Transformer Based Models in Fake News Detection

Sebastian Kula[1,2](✉), Rafał Kozik[1], Michał Choraś[1], and Michał Woźniak[1,3]

[1] UTP University of Science and Technology, Bydgoszcz, Poland
[2] Kazimierz Wielki University, Bydgoszcz, Poland
skula@ukw.edu.pl
[3] Wrocław University of Science and Technology, Wrocław, Poland

Abstract. The article presents models for detecting fake news and the results of the analyzes of the application of these models. The precision, f1-score, recall metrics were proposed as a measure of the model quality assessment. Neural network architectures, based on the state-of-the-art solutions of the Transformer type were applied to create the models. The computing capabilities of the Google Colaboratory remote platform, as well as the Flair library, made it feasible to obtain reliable, robust models for fake news detection. The problem of disinformation and fake news is an important issue for modern societies, which commonly use state-of-the-art telecommunications technologies. Artificial intelligence and deep learning techniques are considered to be effective tools in protection against these undesirable phenomena.

Keywords: Fake news detection · Transformers · Natural language processing · Deep learning · SocialTruth

1 Introduction

The dynamic development of social media, instant messaging, internet information portals, and means of electronic communication resulted in a significant decrease in the influence and importance of the traditional mass media, such as radio, television and paper (printed) press. Modern societies use the technical innovations offered by technology in the ICT area with great enthusiasm.

These dynamic changes are accompanied by new phenomena that may potentially have a destructive impact on society, and may undermine the credibility of institutions, governments and companies. Fake news is one such undesirable phenomenon that is present in the ICT revolution. Fake news is defined as deliberate disinformation, as an action aimed at causing disorder and information chaos through false or partially true messages. This phenomenon, initially unnoticed, is growing. The scale of the impact of this type of practices on society is evidenced by the fact that it is commonly accepted that fake news influenced the results of political elections or referenda in politically significant and important countries. In these countries, despite them having high levels of democratic standards and well-established electoral mechanisms, large social groups

M. Paszynski et al. (Eds.): ICCS 2021, LNCS 12745, pp. 28–38, 2021.
https://doi.org/10.1007/978-3-030-77970-2_3

were successfully manipulated through fake news. Fake news can also be used to manipulate public health, economic marketing, product sales, and public safety. Considering the above, many institutions, governments, authorities responsible for security, as well as companies for which ethical standards and the ethos of the institution are important, significantly increased their interest in the phenomenon of fake news and effective methods of preventing it.

In the fight against fake news, machine learning methods are very successful, allowing the creation of models that are not only accurate and effective, but also practical, ensuring the detection of an undesirable phenomenon almost in real time. Automatic real-time disinformation detection is a must. The enormous amount of information that reaches the average recipient every day means that the models that do not allow for swift detection of disinformation are not applicable.

Among the machine learning methods, the methods based on deep learning show considerable promise. The DL (Deep Learning) methods have been created based on the learning processes taking place in the human brain. It is expected that through DL, it will be possible to create models that will not only detect the already identified types of disinformation, but will be able to act in advance and identify new, currently unknown techniques of disinformation. The big advantage of the DL methods is that they can be trained without feature engineering or with relatively little application of the feature engineering. This allows for the presumption that a neural network based on DL will automatically detect the features, which are characteristic for fake news, and that it will do it better than a human.

Combating disinformation requires the use of NLP (Natural Language Processing) techniques, which replace a journalist, linguist or media expert in the process of evaluating the credibility of information. The NLP procedures imitate human processes; one such process is intuitive reasoning. The intuitive reasoning is considered to be a key element in specialists with extensive professional experience. It is a largely subconscious process resulting from a long process of learning and gaining experience. The phenomenon of intuitive reasoning in humans is reproduced in computer deep learning algorithms, where a relatively large amount of data allows the detection of patterns typical of these data, without the need to engineer the features before training the neural network.

As regards fake news, the DL-based models allow detection of the features that are associated with written texts, such as stylistics, phraseology, syntax, semantics, pragmatics, morphology, i.e., the features that are characteristic of the author of fake news. Thus, if the author is a human, it should be assumed that it will be possible to detect the literary features that are constant for a given author.

