MATHEMATICAL MODELING AND COMPUTING, Vol. 12, No. 3, pp. 972–981 (2025)

Advanced text-based transformer architecture for malicious social bots detection

Ellaky Z., Benabbou F.

Laboratory of Information Technology and Modeling,
Hassan II University of Casablanca Faculty of Sciences Ben M'Sick, Casablanca, Morocco

(Received 28 March 2025; Revised 28 September 2025; Accepted 29 September 2025)

The increasing prevalence of automated social media accounts, or Social Media Bots (SMBs), presents significant challenges in maintaining authentic online discourse and preventing disinformation campaigns on social platforms. This research introduces a novel multiclass classification framework for detecting and categorizing SMBs, leveraging fine-tuned transformer-based models. In this study, we conducted a comprehensive comparative analysis of various transformer variants, including BERT, DistilBERT, RoBERTa, DeBERTa, XLNet, and ALBERT, to evaluate their efficacy in recognizing diverse types of social bots, such as spambots, politically motivated SMBs, Sybil-type accounts, fraudulent and fake accounts, as well as legitimate human users. The empirical findings indicate that the proposed methodology substantially outperforms traditional machine learning and deep learning approaches. Notably, the DistilBERT architecture demonstrated exceptional performance metrics, achieving 96.83% accuracy and 96.85% precision in social bot classification.

Keywords: *social bots; detection; multi-classification; deep learning; transformers; fine-tuning.*

2010 MSC: 68T01, 68T35, 68T50, 91D30 **DOI:** 10.23939/mmc2025.03.972

1. Introduction

Social media platforms are changing people's lives. They are revolutionizing the way people talk to each other and share information. However, rapid digitization is experiencing a growing threat from the proliferation of SMBs. These automated accounts act like real people, spreading misinformation and manipulating online discourse. The proliferation of SMBs creates a multidimensional issue that impacts the trustworthiness of the information faced by Online Social Networks (OSN) stakeholders. In general, SMBs are software programs that perform tasks on OSN to mimic human behaviors. Some of these may be benign but can provide customer services and news. However, malicious SMBs can engage in malicious behaviors that serve malign purposes, from spreading false news to promoting extremist ideologies, committing scams, and conducting political manipulation [1]. Detecting SMBs is critical in protecting the integrity of OSN users' digital interactions and safeguarding the veracity of information encountered online. There exist varied types of SMBs, each designed for a specific type of usage: 1) Spambots: those that pester end-users with spam content, usually promotion messages or advertising links [2]; 2) misinformation bots: these bots are designed to conduct narratives online through automatic commenting or replying to create an impact or affect the public opinion [3]; 3) Cyborgs and Sybils: they impersonate real individuals and generate fake popularity and influence, often with bad intentions in mind [4]; Political Bots: These circulate biased or false information about a particular political viewpoint or ideology [3]. Identifying SMBs ensures a more transparent, trustworthy, and secure online environment. The identification of SMBs presents a crucial challenge due to their continual evolution of strategies to evade detection systems, sophisticated SMBs exhibit increasingly human-like behaviors, complicating their differentiation from authentic users. The primary motivation for proposing a multi classification system for detecting social bots stems from the rapid increase in SMBs, which threaten the integrity and security of online platforms. Traditional classification methods fall short in addressing these complex behaviors. By employing a multi classification system with transformers pre-trained on extensive textual data, we can improve bot detection and gain deeper insights into their characteristics and strategies. Pre-trained transformers, like Bidirectional Encoder Representations from Transformers (BERT), are powerful tools for understanding textual data and distinguishing between human-generated and bot-generated content. They process large-scale datasets efficiently, making them ideal for social media. By fine-tuning transformers, we can adapt them to the domain of social detection, which could improve the accuracy, scalability, and robustness of systems. This research introduces a novel and unique approach to classify various types of social bots beyond the traditional classification of humans versus bots. By employing a multi-class classification method, the proposed system aims to improve the security and integrity of social media platforms. The primary contributions of this research include:

- Leveraged a large dataset for multi-classification with five categories: human, Sybil, fake account, political bots, and spambots, incorporating academic datasets Cresci-2017 [5], Cresci-2015 [6], political bots [3], human accounts, and Twibot-20 [7];
- Addressed potential bias from imbalanced data using the SMOTE technique;
- Fine-tuning six pre-trained transformer models using text-based data, including BERT, DistilBERT, RoBERTa, DeBERTa, ALBERT, and XLNet;
- Evaluated model performance using metrics such as Recall, Precision, Accuracy, and F1-score.

