Design and Evaluation of Deep Learning Models for Real-Time Credibility Assessment in Twitter

Marc-André Kaufhold(✉), Markus Bayer, Daniel Hartung, and Christian Reuter

Science and Technology for Peace and Security (PEASEC), Technical University of Darmstadt, Darmstadt, Germany
{kaufhold,bayer,reuter}@peasec.tu-darmstadt.de, daniel.hartung@student.tu-darmstadt.de

Abstract. Social media have an enormous impact on modern life but are prone to the dissemination of false information. In several domains, such as crisis management or political communication, it is of utmost importance to detect false and to promote credible information. Although educational measures might help individuals to detect false information, the sheer volume of social big data, which sometimes need to be analysed under time-critical constraints, calls for automated and (near) real-time assessment methods. Hence, this paper reviews existing approaches before designing and evaluating three deep learning models (MLP, RNN, BERT) for real-time credibility assessment using the example of Twitter posts. While our BERT implementation achieved best results with an accuracy of up to 87.07% and an F1 score of 0.8764 when using metadata, text, and user features, MLP and RNN showed lower classification quality but better performance for real-time application. Furthermore, the paper contributes with a novel dataset for credibility assessment.

Keywords: Credibility assessment · Social media · Neural networks · Deep learning

1 Introduction

Social media are an integral part of modern everyday life as they allow the creation and exchange of user-generated content. Besides everyday life, social media are used by journalists for reporting, analysing, and collecting information, by organisations to monitor customer feedback and sentiment, but also by citizens and emergency services to gain situational awareness in conflicts and disasters [18]. On the contrary, social media is prone to the dissemination of (potentially) false information, including conspiracy theories, fake news, misinformation, or rumors [35]. While counter-measures such as gatekeeping information, increasing media literacy, or passing new laws seem to be promising approaches [17], the sheer volume of *big social data*, which sometimes needs to be analysed under time-critical constraints, calls for automated and (near) real-time

I. Farkaš et al. (Eds.): ICANN 2021, LNCS 12895, pp. 396–408, 2021.
https://doi.org/10.1007/978-3-030-86383-8_32

credibility assessment methods. Thus, a multitude of different machine learning approaches were established to automatically distinguish false and credible information [27,34,38]. Despite their merits, when reviewing existing deep learning approaches for credibility assessment in social media, we found that most approaches provided binary or multi-class models but did not allow a steady (e.g., percentage) prediction of credibility. Furthermore, most approaches require extensive computations, thus lacking the ability for real-time application in social media, and, to our best knowledge, none of these approaches incorporated previous posts of the user into their analysis. Thus, the paper seeks to answer the following research question: **Which deep learning models and parameters are suitable for real-time credibility assessment in Twitter?**

By answering this research question, the paper makes several contributions. It (i) conducts a review of existing credibility assessment methods (Sect. 2), (ii) presents the design and finetuning of three deep learning models for credibility assessment, (iii) provides a novel dataset for credibility assessment in Twitter (Sect. 3), and (iv) evaluates the quality of the designed models, also examining the usefulness of incorporating previous user posts into credibility assessment (Sect. 4). The paper finishes with a discussion of the findings and implications and highlights possible limitations and potential for future work (Sect. 5).

2 Related Work

Since the study of *credibility* is highly interdisciplinary, there is no universal definition for it [8]. However, it can be understood as a measure which comprises both objective (e.g., useful, good, relevant, reliable, accurate) and subjective (e.g., a perception of the receiver) components. Credible information is characterized by trustworthiness (unbiased, true, good purpose) and expertise (competence, experience, knowledge) [10]. When estimating the credibility of information in social media, users are confronted with different types of harmful information [35] that can be distinguished by the *intention of the publisher* (i.e., intentional or non-intentional) and the *truth of content* (i.e., true or false) [8]. Both disinformation and misinformation are objectively false, but only disinformation (often referred to as fake news [21]) is published intentionally false. Moreover, rumors are statements that cannot be immediately verified as either true or false [28].

