DamascusTeam at CheckThat! 2020: Check Worthiness on Twitter with Hybrid CNN and RNN Models

Ahmad Hussein[1], Abdulkarim Hussein[1], Nada Ghneim[2], and Ammar Joukhadar[1]

[1] Faculty of Information Technology Engineering, Damascus University, Damascus, Syria
{ahmadhussein.ah7, karim.hussein.6.7.0}@gmail.com
ajoukhadar@el-ixir.com
[2] Faculty of Informatics Engineering, Al-Sham Private University, Damascus, Syria
n.ghneim@aspu.edu.sy

Abstract. In recent years, online social networks like Twitter, Facebook, Instagram, and others have revolutionized interpersonal communication and it becomes an important platform to share information about current events. Consequently, the research on the worthiness of posts is becoming more important than ever before. In this paper, we present our approach to analyze the worthiness of Arabic information on Twitter. To train the classification model, we annotated for worthiness a data set of 5000 Arabic tweets -corresponding to 4 high impact news events of 2020 around the world, in addition to a dataset of 1500 tweets provided by CLEF 2020. We propose two models to classify the worthiness of Arabic tweets: BI-LSTM model, and a CNN-LSTM model. Results show that Bi-LSTM model can extract better the worthiness of tweets.

Keywords: fact check-worthiness, neural networks, contrastive ranking.

1 Introduction

With the evolution of online social networks and blogging websites, the internet becomes a treasured source for obtaining news and information about current events and provides a platform for common people to share information and express their opinions. Quick response time and high connectivity speed have fueled the propagation and dissemination of information, by users on online social media services like Facebook, Twitter, and YouTube. The work presented in this paper primarily focuses on Twitter. Twitter is a micro-blogging web service with over 330 million Active Twitter Users per month, and has gained popularity as a major news source and information dissemination agent over the last years. Twitter provides the ground information and helps in reaching out to people in need, thus it plays an important role in aiding crisis management teams as the researchers have shown [1]. With the large scale of data generated

on Twitter, it has a role in spreading rumors and fake news. This would influence the opinions of the readers and can adversely affect thousands of people [2]. A recent study by Allcott and Gentzkow [3] indicated that fake news had an effect on voters during and before the American presidential elections in November 2016. The analysis of the worthiness on Twitter can be very valuable. In this task, we attempt to build a system which can assign a score to an input sentence indicating its check-worthiness. This score can vary from 0 (not check-worthy) to 1 (fully check-worthy). The Damascus-Team participated in Arabic Task 1 [17] of the CLEF2020 - CheckThat! Lab. This paper details our approach and results. The aim of Task 1 is to predict which tweets in a Twitter stream are worth fact-checking. The goal is to produce a ranked list of all tweets based on their worthiness for fact-checking. The organizers of this workshop have provided a data set comprised of 1500 sentences. These sentences are binary labelled, 0 and 1, corresponding to not check-worthy and fully check-worthy respectively.
We built the system using our collected dataset that includes 5000 Arabic annotated tweets, in addition to the dataset of 1500 tweets provided by CLEF 2020. We attempt to build a classifier which assigns a probability score to each sentence and hypothesize that this probability score corresponds to the check-worthiness of the sentence. Our framework classifies the worthiness of Arabic tweets from Twitter posts using a hybrid system of convolutional neural networks and long-short term recurrent neural network classifiers.
This paper is organized as follows: Section 2 describes the related work in this domain; Section 3 gives our methodology in detail; Section 4 discusses the evaluation of our proposed solution and finally, the last section gives the conclusion and describes future works.

2 Related work

There are various techniques used to solve the problem of worthiness on Online Social Media, especially in English content. In [4] ClaimBuster predicts check-worthiness by extracting a set of features (sentiment, statement length, Part-of-Speech (POS) tags, named entities, and tf-idf weighted bag-of-words), and uses a SVM classifier for the prediction. In [5] Patwari et al. presented an approach based on similar features, as well as contextual features based on sentences immediately preceding and succeeding the one being assessed, as well as certain hand-crafted POS patterns. The prediction is made by a multi-classifier system based on a dynamic clustering of the data. A work by Gencheva et al. [6] also extended the features used by ClaimBuster to include more context, such as the sentence's position in the debate segment, segment sizes, similarities between segments, and whether the debate opponent was mentioned. In the CLEF 2019 evaluation lab on check-worthiness detection [7], the best approaches used by the participating teams relied on neural networks for the classification of the instances. For example, Hansen et al. [9] learned domain-specific word embeddings and syntactic dependencies and applied an LSTM classifier. They pre-trained the network with previous Trump and Clinton debates, supervised weakly with the ClaimBuster system. Some efforts were carried out in order to consider the context. Favano et al. [10] trained a

feed-forward neural network, including the two previous sentences as a context. While many approaches relied on embedding representations, feature engineering was also popular [11]. We refer the interested reader to [12] for further details. In the CLEF 2018 evaluation lab on check-worthiness detection [13], Zuo et al. [15] enriched the dataset by producing pseudo-speeches as a concatenation of all interventions by a debater. They used averaged word embeddings and bag of words as representations. Hansen et al. [16] represented the entries with embeddings, part of speech tags, and syntactic dependencies, and used a GRU neural network with attention as a learning model. More details can be found in the task overview paper [13].