The article presents methods based on deep learning models and Transformers architectures. The main goal of this work was to create reliable models allowing the verification of short texts, such as article titles, available on the web. To meet this challenge, the focus was on the analysis of the architectures that are considered to be revolutionary, state-of-the-art methods for many NLP tasks.

The article describes related works in the Sect. 2, the description of Transformers architectures is presented in the Sect. 3, the Sect. 4 is a description of the proposed methods using the Flair library, the conducted experiments are included in the Sect. 5. Results and application of the models for verification of scraped web pages titles are presented in Sects. 6 and 7, respectively.

2 Related Works

Along with the development of AI (Artificial Intelligence) methods, applied in various research and engineering domains, the methods related to NLP are also developing dynamically. Text classifiers, which use better and more precise ML methods, play a special and essential role in defending against fake news. Until recently, the most commonly applied methods used in text classification were algorithms derived from the Naive Bayes, SVM (Support Vector Machines), CNN (Convolution Neural Networks) and RNN (Recurrent Neural Networks) algorithms, which are a type of the DNN (Deep Neural Network). Currently, the methods based on Transformers architectures are achieving outstanding results.

In [21], the authors use the SVM method and the feature selection method to reduce the data size. Their method has been validated using already classified datasets. The authors of the article [5] used the Naive Bayes and SVM methods to classify 327884 pieces of information from Twitter as believable and unbelievable, reporting very high accuracy, up to 99.9%. In another work, an existing classified dataset was used to create an overview of five ML models detecting fake news [3]. The authors analyzed classifiers based on Naive Bayes, Logistic Regression, Linear SVM, Stochastic Gradient Classifier and Random Forest Classifiers. The obtained results showed that SVM and Logistic Regression have the best performance on the applied dataset in the model [3]. The authors of [17] focused on deep learning by applying LSTM (Long Short-Term Memory), CNN and pre-trained GloVe embeddings in their model. A similar hybrid architecture based on the Flair library [4] was presented in [10], where pre-trained GloVe at the word embeddings level and RNN at the document embeddings level were applied. A big change in the classifiers of texts was brought about by the development of Transformer techniques and architectures using attention models. Models based on these techniques are ahead of other, previously used models in terms of the obtained metrics results; simultaneously, they are strongly parallel, which results in the possibility of training on parallel platforms, including the GPUs. This is of essential importance when training the neural network and optimizing the computing cost, which is high in the case of big data. Transformer based methods, specifically BERT, were applied in [9], where BERT base architectures were applied for fake news detection. In [13], BERT was applied to build models relating to the credibility of the texts, based on the database, which contains 20015 news articles, labeled as fake or true [19]; the accuracy of 98% was achieved.

The recent and broad systematic mapping study of the fake news detection techniques is presented in [12].

3 Transformers Architectures

The emergence of the self-attention mechanism and Transformer architectures meant that in NLP algorithms, the context in the sentence is much more important than the words themselves. This mechanism caused it that the binary representation of the word (token) is not constant and changes depending on the surroundings of the words (tokens) in the sentence. The use of Transformers for NLP-related tasks was proposed in [16], where they are presented as departing from recursion in favor of the attention mechanism.

Transformer-based methods are present in many architectures; this article uses the following architectures: BERT [7], RoBERTa [11], DistilBERT [14], xlNet [20], DistilGPT2 [18]. They differ mainly in the size and number of the layers of the neural network applied. The base version of BERT architecture contains 109 million parameters in the case of the cased corpus applied for training, and 110 million parameters for the situation, when it is trained on the uncased corpus; BERT large, in turn, contains 335 million parameters [18]. DistilBERT architectures, trained from the uncased corpus contain 66 million parameters, DistilGPT2 82 million, RoBERTa large 355 million parameters, xlNet trained from the corpus cased 340 million parameters [18]. Table 1 lists the details of the Transformer-based architectures.

