GRAPH CONVOLUTIONAL NETWORKS APPLIED TO FAKENEWS: CORONA VIRUS AND 5G CONSPIRACY

Andrei PARASCHIV[1], George-Eduard ZAHARIA[2],
Dumitru-Clementin CERCEL[3], Mihai DASCALU[4]

The 2020 outbreak of coronavirus pandemic generated a wave of rumours, misinformation, and conspiracy theories; these theories and uninformed speculations gained significant traction through social media platforms. In this paper, we focus on a particular conspiracy theory, related to the unfounded connection between 5G networks and the spread of COVID-19. Several experiments with different types of text classifiers, based on Graph Convolutional Networks enhanced by BERT models, are performed on the MediaEval 2020 5G conspiracy dataset. We show that tweets supporting these theories can be detected with a Matthews Correlation Coefficient score of 0.4975 through state-of-the-art deep learning models. In addition, transfer learning from tasks related to fake news and propaganda improve the performance of our models.

Keywords: Natural Language Processing, Graph Convolutional Networks, Text Classification, Deep Learning, Conspiracy Detection, COVID-19.

1. Introduction

Shortly after the World Health Organization declared the spread of the SARS-CoV-2 virus as a pandemic [1], social media outlets observed a rise in conspiracy theories centered on the origin and the spread of the virus [2]. At times of uncertainty, people tend to consume and spread more rumours and unverified information, bringing forth misleading or outright false claims. In addition, these conspiracy theories and falsehoods gained an accelerated spreading speed in the social media landscape using modern communication tools. Shahsavari et al. [3] identify four main conspiracy theories frequently encountered in online conversations: first, 5G networks and Bill Gates have a role in creating or spreading the virus. Second, the idea that the Chinese

[1] PhD student, Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: `andrei.paraschiv74@stud.acs.pub.ro`

[2] MSc student, Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: `george.zaharia0806@stud.acs.upb.ro`

[3] Lecturer, Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: `dumitru.cercel@upb.ro`

[4] Prof., Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: `mihai.dascalu@upb.ro`

Communist Party has covered up the severity of the virus. Third, theories about the virus being a bio-weapon released by accident or on purpose. Fourth, the virus is no more dangerous than the flu, and its deadliness is exaggerated by a globalist cabal.

A pervasive false claim central to our study is that 5G towers either originate the virus, transmit it, or lower the immune response of the human body. Bruns et al. [4] show a growth of more than ten times of Facebook posts covering COVID-19 in conjunction with 5G from February until April 2020. These claims not only foster an anti-scientific attitude in the public domain, but also triggered attacks on 5G masts in the United Kingdom, Netherlands, Belgium, Italy, Cyprus, and Sweden.

Automated early detection of such conspiracy theories would prove helpful in mitigating the impact of these misinformation campaigns. As such, MediaEval 2020 [5] introduced a dataset based on COVID-19 related Twitter posts [6] labeled into three categories: *Non-conspiracy*, *5G-corona conspiracy*, and *Other conspiracy*. The organizers of the competition proposed a classification task based on the content and/or structure from the tweets.

Our solution considers only the text content of these tweets and relies on Natural Language Processing techniques to classify the messages into the three aforementioned classes. We tackle the problem by deploying two models based on Graph Convolutional Networks (GCNs) [7, 8], combined with Bidirectional Encoder Representations from Transformers (BERT)-based embeddings [9].

The remainder of the paper is structured as follows. The following section presents related work and methodologies focused on conspiracy detection. In the third section, we present the models used in our experiments, together with their implementation details. The fourth section introduces the results and an analysis of our experiments. In the fifth section, we focus on the classification errors and explore avenues of future mitigation. The last section draws the conclusions from our experiments and results.

2. **Related Work**

Previous virus outbreaks sparked conspiratorial talks in social media. For example, Wood [10] analyzed the spread of conspiracy theories on the 2015 Zika virus outbreak through Twitter. Wood [10] noticed defining features in tweets promoting these kinds of misinformation, like rhetorical questions or the ”just asking questions” trope.

Kumar et al. [11] proposed a fine-grained classification dataset containing conspiracy-related tweets, and compared several deep learning models used to classify the messages. They show that RoBERTa [12] based models perform best on their proposed task.