3 Methodology

In this section, we will present our methodology by explaining the different steps: dataset collection and labeling and deep learning models.

3.1 Data Set Collection

We collected our data from Twitter streaming API. For this, we considered four events in 2020, that affected a large population and generated a big number of tweets each. The events are listed in Table 1. We randomly selected 1000-1500 tweets from each event and grouped them to obtain a data set of 5000 tweets, named AWDS (Arabic Worthiness Dataset). In addition, we used a dataset of 1500 tweets provided by CLEF 2020, that includes tweets addressing a wide variety of topics. The dataset includes besides the tweet text, the metadata about the tweet and the tweet author. The next section describes the annotation process.

Table 1. Summary statistics for the studied dataset.

Event	Tweets
Coronavirus - فيروس كورونا	176314
The war in Yemen - الأحداث في اليمن	62682
The war in Libya - الأحداث في ليبيا	38345
Trump peace plan - صفقة القرن	34132
Total tweets	311473

3.2 Data Labeling

In order to create a labeled dataset for our worthiness assessment model, we obtained human labels for around 1000-1500 tweets selected uniformly at random per event. While there exist crowd-sourcing platforms such as Mechanical Turk and CrowdFlower, we relied on our own platform due to limitations imposed by existing platforms

when dealing with Arabic data. We first provided the annotators with the guidelines of the data annotation process provided by the organizers. Then we provided the annotators with a brief description of the event and links from where they can read more about it. We also showed annotators a definition of worthiness and example tweets for each of the annotation options. We provided the annotators with the tweet text and asked them to annotate each tweet with one of these annotations:

- 0 (not check-worthy)
- 1 (fully check-worthy)

To ensure good annotation quality, we used a sample set annotation to check the quality of each annotator before being recruited to the full annotation task. Each tweet was separately annotated by three annotators. The third annotator had also to check the agreement of the three annotations. In case of total agreement, the annotated tweet was added to the dataset, otherwise it was discarded.

3.3 Data Preprocessing

This step manages basic cleaning operations, which consists of removing unimportant or disturbing elements for the next analysis phases. Stop words, hashtags, URLs, mentions, repeated characters, and punctuations were removed.

3.4 Deep learning Models

We implemented two models and compared their results. The first model is a Bidirectional Long Short-Term Memory Units (Bi-LSTM) model, and the second is combination of a CNN and LSTM model.

Bi-LSTM Model:
In our proposed Bi-LSTM model, as shown in Fig. 1, each word is represented by a word embedding. The word embedding is a traditional word2vec model [18] that aims at capturing the semantics of the sentence. For each word in a sentence, the word embedding is concatenated and fed to a recurrent neural network with a Bi-LSTM as memory cells. The output of the Bi-LSTM is fed to a dense layer, with a sigmoid activation function in the output layer.

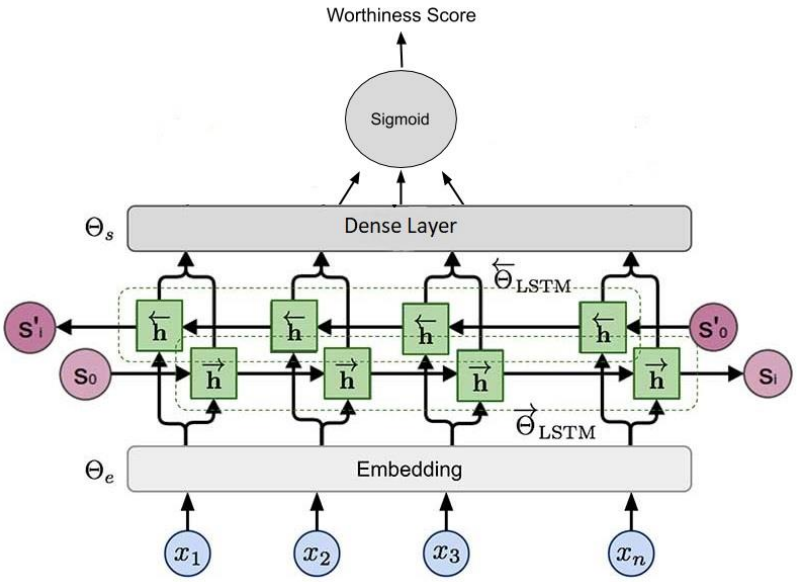

Fig. 1. Bi-LSTM model architecture.

