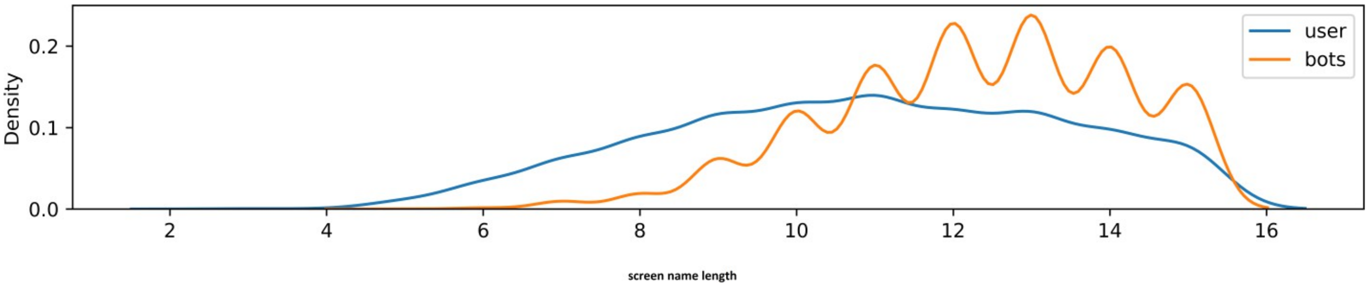


Figure 4: The number of characters in the username

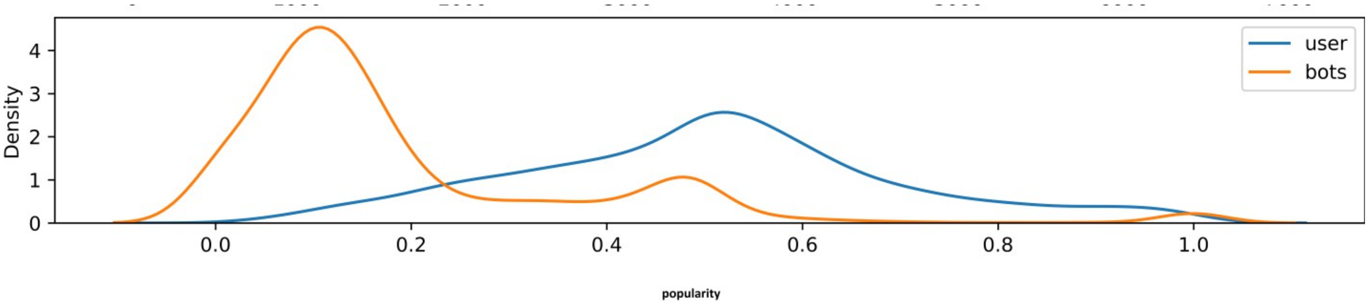


Figure 5: Popularity

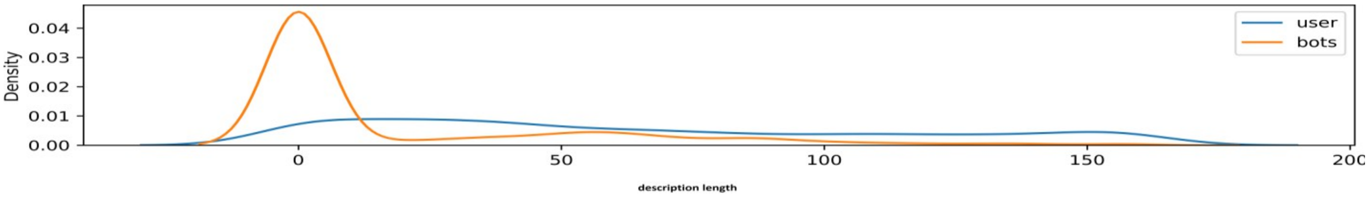


Figure 6: Number of characters describing the username

Figures (7), (8), and (9) show the correlation between each pair of features among bot users, real users, and all users. The correlation coefficient is calculated based on the Pearson correlation, which is the ratio of the covariance to the product of the standard deviations of each pair of features.

