QMUL-SDS at CheckThat! 2020: Determining COVID-19 Tweet Check-Worthiness Using an Enhanced CT-BERT with Numeric Expressions

Rabab Alkhalifa[1,5], Theodore Yoong[4], Elena Kochkina[2,3], Arkaitz Zubiaga[1], and Maria Liakata[1,2,3]

[1] Queen Mary University of London, United Kingdom
[2] University of Warwick, United Kingdom
[3] Alan Turing Institute, United Kingdom
[4] University of Oxford, United Kingdom
[5] Imam Abdulrahman bin Faisal University, Saudi Arabia

Abstract. This paper describes the participation of the QMUL-SDS team for Task 1 of the CLEF 2020 CheckThat! shared task. The purpose of this task is to determine the check-worthiness of tweets about COVID-19 to identify and prioritise tweets that need fact-checking. The overarching aim is to further support ongoing efforts to protect the public from fake news and help people find reliable information. We describe and analyse the results of our submissions. We show that a CNN using COVID-Twitter-BERT (CT-BERT) enhanced with numeric expressions can effectively boost performance from baseline results. We also show results of training data augmentation with rumours on other topics. Our best system ranked fourth in the task with encouraging outcomes showing potential for improved results in the future.

1 Introduction

The vast majority of people seek information online and consider it a touchstone of guidance and authority [1]. In particular, social media has become the key resource to go to for following updates during times of crisis [2]. Any registered user can share posts on social media without content verification, potentially exposing thousands or millions of other users to harmful misinformation. To prevent the undesired consequences of misinformation spread, there is a need to develop tools to assess the validity of social media posts. This problem has been particularly accentuated in light of the COVID-19 pandemic, accompanied by the rising spread of unverified claims and conspiracy theories about the virus and untested dangerous treatments. Compounded with the devastating effects from the virus alone, the social harms from misinformation spread can be particularly injurious [3]. The `CheckThat!` shared task provided a benchmark evaluation lab to develop systems for check-worthiness detection, with the aim of prioritising claims to be provided to fact-checkers.

In this paper, we present our approaches in tackling the check-worthiness detection task as outlined in Task 1 of the CLEF-2020 `CheckThat!` Lab. We evaluated several variants of our Convolutional Neural Network (CNN) model with different pre-processing approaches and several BERT embeddings. We also tested the benefits of including the use of external data to augment the training data provided. We submitted three models that have shown the best performance on the development set. Our best performing model utilised a COVID-Twitter-BERT (CT-BERT) enhanced with numeric expressions, which was ranked fourth in the task.

2 Related Work

We organise the related work into two subsections relevant to our proposed methods and the systems we submitted to the evaluation lab: claim check-worthiness and rumour detection.

2.1 Determination of Claim Check-worthiness

While there is no general streamlined approach to fact-checking, the fact-checking pipeline can be divided into different sub-tasks based on a number of contexts [4]. The first of the tasks and the one concerning our work consists in producing a list of claims ranked by importance (check-worthiness), in an effort to prioritise claims to be fact-checked. Systems for claim detection (as a classification task) and ranking by check-worthiness include (i) ClaimBuster [5], which combines numerous features such as TF-IDF, POS tags and NER on a Support Vector Machine to produce importance scores for each claim, and (ii) ClaimRank [6], which uses a large set of features both from individuals sentences and from surrounding context. More recent methods, such as [7], have made use of embedding-based methods such as InferSent for detecting claims by leveraging contextual features within sentences. In previous editions of CheckThat! [8], the shared task did not involve the detection of claims, as claims were already given as input.

Previous work on claim detection by Konstantinovskiy et al. [7] suggested the use of numeric expressions as a strong baseline for detection of claims. Indeed they showed that the use of numeric expressions as a feature leads to high precision, despite achieving lower recall and overall F1 score than other methods. This is due to the prevailing presence of numeric expressions in check-worthy claims, as opposed to non-check-worthy claims and non-claims. Given the emphasis of the CheckThat! shared task on precision-based evaluation (using mean average precision as a metric), we opted for incorporating numeric expressions in our model.

