Bot detection using a single post on social media

Shad Mohammad
Department of Computer Science
COMSATS UI, Abbottabad Campus
Abbottabad, Pakistan
shadtn@gmail.com

Muhammad U.S. Khan
Department of Computer Science
COMSATS UI, Abbottabad Campus
Abbottabad, Pakistan
ushahid@cuiatd.edu.pk

Mazhar Ali
Department of Computer Science
COMSATS UI, Abbottabad Campus
Abbottabad, Pakistan
mazhar@cuiatd.edu.pk

Leo Liu
Department of Operations,
Technology, Events and
Hospitality Management
Manchester Metropolitan University
Manchester, UK
Leo.Liu@mmu.ac.uk

Matthew Shardlow
Department of Computing
and Maths
Manchester Metropolitan University
Manchester, UK
M.Shardlow@mmu.ac.uk

Raheel Nawaz
Department of Operations,
Technology, Events and
Hospitality Management
Manchester Metropolitan University
Manchester, UK
R.Nawaz@mmu.ac.uk

***Abstract*—Recent studies of social media have made a unanimous conclusion that public opinions can be altered through systematic exploitation of social media using bot accounts. The existing bot detection methodologies utilize features of the accounts to label them as either bot or human. However, in this work, we propose a convolutional neural network (CNN) to identify the bot accounts using a single post on the social media. We have compared our results with an artificial neural network (ANN) trained on the features extracted from the accounts' profiles. Results have shown that bot accounts can be detected with 98.71% accuracy using CNN as compared to the 97.6% of ANN. Moreover, we have also proposed a model that combine both the techniques and have achieved 99.43% accuracy.**

***Index Terms*—Bot detection, Neural networks, social media, cyber security**

I. Introduction

Social media platforms provide the medium for democratic conversations worldwide. Facebook and Twitter provide a forum for debates on civil movements, public health interventions, and elections throughout the world. However, these powerful media platforms have also been used for nefarious purposes, such as radical propaganda by extremist groups or the spread of disinformation such as the anti-vaccination campaign. Researchers have warned about a potential use of social media for political propaganda for more than a decade [1]. Recently, however, a number of incidents have been reported about interventions in elections with the help of social media accounts in USA, UK, and many other countries around the world. One common trait in these events is the use of automated tools to generate large volume of social media posts to support or attack campaigns [2].

The automated social media accounts, or *social bots*, are used and controlled by algorithms, instead of real people. These fake accounts post disinformation while interacting with real people. With the help of social contacts to real people and continuous streams of fake news, social bots influence opinions [2].

Recently, researchers have proposed a series of methods to detect social bots [3] [4] [5]. However, the major challenge to the research on social bots is the lack of ground truth data. The social media companies do not provide complete data for analysis and researchers rely on publicly available datasets. It has been reported that social media companies are also identifying and removing suspicious users. However, it is not clear on what basis they identify the users as a bot or not. Studies in the literature [6] have identified a number of profile features such as friends to followers ratio, number of retweets, number of tweets, favorites, and time of account creation to categorize an account as bot but checking each account is a huge task.

Text mining methods have been extensively used in literature to extract information [7] [8] [9] [10]. In this paper, we propose a novel method to detect bot accounts using the text content of their social media rather than identifying them from their profile features. To our knowledge, no previous attempt has been made to detect a bot using a single social media post. We have used word vectors and a convolutional neural network (CNN) on the publicly available Twitter bot dataset [11] to classify the text as originated from a bot or a human. The technique is similar to the one that identify the sentiment in text [12]. We have compared our technique with one that utilizes the artificial neural network (ANN) trained on profiles' features. Results have shown that utilizing word vectors and CNN for identifying the bots provide better accuracy than using the profiles' features in the ANN. Moreover, we have also proposed a methodology to combine both the techniques to achieve even better accuracy. The main contribution of the paper are as follows:

- A novel convolutional neural network to detect a bot account using a single social media post
- The performance of the CNN based model is compared with the ANN based model that is trained on profile

