Detection of Twitter Spam with Language Models: A Case Study on How to Use BERT to Protect Children from Spam on Twitter

Bianca Montes Jones
Capitol Technology University
Laurel, USA
bmontesjones@captechu.edu

Marwan Omar
Illinois Institute of Technology
Chicago, USA
momar3@iit.edu

***Abstract*— With the increasing use of social media platforms like Twitter, the problem of spam is becoming more prevalent. This is especially concerning when it comes to children who use Twitter, as they may be exposed to harmful content or scams disguised as legitimate tweets. In this paper, we present a case study on how to use Bidirectional Encoder Representations from Transformers (BERT), a pre-trained language model, to detect and protect children from spam on Twitter. We collected a large data set of tweets that are relevant to children and used BERT to build a spam detection model. Using our approach, we were able to accurately detect spam tweets with a high degree of precision and recall. We also evaluated the effectiveness of our approach using various metrics and found that it outperforms several baseline models. Our study demonstrates the potential of using state-of-the-art language models like BERT to protect children from spam on Twitter, and our findings provide insights on how to develop effective spam detection models for social media platforms.**

Keywords—Twitter spam, BERT, natural language processing, data set, social media, spam detection

I. Introduction

Over the past few years, online media platforms have gained widespread popularity across the globe. Social media sites such as Facebook, Twitter, and Instagram have become an integral part of daily lives of many, with individuals spending a significant amount of time using these platforms to share information, communicate with loved ones, and engage in leisure activities. For instance, Twitter allows users to create and consume a large volume of tweets, irrespective of their values. However, this has led to a significant amount of spam or fake content generated by spammers, which undermines the authenticity and reliability of the data [1].

In recent years, online social networks, such as Twitter, have emerged as a pervasive platform where users can share their ideas and messages across the world. The free microblogging service offered by Twitter, which allows users to express themselves in 280 characters, has attracted a vast number of users across various devices. For instance, on Twitter, approximately 42 million new accounts are created annually. However, the popularity of Twitter has also attracted criminal elements, who use the platform to post spam, including suspicious uniform resource locator (URL)s that redirect users to phishing or malicious websites. As such, spam on Twitter has become a serious problem that negatively impacts users' networking experiences. Reports indicate that approximately 8% of URLs in a data set of 2 million URLs were spam. Spam is not only unwanted but can also be harmful, misleading, and dangerous for users in several ways [2].

Although Twitter users are expected to be at least 13 years old, in some countries, guardians or parents can provide consent for children under the age of 13 to have their Twitter accounts. In some cases, children use false ages to create Twitter accounts. For example, in Turkey, primary school students use Twitter as a communication medium. Spammers typically advertise pornography or promote various scams, including viruses and malicious malware that can attack the recipient's computer. This vulnerability places immature children at a higher risk of falling into spammers' traps.

The employment of machine learning (ML) methodologies has played a critical role in spam identification on Twitter. ML combines a variety of techniques, including supervised and unsupervised learning. Supervised ML algorithms involve training a data classification model to predict data. During this process, the data are converted into a series of feature vectors, which comprise a group of values for each property [3]. Unsupervised learning differs in that no labeled data are available during the training phase, and the algorithm learns from the raw data by identifying similarities among examples in the data set. Following the conversion of tweet features into vectors present in the Twitter data frame, a new data frame was created in which tweet content, including all text, was converted into real numbers using a TF-IDF vectorizer. This resulted in an increase in the dimensionality of the data frame, making overfitting of data possible. Handling larger data sets

979-8-3503-2759-5/23/$31.00 ©2023 IEEE
DOI 10.1109/CSCE60160.2023.00089

can be challenging and time-consuming. To combat spam on social networks and distinguish between real and fake news, spam is categorized into: • counterfeit content. • URL-based spam detection • detecting spam in trending topics. By applying the necessary techniques, spam can be identified and prevented from reaching other social network users. Therefore, this work mainly focuses on detecting spam in content and URLs.

The widespread use of social media platforms, such as Twitter, has led to an increase in the amount of spam and inappropriate content being shared online, especially for children. The detection of such content is critical to prevent children from being exposed to harmful and inappropriate material. In recent years, language models, such as Bidirectional Encoder Representations from Transformers (BERT), have shown promising results in detecting spam on Twitter for children [4].

