



Uncovering Coin Market Manipulation on X Using Machine Learning

Zehra Nur ÖZBAY^{1*}, Aybars UĞUR², Mehmet Emin DALKILIÇ³

¹Ege University, International Computer Institute, zehra.nur.ozbay@ege.edu.tr, Orcid No: 0000-0001-5680-1227

²Ege University, Faculty of Computer and Informatics Sciences, Department of Computer Engineering, aybars.ugur@ege.edu.tr, Orcid No: 0000-0003-3622-7672

³Ege University, International Computer Institute, mehmet.emin.dalkilic@ege.edu.tr, Orcid No: 0000-0003-3932-5155

ARTICLE INFO

Article history:

Received 25 September 2025

Received in revised form 25 January 2026

Accepted 30 March 2026

Available online 25 June 2026

Keywords:

social media, spam, coin,
manipulation, disinformation,
machine learning

Doi: 10.24012/dumf.1789337

* Corresponding author

ABSTRACT

Spam-bots have become a widespread problem due to the ease of creating accounts on the X platform. These bots are designed to automatically generate and disseminate unsolicited content, frequently for advertising purposes or events of questionable legality, often contributing to the spread of misleading or deceptive information within the broader context of disinformation. A key challenge in detecting these spam-bots is their ability to create accounts rapidly, post certain volumes of content around an event, and then become dormant. This study examines spam-bot accounts created for coin market manipulation on X. We present a novel approach to obtain a spam-bot dataset for coin-related tweets, including data collection, analysis, and labeling. For data preprocessing, we applied Natural Language Processing techniques, which resulted in an imbalanced dataset. We used BERT embeddings for feature extraction and evaluated the performance of nine machine learning algorithms in detecting the tweets of spam-bot accounts. The SVM model achieves the highest cross-validated accuracy of 98.10% on the balanced dataset, produced using the SMOTE oversampling technique. On the original imbalanced dataset, Logistic Regression has the highest accuracy at 96.46%. The cross-validated accuracy of SVM is further improved through hyperparameter tuning, resulting in 99.70%.

Introduction

Spam is a common term used to describe unsolicited electronic messages. Spam messages demonstrate unidirectional dissemination of information, typically for advertising, phishing, or spreading malware. With the growth of social networks, this type of content is often called Self-Propelled Advertising Material [1]. This has led to the promotion of various products and services, including financial instruments such as cryptocurrencies. Consequently, social media platforms have become a hub for monitoring cryptocurrency trading and exchanges. One of those famous platforms is Twitter (now rebranded as X), which is widely used by crypto traders. Novice investors use X to stay informed and make investment decisions. However, the rapid and easy access to X has also made it a target for market manipulation using some specific techniques. One of them is wash trading [2], a form of market manipulation in which a trader or a group of traders simultaneously buys and sells the same financial asset to create artificial activity in the market, misleading other investors into thinking that the asset is more popular or valuable than it actually is. This activity is accompanied by the creation of fake social media accounts, known as bots, to amplify and spread the manipulation. The detection of

such manipulative activities is crucial for maintaining the integrity of the cryptocurrency market and protecting investors from potential financial losses. Moreover, such coordinated spam-bot activities contribute to the spread of disinformation by amplifying misleading or deceptive narratives within cryptocurrency-related discussions on social media platforms. Thus, studies carried out in this domain are expected to yield dual benefits, namely protecting the integrity of the cryptocurrency market and reducing the dissemination of disinformation on social media platforms.

In this paper, we analyze the activities of spam accounts created to promote or encourage investments in stocks and cryptocurrencies, with the goal of phishing or market manipulation. Considering the widespread impact of manipulative activities in cryptocurrency exchanges, the key contribution of the proposed study is as follows:

- A novel approach for creating, labeling, and analyzing a highly recent spam-bot dataset consisting of coin-related tweets that reflects the current trends.
- Experimental studies on determining the most effective model without complexity about detection of unsolicited advertising tweets sent

for market manipulation using a BERT-based machine learning method.

- Identifying the characteristics of spam accounts which have a very short lifespan

Approximately 16% of the spam-tweeting accounts are automated bot accounts, while the remaining 84% are regular user accounts controlled by human spammers [3]. The low rate of spam-bots compared to real spammers suggests that the identification of spam-bots can provide valuable insights for the field.

Related works

Social media platforms have become a common trending area for manipulative activities carried out by spammers and spam-bots. Researchers have used various methods to detect Twitter spam, such as analyzing content, user behavior, tweet characteristics, network patterns, and combining multiple approaches [4].

The model proposed by Chakraborty et al. [5] used a combination of Graph Representation Learning and Natural Language Processing techniques to distinguish legitimate and spam accounts on the X platform. The authors employed NLP methods to extract attributes from tweets, utilized the followers-following graph to obtain graph centrality metrics, and applied feature engineering to acquire user metadata. The study trained seven different supervised machine learning models, including Naive Bayes, Random Forest, Gradient Boosting Decision Trees, Support Vector Machine (SVM), Logistic Regression, XGBoost, and Neural Network with three layers. The study found that the XGBoost model outperforms the others, achieving an accuracy of 96.99%.