Amongst others, harmful information is disseminated to manipulate political elections and public opinions or to generate financial revenues [1]. Moreover, false information might affect the decision making of emergency services in conflicts or disasters, effectively contributing to the loss of lives. Countermeasures against harmful information comprise the gatekeeping of information by media, increasing the media literacy of citizens, passing new laws and regulations, or detecting harmful information via algorithmic detection approaches [17]. When reviewing literature on credibility assessment in social media (Table 1), we did not only find individual systems but also interesting survey papers comparing different machine learning approaches [27,34,38]. The approaches are primarily based on Twitter data, often attributed to different *domains*, including credibility, fake news, or rumors, and have a different *scope of analysis*.

First, event-based approaches cluster social media messages into events to determine the credibility of the event [2–4,9,12,40]. Second, propagation-based approaches analyse the caused engagement, such as mentions or retweets, of published messages [23,30,31]. Third, message-based approaches assess the credibility of individual messages, using metadata and textual features [11,13,15,28,39]. Especially *methods* based on neural networks achieved high classification performances, e.g., accuracies of 85.20% using NN [15] or 89.20% [31] using RNN and NN. Despite the variety of features involved, to our best knowledge, none of these approaches incorporated previously published messages of the user into their analysis. Further, only two of the approaches allow a near *real-time* application [23,39], i.e., being able to classify tweets directly after their dissemination. This is the case because most approaches are event-based, requiring event detection before classification can take place, or rely on temporal features, such as the number of likes or retweets, which change over the course of time and could lead to a flawed credibility score at retrieval.

In terms of *output*, most approaches allow a binary (i.e., credible or incredible information) or multi-class credibility assessment [11,23,39], although some works outlined that they do not reproduce reality in a sufficient manner [4,5]. Thus, the use of a steady regression seems promising [28] since it allows a percentage-based representation of credibility and better accounts for the subjective component of credibility [8]. Although multiple attempts have been made to establish standard datasets for credibility assessment [25,33,41], almost all publications used their own dataset, probably due to the methodological requirements of their approaches. The lack of standardized datasets is noticed by diverse authors, emphasizing the lack of comparability of the evaluation results of different approaches [27,34,38].

Table 1. Comparison of ML classifiers. Used methods are marked **bold**, sometimes not all methods are listed (*). Abbrev.: Decision Tree (DT), Decision Rule (DR), Bayesian Network (BN), Bayes Classifier (BC), Support Vector Machine (SVM), Random Forrest (RF), Convolutional/Recursive Neural Network (C/R/NN), Naive Bayes (NB).

Ref.	Domain	Scope	Output	Realtime	Methods
[3]	Credibility	Event	binary	no	**DT**, DR, BN, SVM
[4]	Credibility	Event	binary	no	**RF**, LR, *
[2]	Fake News	Event	binary	no	**RF**, *
[15]	Credibility	Message	binary	untested	**NN**
[30]	Astroturfing	Mem	binary	no	**DT**, SVM
[11]	Credibility	Message	5 classes	untested	**SVM-rank**, *
[40]	Crisis Credibility	Event	binary	untested	**SVM**, BN, DT
[12]	Credibility	Event	tertiary	no	**SVM**, DT, NB, RF
[28]	Rumors	Event	steadily	no	**BC**
[13]	Fake News	Message, Source	binary	untested	**NN**, NB, DT, SVM, RF
[31]	Fake News	Event	binary	no	**RNN**, **NN**
[23]	Fake News	Message	4 classes	almost	**RNN**, **CNN**
[39]	False Information	Message	5 classes	yes	**RNN**, **CNN**

3 Concept and Implementation

Based on the outlined research gaps, we seek to implement neural network-based approaches for credibility assessment that (i) work with public Twitter data due to the ease of access, (ii) use regression to allow a steadily (i.e., percentage-based) assessment of credibility, and (iii) allow a near real-time application of the trained models. Furthermore, we intend to (iv) check if the analysis of previously published messages of a user positively impacts the performance of credibility assessment. We also (v) compose a novel dataset for credibility assessment.