2.2 Rumour Detection

A rumour is generally defined as an unverified piece of information that circulates. In the same way that a check-worthiness detection looks at claims to be verified, e.g. in the context of a TV debate, rumour detection consists in detecting pieces of information that are in circulation while they still lack verification, generally in the context of breaking news, making it a time-sensitive task [9]. Rumours differ from check-worthy claims in their nature as well as relevance to the fact-checkers, as not all rumours are necessarily of interest to fact-checkers. Still, both tasks have significant commonalities.

In our approaches to check-worthiness determination, we try to leverage existing data for rumour detection, consisting of rumours and non-rumours, with the aim of providing additional knowledge that would enrich the task (see §3.1).

Rumour detection, as the task of detecting unverified pieces of information, has been studied before, for instance through the RumourEval shared tasks held at SemEval 2019 [10]. Prior to that, Zubiaga et al. [11] introduced a sequential rumour detection model that leveraged Conditional Random Fields (CRF) for leveraging event context, as well as Zhao et al. [12] that looked at evidence from others responding to tweets with comments of the form of *"is this really true?"*, which would be indicative of a tweet containing rumourous content.

3 Task Description

The task we explore was introduced by Barrón-Cedeño et al. [13] and is formulated as follows:

> Given a topic and a stream of potentially-related tweets, rank the tweets according to their check-worthiness for the topic, where a check-worthy tweet is a tweet that includes a claim that is of interest to a large audience (especially journalists) and may have a harmful effect.

For example, consider the target topic–tweet pair:

> **Target topic**: COVID-19
>
> **Tweet**: Doctors in #Italy warn Europe to "get ready" for #coronavirus, saying 10% of #COVID19 patients need ICU care, and hospitals are overwhelmed.
>
> **Label**: Check-worthy

Although Task 1 is available in both English and Arabic, we focussed solely on the English task [14]. Tweets in this dataset for this task exclusively covered the pandemic caused by the Coronavirus Disease 2019 (COVID-19). The ultimate objective of ranking the tweets identified as check-worthy claims is to enable prioritisation of claims to fact-checkers.

The task can be formally described as the following binary classification problem. We define the training set consisting of n labelled tweets as $\mathcal{D} = \{(\mathbf{x}_i, y_i), 1 \leq i \leq n\} \in (\mathcal{X} \times \{0,1\})^n$. Here, $\mathbf{x}_i$ is the i^{th} feature vector in feature space $\mathcal{X}$ which contains the tweet features such as the i^{th} tweet itself $\mathbf{t}_i$, the topic, and whether it is a claim or not; and $y_i \in \{0,1\}$ is the label indicating check-worthiness of $\mathbf{t}_i$. The objective is to obtain a map $h : \mathcal{X} \mapsto \{0,1\}$, based on the class probability measure $\mathbb{P}(y|\mathbf{t})$, which is subsequently used to rank the tweets in the test set.

3.1 Datasets

Dataset	No. of check-worthy tweets (rumours)	No. of non-check-worthy tweets (non-rumours)	Total
CLEF Train	231	441	672
CLEF Development	59	91	150
CLEF Test	80	60	140
PHEME	2402	4023	6425
Twitter 15	1012	362	1374
Twitter 16	536	199	735

Table 1. Number of posts and class distribution in the datasets used

In our experiments, we made use of three datasets of Twitter posts in English, which include the dataset provided by the organisers (CLEF) and two external publicly available datasets (PHEME, Twitter 15 and Twitter 16) to augment the training set. The PHEME, Twitter 15 and Twitter 16 datasets were chosen for augmentation as these are relatively large datasets annotated for rumour detection task, which is very similar to claim check-worthiness as described in section 2. Table 1 shows the number of tweets used in each of the datasets and the class distribution.