Heidari et al. [6] examined how social bots use sentiment analysis and psychological tactics to generate artificial trends or manipulate public opinion. The study found that incorporating sentiment-based features from user posts enhances the accuracy of models such as Neural Network, Random Forest, and SVM in identifying social bots.

The study by Çıtlak et al. [7] used a combination of techniques, including short link analysis, machine learning, and text analysis, to detect spam accounts on social media platforms. The researchers tested their framework on a Twitter dataset and achieved a 95.69% success rate in identifying spam.

The study by Sumathi and Raja [8] compared machine learning methods to detect spam on Twitter, using an unbalanced dataset with a 1:19 ratio of spam and ham instances. The researchers preprocessed the data and then used features such as the number of characters, words, and sentences to train various machine learning classifiers, including Naive Bayes, K- Nearest Neighbors, Decision Trees, Logistic Regression, Random Forest, SVM, XGBoost, Extra Tree Classifier, and Voting Classifier. The ensemble methods, the Voting Classifier and the Extra Trees Classifier, outperformed traditional machine learning techniques in terms of accuracy, precision, recall, and F1 score. Specifically, the Voting Classifier achieved an

accuracy of 97.96%, and the Extra Trees Classifier had an accuracy of 97.77%. Therefore, the researchers recommended using these ensemble methods for the task of spam detection.

Despite the strong performance reported in prior studies, the issue of class imbalance—a well-known challenge in spam and bot detection tasks—has received limited attention. Several existing approaches [9]-[12] evaluate their models on datasets with fixed class distributions, while the issue of class imbalance is not explicitly addressed or systematically analyzed. As a result, performance metrics such as accuracy may provide an overly optimistic assessment of real-world effectiveness. While the study in [8] conducts experiments on an imbalanced dataset, it does not compare model performance across balanced and imbalanced data settings. In contrast to prior work, our study explicitly investigates the effect of class imbalance by conducting experiments on both balanced and imbalanced datasets, thereby providing a clearer understanding of the performance gap encountered in real-world, live environments where imbalanced class distributions are the natural standard.

Data collection

To collect data, we use the popular X social media platform and the Tweepy Python library to interact with the Twitter API. We first subscribed to the X developer platform and generated the necessary user-specific API credentials. We gather a dataset of 350 spam and 350 non-spam (ham) tweets.

For the ham dataset, we selected tweets from legitimate users who have been active on the platform for several years. These accounts are related to the cryptocurrency industry. We selected ham tweets from the following seven accounts whose user names are:

- 1) binance
- 2) BinanceVIP
- 3) BitMartExchange
- 4) coinbase
- 5) CoinMarketCap
- 6) CoinWOfficial
- 7) HotcoinGlobal

The main idea behind selecting ham tweets is to choose tweets that contain URLs, emojis, hashtags and mentions similar to spam tweets to improve the reliability of differentiating between spam and ham tweets.

Data analysis and labeling

Financial spam accounts

Social media platforms are inundated with different forms of unsolicited content, including advertisements, political manipulation, provocation, phishing attempts, etc. This study examines accounts involved in financial market manipulation and stock market scams on social media

platforms. For simplicity, we will refer to these types of spam as financial spam.

Based on the data collected and analyzed in this study, we can illustrate financial spam accounts according to their working strategies, as shown in Fig. 1.

Our data labeling approach

Financial spam messages sent on social media platforms are typically either direct messages or tweets with mention tags. Our data analysis revealed two main types of financial spam accounts:

- Accounts that use their profiles to scam users
- Accounts that post tweets related to financial products and services

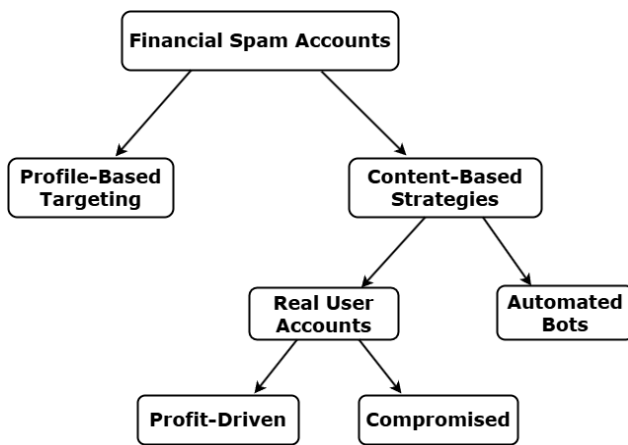


Figure 1. Taxonomy of financial spam account types based on their operational strategies.