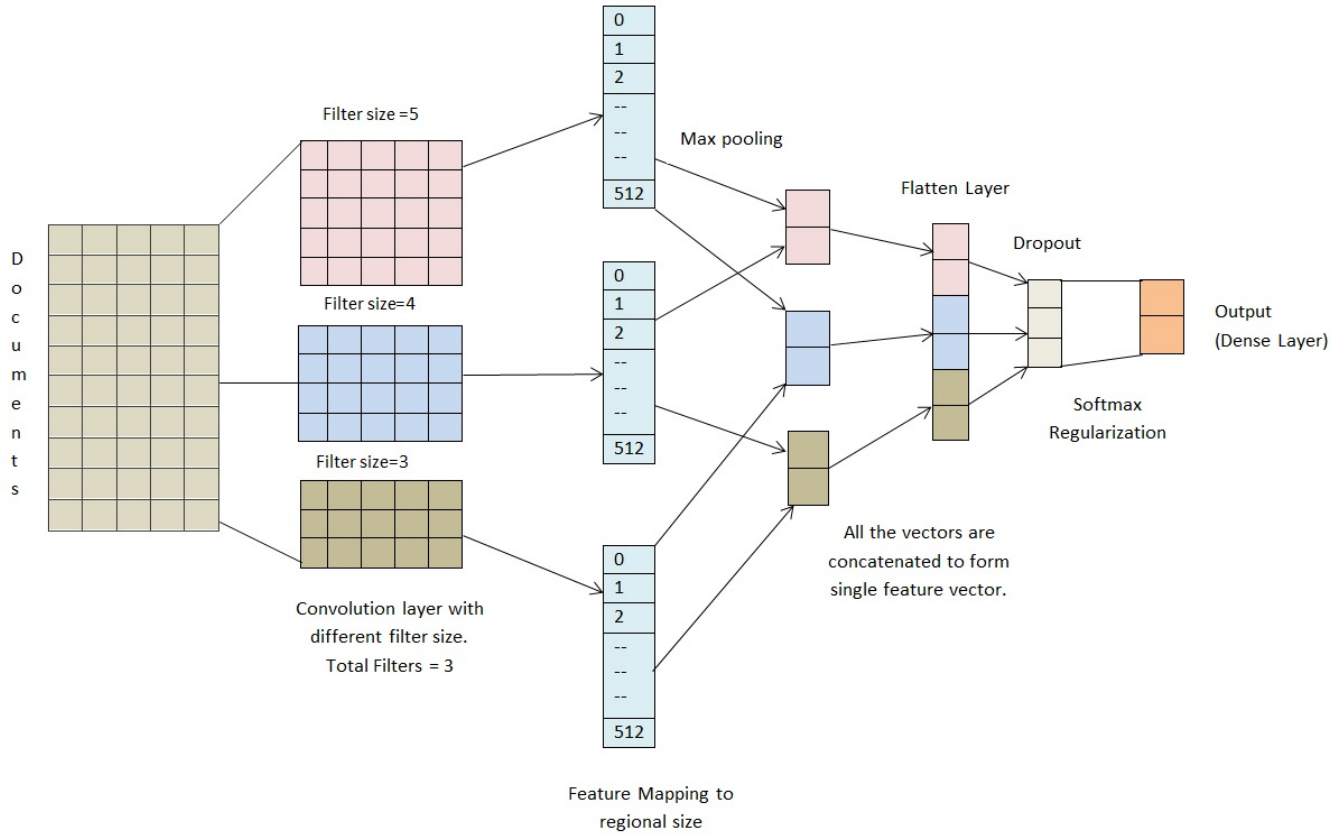

Fig. 1. Proposed CNN Model

features

- A novel method to combine CNN and ANN to detect bot accounts

The rest of the paper is organized as follows. Section II discusses the related works. The proposed model is presented in Section III. The results are presented in Section IV whereas Section V concludes the paper.

II. Related Works

Researchers have been studying malicious contents on social media in the form of spam and phishing for more than a decade. The bots are mostly identified based on their spamming behavior. Shehnepoor et al. [13] proposed a Net-Spam framework that utilizes the user behavior and linguistic features to identify the spamming behavior of the accounts. Moreover, they identified the relative importance of each feature in the spamming behavior. Similarly, Khan et al. [3] proposed a method to separate spammers from genuine users by applying a modified Hyperlink-Induced Topic Search (HITS) on features extracted from users' profiles. Zhu et al. [14] have introduced a template-based spam filtration system named as Tangram. The method extracts templates by using existing spam detection system and uses features such as celebrity name, eye-catching action, and URL.