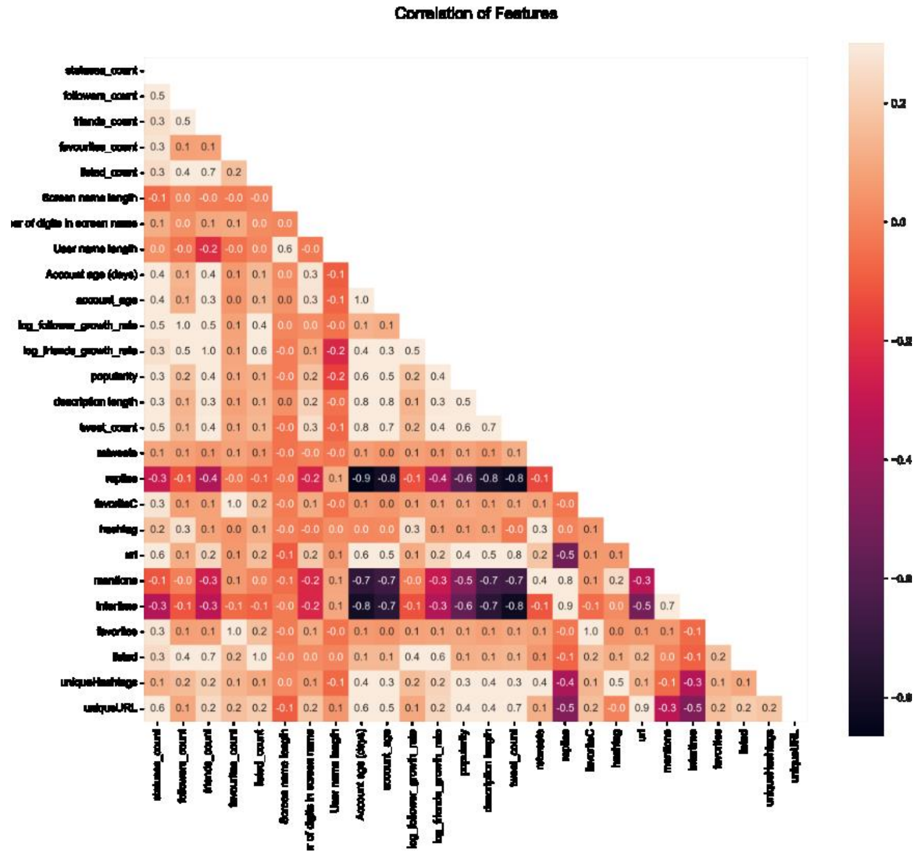


Figure 7: Correlation coefficients between different features among spam users

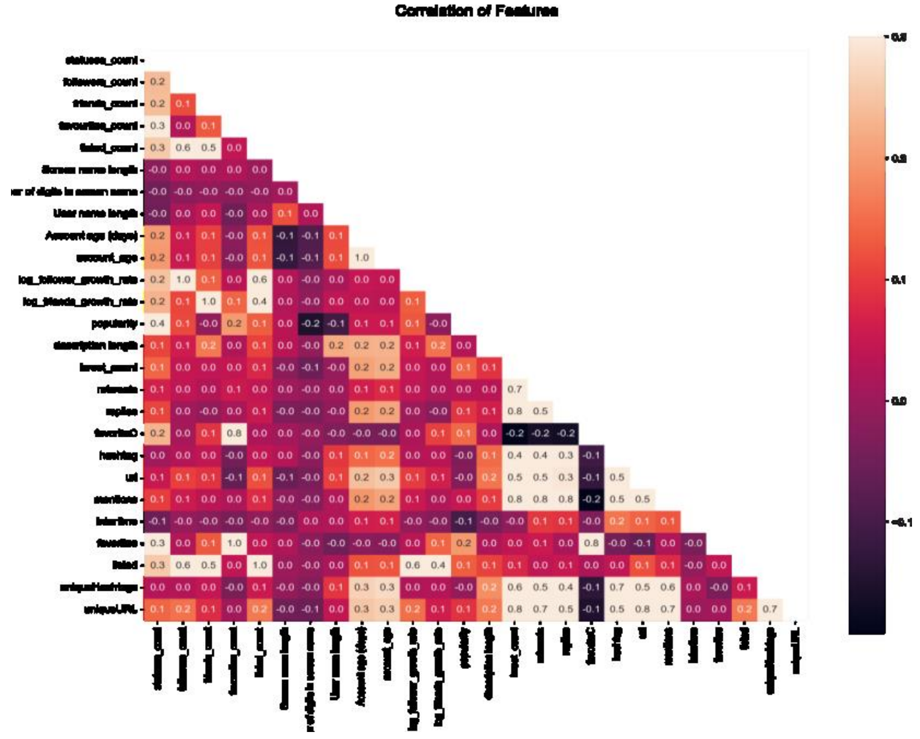


Figure 8: Correlation coefficients between different features among real users

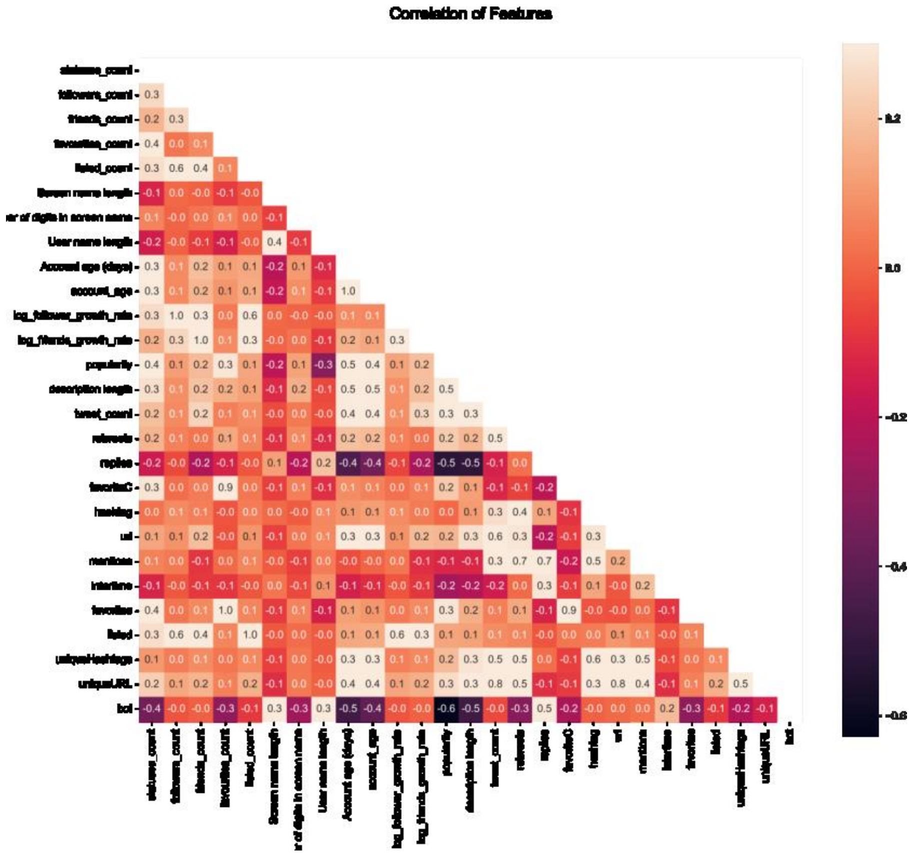


Figure 9: Correlation coefficients between different features among all users

Based on the results, there are correlations between the definitions of different features. For example, the correlation between the number of hashtags used and the number of tweets posted by bots is zero, while it is 0.4 for real users. The results of this analysis will be very useful in categorizing users and published posts. Based on trial and error, a threshold [1] of 0.2 was determined to separate meaningful correlations from meaningless ones. If the correlation coefficient is in the range (+0.2,+1) or (-1, -0.2), it will have a meaningful correlation.

Another way to represent data is also valuable for correlation analysis. For example, Figure (10) shows the difference in the ratio of followers to following between the robot and non-robot categories. Red dots represent the robot category, and blue dots represent the non-robot category.