We identified 39 accounts in the first group with tweets in both Turkish and English. These accounts contain links that invite users to join WhatsApp groups or channels, claiming that they can earn daily returns on American, Turkish, or Indian stock and cryptocurrency exchanges. Here are the key characteristics observed in the first group of accounts:

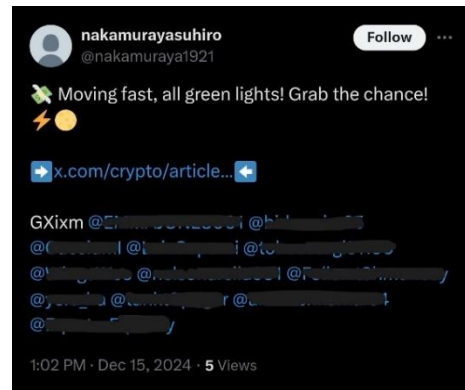
- They either send direct messages or simply follow users without sending any messages, likely to draw attention.
- Their background photos are usually related to finance.
- Their profile pictures depict businesspeople.
- Their profiles include WhatsApp URLs with prefixes like “hai.mewhatsapp”, “wa.me”, or “chat.whatsapp.com”.
- Their profile descriptions are often similar.

Spam accounts in the second group can be divided into two categories: automated bots and real user accounts. The bots were identified by analyzing their user profiles, which exhibited similar behavior patterns and content, such as:

- Accounts being created shortly before spamming activity,

- Similar post content,
- Zero followers and following count,
- The total number of tweets posted in a short period of time of one minute,
- The layout of the tweets,
- The presence of URLs linked to tweets from verified user accounts that promote some cryptocurrency assets,
- Each account is opened to promote only one tweet,
- Limited lifespan, often posting no more than 200 tweets before becoming dormant,
- Exhibiting a lack of genuine engagement with other users and vice versa,
- A large amount of emoji usage,
- One or two meaningful sentences,
- Number of mentions using 10-12 @ symbols in a tweet, and
- Usage of a five-character long random alpha-numeric string.

Fig. 2 shows the similar tweet layout of two different spam-bots advertising and promoting an announcement.



a. First sample spam tweet



b. Second sample spam tweet

Figure 2. Two tweet samples of spam-bot accounts in the dataset

We identified real user accounts by analyzing their profiles. Their account was created years ago but stopped tweeting a few years ago. Their following and follower counts were not zero. However, the interesting thing is that real user accounts also exhibit behavior similar to that of

spam-bots: They send a bulk of tweets in a short period of time of one minute. This leads us to consider these accounts to have been stolen. However, some real user accounts just tweet spam that promotes a verified account's specific tweet. This leads us to consider these accounts to do this to earn money.

In this study, we focus on automated bots. The first group of accounts is frequently suspended for violating platform rules, suggesting that they are more easily detected. Moreover, genuine user accounts often repeatedly post the same content, which would limit the dataset. However, identifying spam-bots presents challenges. One challenge is the requirement for an ad tweet. Another is the need for @ tag mentions to be able to find such accounts. In addition, these accounts have a short lifespan, posting at most 200 tweets before becoming inactive. These spam-bots typically start their activities when a verified known account shares a post about a crypto-related topic. The main purpose of these accounts is to promote the 'opportunities' of cryptocurrency investment through 'pump announcements', coordinated attempts to artificially inflate the price of a specific cryptocurrency through misleading information. Spam-bots direct users to these posts to encourage them to join their "private pump channels".

We identified seven different accounts of spam-bots and scraped their tweets for our analysis. We selected the first 50 spam tweets from each of the seven accounts. The spam dataset contains examples of textual content as shown in Table 1.

Table 1. Text content samples from the spam dataset.

Text
Follow the best crypto analytics in telegram channel...
...Ready for something amazing? Find out more by clicking the link!...
...Web3 takes the next big leap! ... Eligible users:...
...Reach for the Moon! .. Welcome:..
...HOT HOT earn more with our crypto pump!...
...Discover the secret! wow Pump event today!...

Word cloud

To analyze the language used by these financial spam-bots, we created a word-cloud visualization. Figure 3 presents this visualization, highlighting the prevalent terms used in tweets promoting coin market advertisements. The word cloud was generated using Python's WordCloud module, which allowed us to identify the most prominent keywords and phrases in these spam messages. We preprocessed the data before creating the word cloud. This involved converting the text to lowercase, removing mentions, punctuation, and special characters, then tokenizing, lemmatizing, and removing invalid words and stop words from the tweets. We then stored the unique representations of the preprocessed tweets in a set. The resulting word cloud provides valuable insight into the lexical patterns and trends within the coin market spam data. The most frequently occurring words are "eligible", "welcome", "user", "reach",

"today", and "pump". These words suggest that coin-market ads aim to create a sense of urgency and exclusivity, encouraging the reader to act quickly to take advantage of a limited-time offer.