Based on these key elements, this approach provides a valuable contribution to the field of social media security, aiming to develop robust deep-learning models to tackle social bots. The sections of this paper are structured as follows. Section 2 presents a comprehensive review of existing social media bot detection literature, examining various detection techniques. A comparative analysis of state-of-the-art approaches is provided in Section 3. The foundational concepts of transformers are introduced in Section 4. Section 5 outlines the proposed social media bot detection methodology, detailing the model development process and addressing associated challenges. An empirical evaluation of the proposed approach is presented in Section 6. Finally, Section 7 offers concluding remarks and directions for future research.

2. Related works

DeepSBD [8] was implemented to identify Twitter social bots. This approach utilized a Convolutional Neural Network (CNN) in combination with Global Vectors for Word Representation (GloVe) and a Bidirectional Long-Short-Term Memory neural network (BiLSTM). The classification model was developed using five Twitter datasets that incorporated features related to user information, content, network, and temporal patterns. The model demonstrated a bot detection accuracy rate of 97% on Twitter. In a study by [9], a system for detecting social media bots on Twitter was developed using BiLSTM and GloVe word embeddings. The research utilized the Cresci-2017 dataset, encompassing user information (UI), content (CF), network, and temporal features. The proposed model achieved a 92.9% accuracy in identifying Twitter social media bots. While the approach effectively captures semantic relationships within content features, its use has been limited to spam bot detection. Another study [10] focused on detecting fake news related to COVID-19, leveraging BERT for feature extraction and training a feedforward neural network (FFNN) on the Cresci-2017 dataset achieving 86% accuracy. Similarly, research [11] explored the detection of fake Instagram users using five machine-learning algorithms, with the Random Forest model achieving 91.76% accuracy based on UI and CF. Further, an SMB detection system [12] employing FFNN with only UI features from the Cresci-2017 dataset demonstrated a 94% detection accuracy. Combining ELMO and GloVe word embeddings benefited this approach, enhancing the detection of text-based social bots. A notable study [13] combined CNN and LSTM models to analyze textual content and address SMB manipulation on Twitter, particularly after the 2016 US elections. This ensemble approach achieved a 97% accuracy. The BGSRD model [14], integrated Graph Convolutional Networks (GCN) and graph neural networks (GNN), to address misinformation and malicious content. It employed BERT representation datasets and achieved an 80% accuracy, excelling in detecting social bots with minimal content length. The DeeProBot [15], using GloVe word embeddings and LSTMs, was applied to datasets like Midterm-18 and Cresci-Rtbust, achieving an AUC rate of 97%. A study [16] examined vaccine misinformation detection during the COVID-19 pandemic, with BERT outperforming XGBoost and LSTM models, achieving an F1-score of 98%. A system [17] based on the BiLSTM model used tweet content and user metadata and achieved 97% performance across metrics and 99% recall. The study [18] presented a novel method to identify Twitter bots using Deep Learning techniques, focusing on multilingual capabilities. It uses Multilingual Language Models to generate text-based features from user accounts, which are processed using Bot-DenseNet architecture. The model achieved an F1-score of 77%, indicating potential for multilingual bot detection systems. This research [19] examined the use of UI and images to analyze tweets. By converting digital DNA sequences into 3D images, the study employs pre-trained models to enhance the analysis. The proposed methodology is multimodal, integrating TwHIN-BERT for textual representation and VGG16 for visual representation. The study [20] employed GloVe and BiGRU to classify automated accounts on Twitter. The proposed model attained a Precision of 100% and an Accuracy of 99.73% on the Twibot20 dataset.