Based on the design of the experiments, we tested several sets of parameters to select one that gives the experiments the best performance using Grid Search optimization algorithm. These parameters are as follows:

1. Learning rate: the model is trained using the Stochastic Gradient Descent algorithm, while the learning rate is set to 0.01.
2. Network structure: one embedding layer with 32 vector size fed to 50 Bi-LSTM with dropout [21] 0.2 fed to Dense layer with 8 perceptrons fed to Sigmoid function.
3. Number of epochs: 20 epochs.
4. Batch size: 32.

CNN and LSTM Model:

Another popular model is the convolutional neural network (CNN) which has been well known for its application in image processing as well as use in text mining [19]. We propose a new hybrid model that uses a word-embedding layer that is fed to a one-dimensional convolutional neural network followed by a recurrent neural network (RNN) layer then by a Dense layer and a sigmoid activation function in the output layer. Fig. 2 describes the CNN and LSTM model architecture.

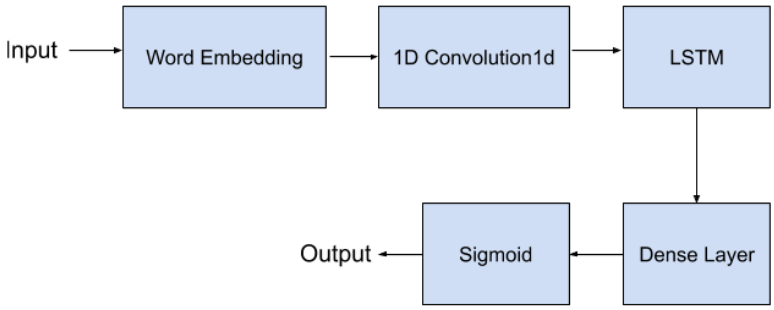

Fig. 2. CNN and LSTM model architecture.

We tested several sets of parameters to select one that gives the experiments the best performance using Grid Search optimization algorithm. The parameters of this model are:

1. Learning rate: the model is trained using the Stochastic Gradient Descent algorithm, while the learning rate is set to 0.01.
2. Network structure: one embedding layer with 32 vector size fed to 1D convolution1d with 3 filters and 2 kernel size fed to one-layer LSTM of size 20 fed to Dense layer with 8 perceptrons fed to Sigmoid function.
3. Number of epochs: 15 epochs.
4. Batch size: 32.

4 Results

In this section, we will introduce the evaluation experiments of our implemented model. For the first experiment, we used the global dataset that contains 6500 tweets (AWDS, our in-house built 5000 tweets dataset, and the CLEF 1500 tweets dataset). The training and the test sets contain 80% and 20% of total samples, respectively. We split the training data set into 80% for training and 20% for validation. In Table 2, we present the evaluation results of our implemented models on the test data.

Table 2. The evaluation results of our models on the test data.

Model	Accuracy
Bi-LSTM	83%
CNN + LSTM	79%

We remark that the accuracy of Bi-LSTM model performs better than the CNN+LSTM model.

Moreover, we tested our system on the test dataset provided by CLEF 2020 organizers, which contains 6000 tweets. In Table 3, we present the evaluation results of our implemented models (trained by the global dataset that contains 6500 tweets) for various metrics: mean precision@k (P@k) (P@10, P@20, P@30), and the mean average precision (MAP).

Table 3. The evaluation results of our models (trained by the global dataset that contains 6500 tweets) on the test dataset provided by CLEF 2020 organizers.

Model	P@10	P@20	P@30	MAP
Bi-LSTM	0.5833	0.5750	0.5472	0.4539
CNN + LSTM	0.4833	0.4875	0.5111	0.4315

We remark that the Bi-LSTM model performs better than the CNN+LSTM model in all used metrics.

In the second experiment, we trained our systems only on the CLEF 1500 tweets dataset. For evaluation, we split the training and the test sets as mentioned in the previous experiment. In Table 4, we present the evaluation results of our best model on the test data.

Table 4. The evaluation results of our models on the test data.

Model	Accuracy
Bi- LSTM	74%

In Table 5, we present the evaluation results of our implemented models (trained by the CLEF 1500 tweets dataset) for various metrics: mean precision@k (P@k) (P@10, P@20, P@30), and the mean average precision (MAP).

Table 5. The evaluation results of our models (trained by the CLEF 1500 tweets dataset) on the test dataset provided by CLEF 2020 organizers.

Model	P@10	P@20	P@30	MAP
Bi-LSTM	0.4000	0.3708	0.3500	0.3172

We remark that the accuracy of the model trained by the global dataset that contains 6500 tweets performs better than the model trained by the CLEF 1500 tweets dataset.