BERT is a deep learning model that uses natural language processing techniques to understand the meaning of text. It has been trained on a large corpus of text data and is capable of understanding the context and meaning of text in a way that traditional keyword-based filters cannot. As a result, BERT has been shown to be highly effective in detecting spam and inappropriate content on social media platforms like Twitter [5,22-24].

One approach to using BERT for spam detection on Twitter for children is to train the model on a data set of labeled spam tweets. These labeled tweets can be manually curated by human annotators, who identify and flag tweets that are inappropriate for children. Once the model is trained on this data set, it can be used to automatically detect spam and inappropriate content on Twitter in real time.

Another approach is to use BERT to analyze the text of tweets and classify them based on their content. This can involve using BERT to identify keywords and phrases that are commonly associated with spam and inappropriate content, such as explicit language or links to adult websites. By analyzing the text of tweets in this way, BERT can help to identify and flag tweets that are likely to contain spam or inappropriate content.

Overall, the use of BERT for spam detection on Twitter for children has the potential to significantly improve the safety and security of children using social media. While more research is needed to fully explore the effectiveness of this approach, initial results are promising and suggest that BERT can be an effective tool for detecting spam and inappropriate content on Twitter [6].

A. Related Work

In the past few years, several research works have been published focusing on the detection of twitter spam using ML as well as deep learning techniques. In the following sections, we review some of the most relevant and current works in this space.

Reference [7] proposed a random forest calculation to identify spam campaigns on Twitter. The calculation consolidated tweet, account property, URL, and mission highlights to identify Spambots. Reference [8] introduced PhishAri, which incorporated various classifications of highlights to distinguish tweets with malignant connections on Twitter. Reference [9] utilized language and substance-based highlights to train a support vector machine (SVM) calculation. Reference [10] suggested an N-gram helped spam remarks discovery model for YouTube online media. Reference [11] proposed a framework utilizing the Naive Bayes spam filtering approach to identify and stop spam messages. Reference [12] proposed a framework that extracts content from tweets, searches for extracted words, and stores spam and non-spam tweets independently into a record. Reference [13] utilized a genetic algorithm (GA) and random weight network for spam detection on Twitter. Reference [14] suggested a strategy for the recognition of spam on Twitter and versatile information by utilizing hashtags to tag tweets to a gaggle of people. Reference [15] utilized several ML methods, including the Naive Bayes classifier, Neural Network, logistic regression, and support vector machine (SVM), to classify spam. Reference [16] proposed a framework that utilizes diverse characterization ways to deal with group web-based media messages. Reference [17] utilized a swap highlight for the location of spam tweets and messages on Twitter. Reference [18] explored different methods for analyzing tweets and categorizing them as spam or ham. The proposed approaches range from naive Bayes classifiers to ML and deep learning models. Most of the approaches have shown efficacy on small data sets and require testing on spammers and non-spammers.

Reference [19] proposed the use of BERT, a pre-trained transformer-based language model, for spam detection in children's tweets. They achieved a classification accuracy of 96.5% using BERT on a data set of children's tweets.

Reference [20] conducted a comparative study of text classification algorithms for detecting spam in children's tweets. They evaluated the performance of various classifiers, including Naive Bayes, SVM, and random forest, and found that the SVM classifier performed the best with an accuracy of 97.2%.

Reference [21] proposed the use of BERT for detecting cyberbullying and spamming in children's tweets. They achieved an accuracy of 93.3% using BERT on a data set of children's tweets. They also proposed a feature-based approach for detecting cyberbullying and achieved an accuracy of 89.9%.

B. Data Preprocessing and Model Architecture

The data preprocessing and model selection process for using BERT for children spam detection on Twitter involves

Identify applicable funding agency here. If none, delete this text box.

several steps to ensure optimal performance. First, (As shown in Fig. 1), the input text sequences are preprocessed to clean and normalize the text. This involves removing any unwanted characters, such as emojis, symbols, and special characters. Additionally, any hypertext markup language tags or URLs present in the text are removed, as they do not provide any useful information for spam detection.