3. Comparative study and analysis

This section presents a comparative analysis of methodologies employed by scholars in classifying. The primary objective is to examine research approaches within this domain. Key criteria have been developed to evaluate the studied articles: 1) Ref: Sources and citations of the reviewed articles. 2) Algorithm: The classification algorithms utilized in each study. 3) Features: The characteristics extracted from social media accounts for classification, such as UI, CF, NF, BF, and TF. 4) Dataset: The data collection used to train and assess the classification models. 5) Embedding: Methods to convert text data into numerical vectors. 6) Classification: or multi-classification. 7) Results: Each model's performance is measured using evaluation metrics, such as Accuracy (A), Precision (P), and Recall (R). By systematically analyzing these criteria, we can understand the strengths and limitations of the current methods used for classifying social media accounts. A detailed comparison of the related work methods is presented in Table 1.

Table 1. Comparative table of the related work methods.

Ref.	Algorithm	Features	Embedding	Classification Results
[8]	BiLSTM, CNN, Attention	CF, TF, UI, NF, BF	GloVe	A: 97%
[9]	BiLSTM	CF, TF, UI, NF, BF	GloVe	A: 92%
[10]	FFNN	CF	BERT	A: 86%
[11]	RF	UI, CF	-	A: 91.76%
[12]	FFNN	UI, CF	GloVe, Elmo	A: 94%
[13]	CNN & LSTM-Attention	CF	BERT	A: 97%
[14]	GNN, GCN	CF, NF	BERT	A: 80%
[15]	DNN and LSTM	CF, UI	GloVe	AUC: 97%
[16]	BERT Transformer	CF	-	F1-score: 97%
[17]	BiLSTM	CF, UI	R	99%
[18]	MLM	CF	RoBERTa	F1-Score: 77%
[19]	Digital DNA	CF	TwiBERT	A: 99.89%
[20]	BiGRU, LSTM	CF	GloVe	P: 100%

An examination of the state-of-the-art articles reveals that the majority of SMB detection models for social media platforms relied on data from Twitter (82%), followed by Sina Weibo (12%) and Instagram (6%). Regarding features, the attributes employed in these models are content and user information. The feature distribution employed in SMB detection research, with CF (95%) dominating, followed by UI (58%), NF (21%), and TF (11%). These findings align with a recent systematic review [21] that identified CF and UI are the impactful features for SMB detection. AI-driven models predominantly leverage deep learning (52%) and neural network-based techniques (29%). ML methods account for the remaining 19%. Within the deep learning paradigm, LSTM and CNN are the most frequently utilized algorithms, with FFNN being the primary neural network architecture. Random Forest is the most used algorithm for machine learning models, achieving a peak accuracy of 99% on the Cresci-2017 dataset. CNN, LSTM, and BiLSTM are the cornerstone algorithms within deep learning for SMB detection. Among the various datasets employed for detecting SMBs, Cresci2017 is widely utilized. However, it does not cover diverse examples of SMBs, and it is limited to spambots. This makes it crucial to provide a pivotal resource for researchers aiming to advance methodologies in identifying all types of automated social media accounts. The GloVe embedding technique is widely used for SMB detection. Deep learning-based methodologies for social media bot detection include LSTM and CNN. FFNN is widely used, with RF being the most prevalent. LSTM models have an F1-score of 97%, while BiLSTM has a recall rate of 99%. Hybrid models combining LSTM and CNN with attention mechanisms and BiGRUs have shown superior performance, achieving 99.44% accuracy. Word embeddings, such as BERT and GloVe, contribute to improved detection capabilities. The examination of the related works unveiled limitations, particularly concerning text presentation, generalization, and multi-classification challenges.

4. Background techniques

4.1. Transformers

Transformers represent a deep learning architecture that has significantly advanced the field of Natural Language Processing (NLP). Their efficacy is particularly notable in handling and generating sequential data, including machine translation, text summarization, and language modeling tasks. The principal innovation of transformers lies in their implementation of attention mechanisms. These mechanisms enable the model to focus selectively on relevant segments of the input sequence when producing each output token, distinguishing them from earlier architectures like RNNs and LSTM networks, which process sequences sequentially. Transformer-based models have markedly advanced the field of NLP. BERT, developed by Google researchers [22], leverages the encoder component of the Transformer architecture. Its principal innovation is the bidirectional attention mechanism, which allows it to interpret words within the full context of a sentence, supported by a sophisticated вЂњmulti-head attentionвЂќ mechanism [23]. BERT is pre-trained on extensive unlabeled text corpora and subsequently fine-tuned for specific NLP tasks. Various adaptations of BERT, such as RoBERTa [24] and DistilBERT [25], transformer-based models have significantly enhanced language understanding and versatility, setting new benchmarks in the field of NLP.