The *CLEF* dataset contains tweets related to the topic of COVID-19. They were annotated by the task organisers as either check-worthy or not check-worthy thus defining a binary classification task.[6] This dataset is rather small and is limited to individual tweets concerned with a single topic. The dataset is imbalanced with the majority of tweets being not check-worthy.

The *PHEME* dataset [9, 15] contains Twitter conversations discussing rumours (defined as unverified check-worthy claims spreading widely on social media) and non-rumours.[7] This dataset contains conversations related to 9 major newsworthy events, such as shooting in Charlie Hebdo, shooting in Ottowa, crash of Germanwings plane. In this work, we use only the source tweets of the conversations in the PHEME dataset (they are conveying the essence of a rumour, rather than the following discussion) in order to have the same input structure as the CLEF dataset. We performed experiments augmenting the training set with both rumours and non-rumours from the PHEME dataset. We found that adding rumours only is more beneficial than adding the full PHEME dataset.

The *Twitter 15 and Twitter 16* datasets [16] contain Twitter conversations, discussing True, False and Unverified rumours as well as non-rumours on various topics. Here, we do not use all 4-class labels, but instead convert True, False and Unverified classes into single check-worthy class. We also use only source tweets to augment the CLEF training set.

3.2 Evaluation

The CLEF dataset is split into training, development and testing sets. The check-worthiness task is evaluated as a ranking task, i.e. the participant systems should produce a list of tweets with the estimated score for check-worthiness. The official evaluation metric is Mean Average Precision (MAP), but the precisions at rank k ($P@5, P@10, P@30$) are also reported. Baseline results provided by the organisers are Random Classification (MAP = 0.35) and SVM with N-gram Prediction (MAP = 0.69).

4 Our Approach

In our approach, we fed a pre-trained word-level vector representation into a CNN model. Using vector representations of words as inputs offers high flexibility, allowing swaps between different pre-trained word vectors during model initialisation to be maintained without additional overhead. For our work, we tested multiple feature representations in order to assess which one would be best for our problem, these include combinations of frequency-based, vector-based representations and authors' profiles. Since most evaluation tweets were not supplied with author's profiles, and frequency-based models may not be of

[6] Annotation rules can be found here: https://github.com/sshaar/clef2020-factchecking-task1#data-annotation-process

[7] https://figshare.com/articles/PHEME_dataset_for_Rumour_Detection_and_Veracity_Classification/6392078

great generalisability for unseen data, we reduced our exploration to two dynamic word embeddings, ELMo [17] and BERT [18]. Since the former did not provide any increase in performance, we chose the latter to be further explored with different pre-processing techniques.

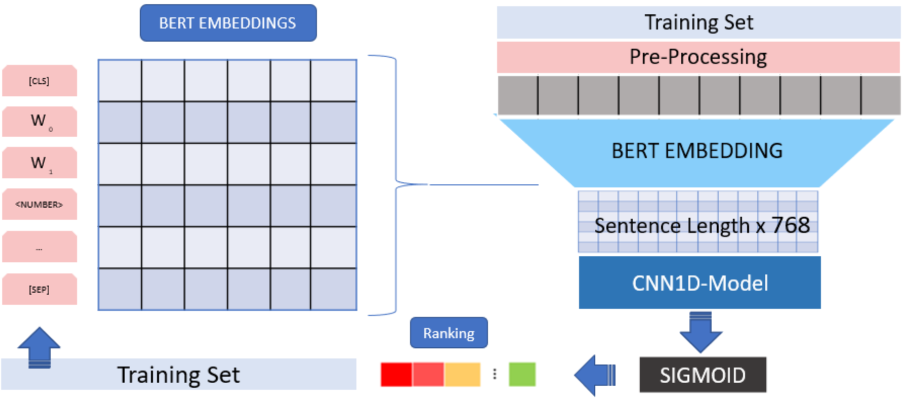

Fig. 1. General Model Architecture.