Several recent studies have used GCN and BERT for text classification tasks. Jeong et al. [13] employed the usage of a BERT-GCN model for context-aware citation recommendation. The role of the context encoder is taken by a

BERT model, with the purpose of extracting embeddings for the input texts. At the same time, a GCN model is applied for the citation encoder module, returning relational learning representations as the embedding vector. Also, a BERT+GCN model was used by Ye et al. [14] for short text classification. The authors added a BiLSTM [15] module on top of the node representations from the GCN layer and the word representations from the BERT encoder, which in turn provided a significant increase in prediction accuracy.

Furthermore, Lu et al. [16] augmented the standard BERT model and improved its classification performance by using a vocabulary graph that considers the lexical relations in a target language. In their approach, Lu et al. [16] constructed their vocabulary graph using the normalized point-wise mutual information (NPMI) [17] between the words in a sliding window, over the entire training corpus. An edge between two words is formed if the NPMI between the words is greater than a certain threshold. In our second model, we build upon their approach, extend and improve their performance.

Moreover, various solutions were introduced to tackle the identification of offensive texts [18, 19, 20, 21], whereas others approach propaganda analysis [22, 23] and fake news detection [24, 25]; all models yielded good performance, while considering state-of-the-art Natural Language Processing models.

3. Method

3.1. Corpus

The used dataset was crawled by Pogorelov et al. [5] using the Twitter search API and includes tweets between January 17th and May 15th 2020 filtered using COVID-19 related keywords and mentions of 5G. The annotators manually labeled a subset of these messages using three classes. The *5G-corona conspiracy* label targets tweets that state a relationship between COVID-19 and 5G networks, with a claimed causal connection between these two. The *Other conspiracy* class contains tweets that propagate conspiracy theories other than the aforementioned one (e.g., other COVID-19 related conspiracies). The third class called *Non-conspiracy* includes all tweets that were not classified in the first two classes. More precisely, these tweets can claim that 5G is not necessary safe, include discussions about the pandemic or even tweets mocking conspiracy theories, but do not incite individuals to have faith in them. The resulting dataset has 6,458 training and 3,230 testing tweets; the distribution the dataset is unbalanced, having ∼68% *Non-conspiracy* records. In our research, we had access only to the training set, which had the distribution presented in Table 1.

3.2. Neural Network Architectures

We employ two models for classifying the tweets, both based on GCNs combined with BERT-based embeddings.

Table 1

Overview of label distributions.

Label	**# Samples**	**Percentage**	**Avg. Tweet Length**
5G-corona conspiracy	1,261	19.53%	220
Other conspiracy	785	12.16%	220
Non-conspiracy	4,412	68.32%	235
Total train set	6,458	100.00%	225
5G-corona conspiracy	533	18.32%	221
Other conspiracy	347	11.93%	240
Non-conspiracy	2,029	69.75%	221
Total test set	2,909	100.00%	227

BERT and Standard GCN (BERT-GCN). First, we employ a state-of-the-art GCN architecture [8] for text classification for which the initial step consists of creating the text graph. The GCN graph contains both document (input entry) and word nodes. The weights of the edges between these types of nodes are determined using the standard algorithm [8], namely: term frequency–inverse document frequency (TF-IDF) [26] for document-word edges and point-wise mutual information (PMI) [27] for word-word edges. Moreover, the inputs were fed to a BERT model, as the graph was computed, for determining the embedding representations of the input tokens based on their different contexts. As expected, several words occurred multiple times in different contexts; thus, their embeddings were average for obtaining a single representation. The adjacency matrix representing the graph, alongside the 768-dimensional embeddings for the words were serialized for faster follow-up computations.

Making use of the previously pre-computed elements, during the training process, we fed the graph and the word representations into a two-layer GCN network, as seen in Figure 1. Considering that our task employed classification, the second GCN layer has the same size as the total number of labels and is followed by a softmax layer to predict the final class for each tweet.

