International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 3, June 2022, pp. 3129~3137
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i3.pp3129-3137

Fake accounts detection system based on bidirectional gated recurrent unit neural network

Faouzia Benabbou, Hanane Boukhouima, Nawal Sael
Faculty of Sciences Ben M'sick, University Hassan II of Casablanca, Casablanca, Morocco

Article Info

Article history:

Received Oct 20, 2020
Revised Jan 16, 2022
Accepted Jan 31, 2022

Keywords:

Bidirectional gated recurrent unit
Convolutional neural networks
Fake account
GloVe
Long short-term memory
Twitter
Word2vec

ABSTRACT

Online social networks have become the most widely used medium to interact with friends and family, share news and important events or publish daily activities. However, this growing popularity has made social networks a target for suspicious exploitation such as the spreading of misleading or malicious information, making them less reliable and less trustworthy. In this paper, a fake account detection system based on the bidirectional gated recurrent unit (BiGRU) model is proposed. The focus has been on the content of users' tweets to classify twitter user profile as legitimate or fake. Tweets are gathered in a single file and are transformed into a vector space using the global vectors (GloVe) word embedding technique in order to preserve the semantic and syntax context. Compared with the baseline models such as long short-term memory (LSTM) and convolutional neural networks (CNN), the results are promising and confirm that using GloVe with BiGRU classifier outperforms with 99.44% for accuracy and 99.25% for precision. To prove the efficiency of our approach the results obtained with GloVe were compared to Word2vec under the same conditions. Results confirm that GloVe with BiGRU classifier performs the best results for detection of fake Twitter accounts using only tweets content feature.

Corresponding Author:

Faouzia Benabbou
Faculty of sciences Ben M'Sick, University Hassan II of Casablanca
Cdt Driss el harti road, B.P: 7955, Ben M'Sick, Casablanca, Morocco
Email: faouzia.benabbou@univh2c.ma

1. INTRODUCTION

Currently, online social networks (OSNs) have a very important role in the daily life of Internet users to carry out their daily actions such as reading news, sharing content, product reviews, posting messages, and discussing events. However, popular social networks are sometimes misused and bring new risks in terms of trust, security, and privacy. Fake accounts are often created to share with genuine users misleading information such as spams, malware, harmful uniform resource locator (URLs), or unsolicited messages. The purpose may be to influence public opinion [1], to manipulate elections [2], to spread rumors [3], to impact the stock market, to influence search engine results, to purchase social media followers, and to destroy real users' reputation. The owner of a fake account can be a person, bot, organization, or company who does not actually exist.

OSNs user profile usually includes two main parts: static data such as the name profile, gender, birthday, picture, while the dynamic data includes a user's activities in the social network. A fake account contains false information, whether it is personal information or information about followers, friends, and comments. In fact, they are designed for a non-legitimate purpose in order to alter opinions with rumors and concepts such as popularity and influence, which could have a significant impact on the economy, politics, and society. With the dangerous character of these accounts and their negative and unfair impact, it is

necessary to design new methodologies to identify and characterize fake accounts. To tackle this problem, many researches have been conducted on different OSNs as Twitter, Facebook, YouTube, LinkedIn, and Weibo. Different tracks have been identified and several solutions have been proposed for fake profile detection in order to classify an account in "fake" and "legitimate" accounts or more deeply as proposed in the work of Wani and Jabin which proposed 5 classes of fake account type [4]: compromise account, cloned profiles, sockpuppets account, sybil account and bot-user account.

Gurajala *et al.* [5] proposed a system to identify a group of fake profiles based on matched profile attributes, screen names similarity and update-time of tweets. User and content features were analyzed for each account and the Shannon entropy and standard deviation were used to decide whether an account belongs to a group. Content features included tweet content, number of duplicate tweets, and tweet postdate. They proved that analyzing temporal behavior of users can reveal interesting indicator to identify fake accounts. Wani *et al.* [6] applied support vector machine (SVM), decision tree (DTree), artificial neural network (ANN), Naïve Bayes (NBayes) and AdaBoost classifiers to predict fake profiles on Facebook social network. Before classification, artificial bee colony (ABC) and ant colony optimization (ACO) techniques have been used in feature selection step. The study confirms that AdaBoost classifier is more efficient when the number of accounts in training dataset increases, but the authors did not provide the features used. Torkyl *et al.* [7] proposed a detection mechanism called fake profiles recognizer (FPR) for detecting fake profiles in OSNs. The methodology is based on the functionality of regular expression (social graph) and deterministic finite automaton (DFA). The friend's list feature is used to define friend pattern and to detect duplicated profile. The proposed detection mechanism achieved 94.93% for accuracy on twitter dataset. Erşahin *et al.* [8] used NBayes classifier for detecting the fake accounts on Twitter social network. They selected sixteen attributes from user and content features and the accuracy achieved is 90.9% against 86% without discretization. Gupta and Kaushal collected their dataset using the Facebook API [9]. The final dataset consisted mainly of seventeen attributes from activity view. A total of twelve supervised machine learning techniques were applied to the dataset and the study showed that «likes» and comments features contribute well to fake account detection task. However, the accuracy of 79% achieved seems not sufficient. Khaled *et al.* [10] proposed a new algorithm SVM-NN which is a hybrid model of SVM and neural network to detect fake Twitter accounts and bots. The principal component analysis (PCA) technique was applied to reduce dimension. The approach relied only on user details (user logs and profiles) and reached an accuracy of 98.3% when using correlation feature selection method.