In order to find the best spam detection algorithm, Chen et al. [15] compared nine different machine learning algorithms and found C5 and Random Forest as the best algorithms to detect the spam behavior of users. Chen et al. [15] studied the continous changing behaviors of spammers to evade detection and proposed a real-time scheme to detect drifted twitter spam.The authors used multiple profile features such as number of hash-tags, tweets, mentions, and URLs. Al-Qurishi et al. [16] proposed a platform to analyze the user behavior to discover malicious activities. The authors used social graph connections, user profile activities as features to understand the user's behavior. The most common behavior of bot accounts is generation of posts in burst(a large number of posts in short span of time) and then getting offline and therefore, bots having such behavior are referred as bursty botnet. Echeverria et al. [4] proposed a scheme to detect bursty botnet at Twitter using a Nave Bayes classifier and features including date and time. Similarly, Star Wars botnets got their name as they have shown similar behavior of posting quotes from the Star Wars novel.Echeverria et al. [5] report the analysis, discovery, and retrieval of Star Wars botnet on Twitter. The authors used location of tweets and calculated tweet quotation count, hashtag, and geographical location as features in their study.

Our work differs from the above-mentioned existing works

as instead of using features driven methodologies to identify common behaviors of the bots, we trained the neural network on text content of posts (tweets) to identify the common behavior in writing among the bots. Neural networks [17] [18] [19] have been shown to outperform other classifiers on a number of NLP tasks [20] [21]. Based on achievements of neural networks in NLP tasks [22], we targeted that our neural network model should be able to identify the bot from only a single post.

III. Proposed Model

The proposed model consists of an embedding layer, three convolution layers, three maxpooling layers, flatten, dropout, and fully dense output layer. The model is presented in Fig.1.

The input to the model is a set of documents (social media posts) and each document is a sequence of words. The embedding layer converts the words in the documents to vectors that are provided to convolution layers as input. In order to have same number of words in all the documents, we padded shorter documents with word PAD.

Features from word vectors trained by embedding layers are extracted using three different convolution layers, each having different size of filters — 3 for first, 4 for second, and 5 for the third layer. The number of filters applied to each layer is 512. Rectified Linear Unit (ReLU) activation function is used in the layers to enforce nonlinearity. The maxpooling layers down samples the convolution layers outputs.

The output of the three maxpooling layers are concatenated before flattening. The output of the flatten layer is a one dimensional vector. To prevent the system from overfitting, a dropout layer is added for regularization. The dropout layer helps prevent co-adaptation of neurons by disabling a proportion of neurons randomly. The last output layer is a dense layer with two neurons having Softmax as activation function. The probability of a document belonging to class yes or no is determined by Softmax function.

A. Models for Comparison

The proposed model is compared with two models. The first model is an artificial neural network (ANN) model trained on 16 different profiles' features shown in Table I. The second model is the combination of the proposed CNN and ANN models.

The ANN model used for comparison contains a total of five layers. The first — input layer — have 75 fully connected neurons and ReLU activation function is used in this layer. The next two layers contains 100 fully connected neurons and uses Tanh activation function. The fourth layer contains 75 neurons and the activation function used in this layer is ReLU. The last layer of ANN model is an output layer which contains 2 neurons and the activation function of this layer is Softmax. This model achieves the best accuracy result among other tried ANN models.The model is trained on 16 different features extracted from profiles and are presented in Table I. The ANN model is presented in Fig. 2.