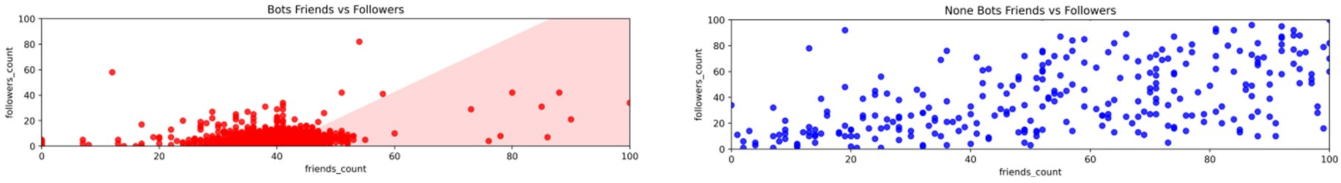


Figure 10: The difference in the ratio of follower to follower in two groups of robots and non-robots (red points are for the robot group and blue points are for the non-robot group)

Table 4: Comparison of classification results using three methods on the available data. In that order, artificial neural networks, naive Bayes, and logistic regression achieved the best classification performance. Precision refers to the percentage of relevant model predictions, while recall refers to the percentage of all predictions the model correctly classifies.

Table 4: Comparison of evaluation metrics for the three methods implemented in the study

Method	Accuracy	Recall	Precision	F1	Time (s)
Logistic Regression	0.879	0.878	0.877	0.878	0.0007
Naive Bayes Gaussian	0.93	0.931	0.934	0.931	0.0013
Artificial Neural Network	0.983	0.982	0.983	0.98	0.1765

Figure 11: The ROC curve of all three methods clearly shows that the performance of the artificial neural network-based model in classification is superior to the other two methods. The classification accuracy is over 98%, which indicates the effectiveness of the proposed method based on integrating user-level and tweet-level features and sentiment analysis using an artificial neural network.

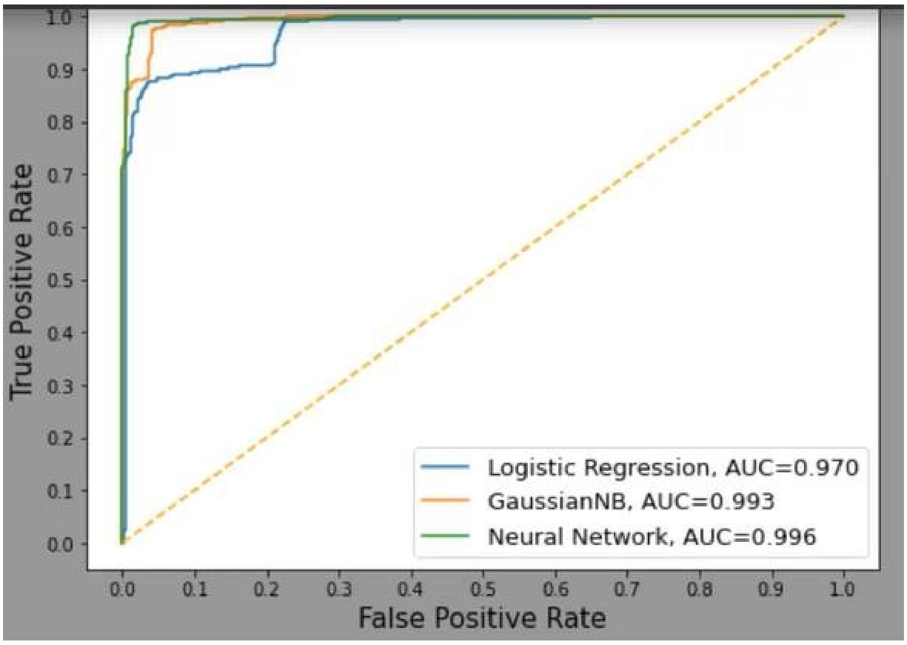
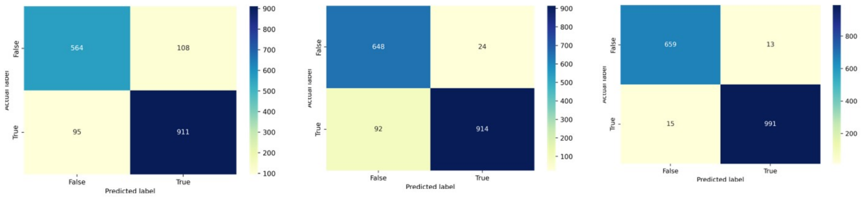


Figure 11: ROC Curve for all three implemented methods



Naive Gaussian Bayes (NGB)/ Naive Bayes Classifier

Artificial Neural Network (ANN)