Recently, Purba *et al.* [11] proposed a system for Instagram fake user's detection based on activity features, and random forest (RForest) outperforms others classifiers with an accuracy score of 91.76%. Swe and Myo used the approach based on user and content feature sets to detect fake accounts on Twitter [12]. Machine learning classifiers such as meta-learner diverse ensemble creation by oppositional relabeling of artificial training examples (DECORATE), RForest, AdaBoost, DTree, and NBayes classifiers were applied to classify fake accounts on Twitter. DECORATE classifier achieved the best accuracy and the detection rate was of 95.1%. In another work, Swe and Myo used a blacklist [13] of words instead of spam words list. Blacklist is created by using topic modeling approach and keyword extraction approach. The user's tweets are deeply analyzed by using topic modeling and bag of word approaches. The term frequency-inverse document frequency (TF-IDF) techniques is applied to extract the most important words to be added in a blacklist. Albayati and Altamimi used data mining techniques to detect fake profiles on Facebook [14]. A set of supervised and unsupervised algorithms were applied to twelve behavioral and non-behavioral discriminative profile attributes from a dataset. The results shown that supervised algorithms outperformed unsupervised algorithms and ID3 had the highest accuracy while k-medoids had the lowest performance. In another paper Albayati and Altamimi proposed FBChecker system to detect fake profiles on Facebook [15]. A set of supervised algorithms (DTree, k-NN, SVM, and NBayes) were applied to the crawled dataset and the focus was on user features. The proposed system shown high efficiency performance for detecting fake profiles with an accuracy rate of 98%. Adikari and Dutta [16] shown that despite the limited dataset size, SVM with polynomial kernel using PCA-selected features performs an accuracy of 87%. Wanda and Jie [17] proposed DeepProfile, a deep neural network (DNN) algorithm, to deal with fake account issues. Notably, they realized a novel pooling layer WalkPool in the hidden layer to optimize the performance and achieved an area under curve (AUC) of 0.9547 which is good result. The features used related to content are message content and URLs and the technique was applied on Facebook dataset.

Zheng *et al.* [18] proposed a spammer detection system for Sina Weibo social networks. They used user features and content features as the number of mentions, URLs, and hashtags. The SVM classifier, trained and tested with 5-fold cross-validation, reached a good performance with 99.1% true positive rate of spammers and 99.9% non-spammers. The top 10 feature ranking list obtained from information gain (Igain) includes number of created days, comments count, URLs count, and fraction of followers per followers. For the same purpose Zhu *et al.* [19] proposed a spammer detection based on logistic regression attribute and

behavior logistic regression (ABLR) over Twitter and Sina Weibo social networks. They considered content and behavior features of users in social network and account features (e.g., the number of friends, followers, and tweets) and they obtained an accuracy rate of 90%. Al-Zoubi *et al*. [20] applied four machine learning classifiers using user and content features for detecting spammers on Twitter. The most influencing features in spam profiles detection was identified with two methods ReliefF and IGain. The experiments shown that NBayes classifier produced the best accuracy of 95.7%. In another paper, with the view of recognizing spammers in Twitter, Al-Zoubi *et al*. [21] presented a hybrid model based on whale optimization algorithm (WOA) and SVM. They used user, content, and activity features. The approach analyzed the linguistic context and its impact on the system performance and the results shown a considerable efficiency for Arabic context. Alom *et al*. [22] extracted a new set of features to detect spammers on Twitter. They considered both graph-based, tweet content features and applied seven machine learning algorithms (k-NN, DTree, NBayes, RForest, logistic regression (LR), SVM, and XGBoost). In the experiment, RForest gives the better result compared to other algorithms, with an accuracy of 91%. The DeepScan proposed by Gong *et al*. [23] focused on user's activity evolution in continuous time intervals. They used time series features through long short-term memory (LSTM) neural network, over real data collected from Dianping, and achieved 0.964 for F1 score. Gong *et al*. [24] proposed another system based on GitHub developer communities using phased LSTM. GitSec distinguishes malicious accounts from legitimate ones based on the user, event, and dynamic activity characteristics. With CatBoost classifier, GitSec achieves an AUC of 0.940. Ahmed and Abulaich proposed an interesting spam detection system for Facebook and twitter social network [25] based on user, activities and contents features. IGain was applied to extract the most relevant features before performing three classifiers: NBayes, J48, and Jrip. They found that Jrip performed better on Twitter dataset with 0.987 for detection rate and no false positive cases. The result proved that activity features were more discriminative than tweet content ones.