Methodology

Our methodology for detecting coin-related tweets from spam-bot accounts on the X platform leverages natural language processing techniques combined with BERT embeddings. The detailed methodology workflow is depicted in Fig. 4

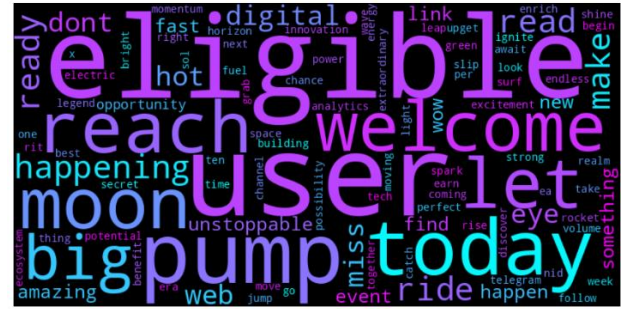


Figure 3. Word cloud of the tweet-based spam-bot dataset

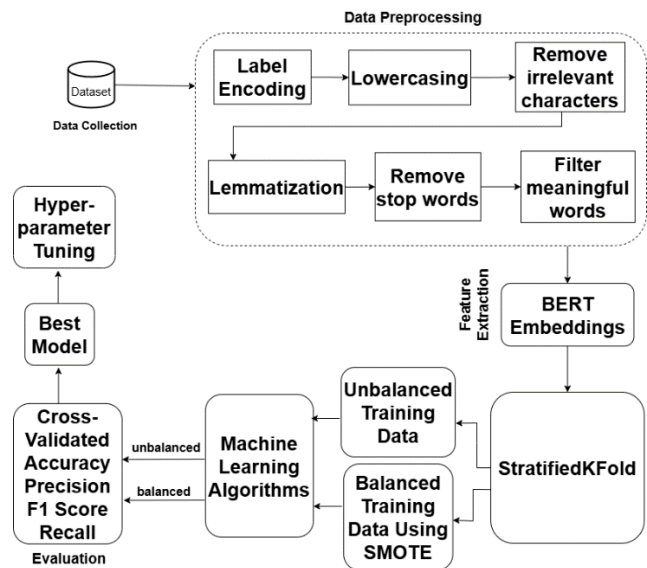


Figure 4. Overview of the proposed workflow for spam-bot tweet detection on the X platform

Data preprocessing

We begin by preprocessing the raw data, assigning labels of 0 to 'ham' tweets and 1 to 'spam' tweets. Next, we perform text preprocessing by converting all characters to lowercase, removing irrelevant elements such as emojis, random alphanumeric text, hashtags, and punctuation marks. Next, we apply lemmatization to obtain the base forms of the words. Then, we remove all stop words. Finally, we retain only the meaningful words. The processed text was then merged back into a single string for each tweet. We utilized a 'set' data structure to identify and remove duplicate tweets. The significant reduction in the

number of spam tweets—from 350 down to 34—is primarily attributed to this step, as the spam bots in our dataset frequently posted identical or near-identical messages. By removing these duplicates, we ensured that the model learns from unique semantic content rather than being biased by repetitive automated posts.

The final dataset contains 368 unique tweets, with 334 classified as ‘ham’ and 34 classified as ‘spam’. The dataset has an uneven class distribution (1:10) after preprocessing, as shown in Fig. 5.

Feature extraction

The tweets with clean and meaningful words were then processed using the BERT tokenizer, which converts the input text into subword tokens, and fed into a pre-trained BERT model to generate contextualized word embeddings.

Imbalanced Dataset After Preprocessing

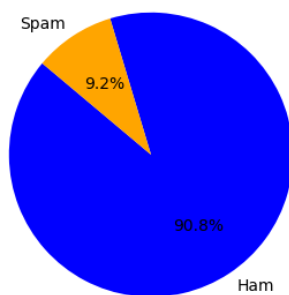


Figure 5. Percentage of ham and spam tweets after preprocessing

We used the BERT-base-uncased model [13] provided by the HuggingFace *transformers* library for feature extraction. The model was employed without fine-tuning and operated in evaluation mode. Input sequences were adjusted to a fixed length of 512 tokens using padding and truncation. Sentence-level representations were obtained by applying mean pooling over the last hidden layer outputs. BERT was utilized as a fixed feature extractor rather than a fine-tuned architecture to ensure a computationally efficient and lightweight solution suitable for the real-time nature of social media. This approach prioritizes lower computational overhead while focusing on the performance of traditional machine learning models on high-quality semantic embeddings.

Training and evaluation

The embeddings are subsequently used as features to train nine different machine learning models. The training is carried out in two different ways:

- Using the original unbalanced dataset,
- Balancing the classes by applying the SMOTE oversampling technique to generate synthetic samples of the minority class (‘spam’ tweets).

The models’ performance is evaluated using stratified 5-fold cross-validation in both scenarios. To prevent data leakage, oversampling is performed separately on the training data within each fold, rather than on the entire dataset.

Performance is assessed using accuracy, precision, recall, and F1-score, all reported as normalized values in the range [0,1]. Precision, recall, and F1-score are computed using weighted averaging in both the original and SMOTE-balanced datasets. Although class balancing reduces the impact of class imbalance, weighted averaging is consistently applied to ensure methodological consistency across experiments.