Logistic Regression

Figure 12: Classification Evaluation Matrix for all three implemented methods

Conclusion

This research addressed methods for identifying spam and its types. Spam or spam in computer science refers to sending or receiving unsolicited or unsolicited electronic messages using email, instant messaging, blogs, newsgroups, social networks, web searches, mobile phones, etc. Spam is now almost unavoidable in all forms of online communication and is recognized as a barrier to the productivity of the environment in which it appears. Various measures have been taken to improve the robustness of various electronic media against a range of spam attacks. These measures are more commonly known as anti-spam techniques or spam fighting techniques. This research provides an overview of spam filtering in social networks, and Twitter reviews spam detection methods and reviews extensive research on sentiment analysis from the text. These methods include account-based spam detection methods, tweet-based spam detection methods, graph-based spam detection methods, hybrid spam detection methods, AI-based detection of multimedia spam (Deepfakes), and machine learning-based methods.

As with the results in the Madisetty & Desarkar [10] paper, with the features of the base paper and the features of the proposed method for accuracy and F-measure criteria, classification results based on the integration of information at the user level as well as at the tweet level based on neural networks perform better. The logistic regression algorithm's performance is better in terms of recall and processing speed. However, the performance of the proposed method in this research outperforms all the methods presented in the Madisetty & Desarkar [10] paper. This shows that the choice of proposed emotional features over the classic user-based or tweet-based features reviewed in the base paper is more effective. The maximum accuracy in the cited research is 93%, while the proposed method's accuracy is estimated to be up to 98%.

On the other hand, the training process was affected by this unbalanced dataset, and because more of the training samples are non-spam and fewer samples are spam, despite the poor results of feature-based classifiers, the models based on the integration of user and tweet information have performed acceptably. The improvement in the performance of the proposed method is due to the use of emotional features in combination with user-level features. This shows that the features used in the base paper have less impact than the combination of emotional features with purely user- and text-based features. However, the presence of emotional features increases the execution time of the proposed method compared to the base paper. In general, the shortest execution time belongs to user account- based features.

Data Availability Statement (DAS)

Dataset generated and/or analyzed during the current study are available from the corresponding author upon request.

References

1. ang AH (2010) Don't follow me: Spam detection in twitter. In 2010 international conference on security and cryptography (SECRYPT), 1 -10.

2. hen C, Zhang J, Chen X, Xiang Y, Zhou W (2015) 6 million spam tweets: A large ground truth for timely Twitter spam detection. In 2015 IEEE international conference on communications (ICC), 7065-70.

3. ee BD Eoff, J Caverlee (2011) Seven months with the devils: A long-term study of content polluters on Twitter, in Proc. ICWSM, 185-92.

4. ndrew S Xiao, Qilian Liang (2024) Spam detection for Youtube video comments using machine learning approaches, Machine Learning with Applications, 16: 100550.

5. hen C, Zhang J, Xie Y, Xiang Y, Zhou W et al. (2015) A performance evaluation of machine learning-based streaming spam tweets detection. IEEE Transactions on Computational social systems, 2: 65-76.

6. edhai S, Sun A (2017) Semi-supervised spam detection in Twitter stream. IEEE Transactions on Computational Social Systems, 5: 169-75.

7. lom Z, Carminati B, Ferrari E (2020) A deep learning model for Twitter spam detection.Online Social Networks and Media, 18: 100079.

8. azzaz Abkenar Sepideh, Mostafa Haghi Kashani, Mohammad Akbari, Ebrahim Mahdipour (2023) Learning textual features for Twitter spam detection: A systematic literature review, Expert Systems with Applications, 228: 120366.

9. e Q, Mikolov T (2014) Distributed representations of sentences and documents. In International conference on machine learning, 1188-96.

10. adisetty S, Desarkar MS (2018) A neural network-based ensemble approach for spam detection in Twitter. IEEE Transactions on Computational Social Systems, 5: 973-84.

11. ang B, Zubiaga A, Liakata M, Procter R (2015) Making the most of tweet-inherent features for social spam detection on Twitter. arXiv preprint arXiv:1503.07405.

12. atil DR, Patil JB (2018) Malicious URLs detection using decision tree classifiers and majority voting technique. Cybernetics and Information Technologies, 18: 11-29.