Few researches conducted a deep analysis on posts contents instate the focus was on the number of words in the tweet, the number of tweets, URLs count, and mention count. Wu *et al*. [26] proposed a technique based on deep learning techniques for twitter spam detection. Word2Vec was applied to pre-process the tweets and the output is given as input to different classifiers and muli-layer perceptron (MLP) achieves the highest performance over all the four datasets with an accuracy of 94.3%. Madisetty and Desarkar [27] proposed an approach based on convolutional neural networks (CNN) combined with different word embeddings techniques as global vectors (GloVe) and Word2Vec. The features used in the model were related to user, content, and n-gram but for CNN model only the tweet contents were used. The performance was of 95.7% for accuracy and 88% for precision, which proves that the model can be improved. Jain *et al*. [28] implemented semantic long short-term memory (SLSTM) for spam classification in SMS and Twitter dataset. The tweet content was indexed with semantic word using WordNet corpus and ConceptNet. The tweets were transformed to a vector of concepts words by Word2Vec. In comparison with KNN, NBayes, RForest, ANN and SVM, SLSTM outperforms with 95.09% for accuracy and 95.54% for precision, which is quite satisfactory. Only the tweet contents were sufficient in this study to detect spammer account with high performance.

Feature selection is an important step for machine learning process that aims to find which set of features is more relevant in fake account detection. Rostani [29] showed that the approach behavior was different from different datasets and features. As an example, for one dataset URL count feature was not important, and for another mention count feature was not relevant which is a little bit confused. An analysis based N-Gram was conducted by Aiyar and Shetty [30] for spam comment detection on YouTube. They used only comment feature, and the result proved the effectiveness of Word-grams method with SVM classifier that performed 97.74% for F1 score. To summarize our finding, we propose a categorization of features used in different papers as shown in Table 1.

Account information such as title, and age are usually fixed. The personal information of the user and the information of the account itself have been separated into two classes: user and account. The other classes represent information that changes and evolves over time. In this category we have the class activity, content, deep content. The class activity reflects the interactions of the user like the mention "like", his friends and his followers. The content and deep content classes specify how the comments are analyzed: statistically or semantically. Deep content class involves the use of some techniques as word embedding that preserve the semantic or/and syntactic form of comments and the most popular techniques utilized were POS, N-Gram, Word2Vec, and GloVe. Analysis of content was widely based on statistical values from content characteristics. One or a combination of classes was used for the detection of fake accounts. In another hand, four classes of detection systems have been identified: fake account or profile, spammers, spam posts, and bots.

The most widely used algorithms are SVM, RForest, NBayes, DNN, DTree, FT, J48, ABLR and ANN. Small attention was paid to deep learning models excluding CNN and LSTM algorithms. However, other deep learning techniques remain interesting to explore as long as they have really proven their

efficiency in several classification problems. Some papers used feature selection methods to find the best features to rely on to detect a fake account and information gain, PCA, correlation, LR, SVM, ReliefF, ABC, and ACO were the most popular. Regarding the importance of characteristics in the detection of fake accounts, the analysis of the literature shows different and sometimes contradictory results depending on the dataset used. The performance of the methods was measured by different metrics as accuracy, precision, recall or F-measure and best accuracy results were up to 99% and performed by SVM, CNN and ensemble classifiers. Twitter social network is the most popular OSN studied ahead of Facebook, LinkedIn, YouTube, GitHub, Instagram, because of its openness and the ease of data extraction.

Based on literature review, we observe that deep learning techniques were not applied enough, only CNN and LSTM were performed. In recent works, the deep learning model bidirectional gated recurrent unit (BiGRU) was applied for the classification concern and performed a good result on different fields, such as dialogue intent classification, intrusion detection and text sentiment classification. On the other side, feature selection techniques applied in different studies showed sometimes opposite results for the relevance of features, we can say that it depends strongly on the dataset type. Using a high number of features did not automatically provide a high performance as showed in [8] and [16]. Other researches focused only on the deep content analysis with a deep learning classifier and outperformed the baselines [26]–[28]. In addition, content characteristics are analyzed from statistics view such as the number of URL, mentions and hashtags. Few papers were interested in semantic issue of comments to discover hidden behavioral patterns. Krombholz and al. showed that behavior of fake users, the amount and the type of information they share in their newsfeed are significantly different from that of the real user [31] which motivated us to further investigate the analysis of the comments and exclude the other features.

