Palestine Technical University Research Journal, 2021, 10(1), 93-102

Identifying Informative Coronavirus Tweets using Recurrent Neural Network Document Embedding

تصنيف الأخبار المسببة حول وباء الكورونا المستجد عبر تويتر باستخدام تضمين الملفات النصية بواسطة خوارزمية الشبكات العصبية المتكررة

Rami Naim Mohammed Yousuf[1*]

رامي نعيم محمد يوسف

Palestine Technical University -PTUK, Tullkarm, Palestine

جامعة فلسطين التقنية خضوري، طولكرم ، فلسطين

Received: 15/08/2021 **Accepted: 20/12/2021** **Published: 30/03/2022**

Abstract: The coronavirus pandemic has led to the spread of tremendous fake news and misleading information through tweets. Hence, an interesting task of classifying tweets into informative and uninformative has motivated researchers to employ machine learning techniques. The state-of-the-art studies showed high dependency on transformers architecture. However, the transformers architecture suffers from the catastrophic forgetting problem where important contextual information is being forgotten by the gradients. Therefore, this paper proposes a document embedding using Recurrent Neural Network. Lastly, three classifiers of LR, SVM and MLP have been used to classify documents into Informative and Uninformative. Using the benchmark dataset of WNUT-2020 at Task 2, LR classifier obtained the highest f-measure of 0.91. This result demonstrates the efficacy of RNN to generate sophisticated document embedding.

Keywords: Coronavirus, Informative Tweets, Document Embedding, Recurrent Neural Network, WNUT-2020.

المستخلص: أدى ظهور جائحة كورونا إلى انتشار عددهائل من أخبار كاذبة ومعلومات مضللة عبر التغريدات. ومن ثم فان هناك حاجة مهمه تتمثل في تصنيف التغريدات إلى تغريدات مسببة اعلاميا و غير مسببة اعلاميا وهذا الامر حفز الباحثين على استخدام تقنيات التعلم الآلي لمعالجة هذا الموضوع . وقدأظهرت الدراسات الحديثة اعتمادا كبيرا على خوارزميات الشبكة العصبية المتحورة. ومع ذلك فهي تعاني من مشكلة تدعى "النسيان الكارثي" حيث تفقد معلومات سياقية خلال عملية التدريب. لذلك هذه الدراسة تهدف الى اقتراح آلية لتضمين النصوص بناءا على الشبكة العصبية المتكررة. أخيرًا ، تم استخدام ثلاثة مصنفات من LR و SVM و MLP لتصنيف المستندات إلى مسببة اعلاميا و غير مسببة اعلاميا. باستخدام مجموعة البيانات المعيارية لـWNUT-2020 في المهمة 2 ، حصل مصنف LR على أعلى مقياس f يبلغ 0.91. توضح هذه النتيجة فعالية RNN في توليد اليات تضمين نصوص مركبة .

الكلمات المفتاحية: فيروس كورونا ، تغريدات مسببة اعلاميا ، تضمين النصوص، الشبكة العصبية المتكررة، بيانات مشوشة من انشاء المستخدم

* Corresponding Author E-mail: r.yousuf@ptuk.edu.ps

INTRODUCTION:

The unprecedented event of the Coronavirus pandemic has led to an exponential use of social network around the world (Orso et al., 2020). This usage growth is referred to the curfew policies that have been implemented by the governments. This has made millions of individuals to stay or work from home which contributed toward the exponential growth of social network usage. Compared to all the social networks, Twitter has caught the attentions of individuals, governments, medical institutions and corporations to exchanged ideas, breaking news, and medical breakthroughs. The reason of such a growth in information exchanging is that Twitter offers a unique experience of mini-blogging where short text is posted through verified accounts of official, governmental and academic entities (Zarocostas, 2020). Tremendous number of tweets have been witnessed through the hashtags '#Coronavirus', '#Covid_19' and '#Covid-19' in which users exchange information about the disease including symptoms, preventive measures, infected and death cases, and vaccination doses. Among this huge information, plenty of prevalent fake news and misleading information are depicted through tweets. The harm of this information does not limit to disobeying preventive policies but rather it affects people lives directly through misleading assumptions that the coronavirus does not exist or that the vaccine would manipulate human genome (Al-Rakhami & Al-Amri, 2020; Stephens, 2020).