4.2. Architecture

While the core architecture of the pre-trained model remains intact, modifications are typically made to the final layers or some intermediate layers to tailor the model to the target task. These adjustments usually involve adding or replacing layers to suit the specific requirements of the new task, with these layers generally being smaller and designed for the task at hand. A transformer model comprises an encoder and a decoder, each consisting of multiple attention layers and feed-forward neural networks. The encoder processes the input sequence to generate a contextual representation for each token, considering its relation to all other tokens. The decoder utilizes this contextual representation and the previously generated tokens to predict the subsequent token in the sequence. These components can be employed independently based on the specific task requirements: 1) Encoder-only models: are suited for tasks that necessitate comprehension of the input, such as sentence classification and named entity recognition; 2) Decoder-only models: are effective for generative tasks, including text generation; 3) Encoder-decoder models: known as sequence-to-sequence models, are appropriate for generative tasks that involve an input, such as translation or summarization. Another fundamental characteristic of the Transformer model is Attention layers that focus on specific words in a sentence, such as “"You like this course” or “this” in French. The French conjugation of “like” depends on the subject, and the model must consider the noun’s gender. This principle applies to any NLP task, as individual words’ contextual significance is influenced by surrounding words.

4.3. Fine-tuning pre-trained models

Fine-tuning involves additional training conducted on a model that has already undergone pretraining. This process begins with a pre-trained language model, subsequently refined using a dataset specific to the target task. Rather than training a model from scratch, the rationale for fine-tuning includes several key considerations: 1) Pretraining Knowledge: The pre-trained model has already been exposed to a large dataset that shares similarities with the fine-tuning dataset. This prior training provides the model with a foundational understanding of the language or domain, which can be leveraged during fine-tuning; 2) Data Efficiency: A pre-trained model requires significantly less data during the fine-tuning phase to achieve effective results due to its initial training on extensive data; 3) Cost: Fine-tuning a pre-trained model is generally less time consuming and resource-intensive than training a model from scratch. It allows for quicker iterations and adjustments in the training process, making it a more practical approach.

5. Proposed methodology

In this research study, the methodology emphasizes using multi-task learning (MTL) as an advanced machine learning strategy [26]. MTL involves training a single model to address multiple tasks concurrently, which enhances generalization and minimizes the need for extensive datasets for each task. The models are trained to classify social accounts into human, spambot, Sybil, fake accounts, or political SMBs. Figure 1 illustrates the proposed methodology, including 1) Data collection, 2) Labeling and Merging Data, 3) Preparing data for Training, 4) Setup of the Training Arguments, 5) Fine-Tuning Transformers, 6) Test and Validation, and 7) Results (see Figure 1).

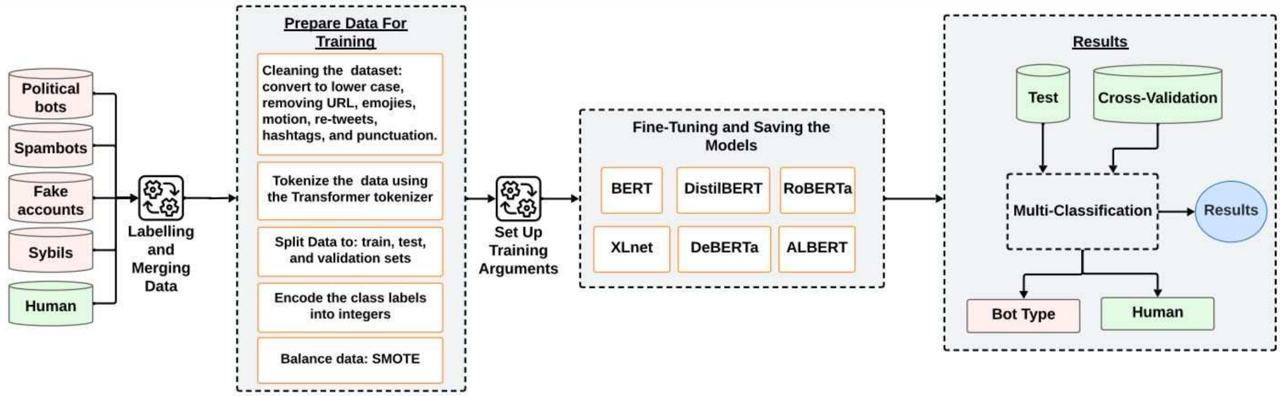
Political bots
Spambots
Fake accounts
Sybils
Human
Labelling and Merging Data
Prepare Data For Training
Cleaning the dataset: convert to lower case, removing URL, emojies, motion, re-tweets, hashtags, and punctuation.
Tokenize the data using the Transformer tokenizer
Split Data to: train, test, and validation sets
Encode the class labels into integers
Balance data: SMOTE
Set Up Training Arguments
Fine-Tuning and Saving the Models
BERT
DistilBERT
RoBERTa
XLnet
DeBERTa
ALBERT
Results
Test
Cross-Validation
Multi-Classification
Results
Bot Type
Human