In this paper a novel approach based on BiGRU deep learning model is proposed to detect legitimate (LA) and fake account (FA) using deep tweet contents analysis. Based on our last work on fraud detection [32], we found that BiGRU is very convenient for classification problems because it has the advantage of capturing local and global contextual information's. The efforts were devoted to tweet contents analysis using GloVe word embedding to capture syntactic and semantic features of tweets. The results showed that using only tweets with GloVe word embedding, BiGRU reached a high performance compared with baseline literature.

Table 1. Category of features

Class of features	Features
User	Civility, profile age, profile image, follower count, following friends count, and statistic values.
Account information	Creation date, geolocation enabled or not, and verified or not.
Account activities	Like count, comments count, time between posts, update-time of posts, reply count, and statistic values, followers.
Content	Tweet count, total number of words in a tweet, hashtag count, mention count, URL count, numeric character count, and statistic values.
Deep content	All user comments

2. BACKGROUND OF THE APPROACH

2.1. Bidirectional gated recurrent unit

This section describes the background of BiGRU technique. The advantage of using deep learning techniques is its ability to learn abstract features while going through multiple hidden layers. GRU model [33] is a variant of LSTM [34], which synthesizes the forgetting gate and input gate to a single update gate and mixes cell state and hidden state. So, the final gated recurrent units (GRU) model is simpler and faster than the standard LSTM algorithm, especially when training big data. It can save a lot of time with small performance difference from that of standard LSTM model. Both LSTM and GRU can retain important features through various gates ensuring that these features will not be lost in long-term transmission. The internal structure of GRU model is shown in Figure 1.

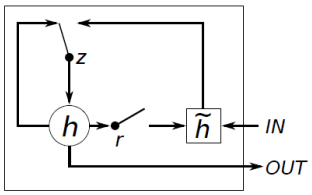

Figure 1. GRU cell architecture

Where Z_t represents update gate and $\mathbf{r_t}$ represents reset gate. At time t, the GRU calculates the new state as:

$$h_t = Z * h_{(t-1)} + (1 - Z) * \tilde{h} \quad (1)$$

This is to compute a linear interpolation between the previous state h_{t-1} and the current candidate state $\tilde{h}_t$ with the new sequence information. The update gate z_t decides to keep how much past information and to add how much new information. It controls the extent to which the information of the previous state is brought into the current state. The larger the value of the state of z_t, the more information of the previous state is brought in. The state of z_t is updated:

$$z_t = \sigma(W_Z . x_t + U_Z . h_{(t-1)} + b_Z) \quad (2)$$

where, x_t is the sample vector at time t and $\tilde{h}_t$ is the candidate state computed in the same way as the hidden layer of traditional RNN.

$$\tilde{h}_t = \tanh(W_h . x_t + r * U_h . h_{(t-1)} + b_h) \quad (3)$$

Where, r_t denotes a reset gate which controls how much the previous state contributes to the current candidate state $\tilde{h}_t$. The smaller the $r_t$value, the smaller the contribution from the previous state. If r_t=0, it will forget the previous state. The reset gate is updated:

$$r_t = \sigma(W_r . x_t + U_r . h_{(t-1)} + b_r) \quad (4)$$

For many sequence modelling tasks, it is beneficial to have access to future as well as past context. However, standard GRU networks processes sequences in temporal order and they ignore future context. Bidirectional GRU networks extend the unidirectional GRU networks by introducing a second layer, where the hidden to hidden connections flow in opposite temporal order. The model is therefore able to exploit information both from the past and the future. Also, GRUs address the vanishing gradient problem, through the use of two gates, the reset gate, and the update gate. Basically, these are two vectors which decide what information should be passed to the output and can be trained to retain information from farther back. This allows it to pass relevant information down a chain of events to make better predictions.

2.2. Global vectors for word representation (GloVe)

Word embedding is a semantic vector space that consists of a representation of a word as a vector of numerical values. The use of word embedding as a standard in language modeling is now prevalent because they can capture both the semantics and the syntactic context using vector arithmetic. The GloVe is a global log-bilinear regression model for the unsupervised learning of word representations [35] that outperforms other models on word analogy and word similarity. GloVe outperforms Word2vec [35], [36] because it captures both global and local co-occurrence counts in a corpus, unlike Word2Vec where only the local context is used. Thanks to GloVe, the tweet vector is formed by concatenating the individual word vectors of the tweet to obtain n*d tweet matrix where d is the dimension of word vector and n the length of the tweet.