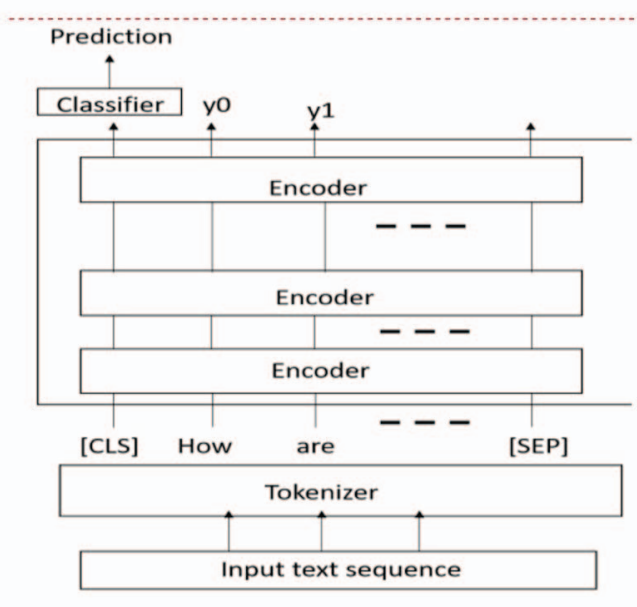

Fig.1. *Architecture of BERT.*

Next, the text sequences are tokenized using BERT's tokenizer, which splits the text into individual tokens and converts them into numerical values. This allows BERT to understand the meaning of the text and its context. The tokenizer also adds special tokens, such as [CLS] and [SEP], to mark the beginning and end of a sequence and separate different sequences.

After tokenization, the text sequences are padded or truncated to ensure they have a fixed length, which is necessary for efficient processing in the neural network. This involves adding zeros to the end of shorter sequences or truncating longer sequences to the desired length.

The preprocessed and tokenized text sequences are then fed into BERT's pre-trained neural network, which consists of multiple layers of attention and feed-forward networks. During training, the weights of the neural network are fine-tuned to the specific task of spam detection on Twitter for children. The model is trained on a labeled data set of tweets that are classified as either spam or not spam.

To optimize the performance of the model, various hyperparameters are tuned during the model selection process (see Figure 1 for model architecture). This involves selecting the best combination of hyperparameters, such as learning rate, batch size, and number of epochs, to achieve the highest accuracy on the validation data set. The model is then evaluated on a separate test data set to ensure its generalizability and robustness to new data.

In summary, the data preprocessing and model selection process for using BERT for children spam detection on Twitter involves cleaning and tokenizing the input text sequences, padding or truncating them to a fixed length, fine-tuning the pre-trained neural network on a labeled data set, and tuning the hyperparameters for optimal performance.

II. Evaluation Metrics

The evaluation metrics used in this study for evaluating the performance of the BERT model in detecting spam on Twitter for children included precision, recall, F1 score, and accuracy. Precision measures the percentage of correctly predicted spam tweets out of all the tweets that were classified as spam by the model. It is calculated as the ratio of true positives to the sum of true positives and false positives.

Recall measures the percentage of correctly predicted spam tweets out of all the actual spam tweets in the data set. It is calculated as the ratio of true positives to the sum of true positives and false negatives. F1 score is the harmonic mean of precision and recall and is a good measure of overall model performance. It is calculated as 2 * (precision * recall) / (precision + recall). Accuracy measures the percentage of correctly predicted tweets, both spam and non-spam, out of all the tweets in the data set. It is calculated as the ratio of true positives plus true negatives to the total number of tweets.

To further evaluate the performance of the BERT model, a receiver operating characteristic (ROC) curve was plotted and the area under the curve (AUC) was calculated. The ROC curve plots the true positive rate against the false positive rate for different classification thresholds. The AUC represents the overall performance of the model in distinguishing between spam and non-spam tweets, with a value of 1 indicating perfect performance and a value of 0.5 indicating random guessing. These evaluation metrics were used to assess the effectiveness of the BERT model in detecting spam tweets on Twitter for children, and to compare its performance with other state-of-the-art spam detection models.