Table 1. Parameters values of transformer architectures [18]

Transformer-based architecture	Number of parameters in millions	Number of layers	Hidden states size	Number of self-attention heads
BERT base cased	109	12	768	12
BERT base uncased	110	12	768	12
BERT large cased	335	24	1024	16
BERT large uncased	336	24	1024	16
DistilBERT uncased	66	6	768	12
DistilGPT2	82	6	768	12
RoBERTa large	355	24	1024	16
xlNet large cased	340	24	1024	16

4 Transformer Based Classifiers

In this work, the Flair [4] library was applied to create text classifier models that detect fake news. This library allows, in addition to choosing the pre-trained architecture, to define the embeddings technique, whether it be at the word, sentence or document level. Selection in the Flair is made with the use of the TransformerDocumentEmbeddings command, which causes the sentence-level embedding to be extracted, and with the use of the DocumentRNNEmbeddings command that document-level embeddings is extracted [4]. Both embeddings techniques were used in the work to create models.

The choice of the embeddings technique modifies the architecture of the neural network. The application of the document level embeddings, i.e., producing vector representations of the entire document, results in adding an additional layer in the architecture [4]. In the presented work, this additional layer is the GRU (Gated Recurrent Units) layer. The second key element modifying the architecture, applied to create the classifier models is the selected pre-trained, Transformer-based embeddings architecture. In the work, selected Transformer-based architectures, listed in Table 1 were applied with maintaining the values of the parameters of these architectures presented in the table. The architecture applied to create the classifier models is therefore dynamic and not the same for all models. In the article, the following designations for the architectures used to create models, based on sentence level embeddings and Transformer-based architectures are introduced: distilbert-base-uncased, distilgpt2, roberta-large, xlnet-large-cased, bert-large-cased_TDE. Architecture based on BERT and document level embeddings is marked as bert-large-cased_DRE. It does not differ from the bert-large-cased_TDE regarding the implemented Transformer-based architecture, which is BERT; the difference between architectures is the additional GRU layer.

The model creation process was as follows: the first step consisted in selecting ready-made databases containing article titles that were already classified as fake and true, next the pre-processing of all data in datasets, setting the dynamic architecture by the selecting the pre-trained Transformer-based architecture and choosing embedding technique, the network training based on the selected version of the dynamic architecture, and the last step was the creation of the model for classification. The described process is shown in Fig. 1. The process is repeatable and depicts all the steps that were performed for all the models created.

5 Experiment Setup

This section details the experiments conducted, including the description of the data applied and the hyperparameters values set in the training routine.

5.1 Collections

Two collections were created for the experiments, based on the data repository FakeNewsNet, KaiDMML [15]. Both collections contain article titles, grouped as fake and true. Collection 1 contains the titles of news content that were collected using the fact-checking website Politifact [8], and collection 2 contains data collected using GossipCop [1]. Collection 1 contains a total of 1056 items, of which 432 are labeled as fake and 624 are labeled as real (true). The titles of the articles in this collection are relatively short, because the maximum length of the word sequence is 340 characters. The primary sources of origin of the articles vary,