3. THE PROPOSED METHOD

In this paper, we propose a fake account detection system in social networks based on BiGRU deep learning model. The OSN studied here is twitter but the technique remains applicable for the other social networks. This system used only comments feature and applied word embedding technique to preserve the context and syntax of comments. For each account, the content of tweets is gathered in a single document and is transformed into a vector space using GloVe. As described in Figure 2, the architecture of our system presents different layers: i) dataset preparation, ii) data preprocessing, iii) embedding, and iv) account classification. The main interest of our proposition is to detect fake account using only a deep analysis of user tweets. In the next sections these layers will be describes with more details.

3.1. Dataset preparation

We utilized a balanced dataset Twitter kindly shared by Cresci *et al.* [37]. The labeled dataset has 2818 accounts/users and their 2,126,962 tweets. As each user publishes one or more tweets, all the tweets of the same user were saved in one document. User who does not have tweets is excluded. The final distribution for our dataset is shown in Table 2.

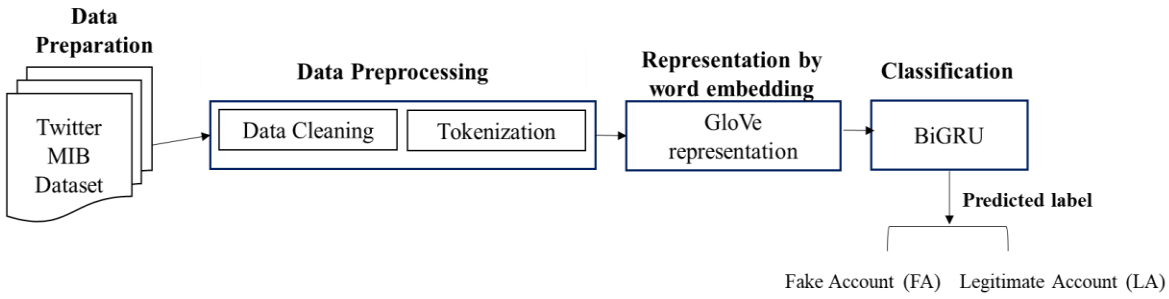

Figure 2. The proposed method for fake account detection

Table 2. Final distribution of the dataset

Fake Users	Legitimate Users	Number of instances
1325 (49.91 %)	1330 (50.09 %)	2655

3.2. Data preprocessing and embedding

The preprocessing step consists of data cleaning and tokenization of all tweets produced by users. We carried out the removal of punctuation and stop words (a, an, the, of, …) in tweets. After all characters are converted to lowercase, and URLs, mention (@username) and hashtag (#) are excluded from tweets. This features were excluded because many researches who were based on them [8], [9], [16] did not produce a high performance, and some works mentioned that content features are not significant [5], [25] or simply did not used them as in [10], [16], [17]. In this study, focus is on semantic and syntactic analysis of tweets provided by word embedding technics to find special patterns in text which can help to distinguish between fake or legitimate profile. For this purpose, the pre-trained word embeddings model GloVe was used. The process of data flow starts from the embedding layer where the data in the form of text was tokenized, then vectorized using the GloVe word embedding which are pre-trained semantic vectors having dimension of 100. This value was chosen after several experiments to find the best that would give the highest performance of our model.

3.3. BiGRU model parameters

A large number of parameters in the deep learning model could increase the overfitting potential. These problems can be overcome using dropping, in our case, we used SpatialDropout1D layer that performs the same function as dropout, however, it drops entire 1D feature maps instead of individual elements. Then a BiGRU layer with recurrent dropout and GlobalMaxPool1D are applied. Since we have a binary classification problem, we have labeled our data by fake accounts data (1) and legitimate accounts (0). To optimize the training network, we used Adam optimization algorithm which combines the best properties of RMSProp and AdaGrad, to cope with noisy or sparse datasets. With mini-batch size of thirty-two examples, the loss function used in our model was binary cross-entropy. The non-linear function rectified linear unit (ReLU) was used for hidden layers and since it is a binary classification problem, a sigmoid function for the output layer was used.

4. RESULTS AND DISCUSSION

In the experiment, we used a machine running Windows 10 with CORE i5 8 th gen processor, 12 GB of RAM, and 1,000 GB hard drive. Keras (2.2.4), Tensorflow (1.14.0) Python (Version 3.7.3), Jupyter (6.0.3) was used to implement our model.