Accordingly, the research community of Natural Language Processing (NLP) and Text Analysis has become more motivated in analyzing and detecting this phenomenon. The presence of API Twitter application has increased such a motivation in which scholars are enabled to surf and gather tweets using various filters such as date, hashtag, location and the written language. Hence, several research studies have been presented to examine the misleading tweets. Some studies have concentrated on the prevalence of fake news and misleading information related to coronavirus within tweets (Yang et al., 2020). Other studies have assessed the credibility reference of URL and official accounts in tweets regarding coronavirus (Gill et al., 2021). The remaining studies have considering different credibility frameworks for tweets (Al-Rakhami & Al-Amri, 2020; Zhou et al., 2020).

Yet, an interesting research effort has been witnessed by the classification of tweets into informative and uninformative. This task aims to employ the Machine Learning (ML) and NLP techniques to classify informative tweets from uninformative tweets. Obviously, informative tweets are defined as the tweets that have real and related information to the coronavirus pandemic, while the uninformative tweets are defined as the containment of either real information but not related to the pandemic or irrelative and misleading information.

Several research efforts have been depicted for classifying tweets into informative and uninformative (Hettiarachchi & Ranasinghe, 2020; Malla & P.J.A, 2021; Møller et al., 2020; Sancheti et al., 2020). The majority of these studies have focused on Transformers architectures. However, such architectures have a remarkable drawback that is represented by forgetting significant information. Therefore, this study aims to propose a document embedding through Recurrent Neural Network (RNN). The proposed RNN architecture will process the term-to-term and document-to-term one-hot encodings of the words within the tweet documents in order to provide a generic embedding vector for each document. Lastly, three

traditional machine learning classifiers including Logistic Regression, Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) will be used to classify the embedding vectors of tweet documents into informative and uninformative.

Tweets are usually containing non-standard or informal language would require using lexicon-based approaches to expand the text with standard medical terminologies. Yet, such expansion might lead to significant change on the sentence or tweet. Therefore, this study will take the advantage of a domain-specific medical pretrained word embedding model based on PubMed resources to accommodate a word-by-word replacement method. Then, the traditional text representation will be utilized to represent the tweets after the replacement task. Consequentially, a Logistic Regression (LR) classifier will be used to extract informative tweets.

The paper has been organized as; Related Work where the state-of-the-art studies and techniques will be highlighted, Proposed Method where the proposed RNN embedding is being described, *Results* where the experimental results are being analyzed, and finally Discussion where the acquired results by the proposed method is being compared against the state-of-the-art.

RELATED WORK:

Hettiarachchi & Ranasinghe (Hettiarachchi & Ranasinghe, 2020) proposed an extraction method for the informative tweets using transformers architectures including BERT, XLNet and RoBERTa. Tran et al. (Tran et al., 2020) have proposed a COVID-Twitter-BERT (CT-BERT) architecture for extracting informative tweets. Sancheti et al. (Sancheti et al., 2020) have proposed different architectures including BERT, FastText and handcrafted feature space for identifying informative tweets. Møller et al. (Møller et al., 2020) showed another COVID-Twitter-BERT architecture that was intended to extract informative tweets. Wadhawan (Wadhawan, 2020) proposed different architectures such as Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) for the task of determining informative tweets. Malla & P.J.A (Malla & P.J.A, 2021) presented different architectures including RoBERTa, BERTweet and CT-BERT for identifying informative tweets.

As noticed from the literature, there is a high dependency on transformers architectures. Basically, the transformers architecture suffers from a specific drawback so-called 'catastrophic forgetting' (Lovón-Melgarejo et al., 2021; Rodriguez et al., 2019). This drawback happens due to the rapid forgetting of important contextual information by the transformers' gradients (Mirzadeh et al., 2021).

The Proposed RNN Document Embedding

As shown in Fig. 1, the proposed method begins with the WNUT-2020 tweet dataset. Then, the proposed document embedding will take a place in which the tweet document will be transformed into one-hot encoding matrix. Such a matrix will be fed into a Recurrent Neural Network (RNN) architecture in order

to generate the document embedding vectors. Lastly, three classifiers including LR, SVM and MLP will be used to classify the tweet document into Informative and Uninformative.