Based on the cross-validated results, the best-performing model is selected, and its hyperparameters are subsequently tuned using GridSearchCV, which exhaustively searches predefined hyperparameter combinations using cross-validation.

Experimental results

This section reports the experimental results for 9 different machine learning models on unbalanced and balanced datasets. The models include linear classifiers (Logistic Regression), distance-based methods (K-Nearest Neighbors), probabilistic models (Naive Bayes), margin-based classifiers (Support Vector Machines), tree-based models (Decision Tree, Random Forest, Extra Trees), and ensemble learning methods (AdaBoost and Gradient Boosting).

Support Vector Machines (SVM) were included due to their strong generalization performance on relatively small datasets [8]. In addition, kernel-based approaches, including SVM, have been extensively adapted to mitigate class imbalance in classification tasks [14]. The remaining models were incorporated primarily for comparative evaluation.

Tables 2 and 3 show the cross-validated results for the unbalanced and balanced datasets, respectively, with performance metrics reported as mean $\pm$ standard deviation based on a 5-fold cross-validation process.

The performance of the models in the two scenarios is visually compared in Figures 6 and 7 as well. The results demonstrate that the performance of all machine learning models, except K-Nearest Neighbor, improves significantly when the datasets are balanced using the SMOTE oversampling technique.

The observed decline in KNN’s performance following SMOTE application may be attributed to the interaction between the oversampling algorithm and the high-dimensional BERT embedding space (768 dimensions). Since SMOTE generates synthetic samples through interpolation, some of these points may introduce local noise or increase class overlap in the embedding space.

Distance-based algorithms such as KNN are particularly sensitive to such changes in local neighborhood structure, which may explain the observed performance decrease.

Table 2. Cross-validation results for unbalanced dataset

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9646 ±0.0139	0.9638 ±0.0152	0.9646 ±0.0139	0.9622 ±0.0156
Random Forest	0.9049 ±0.0082	0.8236 ±0.0094	0.9049 ±0.0082	0.8623 ±0.0084
SVM	0.9103 ±0.0066	0.8450 ±0.0413	0.9103 ±0.0066	0.8696 ±0.0126
Naive Bayes	0.9020 ±0.0420	0.9357 ±0.0204	0.9020 ±0.0420	0.9133 ±0.0326
Decision Tree	0.9292 ±0.0363	0.9186 ±0.0543	0.9292 ±0.0363	0.9236 ±0.0453
Extra Tree Classifier	0.9158 ±0.0098	0.8688 ±0.0516	0.9158 ±0.0098	0.8807 ±0.0200
KNN	0.9457 ±0.0171	0.9465 ±0.0180	0.9457 ±0.0171	0.9344 ±0.0240
AdaBoost	0.9538 ±0.0185	0.9519 ±0.0217	0.9538 ±0.0185	0.9499 ±0.0205
Gradient Boosting	0.9455 ±0.0261	0.9277 ±0.0578	0.9455 ±0.0261	0.9351 ±0.0413

Table 3. Cross-validation results for balanced dataset

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9756 ±0.0101	0.9769 ±0.0099	0.9756 ±0.0101	0.9760 ±0.0099
Random Forest	0.9457 ±0.0192	0.9449 ±0.0236	0.9457 ±0.0192	0.9346 ±0.0258
SVM	0.9810 ±0.0108	0.9812 ±0.0108	0.9810 ±0.0108	0.9804 ±0.0111
Naive Bayes	0.9457 ±0.0193	0.9425 ±0.0247	0.9457 ±0.0193	0.9423 ±0.0224
Decision Tree	0.9375 ±0.0163	0.9340 ±0.0228	0.9375 ±0.0163	0.9335 ±0.0216
Extra Tree Classifier	0.9402 ±0.0203	0.9254 ±0.0536	0.9402 ±0.0203	0.9230 ±0.0348
KNN	0.7555 ±0.0275	0.9292 ±0.0051	0.7555 ±0.0275	0.8057 ±0.0214
AdaBoost	0.9646 ±0.0109	0.9654 ±0.0117	0.9646 ±0.0109	0.9646 ±0.0113
Gradient Boosting	0.9592 ±0.0285	0.9585 ±0.0294	0.9592 ±0.0285	0.9586 ±0.0288

In contrast, classifiers such as SVM and Logistic Regression learn global decision boundaries and appear less affected by these local variations. Nevertheless, the relatively stronger performance of KNN under the original imbalanced setting may still be relevant for real-world social media environments, where spam and non-spam content typically occur in imbalanced proportions.

The SVM model has the highest performance on the balanced dataset compared to the other models. We then analyzed the SVM's training and testing accuracy across folds to determine whether it was overfitting, as shown in Fig. 8. The model achieved very high test accuracy, ranging from 0.972 to 1.0. This small gap between training and test performance shows that the model generalizes well.

Additionally, we tuned the SVM model's hyperparameters to optimize its performance further. With the optimal

hyperparameters of $C=5$, $\gamma=\text{'scale'}$, and $\text{kernel}=\text{'rbf'}$, we were able to achieve an accuracy of 0.9970.