4.1. Performance metrics

To evaluate the performance of our system, the following metrics were used: accuracy, precision, recall (sensitivity), F-measure and AUC. True positive (TP) is the number of fake accounts that are correctly classified as fake account, and false positive (FP) represents the amount of legitimate accounts that are wrongly labeled as fake. Conversely, true negative (TN) refers to the quantity of legitimate account which are exactly considered as legitimate, while false negative (FN) is the number of fake accounts wrongly predicted.

- $Precision = TP/(TP + FP)$: the fraction of detected fake accounts which are really fake.

- $Accuracy = (TP + TN)/(TP + FP + TN + FN)$: means the ratio of accounts identified correctly to all accounts.
- $Recall = TP/(TP + FN)$: the fraction of fake accounts who have been uncovered accurately.
- $F - Measure = 2x(PrecisionxRecall)/(Precision + Recall)$: the harmonic means of precision and recall.
- AUC: the probability that the classifier will rank a randomly chosen fake user higher than a randomly selected legitimate user.

The system was trained and tested using 5-Fold cross validation to split data into a five number of folds where each fold is used as a testing set at some point. The main advantage of data splitting is the robustness of the performance result and overfitting prevention.

4.2. Approach evaluation and comparison

In this section, results of our approach are presented and compared to other deep learning techniques like LSTM and CNN from the literature reviews. Table 3 shows the confusion matrix of the proposed approach using Twitter dataset. In order to confirm the performance of this approach based on BiGRU model, a comparison was conducted, on the same dataset for more rigor, between the baseline models LSTM, CNN, hybrid models LSTM+BiGRU and CNN+BiGRU as shown in the Table 4.

Table 3. Confusion matrix for BiGRU model

		Actual values	
		FA(+)	LA(-)
Predicted values	FA(+)	TP=264	FP=2
	LA(-)	FN=1	TN=264

Table 4. Comparison of model performance on the same dataset

Classifiers	Accuracy	Precision	Recall	AUC
BiGRU	99.44%	99.25%	99.62%	99.44%
LSTM	98.49%	97.41%	99.62%	98.49%
LSTM+BiGRU	98.68%	99.23%	98.11%	98.68%
CNN	98.87%	98.14%	99.62%	98.87%
CNN+BiGRU	98.87%	99.23%	98.49%	98.86%

The results show that BiGRU approach outperforms the other models in all performance measures. For accuracy, BiGRU reached 99.44% followed by CNN and CNN+BiGRU with 98.87% and LSTM+BiGRU with 98.68%, and the last score is 98.49%. For LSTM. The hybrid model based on CNN and BIGRU was not very efficient except for the precision score which has been improved to 99.23%. Also, for the hybrid model LSTM+BiGRU, the precision score has been improved to 99.23%. The proposed approach based on BiGRU for detection of fake Twitter accounts at tweet level using word embedding proves it effectiveness in detecting fake account in twitter social network. The results confirm that analysis of the content timeline of tweets reveals the behaviors of fake OSN user. In addition, Table 4 demonstrates that this model is competitive with state-of-the-art neural language models and outperforms the most relevant ones. The association of the GloVe technique and the BiGRU deep learning model leads to a more efficient system for the detection of fake accounts. However, the best training time achieved is from CNN with 5 min 3 s followed by the LSTM model with 9 min 55 s and finally the BiGRU model with 10 min 25 s.

In a second experiment, Word2vec was used as it represents the most commonly used technique in the papers studied and was compared with the results of this study 'approach. The performance presented in Table 5, confirm that our approach outperforms the approach based on semantic LSTM and Word2vec method over the same dataset MIB [28]. A same experiment with the same input dataset, using Word2vec and BiGRU provides an accuracy of 98.87% which is less than the performance reached using GloVe and BiGRU.

Table 5. Comparison with Word2Vec based approaches

Approach	Model	Word embedding	Accuracy
Our proposition	BiGRU	GloVe	99.44%
BIGRU+Word2vec	BiGRU	Word2Vec	98.87%
[28]	Semantic LSTM	Word2Vec	95.09%

5. CONCLUSION

In this paper, we proposed a system based on BiGRU model and pre-trained model GloVe for the task of fake Twitter accounts detection. This Approach focused on the tweet content level to extract the syntactic and semantic features of comments. The comparison results proved that this approach outperforms LSTM, CNN, LSTM+BiGRU and CNN+BiGRU models with an accuracy of 99.44%. Compared to Word2vec, GloVe captures both global and local features of text which explains why its performance is more significant. However, when it is important to categorize accounts, our system is not able to distinguish between different types of fake account. In a future work, a system that categorizes the different types of fake accounts by considering other features linked to user behavior and comments, to enhance and extend the proposed fake account detection system can be a research proposal project.