Fig. 1. The methodology proposed for multi-classification of SMB.

5.1. Data collection

The data collection process leveraged five distinct datasets. 1) Cresci-2015: This dataset was utilized to identify Sybil accounts, a type of social bot known for artificially inflating influence by creating multiple fake identities [27]; 2) Cresci-2017: Employed to classify spambots, it includes accounts designed to propagate spam and engaged in spam-related activities [5]; 3) Twibot-20: An essential resource to extract recent bots accounts [7]; 4) Fake Followers: This dataset was used to identify and categorize fake accounts, often employed to create deceptive appearances of social influence [28]; 5) Political bots: This dataset was employed to detect SMBs engaged in political discussion and misinformation [3]; 6) Human Accounts: human accounts were extracted from the Cresci-2017 and Twibot-20 to ensure a diverse representation. The overall approach involved filtering and preprocessing these datasets to create a cohesive and balanced dataset of human and bot accounts.

5.2. Labeling and merging data

The final dataset is constructed by merging text-based data from Cresci-2015, Cresci2017, Twibot-20, Fake Followers, and human accounts to form a unified representation encompassing Sybil, spam, fake, political, and human accounts. A standardized labeling scheme was implemented, assigning numerical identifiers to each category. Subsequently, a mapping was established between these numerical labels and their corresponding textual descriptions for reference and interpretation. Figure 2 plots the labels distributions of our final dataset.

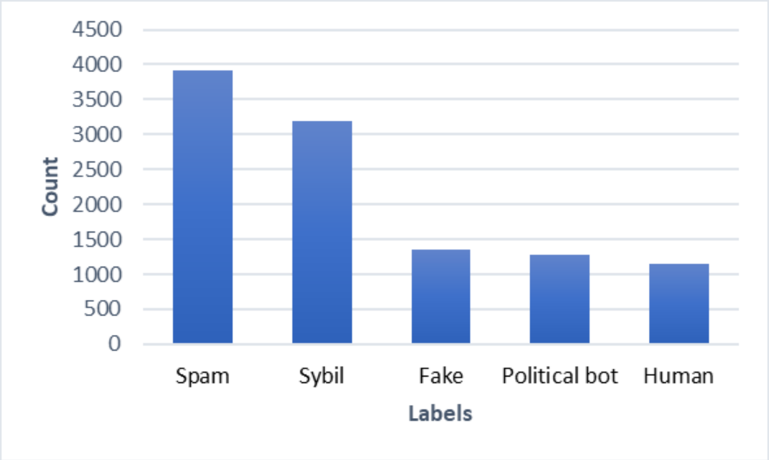

Fig. 2. The classes distributions of the final dataset.

Figure 2 shows that the dataset exhibits a notable class imbalance, particularly with an over representation of spam and Sybil classes. This imbalance may result in model bias and suboptimal performance on less frequent classes. Implementing suitable balancing technique to address this issue can enhance model effectiveness.