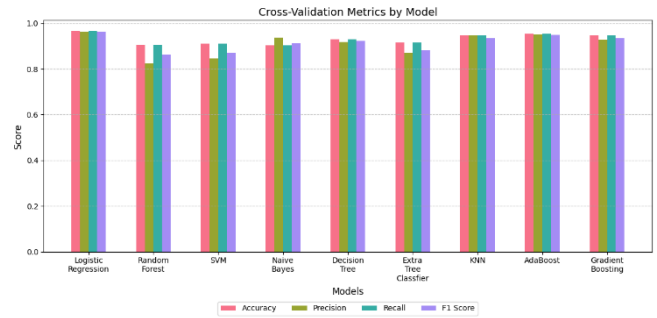


Figure 6. Models' performance with unbalanced original dataset

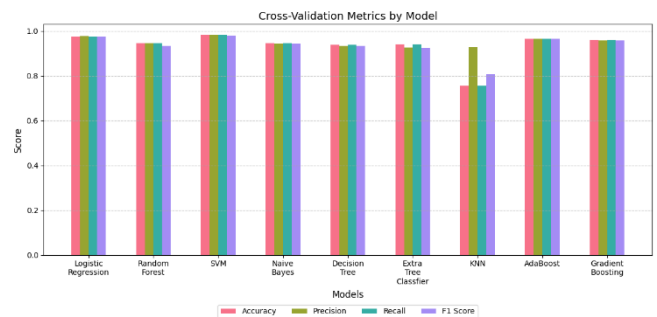


Figure 7. Models' performance with balanced dataset using SMOTE

Finally, we evaluated our methodology on a different dataset, the well-known UCI SMS Spam Collection. We selected this dataset as the data preprocessing procedures enabled us to reframe the tweet spam detection task as an SMS spam detection problem. After preprocessing, we obtained 538 spam messages and 4345 ham messages, resulting in an approximately 1:8 ratio of spam to ham. The SVM model achieved the highest scores across all metrics again: 0.9674 for accuracy, 0.9679 for precision, 0.9674 for recall, and 0.9676 for F1-score. These results validate the effectiveness of our approach. Fig. 9 shows the training and testing accuracy scores across ten folds on balanced training data, which range from 0.9447 to 0.9795, indicating a good generalization. This suggests that the proposed framework can be extended to diverse social media platforms and applications.

Conclusion and discussion

Following the acquisition of Twitter in October 2022, the platform has witnessed a substantial departure of advertisers amid a period characterized by the proliferation of misinformation, policy disputes, and the eroding of consumer trust. According to data from Sensor Tower, as of September 2023, 46 of X's top 100 advertisers had completely stopped their advertising expenditures in the United States [15]. Cryptocurrency assets promoted through unsolicited mass messages undermine trust among not just investors, but also brands.

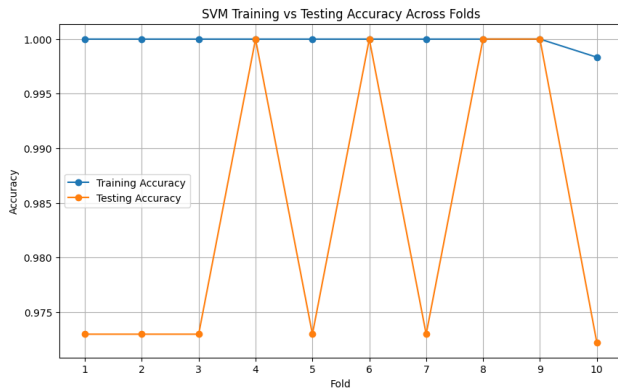


Figure 8. SVM training and testing performance across ten folds for the new financial spam dataset

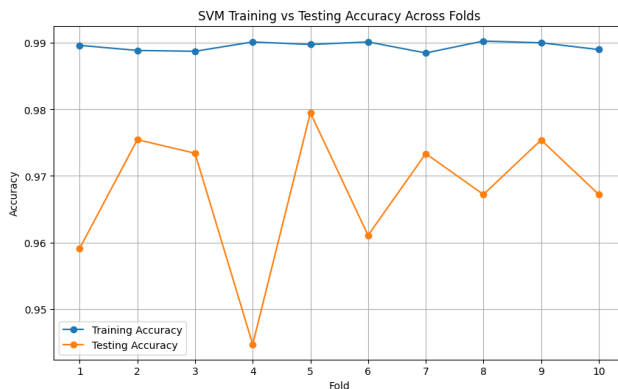


Figure 9. SVM training and testing performance across ten folds for UCI SMS Spam Collection

Although some brands have withdrawn from this platform to protect themselves, it continues to harm brands as users continue to receive ad spam. There is a clear need to develop robust methods for detecting and reducing manipulation on cryptocurrency exchanges. Hence, close monitoring and analysis of emerging spam techniques and formats is critical to restoring investor confidence in the cryptocurrency market.