BERT and Vocabulary GCN (BERT-VGCN). The second model considers a BERT-based model architecture with a Vocabulary GCN (VGCN) - see Figure 2. In contrast to the original paper [16], a larger corpus was used for the VGCN creation by considering conspiracy related corpora from Hagen et al. [28] and from Low et al. [29]. The pre-trained BERT model was enhanced with additional Masked Language Modeling pre-training using over 9.2M news articles consisting mostly of fake and hyperpartisan news [23]. Also, our graph was constructed not from individual words, as the original paper suggested, but from the wordpiece tokens as used by BERT. The word embeddings were provided by the same BERT model used for the final classification. These embeddings were then enriched with the VGCN convolution

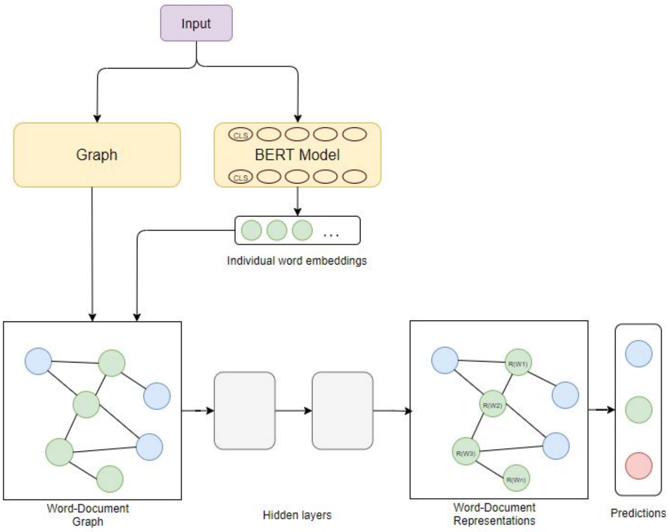

Fig. 1. BERT-GCN architecture.

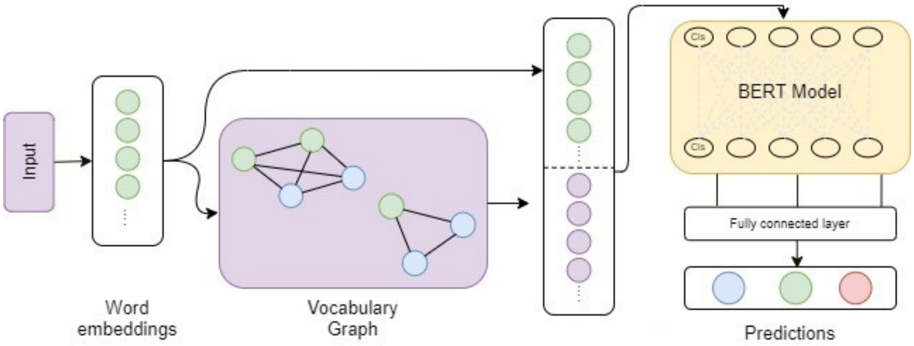

Fig. 2. BERT-VGCN architecture.

outputs by combining words from the input with the related words in the constructed graph.

Our BERT-VGCN model was built by combining text corpora from the training tweets provided by the competition organizers, a large collection of QAnon posts from 4chan and 8chan [28], and data scraped from the r/conspiracy subreddit posted in 2018 and 2019 [29]. The resulting corpus amounted to 206,346 posts and 4.8M individual tokens. The corpus was scanned sequentially, line by line, with a 1000 token window and the normalized point-wise mutual information for each token pair was computed using the Eq. 1 as follows:

$$\mathrm{NPMI}(i,j) = -\frac{1}{\log p(i,j)} \log \frac{p(i,j)}{p(i)p(j)} \tag{1}$$

where *p(i,j)* is the joint probability of a window containing tokens i and j and *p(i)* is the probability of a window containing token i.

Thus, the wordpiece vocabulary for our BERT model consisted of 30,522 graph nodes, one for each possible token, with edges between tokens that have a NPMI larger than 0.3. Also, two special tokens were used for all URLs and Twitter usernames, while a maximum of 10 consecutive username mentions per tweet were considered. The normalized symmetric adjacency matrix $\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ was created, where A is the adjacency matrix of the text graph, D denotes its degree matrix, and $D_{ii} = \sum_j A_{ij}$. Further, we use the adjacency matrix in the graph convolution layer, as shown in the Eq. 2:

$$G_{embedding} = ReLU(X\tilde{A}W_{vh})W_{hg} \tag{2}$$

where X is the BERT embedding matrix for the mini-batch of m documents; W_{vh} and W_{hg} are the hidden weight matrices for the first and second convolution layers; v is the vocabulary size; h is the hidden size, a hyperparameter; g is the graph embedding size, also as a hyperparameter. The following values were set after performing a grid search: h=128 and g=64.