5 Conclusion

In this work, we proposed two different models to predict the check-worthiness of a tweet for CLEF 2020 CheckThat (Task 1). We built and compared two models: Bi-LSTM and CNN-LSTM. The BI-LSTM model, has an embedding layer fed to Bi-LSTM fed to a dense layer followed by a sigmoid activation function. The CNN-LSTM

model, has an embedding layer fed to one-dimensional convolutional neural network fed to LSTM then to a Dense layer followed by sigmoid activation function. To build our models we used our in-house built dataset of 5000 tweets, with the 1500 tweets provided by the organizers. To evaluate the results we used the 6000 tweets provided by the organizers. The Bi-LSTM model gave us better results than the CNN-LSTM one. In our future work, we plan to investigate new approaches and architectures for better modeling check-worthiness. In addition, we intend to investigate the influence of the tweeter on the worthiness of his tweets.

6 References

1. Ntalla, A., Ponis, S.: Twitter as an instrument for crisis response: The Typhoon Haiyan case study. In: 12th Proceedings of the International Conference on Information Systems for Crisis Response and Management (2015).
2. Gupta, A., Lamba, H., Kumaraguru, P.: $1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on Twitter. In: 2013 APWG eCrime Researchers Summit, pp. 1-12 (2013).
3. Allcott, H., Gentzkow, M.: Social Media and Fake News in the 2016 Election. In: Journal of Economic Perspectives, pp. 211-236 (2017).
4. Hassan, N., Arslan, F., Li, C., Tremayne, M.: Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1803-1812 (2017).
5. Patwari, A., Goldwasser, D., Bagchi., Tathya, S.: A multi-classifier system for detecting check-worthy statements in political debates. In: ACM on Conference on Information and Knowledge Management, pp. 2259-2262 (2017).
6. Gencheva, P., Nakov, P., Márquez, L., Barrón-Cedeño, A., Koychev, I.: A context-aware approach for detecting worth-checking claims in political debates. In: International Conference Recent Advances in Natural Language Processing, pp. 267-276 (2017).
7. Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., Atanasova, P.: Overview of the CLEF-2019 CheckThat! Automatic identification and verification of claims. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction, LNCS, Lugano, Switzerland, September (2019).
8. Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) Working Notes of CLEF 2019 Conference and Labs of the Evaluation Forum. CEURWorkshop Proceedings, CEUR-WS.org (2019).
9. Hansen, C., Hansen, C., Simonsen, J., Lioma, C.: Neural weakly supervised fact check-worthiness detection with contrastive sampling-based ranking loss. In: Cappellato et al. [8]
10. Favano, L., Carman, M., Lanzi, P.: TheEarthIsFlat's submission to CLEF'19 CheckThat! Challenge. In: Cappellato et al. [8]
11. Gasior, J., Przybyla, P.: The IPIPAN team participation in the check-worthiness task of the CLEF2019 CheckThat! Lab. In: Cappellato et al. [8]

12. Atanasova, P., Nakov, P., Karadzhov, G., Mohtarami, M., Da San Martino, G.: Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 1: Check-worthiness. In: Cappellato et al. [8]
13. Atanasova, P., Márquez, L., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Zaghouani, W., Kyuchukov, S., Da San Martino, G., Nakov, P.: Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. Task 1: Check-worthiness. In: Cappellato et al. [14]
14. Cappellato, L., Ferro, N., Y. Nie, J., Soulier, L. (eds.) Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org (2018).
15. Zuo, C., Karakas, A., Banerjee, R.: A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In: Cappellato et al. [14]
16. Hansen, C., Hansen, C., Simonsen, J., Lioma, C.: The Copenhagen team participation in the check-worthiness task of the evaluation lab of automatic identification and verification of claims in political debates of the CLEF-2018 fact checking lab. In: Cappellato et al. [14]
17. Hasanain, M., Haouari, F., Suwaileh, R., Ali, Z., Hamdan, B., Elsayed, T., Barrón-Cedeño, A., Da San Martino, G., Nakov, P.: Overview of CheckThat! 2020 Arabic: Automatic Identification and Verification of Claims in Social Media. task 1: Tweet Check-Worthiness.
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111-3119 (2013).
19. Tian Hsu, S., Moon, C., Jones, P., Samatova, N.: A Hybrid CNN-RNN Alignment Model for Phrase-Aware Sentence Classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pp. 443–449 (2017).
20. Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., Shaar, S., Sheikh Ali, Z.: Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. In Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol 12036. Springer, Cham
21. Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdino, R.: Improving neural networks by preventing co-adaptation of feature detectors. In: arXiv preprint arXiv, pp. 1207-0580 (2012).