13. upta S, Khattar A, Gogia A, Kumaraguru P, Chakraborty T (2018) Collective classification of spam campaigners on Twitter: A hierarchical meta-path based approach. In Proceedings of the 2018 World Wide Web Conference, 529 -38.

14. heewala S, Patel R (2018). Machine learning based Twitter Spam account detection: a review. In 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), 79-84. IEEE.

15. eomi Nelin Nicholas, V Nirmalrani (2024) An enhanced mechanism for detection of spam emails by deep learning technique with bio-inspired algorithm, e-Prime - Advances in Electrical Engineering, Electronics and Energy, 8: 100504.

16. orge Rodríguez-Ruiz, Javier Israel Mata-Sánchez, Raúl Monroy, Octavio Loyola-González, Armando López-Cuevas (2020) A one-class classification approach for bot detection on Twitter, Computers & Security, 91: 101715.

17. ang C, Harkreader R, Gu G (2013) Empirical evaluation and new design for fighting evolving twitter spammers. IEEE Trans InfForensics Secur, 8: 1280-93.

18. hu Z, Gianvecchio S, Wang H, Jajodia S. Detecting automation of twitter accounts: are you a human, bot, or cyborg? IEEE Trans Dependable Secure Comput 2012;9(6):811-24

19. akshmana Phaneendra Maguluri, R Ragupathy, Sita Rama Krishna Buddi, Vamshi Ponugoti Tharun Sai Kalimil, Adaptive Prediction of Spam Emails: Using Bayesian Inference, 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC).

20. alerie Niechai (2019) How to Use TF -IDF for SEO, How to use TF-IDF tools for semantic SEO (link-assistant.com)

21. avidson T, Warmsley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. arXiv preprint arXiv: 1703.04009.

22. ustin Hillard, N-gram Language Modeling Tutorial, Lecture notes courtesy of Prof. Mari Ostendorf.

23. aâli Mnasri (2019) How to train word embeddings using small datasets?.

24. rucker H, Wu D, Vapnik VN (1999) Support vector machines for spam categorization. IEEE Transactions on Neural networks, 10: 1048-54.

25. ndroutsopoulos I, Koutsias J, Chandrinos KV, Paliouras G, Spyropoulos CD (2000) An evaluation of naive bayesian anti -spam filtering. arXiv preprint cs/0006013. differentiation for combating link-based Web spam," ACM Trans. Web, 8: 37-40.

26. Zhang, Y Feng, H Shen, W Liang (2015) Differential trust propagation with community discovery for link-based Web spam demotion, in Proc. Int.Conf.Web-Age Inf. Manage. Cham, Switzerland: Springer, 452-6.

27. hen H, Ma F, Zhang X, Zong L, Liu X et al. (2017) Discovering social spammers from multiple views. Neurocomputing, 225: 49-57.

28. artinez-Romo J, Araujo L (2013) Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications, 40: 2992-3000.

29. antos I, Miñambres-Marcos I, Laorden C, Galán-García P, Santamaría-Ibirika A et al. (2014) Twitter content-based spam filtering. In International Joint Conference SOCO', 449-58.

30. erveen N, Missen MMS, Rasool Q, Akhtar N (2016) Sentiment based twitter spam detection. International Journal of Advanced Computer Science and Applications (IJACSA), 7: 568- 73.

31. riented spam research. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 223-32.

32. im Y (2014) Convolutional neural networks for sentence classification. arXiv preprint:1408.5882.

33. rizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60: 84-90.

34. eCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86: 2278-324.

35. accianella, Stefano, Andrea Esuli, Fabrizio Sebastiani (2010) Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." In Lrec, 10: 2200-4.

36. suli A, Sebastiani F (2006) Sentiwordnet: A publicly available lexical resource for opinion mining. In LREC, 6: 417-22.

37. hana B, Tierney B (2009) Sentiment classification of reviews using SentiWordNet. In 9th. it & t conference, 13: 18-30.

38. öscher A, Jahrer M, Bell RM (2009) The bigchaos solution to the netflix grand prize.

39. etflix prize documentation, 1-52.Niculescu-Mizil A, Perlich C, Swirszcz G, Sindhwani V, Liu Y et al. (2009) Winning the KDD cup orange challenge with ensemble selection. In KDD-Cup 2009 Competition.

40. an de Vegte J (1990) Feedback Control Systems (2nd ed.). Prentice Hall.