With these settings, we evaluated our model using two variations of the BERT pre-trained architecture, uncased BERT (uncased-BERT) and COVID-Twitter-BERT (CT-BERT) [19] (see Figure 1). While both are transformer-based models, CT-BERT is pre-trained on COVID-19 Twitter data using whole word-masked modelling and next sentence prediction.

In the following sections we describe the main characteristics of our designed system including pre-processing, feature representation, model architecture and hyperparameters used.

4.1 Pre-processing

We performed standard Twitter data pre-processing in order to improve our system performance. For each model, we implemented different pre-processing variants depending on the vector representation leading to best performance. The pre-processing steps can be summarised as follows.

(i) **Segment2Token**: We split each sentence into tokens considering the type of every segment and breaking it into individual tokens. Digits, URLs, accounts and hashtags were replaced by $\langle number \rangle$, $\langle url \rangle$, $\langle account \rangle$, $\langle hashtag \rangle$. respectively. This was implemented using a simple split function with different expression finding methods. By analysing generated segments, we settled on different treatment for special tokens in every tweet. For example, **hyperlinks** were either completely removed from the dataset or replaced by special tokens. Furthermore, **digits** and all other numerical expressions that contained '%' or '$' were either removed or tokenised. Tokenising numerical expressions allows the model to generalise better (see §5). For example:

> **Tweet**: [$NEWS$] Naver #BAEKHYUN EXO Baekhyun donates 50 million won to prevent the spread of Corona 19 @weareoneEXO #EXO
>
> **Segment2Token**: [$NEWS$] Naver $\langle hashtag \rangle$ EXO Baekhyun donates $\langle number \rangle$ won to prevent the spread of Corona 19 $\langle account \rangle$ $\langle hashtag \rangle$.

(ii) **Segment2Root**: In NLP, the $\tilde{\chi}^2$-statistical measure tests term-dependency of the tweet being about one of the classes as in [20]. We used it to analyse the segments of the tweets. In these settings, for few account handles and hashtags with high $\tilde{\chi}^2$-score, we manually combined them depending on their semantic meaning. For instance:

> **Hashtags**: #coronavirus, #COVID19', #COVID-19, #COVID19, #Coronavirus, #Corona-virus
>
> **Hashtag2Root**: coronavirus

In this case, different hashtags about COVID-19 were all consolidated under the 'coronavirus' umbrella term.

(iii) **Word2id**: As with all BERT models, we included classification embedding tokens for every tweet: `[CLS]` at the beginning and a separator token `[SEP]` at the end. We then decomposed $\mathbf{t}_i$ into a sequence of numerical tokens using BERT tokenisation methods. This was done by mapping each token to a unique integer in the corpus' vocabulary.

(iv) **Padding**: We ensured that the input sequences in every batch was the same length. This was achieved through increasing the length of some of the sequences by adding more tokens. We tried to reduce the padding by allowing our model to decide the padding length based on a given batch size (set to 10) and the longest sequence within the given batch. For example, if the longest sequence length for a given batch is 20, then all other shorter sentences will be padded to match its length.

Finally, a look-up table was used for each token from the generated representation, ready to be fed into the model.

4.2 Model Hyperparameters

The CNN architecture requires tuning various hyperparameters. These include input representations, number of layers and filters, pooling, and activation functions. We utilised a BERT language model in accompaniment to the the general CNN architecture. Within the model design, we propose variations of a three-layer CNN with 32 filters with different window sizes: 2, 4 and 7. The multiple filters act as feature extractors.

Additionally, we used an Adam optimiser with learning rate fixed at $2\mathrm{e}^{-5}$ and number of training epochs set to 8. Our N-gram kernels encompass a Rectified Linear Unit (ReLU) activation function, given by $\max(0, x)$. All pooling layers use a max-pooling operation. For the binary classification, we utilise a sigmoid activation function $\sigma(x)$ for the output layer, defined as

$$\sigma(x) = \frac{1}{1 + \mathrm{e}^{-x}}$$

To determine the final output labels, we classify check-worthiness based on the indicator variable

$$h(x) = \begin{cases} 1 & \text{if } \sigma(x) \geq \frac{1}{2}, \text{ indicating a check-worthy tweet,} \\ \\ 0 & \text{if } \sigma(x) < \frac{1}{2}, \text{ suppressing the tweet as non-check-worthy.} \end{cases}$$

5 Results and Discussion

In the following section, we discuss the selection of the models we tried and ultimately submitted to the shared task. We also evaluate and compare their performance.