The resulting graph embedding $G_{embedding}$ is then concatenated with the original BERT word embeddings, as seen in Figure 2. and passed through the 12 attention headers. Finally, the result is pooled and classified with a softmax layer into one of the three output classes.

A requirement for graph representations of text is the presence of all possible words, from both train and test datasets, as nodes prior to training; however, this is not always possible. We overcame this limitation by using the entire BERT vocabulary for nodes. Also, the probability to encounter never seen, but relevant, token co-occurrences in the test set was reduced by using the aforementioned extended corpus. Thus, the need to re-compute the adjacency matrix for new data was greatly reduced.

4. **Results**

We experimented with five different setups based on the previous models, as follows. The first two experiments used the BERT-GCN architecture with two different strategies. Initially, the BERT-base model was used to extract word embeddings that were directly adopted as representations for the GCN word nodes (i.e., *BERT-GCN-1*). In the second experiment, namely *BERT-GCN-2*, the BERT-base model was fine-tuned alongside the GCN training, and the resulting features were concatenated for final classification. Furthermore, a learning rate decay approach was considered to ensure the optimal weight update.

After that, three different training data sampling methods were employed for BERT-VCGN. All experiments used a 80%/20% train/validation, but with different sampling strategies from the train split. Random samples from the train split were considered in *BERT-VGCN-1*. In contrast to the first random extraction, minority classes were over-sampled inversely proportional to their sizes in *BERT-VGCN-2*, while the majority class (*Non-conspiracy*) was under-sampled during training for *BERT-VGCN-3*. We evaluate the performance of all our models using the multi-class generalization of the Matthews Correlation Coefficient (MCC) [30].

Table 2

Evaluation of our models on the validation set.

Model	Micro-F1	Macro-F1	Weighted-F1	MCC
GCN	72.87%	53.31%	71.53%	0.3746
BERT	76.36%	48.50%	71.56%	0.4392
BERT-GCN-1	75.68%	46.09%	69.84%	0.3907
BERT-GCN-2	76.36%	51.28%	72.80%	0.4452
BERT-VGCN-1	**76.57%**	**58.31%**	**75.35%**	**0.4670**
BERT-VGCN-2	75.98%	57.79%	74.74%	0.4534
BERT-VGCN-3	69.35%	52.25%	70.05%	0.3585

Table 3

Evaluation of our models on the test set.

Model	Micro-F1	Macro-F1	Weighted-F1	MCC
GCN	71.72%	42.56%	65.70%	0.2920
BERT	74.83%	52.16%	71.68%	0.4389
BERT-GCN-1	70.05%	51.74%	69.27%	0.3312
BERT-GCN-2	76.00%	61.49%	76.04%	0.4795
BERT-VGCN-1	**78.30%**	**62.42%**	**77.11%**	0.4973
BERT-VGCN-2	77.31%	61.59%	76.46%	**0.4975**
BERT-VGCN-3	71.08%	55.80%	71.70%	0.3910

5. **Discussion**

The BERT-VGCN model obtains the best results on both datasets (i.e. validation and test), considering the complex structure and therefore its ability to properly grasp features specific to the three classes. As seen in Table 3, the model performs best using its second setup, very closely followed by the first one. However, BERT-VGCN-1 has better results on the validation dataset, surpassing the second variant of the system by a margin of 0.0116 MCC score, as seen in Table 2. The BERT-GCN model stands third in terms of performance, with its second setup achieving a 0.4452 MCC score on the validation dataset and a 0.4795 MCC score on the test set.

Furthermore, the MCC score is generally lower on the validation dataset when compared to the test set for almost all models. This discrepancy can be attributed to the validation dataset, which was generated from the provided training set and may not properly encapsulate all specific aspects present in the test set. Moreover, the latter can contain completely new entries for some classes, in terms of structure or general ideas. As such, similar entries may not be available in the initial training set.