TABLE I
Profile features

Features	Description
favorite_count	Number of likes to a tweet
favourites_count	Number of favorites to a tweet
followers_count	Number of followings
friends_count	Number of followers and followings
geo_enabled	Location of id is enabled or disabled
in_reply_to_status_id	Reply to status ID
in_reply_to_user_id	Reply to user ID
listed_count	Belong to how many lists
num_hashtags	Number of Hashtag use in the tweet
num_mentions	User mentions in the tweet
num_urls	Number of URLs in Tweet
profile_use_background_image	Background image used in twitter id
reply_count	Number of replies to a tweet
retweet_count	Number of shares to a tweet
retweeted_status_id	Status of the id who shares the tweet
statuses_count	Number of statuses

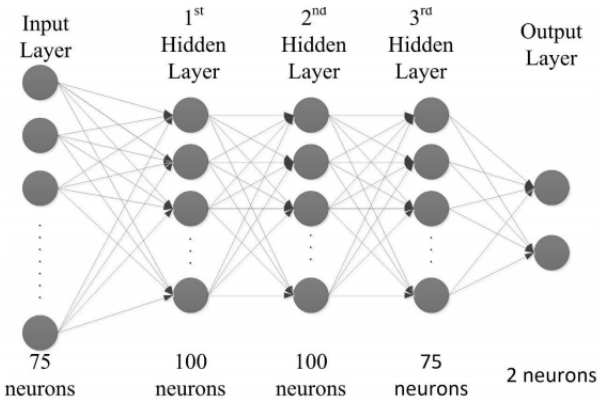

Fig. 2. ANN Model

The combined model is developed by merging the CNN and ANN models. Tweets along with the profile features of the account posting the tweets are the input to the model. The combined model comprises of an embedding layer, three convolution layers, three maxpooling layers, flatten layer, dropout layer, and five dense layers. The embedding layer takes the tweets as input whereas the dense layer takes the profile features as input. The output of the embedding layer is passed on to the convolution layers and further to the maxpooling layers. The output of the maxpooling layers are concatenated together before being flattened out by the flatten layer. The one dimensional output of the flatten layer is taken as input in the dropout layer. The setting of the layers is similar to the ones in the CNN model. The output of the dense layer that took the profile features as input is passed on to two other dense layers, similar to the ANN model. The output of the dropout layer and last dense layer are merged together and passed on to another dense layer having 50 neurons. The last layer of the model has two neurons with a Softmax activation function.

IV. Results and Discussion

In this section, we present the evaluation results of the proposed model. We have carried out experiments on P3