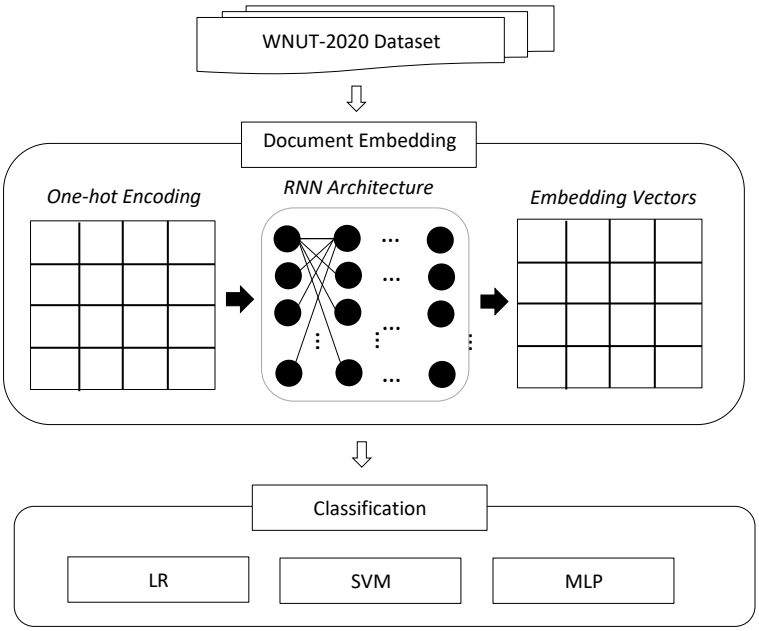

Fig. 1. Proposed RNN document embedding

1.1 Tweets Dataset

The dataset used in this study is the WNUT-2020 Task 2 of COVID-19 English Tweets which has been introduced by the study of Nguyen et al. (Nguyen et al., 2020). The data consists of tweets that have been labeled as either 'Informative' or 'Uninformative'. Table 1 shows the statistics of the dataset.

Table (1). Dataset details

Classes	Training	Validation	Testing	Total
Informative	3303	472	944	4719
Uninformative	3697	528	1056	5281
Total	7000	1000	2000	10000

1.2 Document Embedding

In order to generate the document embedding vectors using RNN, it is necessary to articulate two paradigms of one-hot encoding. The first paradigm is known as term-to-term one-hot encoding. Such encoding is intended to represent the co-occurrence of the distinctive terms within all the documents and generate a sparse matrix (Seger, 2018). This kind of encoding has been examined by the traditional word embedding using the Word2Vec architecture. However, the second paradigm has been depicted by the RNN document embedding where the one-hot encoding will be also articulated for document-to-term.

In such a case, each document will be addressed in terms of distinctive terms where the population of such a matrix is through the term occurrence. Assume three documents as follows:

Document 1 = "Moroccan coronavirus cases rise to 115"

Document 2 = "Canada coronavirus cases rise to 51"

Document 3 = "Singapore coronavirus cases rise to 266"

The term-to-term one-hot encoding matrix can be depicted in Table 2. Note that, insignificant terms such as the stop word of 'to' and the numbers will be discarded.

Table (2). Term-to-term one-hot encoding

	moroccan	**canada**	**singapore**	**coronavirus**	**case**	**rise**
moroccan	1	0	0	0	0	0
canada	0	1	0	0	0	0
singapore	0	0	1	0	0	0
coronavirus	0	0	0	1	0	0
case	0	0	0	0	1	0
rise	0	0	0	0	0	1

As shown in Table 2, each correspondence among the distinctive terms has been represented as '1' while, the remaining values have been set to '0'. On the other hand, the document-to-term one-hot encoding can be depicted in Table 3.

Table (2). Document-to-term one-hot encoding

	moroccan	**canada**	**singapore**	**coronavirus**	**case**	**rise**
D1	1	0	0	1	1	1
D2	0	1	0	1	1	1
D3	0	0	1	1	1	1

As shown in Table 3, the distinctive terms will be examined in terms of each document in which the occurrence will be represented as '1' otherwise is '0'. Now, in order to feed the one-hot matrices into the RNN architecture, each document will be handled separately where the terms of such a document will be represented by their vectors from the term-to-term matrix along with the document vector from the document-to-term matrix. Fig. 2 shows an example of feeding the first document into the RNN architecture.