Most misclassified entries by the BERT-GCN model are part of either *5G-corona conspiracy* or *Other conspiracy* classes. Misclassifications occur because of the structural similarities between entries from the two classes, considering that they both are subclasses of *conspiracy*. Moreover, the reduced number of training entries depicting *5G-corona conspiracies* or *Other conspiracies* leads to the inability of the models to properly grasp features that support a proper identification of classes, when presented with new test examples.

The confusion matrix in Figure 3 emphasizes that the predominant presence of *Non-conspiracy* entries is detrimental for the considered architectures, inasmuch as, during the training process, the system has a strong bias towards considering conspiracy entries as part of the *Non-conspiracy* class. The features specific to the former are important for the model's performance, and the lack of examples can only be a negative aspect. A common practice employed for solving the previously mentioned issue is dataset augmentation. However, considering that the 5G-coronavirus topic is relatively new, there are no proper augmentation datasets that can be used to enhance feature extraction and prevent overfitting.

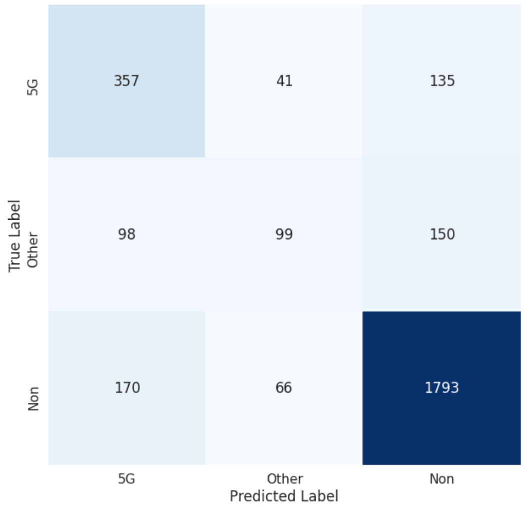

Fig. 3. Confusion matrix for BERT-VGCN-2.
(*5G*-corona conspiracy; *Other* conspiracies; *Non*-conspiracy)

At the same time, the dataset is prone to the annotators' subjectivity, especially in situations when the class cannot be properly identified (e.g., the tweet considered a conspiracy is in fact a *Non-conspiracy*, or a mixture of *5G-corona conspiracy* and *Other conspiracy*). These situations can lead to prediction errors, as seen in Table 4.

Table 4

Misclassified entries by the BERT-GCN-2 model.

Entry	**True label**	**Predicted label**
#Qannon #5G #COVID_19 There is almost 8 billion of us and we don't have the power against the few? The greatest illusion is the illusion of powerlessness. And if we don't do something the whole world is gonna look like this.	Non-conspiracy	5G-corona conspiracy
I never said that David believes that "5G is causing COVID-19". I said that he thinks the virus doesn't exist and that 5G is causing corona-like symptoms your fruitcake! https://t.co/CIgLXctVzI	5G-corona conspiracy	Non-conspiracy
And the " lack of Adrenochrome supply"...isn't it interesting that Wuhan has a Pharma co that actually manufactures #Adrenochrome???	Other conspiracy	Non-conspiracy

The main token that creates confusion between the first and the second class is represented by *5G*, a concept recurrent in both conspiracy classes that thus blurs the border between them. Moreover, the same token is predominant in the third class, which can confuse both automated models and annotators.

6. Conclusions and Future Work

Assessing the veracity of a tweet without any background or context is a difficult task, even for humans. With regards to a pandemic or a global crisis, access to accurate information and the limitation of disinformation is critical. One of the most shared conspiracies regarding COVID-19 was related to the possible correlation between 5G networks and the spread of the SARS-COV-2 virus. Considering the extensive number of such conspiracies that occurred during the past months, we consider that, when deployed in a production environment, our methods could help counter the misinformation assault and reduce the spread of related conspiracy theories. In this paper, various automated models considering BERT augmented with GCNs were employed and obtained promising results. Through GCNs, we obtained document-level representations that paired with BERT encoder embeddings created a feature vector that can be properly used for classification.