5.1 Model Selection on the CLEF Development Set

We performed our model selection using the development set. Details of the pre-processing steps and embeddings applied to each of the eight models we tested are given in Table 2. The performance of the models according to the various precision metrics are shown in Table 3.

CLEF Benchmark Data Experiments: Text distortion has been used by [21], where their methods were more successful than using the full text in the classification process. Taking inspiration from their work, we used tokenisation for different segments of the tweet where account handles, hashtags, URLs and digits were assigned special tokens and added to the model vocabulary. The goal of this step was to avoid over-fitting the training data.

In Model 7, we experimented with ELMo embeddings, which gave the best performance in terms of MAP, in our tests without additional pre-processing and outperforms random baseline (MAP = 0.35). However, Model 7 did not outperform the N-gram baseline (MAP = 0.69), and thus we did not choose it for the test set submission.

In Model 4, we trained word embeddings on the training set along with the model and combined them with TF-IDF representations. This led to improvements over Models 5-7 and over the N-gram baseline.

For Model 1, we used intensive pre-processing in tandem with a CNN model with three filters of sizes 2, 4 and 7. On the other hand, in Model 2, only numeric expressions were tokenised and the CNN model only had two filters of sizes 2 and 4. Models 1 and 2 displayed the best performance on the development set, and were hence chosen for submission on the testing set.

Model No.	Pre-processing	Embeddings
1	Frequently mentioned entities replaced with their account name. Hashtags with repeated topics combined using the $\tilde{\chi}^2$-score, other URLs and hashtags tokenised with special tokens.	CT-BERT
2	Special tokens for digits. Account handles, URLs and hashtags removed.	CT-BERT
3	Training set merged with rumours from PHEME. Special tokens for digits.	BERT-EN (Uncased)
4	Digits, account handles, URLs and hashtags removed.	CLEF Train Embeddings + TF-IDF
5	Training set merged with Twitter 15 and Twitter 16 datasets. Special tokens for digits.	BERT-EN (Uncased)
6	Training set merged with PHEME, Twitter 15 and Twitter 16 datasets. Special tokens for digits.	BERT-EN (Uncased)
7	No pre-processing was applied.	ELMo
8	Trained using PHEME, Twitter 15 and Twitter 16 datasets only, **without** CLEF training data.	BERT-EN (Uncased)

Table 2. Description of the models tested.

Data Augmentation Experiments: While the task definition states the presence of the target topic when identifying tweet check-worthiness, the dataset provided only covers a single topic, COVID-19. We experimented with training data augmentation using check-worthy tweets from external datasets (PHEME and Twitter15, Twitter 16 as described in section 3.1) (see models 3, 5, 6 in Table 2). These datasets cover different topics, accounts and vocabulary, so incorporating them could contribute to future generalisability of the model. We also performed experiments using only existing external datasets of rumours and non-rumours and omitting the CLEF training data to test the generalisability of the currently available datasets and models to the emergence of new rumour topics (see model 8 in Table 2). The results are presented in Table 3.