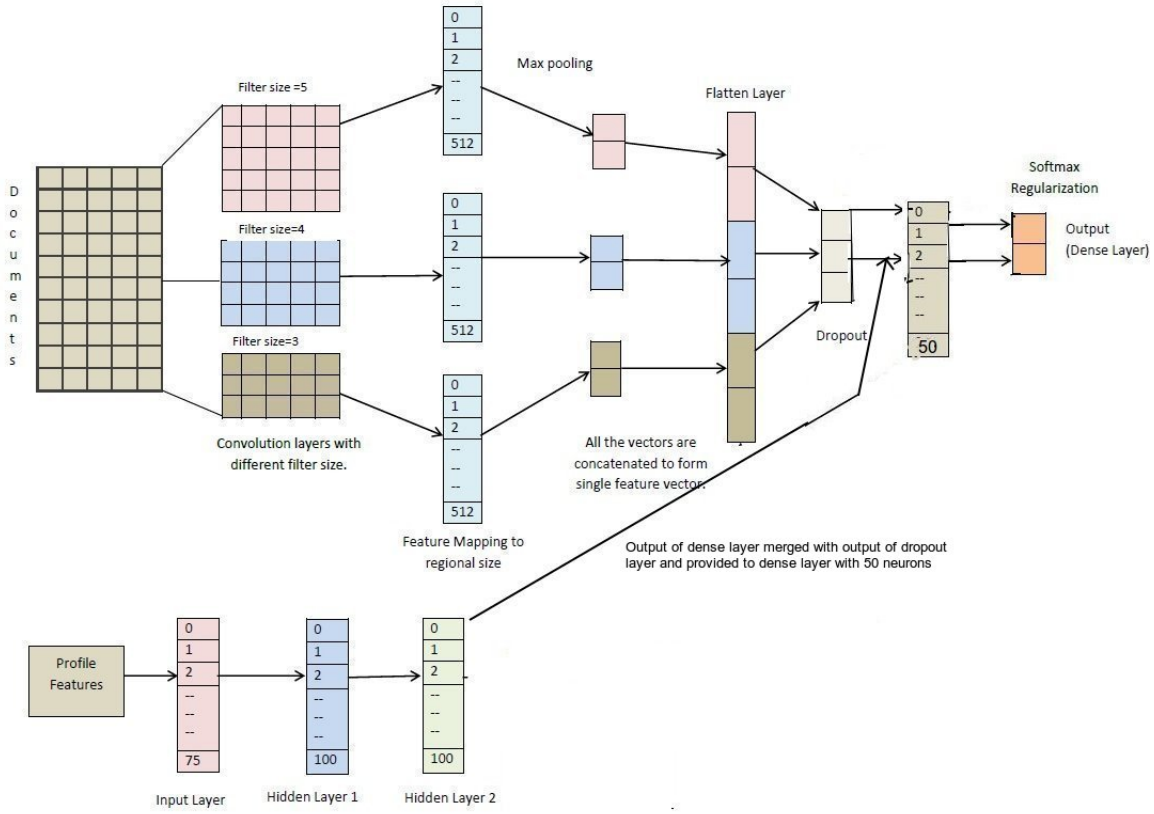

Fig. 3. Combined CNN and ANN model

instance at Amazon EC2. The instance contains a Tesla V100 GPU, with 61 GB memory. Python is used as the programming language and the main libraries used in our code are keras and tensorflow.

We have used a publicly available dataset [11] for our experiments. The dataset comprises of 8,377,522 tweets from 3474 genuine accounts, 1,610,176 tweets from 991 bot accounts, termed as spambot1 set,and 428,542 tweets from 3457 bot accounts, termed as spambot2. For experimentation purpose, the models are trained and tested on first using genuine and spambot1 data and second time using genuine and spambot2 data. The 70 to 30 percent ratio is taken for testing and training in our experiments. The results are presented as average of 10 different trials using 10 epochs for each trial.

To evaluate the performance of the models, we have used the following measures: (1) Accuracy, (2) Precision, (3) Recall, and (4) F-measure [23] [24]. Accuracy presents the ratio of total correct predicted results (both bots and human) to the total tested tweets. The accuracy can be calculated as:

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{1}$$

Precision gives the average quality of predicted bots by the models and can be calculated as:

$$Precision = \frac{tp}{tp + fp} \tag{2}$$

Recall is defined as a ratio of hit set size to the total size of test set, and is the measure of the prediction coverage by a detection system, given as:

$$Recall = \frac{tp}{tp + fn} \tag{3}$$

F-measure is the harmonic mean of precision and recall

$$F - measure = \frac{2 \times Precison \times Recall}{Precision + Recall} \tag{4}$$

The accuracy,precision, recall, and f-measure scores of the CNN, ANN, and the combined model are presented in Fig. 4, Fig. 5, and Fig. 6. The CNN model has achieved the accuracy of 98.71 % for spambot1 and 93.01 % for spambot2 detection. The results are better than the ANN model (Fig.5) that have shown accuracy of 97.60% and 87% for spambot1 and spambot2 detection respectively. However, the combined model (Fig. 6) has shown even better results, 99.43 % and 93.69% accuracy for spambot1 and spambot2 detection.

The results of all the three models have shown that spambot1 can be detected more accurately as compared to spambot2. On the analysis of the data, it is observed that most tweets in the spambot1 are broken sentences as compared to the more human like sentences in the spambot2 dataset. In experimentation, it is observed that model that has learned on profile features tend to drop the accuracy more faster as