3.1 Features and Model

In order to train our models, we reviewed the features used by previous approaches. The used features can be roughly categorized into four types: (i) **metadata features** (n = 10) of a tweet provided by the Twitter API, such as hashtags, links, or mentions, (ii) computationally extracted **text features** (n = 25) from the tweet's body, such as number of words, text length, or sentiment, (iii) **user features** (n = 17) provided by the Twitter API, such as the number of followers or published tweets, and (iv) **timeline features** (n = 140), including the maximum, minimum, arithmetic mean, and standard deviation (4 *) of both the metadata and text features (10 + 25) of the last 40 tweets of the user.

First, the baseline model is a simple **multilayer perceptron (MLP)** that consists of an input layer with 192 neurons for the features described before. These are projected into a hidden layer with 32 neurons with tanh activations. Since the problem to solve is a regression task, a sigmoid function was selected for the output and the entire network is trained with mean square error (MSE). The layers are fully connected with a dropout rate of 0.3. Further hyperparameters of the learning process are a learning rate of 0.01, a batch size of 256, and the maximum number of epochs of 10,000, which is contained by early stopping on the development set. We choose ADAM [20] as optimizer.

Second, to extend the baseline, we embed the sentiment and textual content of a tweet with a **recurrent neural network (RNN)** and feed those tweet embeddings into the baseline model. As a first step, GloVe [26] pretrained Twitter embeddings (dimension of 50) are utilized to create word embeddings of each word in the tweet. These word embeddings are enriched by another dimension that represents the VADER [14] sentiment value. Every embedding is then processed by a RNN that produces a hidden state (tweet embedding) that serves as another input into the baseline MLP.

Third, another approach to extend the baseline is to use finetuned **BERT embeddings** [6]. We replaced user mentions, URLs, and emoticons with special tokens. Then we finetuned the base BERT model with its CLS token as output with the training data (batch size: 16, learning rate: $5 \cdot 10^{-5}$ and 3 epochs).

The finetuned model is then used to produce the additional input for the baseline. The BERT connection to the baseline is regularized by a dropout connection with rate 0.3. As the dimensionality of the BERT embeddings is substantially higher, we increased the number of hidden neurons of the baseline to 128.

3.2 Automatic Dataset Composition

The task of credibility assessment requires much data from different topics and time frames to make the model invariant to these patterns. Accordingly, we searched for Twitter datasets that can be combined into a larger set. The **PHEME** [41] dataset contains 300 binarily annotated posts from which both the classes "true" and "false" are mapped to a credibility score of 1 and 0 respectively in our coding schema. In contrast, the **Twitter15** [22] and **Twitter16** [24] datasets are categorized into four classes. The classes "true" and "false" are mapped analogously to the PHEME dataset. For the tweets of the class "unverified", we decided for a more uncertain score of 0.3 that reflects a tendency towards dubious content. After the manual inspection of the last class "no rumor", we chose a score of 0.9 as instances of this class seem to be primarily true. This way, additional 2,308 instances were added to the corpus of this paper.

Then, we implemented an automatic coding scheme for the **FakeNewsNet** [33] dataset. It contains several topics to which a large number of tweets are assigned. For each topic, an associated headline was labelled as true or false. Since the assignment of tweets to the topics was carried out using keywords, posts may also be incorrectly assigned to a topic and a mapping from the headline label to the label of tweets in it is not possible. It also allows a topic that is annotated as "false" to contain posts that expose the topic as wrong, which is especially important in our use case. To compensate for this, we perform a temporal, keyword-based, similarity, and topic filtering, which is described in detail in the appendix [16]. Thus, 1,378 credible and 729 implausible tweets were retrieved and mapped to a target score of 0.9 and 0.1, respectively.

We also used the **Twitter20** [16] dataset, where various German tweets of the COVID-19 pandemic were individually labelled. The following assumptions have to be true so that a translated version can be included into the dataset of our paper: (i) incorrect information in German and English are syntactically the same, (ii) the dataset contains only a few posts with misinformation or satire, and (iii) during the translation of the articles no linguistic properties (e.g., rhetorical stylistic devices) that are a characteristic of misinformation are lost. We used the Google Translate API and automatically corrected wrong @ and # placements, to preserve the general syntax of tweets. Since the assumptions do not necessarily have to apply, we have decided to create a "default" dataset (Fig. 1) without and a "large" dataset (Fig. 2) with these translated instances.