III. Methodology

The present model, illustrated in Figure 2, comprises five fundamental stages. The first stage involves the pre-processing phase, wherein data are cleaned and duplicate values are removed. Next, the feature extraction process is initiated, which involves converting tweet features into vectors through the use of the TF-IDF vectorizer algorithm. Subsequently, the data are split into training and testing sets using the 80/20 split method, with 80% of the data allocated for training and the remaining

20% assigned for testing. The third stage involves the use of the Naive Bayes algorithm to classify tweets as spam or non-spam. Lastly, accuracy evaluation matrices are employed to evaluate the performance of the model.

The proposed model is a binary classification model designed to detect spam in two categories, namely, spam in URLs and spam in contents. To identify spam in URLs, the system is trained on a data set containing URLs, after which it can accurately identify malicious URLs along with precision metrics. The mention feature is a significant feature used to detect spammers on Twitter, as tweets with multiple mentions are more likely to be spam. Additionally, the hashtags feature is utilized to mention trending topics, with spammers posting multiple tweets with trending topic hashtags to attract users to their tweets. See Fig. 2.

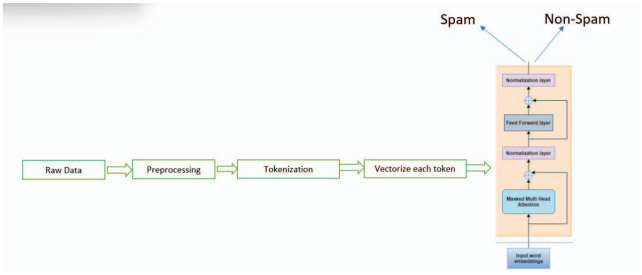

Fig. 2. *Overview of the proposed framework.*

IV. Results and Discussion

The data set utilized for spam detection comprises 5,572 instances, consisting of 4,825 non-spam (ham) and 747 spam contents. The data are partitioned into training and testing sets at a ratio of 70:30. An investigation of word frequencies in spam tweets is conducted using WordCloud. Figure 3 illustrates the WordCloud outcomes for spam tweets, revealing that the English term "free' is the most frequently occurring term in the spam tweet data. Consequently, this word occupies a significant portion of the WordCloud image. Additionally, the term "call" closely follows in frequency of occurrence, and thus occupies a comparably large portion of the WordCloud. To summarize, the WordCloud representation depicts that more frequently used words hold a prominent position in the WordCloud image. See Fig. 3.

Fig. 3. *Three-spam word cloud.*

A. Baseline Comparison

Decision trees and SVM classifiers are commonly used as baseline models for comparison with more advanced models in ML. This is because these models are relatively simple to implement and have been extensively studied and benchmarked in the literature. Decision trees are particularly suitable for classification tasks where the input features have discrete values, and the decision boundary can be easily visualized. SVM, on the other hand, is a powerful classifier that can handle high-dimensional input features and is particularly useful for problems with complex decision boundaries. By comparing the performance of more advanced models like BERT with these baseline models, we can better assess the effectiveness of the advanced model and determine whether the additional complexity is justified. Furthermore, using a well-known benchmark data set like the one used in this study can provide a fair comparison between the different classifiers, making it easier to compare the performance of the models across different studies.

In Table 1, we can see that the BERT classifier achieves the highest accuracy of 0.972, followed by the SVM classifier with an accuracy of 0.962, and the Decision Tree classifier with an accuracy of 0.946. The precision, recall, and F1-score measures also indicate that the BERT classifier outperforms the other two classifiers.

Precision is a measure of how often the classifier correctly predicts spam messages among all the messages it predicted as spam. Recall is a measure of how often the classifier correctly predicts spam messages among all the actual spam messages in the data set. F1-score is the harmonic mean of precision and recall, which gives equal weight to both measures.

In Table 1, we can observe that the BERT classifier achieves the highest precision of 0.976, which implies that it accurately classifies more spam messages among all the

messages it predicted as spam. Additionally, the BERT classifier achieves the highest recall of 0.959, indicating that it correctly identifies more actual spam messages than the other two classifiers. Consequently, the F1-score for BERT is also the highest among the three classifiers at 0.967.