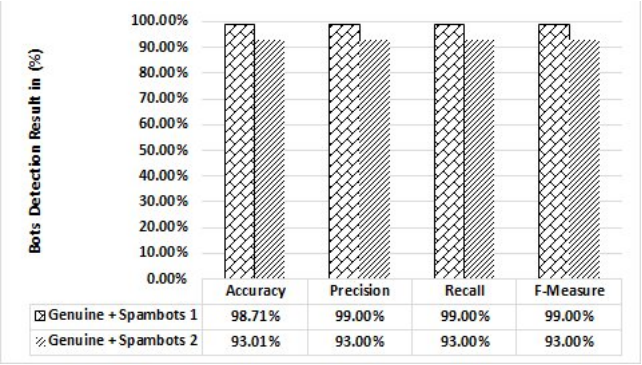

Fig. 4. CNN Model

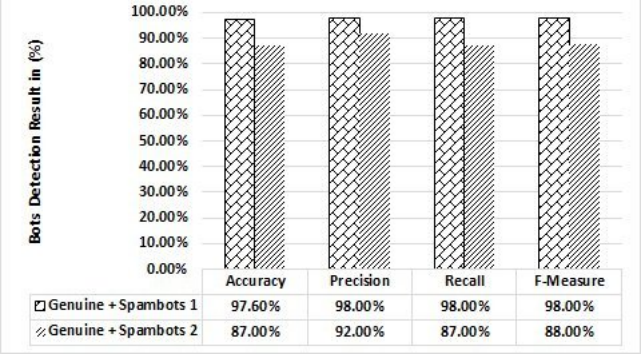

Fig. 5. ANN Model

compared to the CNN model when they are used to detect more human-like tweets.The authors have concluded that convolutional neural network models can learn the difference between the text contents of tweets by bots and humans better than ANN. However, these CNN models are still dependent on the ground truth provided by a human.

Although, the combined model has shown slightly better results than the CNN model, the authors still prefer the CNN model as detecting bot from single post require much less preprocessing as compared to selecting the profile and extracting features from that profile.

V. Conclusions

In this paper, we propose a novel convolutional nerual newtwork model (CNN) to detect bot accounts using the text content of posts on social media rather than identifying them using profile features. We have compared our results with an artificial neural network (ANN) trained on the features extracted from the accounts' profiles and a combined approach that merge both the CNN and ANN models. Results have shown that the CNN model performs better than ANN model and achieves above 90 % accuracy. However, the combined model performs slightly better but requires more preprocessing and kills the advantage of detecting the bot using only a single social media post. In the future works, we are aiming at application of bot detection in detection of fake news and enriching the quality of news events [25] [26].

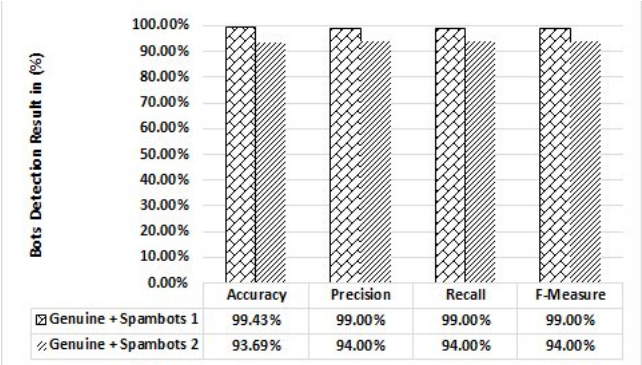

Fig. 6. Combined CNN and ANN Model

References

[1] K. Kyriakopoulou, "Authoritarian states and internet social media: Instruments of democratisation or instruments of control?," *Human Affairs*, vol. 21, no. 1, pp. 18–26, 2011.

[2] E. Ferrara, "Disinformation and social bot operations in the run up to the 2017 french presidential election," *First Monday*, vol. 22, no. 8, 2017.

[3] M. U. S. Khan, M. Ali, A. Abbas, S. U. Khan, and A. Y. Zomaya, "Segregating spammers and unsolicited bloggers from genuine experts on twitter," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 551–560, 2016.

[4] J. Echeverria, C. Besel, and S. Zhou, "Discovery of the twitter bursty botnet," *Data Science for Cyber-Security*, 2017.

[5] J. Echeverria and S. Zhou, "Discovery, retrieval, and analysis of the'star wars' botnet in twitter," in *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pp. 1–8, ACM, 2017.

[6] M. Tsikerdekis and S. Zeadally, "Multiple account identity deception detection in social media using nonverbal behavior," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 8, pp. 1311–1321, 2014.

[7] M. Shardlow, R. Batista-Navarro, P. Thompson, R. Nawaz, J. McNaught, and S. Ananiadou, "Identification of research hypotheses and new knowledge from scientific literature," *BMC medical informatics and decision making*, vol. 18, no. 1, p. 46, 2018.