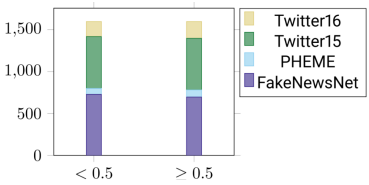

Fig. 1. Default dataset (N = 3,178, whereof $n_{credible} = 1{,}589$).

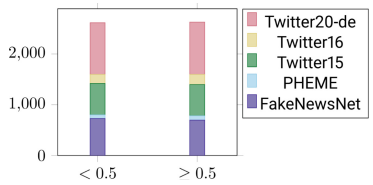

Fig. 2. Large dataset (N = 5,225, whereof $n_{credible} = 2{,}619$).

4 Evaluation

For the evaluation of our models, we use metrics based on regression fitting to the dataset development, classification for comparisons with past and future work, and execution time for insights related to real-time application. For the classification, tweets with a score of less than 0.5 are classified as 0 and 1 otherwise. We split 80%, 10% and 10% of the posts into training, development and testing sets. We also ensure that all posts by a user who is represented more than once in the dataset are included in the training set so that no information of the other sets is already seen during training. For the implementation we used PyTorch, Huggingface Transformers and NLTK. The system used for the evaluation has an Intel i7-9750 with 6 cores and 2.6 GHz, a NVIDIA GeForce RTX 2070 graphics card with 8 GB Graphics memory and 32 GB of RAM.

4.1 Evaluation of Model Quality

For the evaluation of the performance of the different model architectures and feature combinations, we first tuned the hyperparameters (e.g., batch size and dropout-rate) on the development set of both datasets. We chose to proceed with the models that have the lowest MSE. The evaluation results regarding the development set can be found in the appendix [16].

The testing set results on both datasets are shown in Table 2. It is clear to see that the adaption of the standard MLP model is very beneficial. Especially the BERT model can gain additive accuracy improvements of up to 21.63%. Looking at the feature constellations in the MLP network, it is evident that they are suitable for distinguishing credible and implausible posts without the addition of sentence or BERT embeddings (reaching up to 66.77% accuracy and 0.6513 F1 score). The Tweet, user and text feature constellation even reaches a slightly better MSE than the RNN basis model on the default dataset.

The more sophisticated BERT-based model, however, draws less benefit from the additional feature inputs. With the default set, a minimal improvement of less than 1 accuracy point is achieved, while the features for the large set even degrade the BERT model. Sometimes we noticed improvements from the timeline feature, but no significant results were found when the test set evaluation was

performed. When inspecting relative changes with regard to both datasets, it becomes apparent that the BERT-based model has a greater positive impact on the first set. The additive accuracy improvements on this dataset are of up to 21.63% compared to just up to 14.12% on the large dataset.

Table 2. Results of the quality analysis per model on both datasets. Abbrev.: Tweet Features (Tw), User Features (Us), Text Features (TX), Advanced timeline features (ATi).