Model No.	Average precision	R-precision ($R = 59$)	Precision@k					
			@1	@3	@5	@10	@20	@50
1	0.81	0.71	1.00	1.00	1.00	1.00	0.95	0.74
2	0.80	0.71	1.00	1.00	1.00	1.00	0.95	0.76
3	0.75	0.69	1.00	0.67	0.60	0.80	0.85	0.74
4	0.74	0.63	1.00	1.00	0.80	0.90	0.85	0.68
5	0.65	0.59	1.00	1.00	0.80	0.80	0.75	0.62
6	0.56	0.57	0.00	0.33	0.20	0.50	0.70	0.58
7	0.53	0.56	0.00	0.33	0.40	0.50	0.55	0.58
8	0.45	0.44	1.00	0.33	0.40	0.40	0.50	0.44

Table 3. Model performance on the development set

As expected, Model 8, which did not use the CLEF training data, performed worse compared to the models that did make use of the training data provided. However, it outperformed the random baseline (MAP = 0.35) by 10%, showing that there is enough overlap in the task definitions and inherent nature of rumours/check-worthy claims to provide meaningful signal for model training. Models 3, 5 and 6, which augmented the training data, did not perform as well as the models using only the training data. Models 5 and 6 did not outperform the

N-gram baseline (MAP = 0.69) provided by the organisers[8]. Model 3 only adds rumours to the training data, thus shifting the class balance in the dataset, and performed better than adding both rumours and non-rumours from PHEME to the training set.

These results show the importance of model training or fine-tuning on the evaluation domain. The lack of performance improvement could be also due to the differences in the definitions of the tasks and rumours/check-worthy claims by each of the datasets. Moreover, different fact-checking organisations would naturally make different choices when analysing the same data. In [7], they found that educational background can lead to bias in annotation efforts for fact-checking. The subjectiveness that underlies check-worthiness thereby adds further complications to the task of ranking by importance. These results also highlight the especially challenging aspects of the need for generalising to new unseen topics, as well as leveraging data from a related task such as that of rumour detection, in detecting tweet check-worthiness.

5.2 Results on the CLEF Test

We selected the best three models based on MAP (see Table 3). Table 4 shows the official results obtained by our systems on the testing set.

Model No.	Average precision	R-precision ($R = 59$)	Precision@k					
			@1	@3	@5	@10	@20	@50
1	0.71	0.63	**1.00**	**1.00**	**1.00**	0.90	0.80	0.64
2	**0.78**	**0.70**	**1.00**	**1.00**	**1.00**	**1.00**	**0.85**	**0.70**
3	0.73	0.63	**1.00**	**1.00**	**1.00**	0.90	**0.85**	0.68

Table 4. Performance of the submitted models on the testing set

Models 1 and 2: We found that avoiding extensive tokenisation (Model 1) while merely tokenising numeric expressions yields better results in the test set and allows the model to learn more general patterns from the training set (Model 2). For example:

[8] `https://github.com/sshaar/clef2020-factchecking-task1#baseline`

Tweet: France, Spain and Germany are about 9 to 10 days behind Italy in #$COVID$19 progression; the UK and the US follow at 13 to 16 days.

Digit2Token: France, Spain and Germany are about $\langle number \rangle$ to $\langle number \rangle$ days behind Italy in corona virus progression; the UK and the US follow at $\langle number \rangle$ to $\langle number \rangle$ days.

Moreover, our approach uses a less complex model which keeps the weights small and results in better overall performance. Therefore, in order for our model to generalise to unseen tweets in the test set, numeric expressions should be unified and model complexity needs to be maintained.

Model 3: In this submission, we augmented the training data with rumours from the PHEME dataset. While the results of this submission on the development set were the lowest out of the selected three, on the testing set it outperforms Model 1. This shows that external data from a related task adds meaningful signal for model training and contributes to system generalisability.

6 Conclusion

This paper describes our efforts as participants of the Task 1 of the `CheckThat!` 2020 evaluation lab, in which we ranked fourth, which was held in conjunction with the CLEF conference. We describe our proposed model that leverages the COVID-Twitter-19 BERT (CT-BERT) word embeddings and performs a special treatment for rare tokens with a CNN relying on the tweet alone.

The experimental results show that the performance of our model increases significantly by tokenising numerical expressions. The present work is restricted in choosing the best feature representation. In the future, this work can be enhanced in different possible directions. For example, incorporating pragmatic information related to author's profile information. Thus, simulate actual users' behaviour in verifying claims in social media.