[8] S. Ananiadou, P. Thompson, and R. Nawaz, "Enhancing search: Events and their discourse context," in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 318–334, Springer, 2013.

[9] R. Nawaz, P. Thompson, and S. Ananiadou, "Identification of manner in bio-events.," in *LREC*, pp. 3505–3510, 2012.

[10] A. Abbas, M. U. Khan, M. Ali, S. U. Khan, and L. T. Yang, "A cloud based framework for identification of influential health experts from twitter," in *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, pp. 831–838, IEEE, 2015.

[11] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 963–972, International World Wide Web Conferences Steering Committee, 2017.

[12] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[13] S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, "Netspam: A network-based spam detection framework for reviews in online social media," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1585–1595, 2017.

[14] T. Zhu, H. Gao, Y. Yang, K. Bu, Y. Chen, D. Downey, K. Lee, and A. N. Choudhary, "Beating the artificial chaos: fighting osn spam using its own templates," *IEEE/ACM Transactions on Networking*, vol. 24, no. 6, pp. 3856–3869, 2016.

[15] C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou, and G. Min, "Statistical features-based real-time detection of drifted twitter spam," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 914–925, 2016.

[16] M. Al-Qurishi, M. S. Hossain, M. Alrubaian, S. M. M. Rahman, and A. Alamri, "Leveraging analysis of user behavior to identify malicious activities in large-scale social networks," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 799–813, 2017.

[17] M. Jahangir, H. Afzal, M. Ahmed, K. Khurshid, and R. Nawaz, "An expert system for diabetes prediction using auto tuned multi-layer perceptron," in *2017 Intelligent Systems Conference (IntelliSys)*, pp. 722–728, IEEE, 2017.

[18] M. U. Khan, A. Abbas, M. Ali, M. Jawad, and S. U. Khan, "Convolutional neural networks as means to identify apposite sensor combination for human activity recognition," in *2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 45–50, IEEE, 2018.

[19] M. U. S. Khan, A. Abbas, M. Ali, M. Jawad, S. U. Khan, K. Li, and A. Y. Zomaya, "On the correlation of sensor location and human activity recognition in body area networks (bans)," *IEEE Systems Journal*, vol. 12, no. 1, pp. 82–91, 2016.

[20] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[21] R. Yunus, O. Arif, H. Afzal, M. F. Amjad, H. Abbas, H. N. Bokhari, S. T. Haider, N. Zafar, and R. Nawaz, "A framework to estimate the nutritional value of food in real time using deep learning techniques," *IEEE Access*, vol. 7, pp. 2643–2652, 2019.

[22] R. T. Batista-Navarro, G. Kontonatsios, C. Mihăilă, P. Thompson, R. Rak, R. Nawaz, I. Korkontzelos, and S. Ananiadou, "Facilitating the analysis of discourse phenomena in an interoperable nlp platform," in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 559–571, Springer, 2013.

[23] M. U. S. Khan, O. Khalid, Y. Huang, R. Ranjan, F. Zhang, J. Cao, B. Veeravalli, S. U. Khan, K. Li, and A. Y. Zomaya, "Macroserv: A route recommendation service for large-scale evacuations," *IEEE Transactions on Services Computing*, vol. 10, no. 4, pp. 589–602, 2015.

[24] H. Qadir, O. Khalid, M. U. Khan, A. U. R. Khan, and R. Nawaz, "An optimal ride sharing recommendation framework for carpooling services," *IEEE Access*, vol. 6, pp. 62296–62313, 2018.

[25] P. Thompson, R. Nawaz, J. McNaught, and S. Ananiadou, "Enriching news events with meta-knowledge information," *Language Resources and Evaluation*, vol. 51, no. 2, pp. 409–438, 2017.

[26] P. Thompson, R. Nawaz, I. Korkontzelos, W. Black, J. McNaught, and S. Ananiadou, "News search using discourse analytics," in *2013 Digital Heritage International Congress (DigitalHeritage)*, vol. 1, pp. 597–604, IEEE, 2013.