Although automated spam accounts were previously observed to generate an average of 12 tweets per day at regular intervals [16], the characteristics of spam-bots have changed significantly over time. Spam-bots, having new characteristics, are not consistently active, but rather become active only for a limited duration when a particular event is imminent. This makes it difficult for X to detect them promptly. The spam-bot accounts addressed in this study took Twitter weeks to suspend. Furthermore, the challenge in detecting spam activity is compounded by the selective nature of these spam-bot communities, which target specific accounts. In this context, it is crucial to establish credibility and demonstrate genuine interest in the domain through active participation, which makes acquiring sufficient data for analysis a time-consuming endeavor.

In light of these challenges in ensuring data quality, the collection of 'ham' (non-spam) tweets for this study focused on seven major and verified cryptocurrency

exchange accounts. This selection was intended to establish a reliable baseline of legitimate financial content from authoritative sources. However, it is important to acknowledge that focusing on a limited number of high-profile corporate accounts may introduce a degree of selection bias. The formal and structured linguistic style of official exchange communications may not fully reflect the diverse or casual tone used by individual retail investors on the platform. While the deep semantic capabilities of the BERT-based model help to minimize the impact of such stylistic variations, this remains a limitation of the current sampling strategy. Future research could enhance the generalizability of the proposed method by incorporating a broader spectrum of ham data, including verified individual traders and independent financial news outlets, to represent a wider array of communication styles.

In this paper, we outlined a methodological approach to identify financial spam-bots in X. We presented the characteristics of these newly emerging bots and categorized the financial spam accounts. The created dataset in this work became unbalanced after applying data preprocessing techniques. To address the issue of class imbalance, we employed the SMOTE oversampling technique to balance the dataset. We compared nine different machine learning models on the original and balanced datasets, and the results showed that the Support Vector Machine with a balanced dataset, utilizing BERT embeddings and data preprocessing techniques, outperformed the other models with an accuracy of 0.9810. The SVM performance was further improved through hyperparameter tuning, achieving an accuracy of 0.9970. Ultimately, the high effectiveness of the proposed model in detecting these automated tweets provides a crucial tool for addressing market manipulation tactics, such as wash trading, which rely on artificial activity to mislead investors on social media.

Future work

Beyond the technical performance of the proposed approach, the observed spam-bot activities can also be framed within the broader concept of *disinformation*. According to the High Level Expert Group (HLEG), disinformation encompasses all forms of false, inaccurate, or misleading information that are deliberately designed, presented, and promoted with the intention of causing public harm or generating profit [17]. Cryptocurrency-related spam messages disseminated by automated accounts align with this definition, as they intentionally promote misleading content for financial gain while exposing users and investors to potential economic harm. From this perspective, such activities may be regarded as a form of *financial disinformation*, where coordinated spam campaigns exploit social media platforms to manipulate perceptions of value, legitimacy, or investment opportunity. Therefore, detecting and combating spam-bot-driven content is not only essential for protecting investors and market integrity, but also for limiting the spread of financially motivated disinformation within the broader digital ecosystem.

In today's digital ecosystem, information spreads at an unprecedented speed, while access to large-scale, up-to-date, and reliably labeled data remains limited. Moreover, the widespread availability of large language models (LLMs) has significantly lowered the barrier to generating persuasive spam and disinformation content, making such messages more abundant and increasingly difficult to distinguish from legitimate information. Although large-scale neural architectures—such as recurrent neural networks (e.g., LSTM) and graph-based models (e.g., Graph Neural Networks)—are well suited for modeling complex temporal and relational patterns in social media data, they were not explored in the present study due to dataset size and labeling constraints. Nevertheless, considering the structural complexity and dynamic nature of social networks, investigating such deep learning models represents a promising direction for future work. Advancing research in this area will require the availability of substantially larger, professionally annotated, bias-aware, open-source, and up-to-date benchmark datasets to enable robust and reproducible evaluation.

In addition to content volume and structural complexity, recent policy changes on the X platform have introduced new challenges for manipulation detection. The platform's monetization model, which enables verified accounts to earn revenue from user engagement, has incentivized the dissemination of emotionally charged content designed to influence users' beliefs and attitudes. Such content often relies on strong emotional appeals—such as anger, fear, hatred, or their opposites—to maximize interaction and visibility. Notably, even verified accounts have been reported, through Community Notes, to engage in deceptive or scam-oriented behavior. These developments suggest that future research should also focus on detecting emotion-driven manipulative activities, integrating affective features and emotional dynamics into spam and disinformation detection frameworks.

Ethics committee approval and conflict of interest statement

There is no need to obtain permission from the ethics committee for the article prepared. There is no conflict of interest with any person / institution in the article prepared.