5.3. Preparing data for the training

To prepare the data for training using the Hugging Face Transformers library we followed these steps: 1) Loading the Dataset: Converting data into the Hugging Face Dataset format to handle large-scale datasets and get access to advanced preprocessing capabilities and accelerating model training; 2) Tokenization: segmenting text into individual tokens, incorporating tokens, and standardizing sequence length through padding or truncation; 3) Data Formatting: To optimize transformer model training, data are formatted in a TensorFlow-compatible structure; 4) Split data: The dataset is divided into training, validation, and test sets. A common partitioning strategy is a 7020%, and 10% ratio for training, validation, and test sets. 5) Balancing data: Given the imbalance within the training dataset (as shown in Figure 2), the Synthetic Minority Oversampling Technique (SMOTE) combined with Edited Nearest Neighbors (SMOTE-ENN) was applied to mitigate this challenge. This approach was selected based on prior experimental studies we conducted [21], where SMOTE-ENN demonstrated its efficacy in handling imbalanced data, reducing overfitting, enhancing model performance, and narrowing disparities between evaluation metrics.

5.4. Fine-tuning pre-trained transformers

Due to their robust architectures, Bert, DistilBERT, ROBETA, DeBERTa, ALBERT, and XLNet were selected as foundational models. BERT leverages bidirectional transformers to comprehensively capture contextual information, while RoBERTa refines BERT's architecture through hyperparameter optimization and architectural simplifications for enhanced performance [19]. DeBERTa introduces disentangled attention mechanisms and advanced training strategies to improve model capabilities further.

6. Experiment and results

This section comprehensively evaluates fine-tuned transformer architectures for the multi-class classification of social media bots. Six transformer models were fine-tuned, and the implementation, training, and evaluation were conducted using the Hugging Face Transformers platform.

6.1. Evaluation metrics

The research used accuracy, precision, recall, and F1-score as evaluation metrics. Accuracy measures the proportion of correct predictions across all classes, while precision measures the proportion of True Positive (TP) predictions. Precision measures the proportion of TP predictions out of all instances predicted as positive. It is beneficial when focusing on minimizing false positives (FP). Recall quantifies the proportion of positive cases that the model correctly identifies. The F1-score is the harmonic mean of precision and recall and is particularly valuable for evaluating models on imbalanced datasets [20].

6.2. Model architecture and fine-tuning

The selected transformer models underwent hyper-parameter tuning to adapt their parameters to the social media bot detection task. We employed a transfer learning paradigm, leveraging pre-trained language models as a foundation and adjusting model To optimize the performance on our dataset, we adjusted the weights. The main hyper parameters are:

- Evaluation_strategy : Epoch,
- Save_strategy : Epoch,
- Learning_rate : 2×10^{-5},
- Num_train_epochs : 10,
- Weight_decay : 0.01,
- Warmup_steps : 500,
- Gradient_accumulation_steps : 2,
- Per_device_train_batch_size : 16,
- Per_device_eval_batch_size : 16,
- Logging_steps : 100,
- Save_steps : 1000,
- Save_total_limit : 2.

The results achieved through fine-tuning various transformer architectures for the multi-class classification of SMBs offer insights into the efficacy of different model configurations and hyperparameter settings. Table 2 provides the performance achieved by the fine-tuned transformers.

Table 2. Fine-tuning results.

Transformer	Accuracy	Precision	Recall	F1-score
BERT	96.29%	96.25%	96.20%	96.22%
DistilBERT	96.83%	96.85%	96.83%	96.84%
RoBERTa	61.20%	61.25%	61.20%	61.22%
DeBERTa	59.33%	59.36%	59.33%	59.34%
XLNet	95%	95.10%	95%	95.05%
ALBERT	56.13%	56.15%	56.13%	56.14%

The findings detailed in Table 2 regarding transformer effectiveness in classifying social media bots indicate that DistilBERT is the most proficient model, achieving an accuracy, precision, recall, and F1-score over 96.80%. Models such as BERT and XLNet also demonstrate good performances. Nevertheless, there is a notable disparity between the leading models and others like RoBERTa, DeBERTa, and ALBERT. The differences in performance can be attributed to factors like model architecture, the quality of pre-training data, fine-tuning hyperparameters, and the characteristics of the dataset. Table 3 presents the results achieved by DistilBERT for each class. This would provide a more comprehensive understanding of how well the model performs on each specific social bot type.

Table 3. The Classification results achieved by DistilBERT for each class.

Class	Accuracy	Precision	Recall	F1-Score
Spam	100%	98%	100%	99%
Fake	99%	86%	99%	92%
Political Bot	81%	99%	81%	89%
Human	98%	97%	98%	98%
Sybil	100%	100%	100%	100%

The provided classification report indicates strong performance across all classes, with high accuracy, precision, recall, and F1-scores. This suggests that the model is effective in classifying different types of social media accounts, including human users and various types of bots.