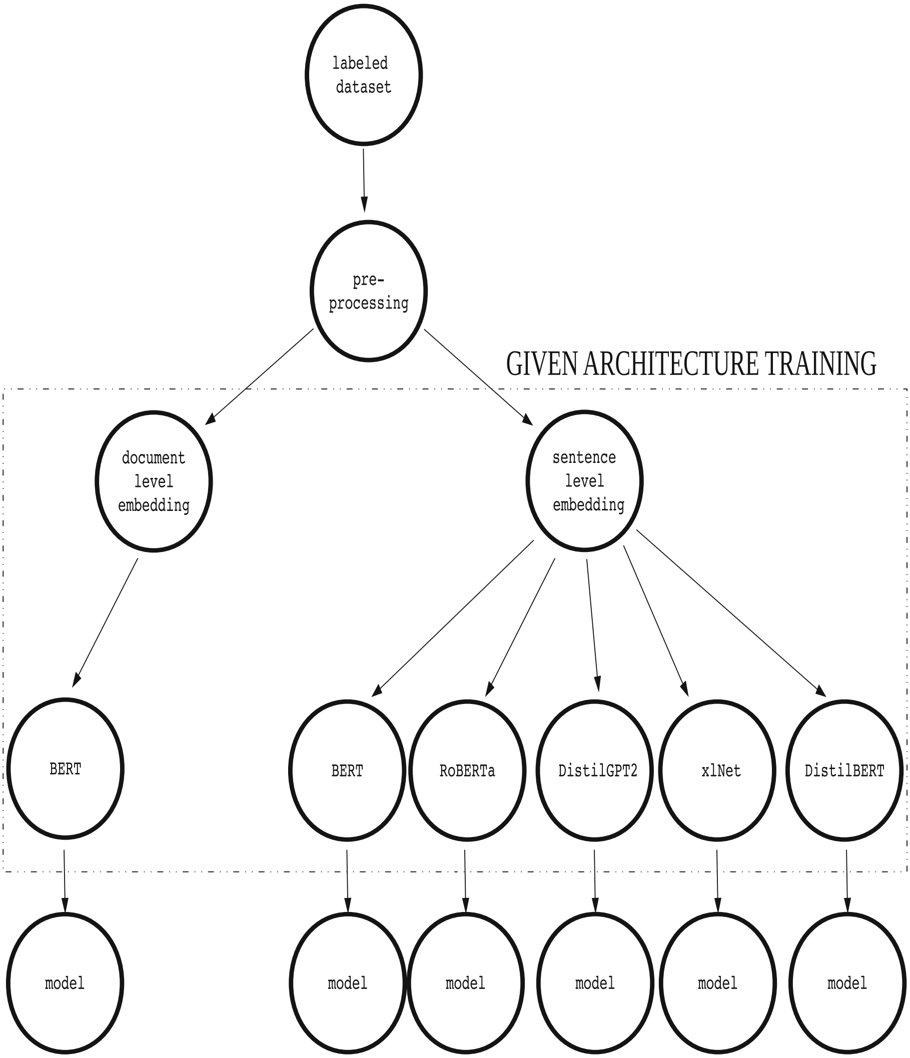

Fig. 1. The proposed processing pipeline

but 7.5% of the items labeled as true came from the www.youtube website, and 3.5% of the items classified as fake came from the www.yournewswire website.

Collection 2 is larger, with 16,817 items classified as true and 5323 items classified as fake. The number of article titles in the fake category is over three times smaller than the number of titles in the true category. Like in the collection 1, the web sources are differentiated with the leading portals; in the case of the titles of articles in the true category, 9.3% are derived from people.com portal, and in the case of fake articles, 8.6% are derived from the hollywoodlife.com portal.

5.2 Models Training

Before the training, pre-processing was performed and hyperparameters were set. There are different approaches to pre-processing; in this paper the maximum reduction was chosen, assuming that the titles of the articles mainly have an informative function and the linguistic correctness or punctuation were considered as secondary elements in this context. The assumption was based on the premise that the linguistic correctness is of secondary importance in spoken and colloquial language. Social media and internet portals often use simplified, abbreviated or even colloquial speech. As part of the pre-processing, punctuation marks, possible emoticons, website addresses, http and e-mail addresses were removed from the collections. As a result of the reduction, pure text was obtained, which required less computing power while training the neural network.

The adopted values of hyperparameters vary, depending on the collection and the adopted embeddings technique (either sentence-level or document-level embeddings). The hyperparameters adopted for the experiments related to collection 1 are presented in Table 2, and the parameters adopted for the experiments related to collection 2 are presented in Table 3.

Table 2. Hyperparameters values of experiments for the collection 1

Hyperparameter	Hyperparameter values for document-level embeddings	Hyperparameter values for sentence-level embeddings
Max number of epochs	5	5
Patience	5	3
Anneal factor	0.5	0.5
Batch size	32	32
Learning rate	0.1	3e-05

Collections 1 and 2 were prepared for the cross validation procedure by dividing into training, validation and testing parts in the proportion 0.8/0.1/0.1. However, the collection 2 has been reduced to 8743 items for the training, 1094 items for the validation and 1094 items for the testing, for the purpose of balancing. The adopted cross validation procedure is the default procedure built into the Flair library [4]. This procedure requires dividing the corpus into training, testing and validation sets. The testing set is used only to conduct tests and to calculate metrics on the trained and selected as the best one model. The set of validation is used to indicate the best model from all the models obtained after each training epoch. The training set is used to train, i.e. modify the parameters of the neural network. During each training epoch, 10 iterations were performed using the training set. Five models were made for each collection; for the collection 1 models based on the following architectures were prepared: roberta-large, bert-large-cased_TDE, distilgpt2, xlnet-large-cased

Table 3. Hyperparameters values of experiments for the collection 2

Hyperparameter	Hyperparameter values for document-level embeddings	Hyperparameter values for sentence-level embeddings
Max number of epochs	15	15
Patience	5	3
Anneal factor	0.5	0.5
Batch size	32	32
Learning rate	0.1	3e-05

using the sentence-level embeddings technique and bert-large-cased_DRE with the document-level embeddings technique. Models based on the following architectures were created for the collection 2: bert-large-cased_TDE, distilbert-base-uncased, distilgpt2, roberta-large with the sentence-level embeddings technique and bert-large-cased_DRE with the document-level embeddings technique.