REFERENCES

[1] A. Alarifi, M. Alsaleh, and A. Al-Salman, "Twitter turing test: Identifying social machines," *Information Sciences*, vol. 372, pp. 332–346, Dec. 2016, doi: 10.1016/j.ins.2016.08.036.

[2] P. G. Pratama and N. A. Rakhmawati, "Social bot detection on 2019 Indonesia president candidate's supporter's tweets," *Procedia Computer Science*, vol. 161, pp. 813–820, 2019, doi: 10.1016/j.procs.2019.11.187.

[3] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018, doi: 10.1126/science.aap9559.

[4] M. A. Wani and S. Jabin, "A sneak into the devil's colony-fake profiles in online social networks," May 2017, eprint arXiv:1803.08810 [cs]. Accessed: Aug. 12 2020. [Online]. Available: http://arxiv.org/abs/1705.09929.

[5] S. Gurajala, J. S. White, B. Hudson, B. R. Voter, and J. N. Matthews, "Profile characteristics of fake Twitter accounts," *Big Data and Society*, vol. 3, no. 2, Art. no. 205395171667423, Dec. 2016, doi: 10.1177/2053951716674236.

[6] S. Y. Wani, M. M. Kirmani, and S. I. Ansarulla, "Prediction of fake profiles on Facebook using supervised machine learning techniques-a theoretical model," *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 7, no. 4, pp. 1735–1738, 2016.

[7] M. Torky, A. Meligy, and H. Ibrahim, "Recognizing fake identities in online social networks based on a finite automaton approach," in *2016 12th International Computer Engineering Conference (ICENCO)*, Dec. 2016, pp. 1–7, doi: 10.1109/ICENCO.2016.7856436.

[8] B. Ersahin, O. Aktas, D. Kilinc, and C. Akyol, "Twitter fake account detection," in *2017 International Conference on Computer Science and Engineering (UBMK)*, Oct. 2017, pp. 388–392, doi: 10.1109/UBMK.2017.8093420.

[9] A. Gupta and R. Kaushal, "Towards detecting fake user accounts in facebook," in *2017 ISEA Asia Security and Privacy (ISEASP)*, Jan. 2017, pp. 1–6, doi: 10.1109/ISEASP.2017.7976996.

[10] S. Khaled, N. El-Tazi, and H. M. O. Mokhtar, "Detecting fake accounts on social media," in *2018 IEEE International Conference on Big Data (Big Data)*, Dec. 2018, pp. 3672–3681, doi: 10.1109/BigData.2018.8621913.

[11] K. R. Purba, D. Asirvatham, and R. K. Murugesan, "Classification of instagram fake users using supervised machine learning algorithms," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 3, pp. 2763–2772, Jun. 2020, doi: 10.11591/ijece.v10i3.pp2763-2772.

[12] M. M. Swe and N. N. Myo, "Fake accounts classification on Twitter," *International Journal of Latest Engineering and Management Research (IJLEMR)*, vol. 3, no. 6, 2018.

[13] M. M. Swe and N. N. Myo, "Fake accounts detection on twitter using blacklist," in *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, Jun. 2018, pp. 562–566, doi: 10.1109/ICIS.2018.8466499.

[14] M. B. Albayati and A. M. Altamimi, "Identifying fake Facebook profiles using data mining techniques," *Journal of ICT Research and Applications*, vol. 13, no. 2, pp. 107–117, Sep. 2019, doi: 10.5614/itbj.ict.res.appl.2019.13.2.2.

[15] M. B. Albayati and A. M. Altamimi, "An empirical study for detecting fake Facebook profiles using supervised mining techniques," *Informatica*, vol. 43, no. 1, Mar. 2019, doi: 10.31449/inf.v43i1.2319.

[16] S. Adikari and K. Dutta, "Identifying fake profiles in linkedIn." Jun. 2020, eprint arXiv:2006.01381 [cs]. Accessed: Jun. 20 2020. [Online]. Available: http://arxiv.org/abs/2006.01381.

[17] P. Wanda and H. J. Jie, "DeepProfile: Finding fake profile in online social network using dynamic CNN," *Journal of Information Security and Applications*, vol. 52, p. 102465, Jun. 2020, doi: 10.1016/j.jisa.2020.102465.

[18] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing*, vol. 159, pp. 27–34, Jul. 2015, doi: 10.1016/j.neucom.2015.02.047.

[19] X. Zhu, Y. Nie, S. Jin, A. Li, and Y. Jia, "Spammer detection on Online social networks based on logistic regression," in *Web-Age Information Management*, Springer International Publishing, 2015, pp. 29–40.

[20] A. M. Al-Zoubi, J. Alqatawna, and H. Paris, "Spam profile detection in social networks based on public features," in *2017 8th International Conference on Information and Communication Systems (ICICS)*, Apr. 2017, pp. 130–135, doi: 10.1109/IACS.2017.7921959.