7. Conclusion

In conclusion, this study highlights the significant role that transformer-based models play in detecting and classifying SMBs, which is a growing concern for safeguarding the integrity of online platforms. By fine-tuning models such as BERT, DistilBERT, RoBERTa, DeBERTa, ALBERT, and XLNet, the research demonstrates the superior effectiveness of transformer models in distinguishing between different types of social bots, outperforming traditional machine learning methods. The proposed multi-class classification framework not only enhances accuracy but also provides a robust solution to improve social media security based on the generated text data by users. For future directions, of this research, we suggest investigating alternative multi-model architectures, optimizing hyperparameters, and incorporating diverse data types. Given the continuous evolution of SMB tactics, it will be essential to adopt a multi-model approach that integrates various attributes, including images and multilingual text data. Furthermore, regularly updating and retraining these models will be crucial in staying ahead of new bot strategies. Methods such as feature importance analysis and attention mechanisms can also offer deeper insights into the model's decision-making process. In addition, this study demonstrates the effectiveness of detecting SMBs using Twitter data; however, it encounters limitations in generalizability to other social media platforms due to their differing content formats and user interactions. Future research should focus on developing strategies independent of specific platforms by exploring the unique characteristics of each platform, designing adjusted feature engineering techniques, and employing transfer learning to extend insights gained from Twitter to other platforms. To validate the robustness and generalizability of any proposed approach, rigorous empirical testing is essential across a diverse range of datasets.

[1] Ellaky Z., Benabbou F., Ouahabi S., Sael N. Word Embedding for Social Bot Detection Systems. 2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS). 1–8 (2021).

[2] Ellaky Z., Benabbou F., Ouahabi S., Sael N. A Survey of Spam Bots Detection in Online Social Networks. 2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA). 58–65 (2021).

[3] Ellaky Z., Benabbou F. Political social media bot detection: Unveiling cutting-edge feature selection and engineering strategies in machine learning model development. Scientific African. **25**, e02269 (2024).

[4] Goyal B., Gill N. S., Gulia P., Prakash O., Priyadarshini I., Sharma R., Obaid A. J., Yadav K. Detection of Fake Accounts on Social Media Using Multimodal Data With Deep Learning. IEEE Transactions on Computational Social Systems. 1–12 (2024).

[5] Cresci S., Pietro R. D., Petrocchi M., Spognardi A., Tesconi M. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion. 963–972 (2017).

[6] Benabbou F., Boukhouima H., Sael N. Fake accounts detection system based on bidirectional gated recurrent unit neural network International Journal of Electrical and Computer Engineering. **12** (3), 3129 (2022).

[7] Feng S., Wan H., Wang N., Li J., Luo M. TwiBot-20: A Comprehensive Twitter Bot Detection Benchmark. CIKM '21: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 4485–4494 (2021).

[8] Fazil M., Sah A. K., Abulaish M. DeepSBD: A Deep Neural Network Model With Attention Mechanism for SocialBot Detection. IEEE Transactions on Information Forensics and Security. **16**, 4211–4223 (2021).

[9] Wei F., Nguyen U. T. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. 2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). 101–109 (2019).

[10] Heidari M., Zad S., Hajibabaee P., Malekzadeh M., HekmatiAthar S., Uzuner O., Jones J. H. Bert model for fake news detection based on social bot activities in the Covid-19 pandemic. 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). 0103–0109 (2021).

[11] Purba K. R., Asirvatham D., Murugesan R. K. Classification of instagram fake users using supervised machine learning algorithms. International Journal of Electrical and Computer Engineering. **10** (3), 2763–2772 (2020).

[12] Heidari M., Jones J. H., Uzuner O. Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter. 2020 International Conference on Data Mining Workshops (ICDMW). 480–487 (2020).

[13] Kumar S., Garg S., Vats Y., Parihar A. S. Content Based Bot Detection using Bot Language Model and BERT Embeddings. 2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP). 285–289 (2021).

[14] Guo Q., Xie H., Li Y., Ma W., Zhang C. Social bots detection via fusing bert and graph convolutional networks. Symmetry. **14** (1), 30 (2021).