Model	Features	MSE	Acc	Pre	Rec	F1
MLP (default)	Base	–	–	–	–	–
	Tw, Us, Tx	0.1367	0.6552	0.6323	0.6490	0.6405
	Tw, Us, Tx & ATi	0.1474	0.6677	0.6471	0.6556	0.6513
RNN (default)	Base	0.1380	0.7116	0.6879	0.7152	0.7116
	Tw, Us, Tx	0.1202	0.7367	0.7190	0.7285	0.7237
	Tw, Us, Tx & ATi	0.1199	0.7429	0.7226	0.7417	0.7320
BERT (default)	Base	0.0806	0.8621	0.8497	0.8609	0.8553
	Tw, Us, Tx	**0.0794**	**0.8715**	**0.8618**	**0.8675**	**0.8674**
	Tw, Us, Tx & ATi	0.0805	0.8621	0.8591	0.8477	0.8533
MLP (large)	Base	–	–	–	–	–
	Tw, Us, Tx	0.1803	0.6469	0.6734	0.6162	0.6435
	Tw, Us, Tx & ATi	0.1841	0.6412	0.6895	0.5572	0.6163
RNN (large)	Base	0.1720	0.7042	0.7266	0.6863	0.7059
	Tw, Us, Tx	0.1639	0.7118	0.7143	0.7380	0.7260
	Tw, Us, Tx & ATi	0.1603	0.7042	0.7538	0.6679	0.7002
BERT (large)	Base	0.1347	**0.7844**	0.7883	**0.7970**	**0.7927**
	Tw, Us, Tx	0.1339	0.7824	0.7897	0.7897	0.7897
	Tw, Us, Tx & ATi	**0.1319**	0.7824	**0.7962**	0.7786	0.7873

From a dataset development perspective, one might think that the larger dataset contains more false annotated data, since the classifier scores are worse on this dataset. This can apply, e.g., if one of the assumptions given in Sect. 3.2 is incorrect and significant linguistic properties were lost during the translation of the Twitter20 dataset. Another consideration could be that the large dataset covers more different or domain-specific tweets; this makes classification more difficult but increases the generalizability and practicality of a classifier. When inspecting the translated posts in the Twitter20 dataset, we noticed some mistakes in the translated text. However, we tend to the second explanation as the actual content was preserved most of the time and we were still able to identify the credibility.

This consideration comes also into play when inspecting the stronger impact of the BERT-based model on the default dataset. BERT can have a major impact

when it is applied to a dataset with fewer data instances as it can transfer knowledge from its previously learned tasks. The other algorithms can only get closer to the evaluation results if the dataset grows in its size, since they do not have this initial capacity.

Furthermore, the BERT model just slightly improves with the feature engineering process while the features seem more useful when applied to the other models. This might be due to the high capabilities of the pre-trained model. The incorporation of textual features might be redundant as the language model is able to identify some of these by itself. Some of the features might even be misleading and in this low data regime unwanted statistics are more likely to appear during the training process leading to better scores for the other models that do not have the generalization capabilities of BERT. The RNN model builds upon GloVe embeddings which also impose a certain generalization that is reflected in the results. However, with this model we still expect a bias towards unwanted statistics.

4.2 Evaluation of Model Execution Time

To measure the execution time of our models, all tweets and previous posts of the large dataset were loaded to the RAM in order to reduce variances of HDD memory access. The individual models were executed using the whole dataset to measure the execution time of different steps, such as the model initialization, the processing time per tweet, and the processed tweets per second (see Table 3). For the RNN modell, there are two options to read the required embeddings: (i) a filesystem-based approach that reads and indexes the embedding file once (fs) and (ii) a memory-based approach, where the whole file is loaded into RAM. While the first approach consumes less memory and has a shorter initialization time, the second approach offers a faster access to the embeddings. For the first approach, the use of an SSD ($\approx 3,500\,\text{MB/s}$ reading speed) or HDD ($\approx 100\,\text{MB/s}$ reading speed) did not yield measurable differences in execution time.

The BERT model strongly benefits from GPU acceleration. For comparison, Table 3 highlights the execution times with (gpu) and without (cpu) acceleration by a graphics card. For other models, the use of a GPU did not yield measurable performance improvements. Generally spoken, complex models require longer execution times than simple models. The base RNN model is seven times faster than BERT using a GPU and more than 110 times faster than BERT without a GPU. Furthermore, the processing of timeline features, i.e., incorporating up to 40 previous posts of a tweet, requires significantly more time. While the RAM-based RNN model is able to classify up to 5.5k tweets per second, BERT processes up to 133 tweets per second with a GPU, but only 6.6 without a GPU. In that model, additional features show negligible impact on the overall execution time.

Table 3. Results of the temporal analysis per model. The column *tweets/second* ignores the initialization time.