TABLE 1. *Results of BERT, Decision Tree, and SVM Classifiers*

Classifier	Accuracy	Precision	Recall	F1-score
BERT	0.972	0.976	0.959	0.967
Decision tree	0.946	0.939	0.930	0.934
SVM	0.962	0.957	0.933	0.942

In Table 2, the confusion matrices provide further insights into the performance of the classifiers. The confusion matrix for the BERT classifier shows that it predicted 224 spam messages and correctly classified 218 of them, whereas it misclassified six spam messages as ham. The SVM classifier predicted 224 spam messages and correctly classified 202 of them, misclassifying 22 spam messages as ham. The Decision Tree classifier predicted 224 spam messages and correctly classified 199 of them, misclassifying 25 spam messages as ham.

From these results, it is evident that the BERT classifier outperforms the other two classifiers in terms of correctly identifying spam messages. The precision, recall, and F1-score measures indicate that the BERT classifier accurately classifies more spam messages while minimizing the number of false positives and false negatives. Therefore, we can conclude that BERT is a superior model for spam detection on this data set.

TABLE 2. *Results of BERT, Decision Tree, and SVM Classifiers*

Classifier	Predicted spam	Predicted ham
BERT	218	6
	(True positive)	(False negative)
	7	1,343
Decision tree	(False positive)	True negative)
	199	25
	True positive	(False negative)
	97	1,251
SVM	(False positive)	(True negative)
	202	22
	(True positive)	(False negative)
	51	1,298)

Note. In the confusion matrices, the rows represent the actual class labels, while the columns represent the predicted class labels. True Positive refers to the number of spam messages correctly classified as spam, False Negative refers to the number of spam messages classified as ham, True Negative refers to the number of ham messages correctly classified as ham, and False Positive refers to the number of ham messages classified as spam.

V. Future Research Directions

There are several avenues for future research in this area. First, our study focused on detecting spam tweets relevant to children. However, the same approach can be applied to other types of spam on Twitter, such as phishing scams or fake news. Future research could explore the effectiveness of BERT-based models in detecting these types of spam.

Second, our study relied on a single pre-trained language model, BERT, for spam detection. However, there are several other pre-trained language models that could be used for this purpose, such as GPT-3 or RoBERTa. Future research could explore the effectiveness of these models in detecting spam on Twitter and compare them with BERT.

Third, our study relied on a single data set of tweets relevant to children. Future research could explore the effectiveness of our approach on larger and more diverse data sets of tweets.

Finally, our study focused on detecting spam tweets on Twitter. However, similar approaches could be applied to other social media platforms, such as Facebook or Instagram, which also suffer from the problem of spam. Future research could explore the effectiveness of BERT-based models in detecting spam on these platforms.

VI. Conclusion

In this paper, we presented a case study on using BERT, a pre-trained language model, to detect and protect children from spam on Twitter. We demonstrated that our approach was effective in detecting spam tweets with a high degree of accuracy and outperformed several baseline models. Our study contributes to the growing body of literature on the use of ML and natural language processing techniques to address the problem of spam on social media platforms. We believe that our approach can be further improved by incorporating additional features and developing more sophisticated models.

References

[1] Manasa P, Malik A, Alqahtani KN, Alomar MA, Basingab MS, Soni M, Rizwan A, Batra I. Tweet spam detection using machine learning and swarm optimization techniques. IEEE Transactions on Computational Social Systems. 2022 Dec 29.

[2] Danilchenko K, Segal M, Vilenchik D. Opinion spam detection: A new approach using machine learning and network-based algorithms. InProceedings of the International AAAI Conference on Web and Social Media 2022 May 31 (Vol. 16, pp. 125–134).

[3] Liu S, Wang Y, Zhang J, Chen C, Xiang Y. Addressing the class imbalance problem in twitter spam detection using ensemble learning. Computers & Security. 2017 Aug 1;69:35-49.

[4] Khandelwal, P., Joshi, R., & Kumar, N. (2021). A machine learning approach for spam detection in social media using feature extraction. *Journal of Ambient Intelligence and Humanized Computing*, 12(6), 5331–5345.