[15] Hayawi K., Mathew S., Venugopal N., Masud M. M., Ho P.-H. DeeProBot: a hybrid deep neural network model for social bot detection based on user profile data. Social Network Analysis and Mining. **12** (1), 43 (2022).

[16] Hayawi K., Shahriar S., Serhani M. A., Taleb I., Mathew S. S. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. Public Health. **203**, 23–30 (2022).

[17] Messai A., Hamida Z. F., Drif A., Giordano S. Multi-input BiLSTM deep learning model for social bot detection. 2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS). 1–6 (2023).

[18] Martin-Gutierrez D., Hernandez-Penaloza G., Hernandez A. B., Lozano-Diez A., Alvarez F. A Deep Learning Approach for Robust Detection of Bots in Twitter Using Transformers. IEEE Access. **9**, 54591–54601 (2021).

[19] Ilias L., Kazelidis I. M., Askounis D. Multimodal Detection of Bots on X (Twitter) Using Transformers. IEEE Transactions on Information Forensics and Security. **19**, 7320–7334 (2024).

[20] Ellaky Z., Benabbou F., Matrane Y., Qaqa S. A Hybrid Deep Learning Architecture for Social Media Bots Detection Based on BiGRU-LSTM and GloVe Word Embedding. IEEE Access. **12**, 100278–100294 (2024).

[21] Ellaky Z., Benabbou F., Ouahabi S. Systematic Literature Review of Social Media Bots Detection Systems. Journal of King Saud University – Computer and Information Sciences. **35** (5), 101551 (2023).

[22] Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Preprint arXiv:1810.04805 (2018).

[23] Rothman D. Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more. Packt Publishing Ltd (2021).

[24] Delobelle P., Winters T., Berendt B. RobBERT: a Dutch RoBERTa-based Language Model. Findings of the Association for Computational Linguistics: EMNLP 2020. 3255–3265 (2020).

[25] Sanh V., Debut L., Chaumond J., Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Preprint arXiv:1910.01108 (2019).

[26] Pujari S. C., Friedrich A., Strotgen J. A Multi-task Approach to Neural Multi-label Hierarchical Patent Classification Using Transformers. Advances in Information Retrieval. 513–528 (2021).

[27] Cresci S., Pietro R. D., Petrocchi M., Spognardi A., Tesconi M. Fame for sale: Efficient detection of fake Twitter followers. Decision Support Systems. **80**, 56–71 (2015).

[28] Yang K., Varol O., Davis C. A., Ferrara E., Flammini A., Menczer F. Arming the public with artificial intelligence to counter social bots. Human Behavior and Emerging Technologies. **1** (1), 48–61 (2019).

Розширена текстова архітектура трансформера для виявлення шкідливих соціальних ботів

Еллакі З., Бенаббу Ф.

Лабораторія інформаційних технологій та моделювання,
Університет Хасана II Касабланки, Факультет наук Бен М'Сік, Касабланка, Марокко

Зростаюча поширеність автоматизованих облікових записів у соціальних мережах, або ботів соціальних мереж (SMB), створює значні проблеми для підтримки автентичного онлайн-дискурсу та запобігання дезінформаційним кампаніям на соціальних платформах. Це дослідження представляє нову багатокласову систему класифікації для виявлення та категоризації SMB, використовуючи точно налаштовані моделі на основі трансформерів. У цьому дослідженні проведено комплексний порівняльний аналіз різних варіантів трансформерів, включаючи BERT, DistilBERT, RoBERTa, DeBERTa, XLNet та ALBERT, щоб оцінити їхню ефективність у розпізнаванні різних типів соціальних ботів, таких як спам-боти, політично мотивовані SMB, облікові записи типу Sybil, шахрайські та підроблені облікові записи, а також легітимних користувачів-людей. Емпіричні результати показують, що запропонована методологія значно перевершує традиційні підходи машинного навчання та глибокого навчання. Зокрема, архітектура DistilBERT продемонструвала виняткові показники продуктивності, досягнувши точності 96.83% та точності 96.85% у класифікації соціальних ботів.

Ключові слова: *соціальні боти; виявлення; мультикласифікація; глибоке навчання; трансформери; точне налаштування.*