Authors' Contributions

- Study conception and design: Özbay
- Acquisition of data: Özbay
- Analysis and interpretation of data: Özbay
- Drafting of manuscript: Özbay, Uğur
- Critical revision: Özbay, Uğur, Dalkılıç

References

[1] E. H. Tusher, M. A. Ismail, M. A. Rahman, A. H. Alenezi, and M. Uddin, "Email Spam: A Comprehensive Review of Optimize Detection Methods, Challenges, and Open Research Problems," *IEEE Access*, vol. 12, pp.

143627–143657, 2024, doi: 10.1109/ACCESS.2024.3467996.

[2] L. Cong, X. Li, K. Tang, and Y. Yang, "Crypto Wash Trading", *SSRN Journal*, 2019, doi: 10.2139/ssrn.3530220.

[3] P. Manasa et al., "Tweet Spam Detection Using Machine Learning and Swarm Optimization Techniques", *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 4870–4877, Aug. 2024, doi: 10.1109/TCSS.2022.3230823.

[4] S. Bazzaz Abkenar, M. Haghi Kashani, M. Akbari, and E. Mahdipour, "Learning textual features for Twitter spam detection: A systematic literature review", *Expert Systems with Applications*, vol. 228, p. 120366, Oct. 2023, doi: 10.1016/j.eswa.2023.120366.

[5] M. Chakraborty, S. Das, and R. Mamidi, "Detection of Fake Users in SMPs Using NLP and Graph Embeddings", in *2022 14th International Conference on COMMunication Systems & NETWORKS (COMSNETS)*, Jan. 2022, pp. 754–758. doi: 10.1109/COMSNETS53615.2022.9668371.

[6] M. Heidari, J. H. J. Jones, and O. Uzuner, "An Empirical Study of Machine learning Algorithms for Social Media Bot Detection", in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, Apr. 2021, pp. 1–5. doi: 10.1109/IEMTRONICS52119.2021.9422605.

[7] O. Çıtlak, M. Dörterler, and İ. Dogru, "A Hybrid Spam Detection Framework for Social Networks," *Politeknik Dergisi*, vol. 26, no. 2, pp. 823–837, Jul. 2023, doi: 10.2339/politeknik.933785.

[8] M. Sumathi and S. P. Raja, "Machine learning algorithm-based spam detection in social networks," *Soc. Netw. Anal. Min.*, vol. 13, no. 1, p.104, Aug. 2023, doi: 10.1007/s13278-023-01108-6.

[9] Y. Li, Y. Xiao, Y. Huang, and R. Ma, "Analysis of social networks content to identify fake news using stacked combination of deep neural networks," *Egyptian Informatics Journal*, vol. 30, p. 100707, Jun. 2025, doi: 10.1016/j.eij.2025.100707.

[10] V. Vysotska, N. Babkova, D. Huliieva, Z. Kochuieva, N. Ugolnikova, and M. Kozulia, Identifying fake news in political discourse behavior using machine learning methods. 2025. doi: 10.31110/COLINS/2025-3/012.

[11] A. A. Alhaj, O. Safwat, Y. Ghoneim, I. Zualkernan, and A. R. Sajun, "Investigating Fake News Detection Using BERT/RoBERTa LLMs," in *Proceedings of Tenth International Congress on Information and Communication Technology*, X.-S. Yang, R. S. Sherratt, N. Dey, and A. Joshi, Eds., Singapore: Springer Nature, 2025, pp. 241–251. doi: 10.1007/978-981-96-6929-5_19.

- [12] E.-S. Apostol, C.-O. Truică, and A. Paschke, "ContCommRTD: A Distributed Content-Based Misinformation-Aware Community Detection System for Real-Time Disaster Reporting," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 5811–5822, Nov. 2024, doi: [10.1109/TKDE.2024.3417232](https://doi.org/10.1109/TKDE.2024.3417232).
- [13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, *arXiv*: arXiv:1810.04805. doi: [10.48550/arXiv.1810.04805](https://doi.org/10.48550/arXiv.1810.04805).
- [14] M. Q. Alnabhan and P. Branco, "Fake News Detection Using Deep Learning: A Systematic Literature Review," *IEEE Access*, vol. 12, pp. 114435–114459, 2024, doi: [10.1109/ACCESS.2024.3435497](https://doi.org/10.1109/ACCESS.2024.3435497).
- [15] "Where's the 'X' Factor? X (Twitter) One Year Post-Acquisition Recap," *Sensor Tower Blog*. Accessed on: Sept. 22, 2025, [Online] Available: <https://sensortower.com/blog/wheres-the-x-factor-x-twitter-one-year-post-acquisition-recap>.
- [16] I. Inuwa-Dutse, M. Liptrott, and I. Korkontzelos, "Detection of spam-posting accounts on Twitter", *Neurocomputing*, vol. 315, pp. 496–511, Nov. 2018, doi: [10.1016/j.neucom.2018.07.044](https://doi.org/10.1016/j.neucom.2018.07.044).
- [17] European Commission, Ed., A multi-dimensional approach to disinformation: report of the independent High level Group on fake news and online disinformation. Luxembourg: Publications Office, 2018. doi: [10.2759/0156](https://doi.org/10.2759/0156).