Given the small size of the training data provided by the organisers, we also performed additional experiments leveraging external datasets with the aim of augmenting the training data. External data we incorporated had the challenge of being datasets pertaining to rumour detection and on different topics, hence with slight differences with the task and domain at hand. Our experiments with data augmentation did not lead to improved performance, highlighting that inclusion of external data of a different nature (i.e. in terms of task and domain) is particularly challenging and, if they can provide an improvement to the check-worthiness detection task, more careful integration and adaptation will be necessary.

7 Acknowledgments

This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT. http://doi.org/10.5281/zenodo.438045. This work was also partially supported by The Alan Turing Institute under the EPSRC grant EP/N510-129/1.

References

1. L. M. S. Miller and R. A. Bell, "Online health information seeking: the influence of age, information trustworthiness, and search challenges," *Journal of aging and health*, vol. 24, no. 3, pp. 525–541, 2012.
2. L. Palen, "Online social media in crisis events," *Educause quarterly*, vol. 31, no. 3, pp. 76–78, 2008.
3. J. Y. Cuan-Baltazar, M. J. Muñoz-Perez, C. Robledo-Vega, M. F. Pérez-Zepeda, and E. Soto-Vega, "Misinformation of covid-19 on the internet: infodemiology study," *JMIR public health and surveillance*, vol. 6, no. 2, p. e18444, 2020.
4. M. Babakar and W. Moy, "The State of Automated Factchecking," Full Fact, London, UK, Tech. Rep., 2016.
5. N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak *et al.*, "ClaimBuster: the first-ever end-to-end fact-checking system," *Proceedings of the VLDB Endowment*, vol. 10, no. 12, pp. 1945–1948, 2017.
6. P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, and I. Koychev, "A context-aware approach for detecting worth-checking claims in political debates," in *Proceedings of the International Conference Recent Advances in Natural Language Processing.* INCOMA Ltd., 2017, pp. 267–276.
7. L. Konstantinovskiy, O. Price, M. Babakar, and A. Zubiaga, "Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection," *ACM Digital Threats: Research and Practice*, 2020.
8. M. Hasanain, R. Suwaileh, T. Elsayed, A. Barrón-Cedeno, and P. Nakov, "Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims. task 2: Evidence and factuality." in *CLEF (Working Notes)*, 2019.
9. A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, pp. 1–36, 2018.
10. G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, and L. Derczynski, "Semeval-2019 task 7: Rumoureval, determining rumour veracity and support for rumours," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 845–854.
11. A. Zubiaga, M. Liakata, and R. Procter, "Exploiting context for rumour detection in social media," in *International Conference on Social Informatics.* Springer, 2017, pp. 109–123.
12. Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 1395–1405.
13. A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, and F. Haouari, "Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media," in *European Conference on Information Retrieval.* Springer, 2020, pp. 499–507.

14. A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, S. Shaar, and Z. Sheikh Ali, "Overview of CheckThat! 2020: Automatic identification and verification of claims in social media," ser. LNCS (12260), A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, and N. Ferro, Eds. Springer, 2020.
15. E. Kochkina, M. Liakata, and A. Zubiaga, "All-in-one: Multi-task learning for rumour verification," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3402–3413.
16. J. Ma, W. Gao, and K.-F. Wong, "Detect rumors in microblog posts using propagation structure via kernel learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2017, pp. 708–717.
17. M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 2227–2237.
18. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
19. M. Müller, M. Salathé, and P. Kummervold, "Covid-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter," *arXiv preprint arXiv:2005.07503*, 2020.
20. A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, "Crisislex: A lexicon for collecting and filtering microblogged communications in crises," in *Eighth international AAAI conference on weblogs and social media*, 2014.
21. B. Ghanem, M. Montes-y Gómez, F. Rangel, and P. Rosso, "Upv-inaoe-autoritas-check that: Preliminary approach for checking worthiness of claims," in *CLEF*, 2018.