	Configuration	Initialization	Time/Tweet	Tweets/Second
Features	Text	203 ms	914 µs	1094
	Tweet	0 s	144 µs	6944
	User	60 ms	130 µs	7692
	Timeline	265 ms	40,000 µs	25
MLP	Basis	1,400 ms	44 µs	22,727
	Adv. Timeline	1,430 ms	52,000 µs	19.2
RNN	Basis (fs)	7,653 ms	1,354 µs	738
	Basis (ram)	32 s	179 µs	5586
	Adv. Timeline (fs)	7,653 ms	107,000 µs	9.3
	Adv. Timeline (ram)	32 s	88,000 µs	11.3
BERT	Basis (gpu)	3,706 ms	7,495 µs	133
	Basis (cpu)	3,720 ms	150,000 µs	6.6
	Adv. Timeline (gpu)	3,680 ms	224,000 µs	4.4
	Adv. Timeline (cpu)	3,695 ms	>4 s	<¼

5 Discussion and Conclusions

Nowadays, social media is widely used for multiple purposes, such as relationship maintenance, journalism, customer interactions but also for crisis management. However, these activities can be severely impeded by the propagation of false information. Hence, it is important to promote credible and to counter implausible information. In this work, we reviewed existing approaches before designing and evaluating three neural network models capable of near real-time credibility assessment in Twitter to answer the following research question: **Which deep learning models and parameters are suitable for real-time credibility assessment in Twitter?**

Our findings indicate that our BERT-based model achieves the best results when using metadata, text, and user features, reaching an accuracy of 87.07% and F1 score of 0.8764 on the default dataset. In comparison to existing works, the results appear to be promising. While Helmstetter and Paulheim [13] reached an F1 score of 0.7699, Iftene et al. [15] achieved an accuracy of 85.20%. Although Ruchansky, Seo, and Liu [31] reached an accuracy of 89.20% and F1 score of 0.9840, their approach is propagation-based, thus having limited real-time capability, and classifies events instead of individual tweets. Similarly, Liu and Wu [23] reach an F1 score of 0.8980; however, their approach focuses on the detection of disinformation and also incorporates propagation-based features.

Furthermore, we compared the real-time capabilities of our three models. While our MLP baseline is capable of processing high volumes of data (>20k tweets/sec) with a low resource demand, the accuracy of up to 66.77% does not

allow for a reliable classification. In contrast, our RNN model is still capable of processing high volumes of data (>5k tweets/sec when used in RAM) for classification while reaching more promising accuracies of up to 74.29%. Finally, BERT reached accuracies of 87.07% but was only able to process a considerably lower amount of data (>0,1k tweets/sec when used with a GPU). When including the previous posts of the user into computation, we did not achieve consistent improvements of classification performance; however, the real-time capability of all feature and model combinations was lost.

5.1 Practical and Theoretical Implications

We compared existing datasets and combined suitable ones into a **novel dataset to increase the amount of available data for model training (C1)**. Since available datasets are used for varying credibility classification tasks, several steps of transformation were required to convert them into a unified structure. Due to the combination of datasets, it comprises a richer number of users, topics, and message characteristics. Our future work will include the application of data augmentation techniques to increase the size and richness of the dataset.

We provided a **review of existing machine learning approaches for credibility assessment in Twitter (C2)**. In contrast to other works, we critically examined and compared approaches for credibility assessment in Twitter. Many of the reviewed approaches did not use a development set, relied on a small dataset, or conducted many hyperparameter optimizations, which entails the risk of overfitting. While difficult to compare, it seems that propagation-based models achieve the best results [31]; however, they lack the ability of real-time application. As the engagement based on tweets unfolds over time, propagation-based models seem promising when no time constraints are present.

In addition, our work contributes with **insights into the real-time capability of neural networks for credibility assessment (C3)**. Comparing our models in terms of real-time capability, their usefulness seems to be dependent on many factors. While our MLP baseline shows excellent execution times for large-volume data processing, the lack of classification performance disqualifies its real-world applicability. In contrast, our RNN model still offers suitable execution times and maintains a better classification performance. Finally, despite achieving the best classification results, BERT offers limited realtime capability when used for large-scale data analysis unless considerable GPU power is used for processing. In the end, we would still advise to do further research regarding the BERT model, as it has the best generalization capabilities. The credibility research shows that simple algorithms tend to be very biased towards the topics and domains in the dataset and often bahave more like a topic classifier.