The work related to pre-processing and training of the neural network was performed on the Google Colaboratory remote platform by applying the GPU TeslaT4, CUDA version 10.1, RAM 12.72 GB, 68.4 GB HDD, pandas version 1.1.5 and the Flair version 0.6.1.

The computational times required to train the neural network for models for the collection 1 are from a minimum of 91 s to a maximum of 437 s, and for the collection 2 from 1245 s to 9909 s. Detailed data on neural network training times are shown in Table 4 and 5. On their basis, it was noticed that more computation time is required for the architectures with more parameters, and simultaneously, a significant reduction in the computational time was obtained for the hybrid method, with the use of the document level embeddings technique.

6 Results

The obtained models were verified by analyzing the following metrics: precision, recall, f1-score, obtained in the model testing procedure. The conducted analysis proved, that in the vast majority of cases, the obtained results significantly outperformed the results presented in the original work for the subject of the KaiDMML dataset [15]. Table 6 depicts the results for the collection 1, and Tab. 7 for the collection 2; in both tables the results which outperform the results in [15] are marked in bold.

In order to confirm the effectiveness of the created models in detecting false information and the usefulness of the models in practical applications, additional practical tests were carried out. They consisted in downloading the titles of articles by applying the web scraper technique from selected websites, and their verification in terms of the content of disinformation or fake news. Sixty-seven titles were downloaded from the newyorker.com, borowitz-report webpage [6], and 141 article titles from the Deutsche Welle webpage [2]. The created model,

Table 4. Computation time needed for models training, based on collection 1; the comparison between various architectures applied for the training

Architecture	Training time [s]
xlnet-large-cased	437
roberta-large	336
distilgpt2	93
bert-large-cased_TDE	353
bert-large-cased_DRE	91

Table 5. Computation time needed for models training, based on collection 2; the comparison between various architectures applied for the training

Architecture	Training time [s]
roberta-large	8337
distilgpt2	2309
distilbert-base-uncased	3384
bert-large-cased_TDE	9909
bert-large-cased_DRE	1245

based on the pre-trained roberta-large architecture and sentence-level embedding technique classified 83.58% of the titles on the website of borowitz-report as untrue, and 4.96% of the titles on the website of Deutsche Welle also as untrue. Such a large difference in the results confirms the practical effectiveness of the created model, which clearly indicated the satirical website as a source of information classified as fake news. The obtained value of the precision metric of the test was 0.8889.

Table 6. Resulted metrics for testing of models based on the collection 1 for the label fake (the comparison between architectures, xlnet-large-cased, roberta-large, distilgpt2, bert-large-cased_TDE, bert-large-cased_DRE); the results which outperform the values in [15] are marked in bold

Architecture	Precision	Recall	f1-score
xlnet-large-cased	**0.8444**	**0.8837**	**0.8636**
roberta-large	**0.8889**	**0.9302**	**0.9091**
distilgpt2	**0.8788**	0.6744	**0.7632**
bert-large-cased_TDE	**0.9412**	0.7442	**0.8312**
bert-large-cased_DRE	**0.8780**	**0.8372**	**0.8571**

Table 7. Resulted metrics for testing of models based on the collection 2 for the label fake (the comparison between architectures, roberta-large, distilgpt2, distilbert-base-uncased, bert-large-cased_TDE, bert-large-cased_DRE); the results which outperform the values in [15] are marked in bold

Architecture	Precision	Recall	f1-score
roberta-large	**0.7823**	0.8308	**0.8058**
distilgpt2	**0.7968**	0.7519	**0.7737**
distilbert-base-uncased	**0.7914**	0.7914	**0.7914**
bert-large-cased_TDE	**0.7939**	0.7594	**0.7762**
bert-large-cased_DRE	**0.7637**	0.7350	**0.7490**

7 Conclusion

The paper presents effective methods of detecting fake news and disinformation, based on Transformer architectures. The major contribution is the creation of disinformation detection models based on the Flair library, which enables to design various architectures to train neural networks. The architectures were created through the implementation of pre-trained Transformers based architectures and the application of the sentence level or the document level embeddings. The presented experiments have proved that applying remote platforms and state-of-the-art NLP approaches can successfully detect disinformation. The results of the experiments showed that the Transformer based models outperform the models reported so far.