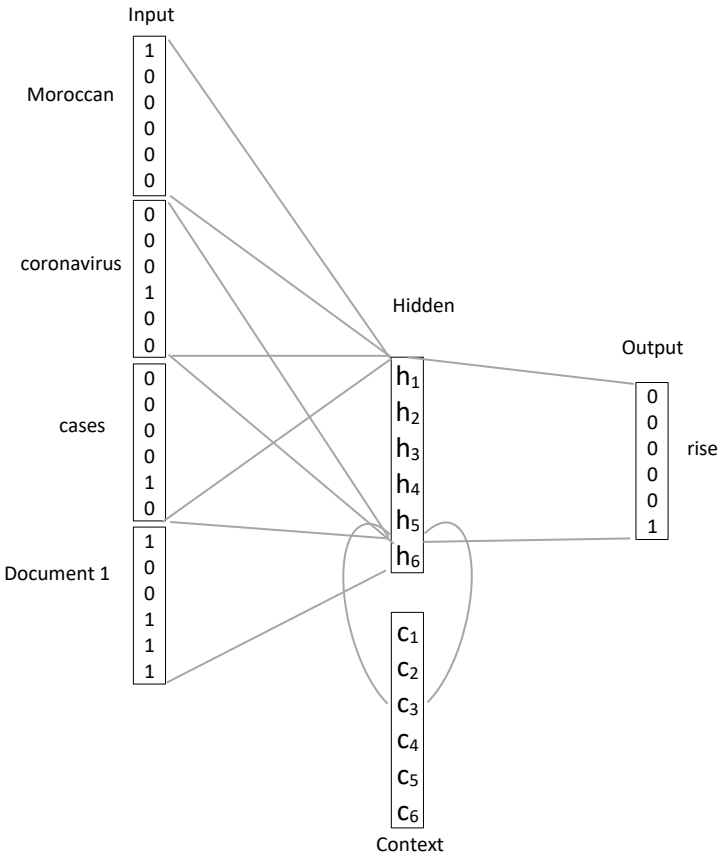

Fig. (2) Feeding document 1 into the RNN architecture

As in the classical neural network, the RNN architecture is composed of three layers input, hidden and output. Yet, the RNN contains an additional layer called context layer which aims to memorize information related to the current input where next input can utilize this information (Zaremba et al., 2014). As depicted in Fig. 2, every term-to-term one-hot vector of the distinctive words shown in Table 2 will be processed as input. Additionally, the document-to-term vector of the specified document will be also processed within the input. In this regard, the output would represent a term-to-term one-hot vector of a target word that is intended to be predicted. Similar to the traditional NN, the weights will be generated and the equations of computing the hidden neurons will be applied.

Additionally, in RNN the hidden layer's outcome will be passed into the context layer where its outcome will be passed recursively to the hidden layer as well. Such a recursive process is intended to discard any potential vanishing or exploding gradients (Rikukawa et al., 2020; Srivastava et al., 2014).

Once the training is finished and the error rate is minimized, the hidden neurons will articulate the document embedding. To this end, the Doc2Vec model existed in the Python library of Gensim will be utilized (Srinivasa-Desikan, 2018).

1.3 Informative and Uninformative Tweet Classification

Once the document embedding model is built where every tweet document would have an embedding vector of 300 dimension, the classification of Informative and Uninformative tweet will take a place. For this purpose, three traditional machine learning classifiers will be used including LR, SVM and MLP.

RESULTS:

Prior to the results analysis, it is worth mentioning the experiment settings. Table 4 depicts the hyper parameters used to build the document embedding model.

Table (4). The proposed RNN hyper parameters

Parameter	Value
Vector dimension	300
Number of epochs	100
Alpha	0.025
Minimum count	5

The results of classification have been computed based on Precision, Recall and F-measure. For the three classifiers, the classification has been conducted based on the document embedding generated by the proposed RNN. Table 5 depicts the classification results of the three classifiers.

Table (5). Classification results

Classifier	Precision	Recall	F-measure
SVM	1.0	0.80	0.89
LR	0.91	0.91	0.91
MLP	0.95	0.86	0.89

As shown in Table 5, SVM showed a precision of 1.0, a recall of 0.80 and an f-measure of 0.89. whereas, LR showed 0.91 for precision, recall and f-measure. Lastly, the MLP classifier showed a precision of 0.95, a recall of 0.85 and an f-measure of 0.89.

It is obvious that the LR classifier has outperformed the other classifiers. The reason of such a superiority is due to the classifier's ability to linearly learn the embedding features.

DISCUSSION:

Generally speaking, the best results of f-measure obtained by the LR was 0.91 which is superior than some state-of-the-art such as (Hettiarachchi & Ranasinghe, 2020) who obtained 0.90 of f1-score and (Tran et al., 2020) who obtained 0.90 of f1-score. This proves the efficacy of using the proposed document embedding through RNN.

CONCLUSION:

In this paper, a document embedding approach through RNN has been presented for the identification of informative and uninformative tweets related to the coronavirus pandemic. Two one-hot encodings have been utilized including term-to-term and document-to-term for generating the embedding vectors of tweet documents. Once the document embedding vectors are generated, three classifiers of LR, SVM and MLP will be trained and tested in terms of classifying the tweet document into informative and uninformative. Using a benchmark dataset of labelled tweets, the LR classifier has obtained the highest f-measure of 0.91 which outperformed the state-of-the-art studies. for future directions, the use of deep learning classifier might yield promising results in terms of the classification accuracy.