5.2 Limitations and Outlook

While this work is subject limitations, they also offer potentials for future research. First, although BERT achieved the best classification results, it was also the slowest classifier. Variations of BERT, such as DistilBERT [32], provide

smaller models or shared weightings within the model to achieve a lower memory usage and faster execution time. Thus, future work could examine if they achieve comparable classification results for credibility assessment. Second, the classifier is limited by merely using a dataset based on textual Twitter data. Although it can be used for other social media, it might perform worse due to different linguistic features. Thus, the exploration of a cross-platform dataset, supported by active learning and data augmentation techniques, could be worthwhile for future research [19]. Furthermore, pictures displaying text messages (requiring optical character recognition techniques) or external sources could be incorporated into the credibility assessment concept. Third, novel but similar publications emerged during the implementation of our study. For instance, Tian et al. [37] contributed with a rumor detection algorithm that achieves an F1 score of 0.862 but does not provide a steady regression of findings. A further work used ALBERT to reduce the memory usage of BERT, reaching an F1 score of 0.795 compared to 0.71 of the original BERT model [36]. Furthermore, additional research was conducted to detect fake news spreaders by analyzing their previous posts [7,29].

Acknowledgements. This work has been co-funded by the German Federal Ministry of Education and Research (BMBF) in the project CYWARN (No. 13N15407) and by the BMBF and the Hessen State Ministry for Higher Education, Research and Arts (HMKW) within the SecUrban mission of the National Research Center for Applied Cybersecurity ATHENE.

References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. J. Econ. Perspect. **31**(2), 211–236 (2017)
2. Buntain, C., Golbeck, J.: Automatically identifying fake news in popular twitter threads. In: IEEE Proceedings of (SmartCloud), pp. 208–215 (2017)
3. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of WWW, p. 675 (2011)
4. Castillo, C., Mendoza, M., Poblete, B.: Predicting information credibility in time-sensitive social media. Internet Res. **23**(5), 560–588 (2013)
5. Conroy, N.K., Rubin, V.L., Chen, Y.: Automatic deception detection: methods for finding fake news. Proc. ASIS&T **52**(1), 1–4 (2015)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
7. Duan, X., Naghizade, E., Spina, D., Zhang, X.: RMIT at PAN-CLEF 2020: proling fake news spreaders on twitter. In: CLEF 2020 (2020)
8. Flanagin, A.J., Metzger, M.J.: Digital media and youth: unparalled opportunity and unprecedented responsibility. In: Flanagin, A.J., Metzger, M.J. (eds.) Digital Media, Youth, and Credibility, pp. 5–28 (2008)
9. Floria, S.A., Leon, F., Logofătu, D.: A credibility-based analysis of information diffusion in social networks. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) Proceedings of ICANN, pp. 828–838 (2018)
10. Fogg, B.J., Tseng, H.: The elements of computer credibility. In: Proceedings of CHI, pp. 80–87 (1999)

11. Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: TweetCred: real-time credibility assessment of content on twitter. In: Aiello, L.M., McFarland, D. (eds.) Social Informatics, vol. 8851, pp. 228–243 (2014)
12. Hassan, D.: A text mining approach for evaluating event credibility on twitter. In: Proceedings of WETICE, pp. 171–174 (2018)
13. Helmstetter, S., Paulheim, H.: Weakly supervised learning for fake news detection on twitter. In: Proceedings of ASONAM, pp. 274–277. IEEE (2018)
14. Hutto, C., Gilbert, E.: VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of ICWSM (2015)
15. Iftene, A., Gifu, D., Miron, A.R., Dudu, M.S.: A real-time system for credibility on twitter. In: Proceedings of LREC, pp. 6166–6173 (2020)
16. Kaufhold, M.A., Bayer, M., Hartung, D., Reuter, C.: Paper Appendix (2021). https://github.com/mkx89-sci/KaufholdBayerHartungReuter2021_ICANN
17. Kaufhold, M.A., Reuter, C.: Cultural violence and peace in social media. In: Reuter, C. (ed.) Information Technology for Peace and Security - IT-Applications and Infrastructures in Conflicts, Crises, War, and Peace, pp. 361–381 (2019)
18. Kaufhold, M.A.: Information Refinement Technologies for Crisis Informatics: User Expectations and Design Principles for Social Media and Mobile Apps (2021)
19. Kaufhold, M.A., Bayer, M., Reuter, C.: Rapid relevance classification of social media posts in disasters and emergencies: a system and evaluation featuring active, incremental and online learning. IP&M **57**(1), 102132 (2020)
20. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2017)
21. Lazer, D.M.J., et al.: The science of fake news. Science **359**(6380), 1094–1096 (2018)
22. Liu, X., Nourbakhsh, A., Li, Q., Fang, R., Shah, S.: Real-time rumor debunking on twitter. In: Proceedings of CIKM, pp. 1867–1870 (2015)
23. Liu, Y., Wu, Y.F.B.: Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
24. Ma, J., et al.: Detecting rumors from microblogs with recurrent neural networks. In: IJCAI International Joint Conference on Artificial Intelligence, pp. 3818–3824 (2016)
25. Mitra, T., Gilbert, E.: CREDBANK: a large-scale social media corpus with associated credibility annotations. In: Proceedings of ICWSM (2015)
26. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of EMNLP, pp. 1532–1543 (2014)
27. Pierri, F., Ceri, S.: False news on social media: a data-driven Survey. ACM SIGMOD Record **48**(2), 18–27 (2019)
28. Qazvinian, V., Rosengren, E., Radev, D.R., Mei, Q.: Rumor has it: identifying misinformation in microblogs. In: Proceedings of EMNLP, pp. 1589–1599 (2011)
29. Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th Author Proling Task at PAN 2020: Proling Fake News Spreaders on Twitter (2020)
30. Ratkiewicz, J., et al.: Truthy: mapping the spread of astroturf in microblog streams. In: Proceedings of WWW, p. 249 (2011)
31. Ruchansky, N., Seo, S., Liu, Y.: CSI: a hybrid deep model for fake news detection. In: Proceedings of CIKM, pp. 797–806 (2017)
32. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (2020)
33. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: FakeNewsNet: a data repository with news content, social context and dynamic information for studying fake news on social media. Big Data **8**(3) (2018)

34. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
35. Tandoc, E.C., Lim, Z.W., Ling, R.: Defining "fake news": a typology of scholarly definitions. Digit. Journal. **6**(2), 137–153 (2018)
36. Tian, L., Zhang, X., Peng, M.: FakeFinder: twitter fake news detection on mobile. In: Companion Proceedings of the Web Conference, vol. 2020, pp. 79–80 (2020)
37. Tian, L., Zhang, X., Wang, Y., Liu, H.: Early detection of rumours on twitter via stance transfer learning. In: Jose, J.M., et al. (eds.) Advances in Information Retrieval, vol. 12035, pp. 575–588 (2020)
38. Viviani, M., Pasi, G.: Credibility in social media: opinions, news, and health information-a survey: credibility in social media. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. **7**(5), e1209 (2017)
39. Wu, L., Rao, Y., Yu, H., Wang, Y., Nazir, A.: False information detection on social media via a hybrid deep model. In: Staab, S., Koltsova, O., Ignatov, D.I. (eds.) SocInfo 2018. LNCS, vol. 11186, pp. 323–333. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01159-8_31
40. Xia, X., Yang, X., Wu, C., Li, S., Bao, L.: Information credibility on twitter in emergency situation. In: Chau, M., Wang, G.A., Yue, W.T., Chen, H. (eds.) PAISI 2012. LNCS, vol. 7299, pp. 45–59. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30428-6_4
41. Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., Tolmie, P.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE **11**(3), e0150989 (2016)