For the future work, experiments are planned on much larger databases, which will be created from scratch by obtaining contents from publicly available websites using the web scraper techniques. Data classification will be made on the basis of the prevailing opinions about sources, i.e., the addresses of websites.

Acknowledgement. This work is supported by SocialTruth project (http://socialtruth.eu), which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825477.

References

1. Gossip Cop. https://www.gossipcop.com/. Accessed 03 Jan 2021
2. TOP STORIES. https://www.dw.com/en/. Accessed 03 Jan 2021
3. Agarwal, V., Sultana, H.P., Malhotra, S., Sarkar, A.: Analysis of classifiers for fake news detection. Procedia Comput. Sci. **165**, 377–383 (2019). 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION, 11–12 November 2019
4. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: an easy-to-use framework for state-of-the-art NLP. In: Ammar, W., Louis, A., Mostafazadeh, N. (eds.) NAACL-HLT (Demonstrations), pp. 54–59. Association for Computational Linguistics (2019)

5. Aphiwongsophon, S., Chongstitvatana, P.: Detecting fake news with machine learning method. In: 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp. 528–531 (2018)
6. Borowitz, A.: Satire from the Borowitz Report. https://www.newyorker.com/humor/borowitz-report/. Accessed 03 Jan 2021
7. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: 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, vol. 1 (Long and Short Papers), Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics, June 2019
8. Poynter Institute. Politifact. https://www.politifact.com/. Accessed 03 Jan 2021
9. Kula, S., Choraś, M., Kozik, R.: Application of the BERT-based architecture in fake news detection. In: Herrero, Á., Cambra, C., Urda, D., Sedano, J., Quintián, H., Corchado, E. (eds.) CISIS 2019. AISC, vol. 1267, pp. 239–249. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-57805-3_23
10. Kula, S., Choraś, M., Kozik, R., Ksieniewicz, P., Woźniak, M.: Sentiment analysis for fake news detection by means of neural networks. In: Krzhizhanovskaya, V.V., et al. (eds.) ICCS 2020. LNCS, vol. 12140, pp. 653–666. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50423-6_49
11. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR, abs/1907.11692 (2019)
12. Choraś, M., et al.: Advanced machine learning techniques for fake news (online disinformation) detection: a systematic mapping study. Appl. Soft Comput. **101**, 107050 (2021)
13. Rodríguez, Á.I., Iglesias, L.L.: Fake news detection using deep learning. CoRR, abs/1910.03496 (2019)
14. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR, abs/1910.01108 (2019)
15. 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. arXiv preprint arXiv:1809.01286 (2018)
16. Vaswani, A., et al.: Attention is all you need. arxiv:1706.03762Comment, 15 pages, 5 figures (2017)
17. Volkova, S., Shaffer, K., Jang, J.Y., Hodas, N.O.: Separating facts from fiction: linguistic models to classify suspicious and trusted news posts on twitter. In: Barzilay, R., Kan, M.-Y. (eds.) ACL (2), pp. 647–653. Association for Computational Linguistics (2017)
18. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Liu, Q., Schlangen, D. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, 16–20 November 2020, pp. 38–45. Association for Computational Linguistics (2020)
19. Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., Yu, P.S.: TI-CNN: convolutional neural networks for fake news detection. CoRR, abs/1806.00749 (2018)
20. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: Xlnet: generalized autoregressive pretraining for language understanding. CoRR, abs/1906.08237 (2019)
21. Yazdi, K.M., Yazdi, A.M., Khodayi, S., Hou, J., Zhou, W., Saedy, S.: Improving fake news detection using k-means and support vector machine approaches. Int. J. Electrical Electronic Commun. Sci. 13.0(2) (2020)