ACKNOWLEDGEMENTS:

The author would like to thank the Palestine Technical University-Kadoorie for their financial support to conduct this research.

REFERENCES:

Al-Rakhami, M. S., & Al-Amri, A. M. (2020). Lies kill, facts save: detecting COVID-19 misinformation in twitter. IEEE Access, 8, 155961-155970

Gill, S., Kinslow, K., McKenney, M., & Elkbuli, A. (2021). Twitter and the Credibility of Disseminated Medical Information During the COVID-19 Pandemic: SAGE Publications Sage CA: Los Angeles, CA

Hettiarachchi, H., & Ranasinghe, T. (2020). Infominer at wnut-2020 task 2: Transformer-based covid-19 informative tweet extraction. arXiv preprint arXiv:2010.05327.

Lovón-Melgarejo, J., Soulier, L., Pinel-Sauvagnat, K., & Tamine, L. (2021). Studying Catastrophic Forgetting in Neural Ranking Models. arXiv preprint arXiv:2101.06984.

Malla, S., & P.J.A, A. (2021). COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets. Applied Soft Computing, 107, 107495. doi:https://doi.org/10.1016/j.asoc.2021.107495

Mirzadeh, S. I., Chaudhry, A., Hu, H., Pascanu, R., Gorur, D., & Farajtabar, M. (2021). Wide Neural Networks Forget Less Catastrophically. arXiv preprint arXiv:2110.11526.

Møller, A. G., Van Der Goot, R., & Plank, B. (2020). NLP North at WNUT-2020 Task 2: Pre-training versus Ensembling for Detection of Informative COVID-19 English Tweets. Paper presented at the Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020).

Nguyen, D. Q., Vu, T., Rahimi, A., Dao, M. H., Nguyen, L. T., & Doan, L. (2020). WNUT-2020 task 2: identification of informative COVID-19 english tweets. arXiv preprint arXiv:2010.08232.

Orso, D., Federici, N., Copetti, R., Vetrugno, L., & Bove, T. (2020). Infodemic and the spread of fake news in the COVID-19-era. European Journal of Emergency Medicine.

Rikukawa, S., Mori, H., & Harada, T. (2020). Recurrent neural network based stock price prediction using multiple stock brands. International Journal of Innovative Computing, Information and Control, 16(3), 1093-1099.

Rodriguez, P. U., Jafari, A., & Ormerod, C. M. (2019). Language models and Automated Essay Scoring. arXiv preprint arXiv:1909.09482.

Sancheti, A., Chawla, K., & Verma, G. (2020). LynyrdSkynyrd at WNUT-2020 task 2: semi-supervised learning for identification of informative COVID-19 english tweets. arXiv preprint arXiv:2009.03849.

Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. (Independent thesis Basic level (degree of Bachelor)), Royal Institute of Technology, Stockholm, Sweden. (TRITA-EECS-EX ; 2018:596)

Srinivasa-Desikan, B. (2018). Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras: Packt Publishing Ltd.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), 1929-1958.

Stephens, M. (2020). A geospatial infodemic: Mapping Twitter conspiracy theories of COVID-19. Dialogues in Human Geography, 10(2), 276-281.

Tran, K. V., Phan, H. P., Van Nguyen, K., & Nguyen, N. L.-T. (2020). UIT-HSE at WNUT-2020 Task 2: Exploiting CT-BERT for Identifying COVID-19 Information on the Twitter Social Network. arXiv preprint arXiv:2009.02935.

Wadhawan, A. (2020). Phonemer at WNUT-2020 Task 2: Sequence Classification Using COVID Twitter BERT and Bagging Ensemble Technique based on Plurality Voting. arXiv preprint arXiv:2010.00294.

Yang, K.-C., Torres-Lugo, C., & Menczer, F. (2020). Prevalence of low-credibility information on twitter during the covid-19 outbreak. arXiv preprint arXiv:2004.14484.

Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. arXiv preprint arXiv:1409.2329.

Zarocostas, J. (2020). How to fight an infodemic. The lancet, 395(10225), 676.

Zhou, X., Mulay, A., Ferrara, E., & Zafarani, R. (2020). Recovery: A multimodal repository for covid-19 news credibility research. Paper presented at the Proceedings of the 29th ACM International Conference on Information & Knowledge Management.