[21] A. M. Al-Zoubi, H. Faris, J. Alqatawna, and M. A. Hassonah, "Evolving support vector machines using whale optimization algorithm for spam profiles detection on online social networks in different lingual contexts," *Knowledge-Based Systems*, vol. 153, pp. 91–104, Aug. 2018, doi: 10.1016/j.knosys.2018.04.025.

[22] Z. Alom, B. Carminati, and E. Ferrari, "Detecting spam accounts on Twitter," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2018, pp. 1191–1198, doi: 10.1109/ASONAM.2018.8508495.

[23] Q. Gong *et al.*, "DeepScan: exploiting deep learning for malicious account detection in location-based social networks," *IEEE Communications Magazine*, vol. 56, no. 11, pp. 21–27, Nov. 2018, doi: 10.1109/MCOM.2018.1700575.

[24] Q. Gong *et al.*, "Detecting malicious accounts in online developer communities using deep learning," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, Nov. 2019, pp. 1251–1260, doi: 10.1145/3357384.3357971.

[25] F. Ahmed and M. Abulaish, "A generic statistical approach for spam detection in Online Social Networks," *Computer Communications*, vol. 36, no. 10–11, pp. 1120–1129, Jun. 2013, doi: 10.1016/j.comcom.2013.04.004.

[26] T. Wu, S. Liu, J. Zhang, and Y. Xiang, "Twitter spam detection based on deep learning," in *Proceedings of the Australasian Computer Science Week Multiconference*, Jan. 2017, pp. 1–8, doi: 10.1145/3014812.3014815.
[27] S. Madisetty and M. S. Desarkar, "A neural network-based ensemble approach for spam detection in Twitter," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 4, pp. 973–984, Dec. 2018, doi: 10.1109/TCSS.2018.2878852.
[28] G. Jain, M. Sharma, and B. Agarwal, "Optimizing semantic LSTM for spam detection," *International Journal of Information Technology*, vol. 11, no. 2, pp. 239–250, Jun. 2019, doi: 10.1007/s41870-018-0157-5.
[29] R. R. Rostami, "Detecting fake accounts on Twitter social network using multi-objective hybrid feature selection approach," *Webology*, vol. 17, no. 1, pp. 1–18, May 2020, doi: 10.14704/WEB/V17I1/a204.
[30] S. Aiyar and N. P. Shetty, "N-Gram assisted youtube spam comment detection," *Procedia Computer Science*, vol. 132, pp. 174–182, 2018, doi: 10.1016/j.procs.2018.05.181.
[31] K. Krombholz, D. Merkl, and E. Weippl, "Fake identities in social media: A case study on the sustainability of the Facebook business model," *Journal of Service Science Research*, vol. 4, no. 2, pp. 175–212, Dec. 2012, doi: 10.1007/s12927-012-0008-z.
[32] I. Sadgali, N. Sael, and F. Benabbou, "Bidirectional gated recurrent unit for improving classification in credit card fraud detection," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 21, no. 3, pp. 1704–1712, Mar. 2021, doi: 10.11591/ijeecs.v21.i3.pp1704-1712.
[33] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling", Dec. 2014, eprint arXiv:1412.3555 [cs.NE]. Accessed: Jul. 28 2020. [Online]. Available: http://arxiv.org/abs/1412.3555.
[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[35] J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.
[36] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. pp. 3111–3119, 2013.
[37] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots," in *Proceedings of the 26th International Conference on World Wide Web Companion- WW '17 Companion*, 2017, pp. 963–972, doi: 10.1145/3041021.3055135.

BIOGRAPHIES OF AUTHORS

Faouzia Benabbou is an Associate Professor of Computer Science and member of Compute Science and Information Processing laboratory of science Ben M'sick. She is Head of the team "Cloud Computing, Network and Systems Engineering (CCNSE)". Her research areas include cloud Computing, deep learning, machine learning, and Natural Language Processing. She can be contacted at email: faouzia.benabbou@univh2c.ma.

Hanane Boukhouima obtained a master's degree in Data Science and Big Data from Faculty of sciences Ben M'Sick, Morocco, in 2020. Currently she is preparing her PhD in Computer Science in Faculty of Science Ben M'sik. Her research interests include machine learning and OSNs analysis. She can be contacted at email: hanane.boukhouima@gmail.com.

Nawal Sael is an Associate Professor of Computer Science and member of Computer Science and Information Processing laboratory at faculty of science Ben M'sick. Here research interests include data mining, educational data mining, machine learning, deep learning, and Internet of things. She can be contacted at email: saelnawal@hotmail.com.