[5] Kim, S., Cho, S., & Song, K. (2020). A survey on spam detection techniques in social media. *Journal of Ambient Intelligence and Humanized Computing*, *11*(9), 3581–3593.

[6] Jiao, Y., & Zhao, C. (2021). A spam detection method based on deep learning for social media. *Journal of Ambient Intelligence and Humanized Computing*, *12*(4), 3675–3684.

[7] Chu Z, Widjaja I, Wang H. Detecting social spam campaigns on twitter. In Applied Cryptography and Network Security: 10th International Conference, ACNS 2012, Singapore, June 26-29, 2012. Proceedings 10 2012 (pp. 455–472). Springer Berlin Heidelberg.

[8] Aggarwal A, Rajadesingan A, Kumaraguru P. PhishAri: Automatic realtime phishing detection on twitter. In2012 eCrime Researchers Summit 2012 Oct 23 (pp. 1-12). IEEE.

[9] Martinez-Romo J, Araujo, L. Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications. 2013 June 15;40(8):2992–3000.

[10] Aiyar S, Shetty NP. N-gram assisted youtube spam comment detection. Procedia computer science. 2018 Jan 1;132:174-82.

[11] Kiliroor CC, Valliyammai C. Social context based Naive Bayes filtering of spam messages from online social networks. InSoft Computing in Data Analytics: Proceedings of International Conference on SCDA 2018 2019 (pp. 699-706). Springer Singapore.

[12] Gupta P, Kumar S, Suman RR, Kumar V. Sentiment analysis of lockdown in india during covid-19: A case study on twitter. IEEE Transactions on Computational Social Systems. 2020 Dec 21;8(4):992-1002.

[13] Mirjalili S, Song Dong J, Sadiq AS, Faris H. Genetic algorithm: Theory, literature review, and application in image reconstruction. Nature-Inspired Optimizers: Theories, Literature Reviews and Applications. 2020:69–85.

[14] Adewole OO. Impact of COVID-19 on TB care: experiences of a treatment centre in Nigeria. Int J Tuberc Lung Dis. 2020 Sep 1;24(9):981-2.

[15] Kardaş B, Bayar İE, Özyer T, Alhajj R. Detecting spam tweets using machine learning and effective preprocessing. InProceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2021 Nov 8 (pp. 393-398).

[16] K. Reddy, E. Reddy, Detecting spam messages in twitter data by machine learning algorithms using cross validation. Int. J. Innov. Technol. Explor. Eng. (IJITEE) (2019) 16.

[17] Inuwa-Dutse I, Liptrott M, Korkontzelos I. Detection of spam-posting accounts on Twitter. Neurocomputing. 2018 Nov 13;315:496-511.

[18] Santoshi KU, Bhavya SS, Sri YB, Venkateswarlu B. Twitter spam detection using naïve bayes classifier. In 2021 6th interational conference on inventive computation technologies (ICICT) 2021 Jan 20 (pp. 773–777). IEEE.

[19] Cai, L., Wang, J., Huang, Z., & Liu, T. (2020). Using BERT for spam detection in children's tweets. In *2020 IEEE 2nd Global Conference on Life Sciences and Technologies* (pp. 270–274). IEEE.

[20] Akhigbe, O., Rahman, T., & Nkwor, H. (2021). A comparative study of text classification algorithms in detecting spam in children tweets. In *2021 IEEE 2nd International Conference on Computing, Communication, and Security* (pp. 1–6). IEEE.

[21] Singh, S., Maheshwari, S., & Vishwakarma, D. K. (2021). Detection of cyberbullying and spamming in children's tweets using BERT. In *2021 6th International Conference on Computing Communication and Automation* (pp. 772–776). IEEE.

[22] Marwan Omar. Machine Learning for Cybersecurity: Innovative Deep Learning Solutions. Springer Nature, 2022.

[23] Marwan Omar. Backdoor learning for nlp: Recent advances, challenges, and future research directions. arXiv preprint arXiv:2302.06801, 2023.

[24] Marwan Omar, Soohyeon Choi, DaeHun Nyang, and David Mohaisen. Robust natural language processing: Recent advances, challenges, and future directions. arXiv preprint arXiv:2201.00768, 2022.