Conspiracy and fake news domains have a high affinity, seeing that both contain statements or ideas that are not based on real facts and are therefore unfounded. The 5G-coronavirus conspiracies are also spread through fake news, while people express their ideologies through deceitful posts and propaganda. Considering their close relationship, we extended our training dataset with entries belonging to fake news, thus empowering transfer learning [31] between the two domains and improving the performance of part of our models.

In terms of future work, we intend to further exploit the power of Transformer-based models [32] and GCN variants by using their features at different levels in our architectures. Moreover, the conspiracy entries can enclose features of different specificities, situated at various levels, and therefore capsule networks [33] can become helpful for tackling the conspiracy identification issue.

Acknowledgement

The work has been funded by the Operational Programme Human Capital of the Ministry of European Funds through the Financial Agreement 51675/09.07.2019, SMIS code 125125.

REFERENCES

[1] *B. News*, "Coronavirus confirmed as pandemic by world health organization." `https://www.bbc.com/news/world-51839944/`, 2020. [Online; accessed November 25, 2020].

[2] *M. S. Islam, T. Sarkar, S. H. Khan, A.-H. M. Kamal, S. M. Hasan, A. Kabir, D. Yeasmin, M. A. Islam, K. I. A. Chowdhury, K. S. Anwar,* et al., "Covid-19–related infodemic and its impact on public health: A global social media analysis," *The American Journal of Tropical Medicine and Hygiene*, vol. 103, no. 4, 1621–1629, 2020.

[3] *S. Shahsavari, P. Holur, T. R. Tangherlini, and V. Roychowdhury*, "Conspiracy in the time of corona: Automatic detection of covid-19 conspiracy theories in social media and the news," *arXiv preprint arXiv:2004.13783*, 2020.

[4] *A. Bruns, S. Harrington, and E. Hurcombe*, "'covid19? corona? 5g? or both?': the dynamics of covid-19/5g conspiracy theories on facebook," *Media International Australia*, vol. 177, no. 1, 12–29, 2020.

[5] *K. Pogorelov, D. T. Schroeder, L. Burchard, J. Moe, S. Brenner, P. Filkukova, and J. Langguth*, "Fakenews: Corona virus and 5g conspiracy task at mediaeval 2020," in *MediaEval 2020 Workshop* (2020).

[6] *D. T. Schroeder, K. Pogorelov, and J. Langguth*, "Fact: a framework for analysis and capture of twitter graphs," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (2019), 134–141, IEEE.

[7] *T. N. Kipf and M. Welling*, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[8] *L. Yao, C. Mao, and Y. Luo*, "Graph convolutional networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (2019), 7370–7377.

[9] *J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova*, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] *M. J. Wood*, "Propagating and debunking conspiracy theories on twitter during the 2015–2016 zika virus outbreak," *Cyberpsychology, behavior, and social networking*, vol. 21, no. 8, 485–490, 2018.

[11] *S. Kumar*, "A fine-grained analysis of misinformation in covid-19 tweets," in *Virtual Conference on Social-Cybersecurity in Times of Crisis and Change, Center for Informed Democracy & Social-Cybersecurity (IDeaS), Carnegie Mellon University* (2020). November 19.

[12] *Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[13] *C. Jeong, S. Jang, E. Park, and S. Choi*, "A context-aware citation recommendation model with bert and graph convolutional networks," *Scientometrics*, vol. 124, no. 3, 1907–1922, 2020.

[14] *Z. Ye, G. Jiang, Y. Liu, Z. Li, and J. Yuan*, "Document and word representations generated by graph convolutional network and bert for short text classification," in *Proceedings of the 24th European Conference on Artificial Intelligence - ECAI 2020* (2020), ECAI.

[15] *S. Hochreiter and J. Schmidhuber*, "Long short-term memory," *Neural computation*, vol. 9, no. 8, 1735–1780, 1997.

[16] *Z. Lu, P. Du, and J.-Y. Nie*, "Vgcn-bert: Augmenting bert with graph embedding for text classification," in *European Conference on Information Retrieval* (2020), 369–382, Springer.

[17] *G. Bouma*, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of GSCL*, 31–40, 2009.

[18] *A. Paraschiv and D.-C. Cercel*, "Upb at germeval-2019 task 2: Bert-based offensive language classification of german tweets.," in *KONVENS* (2019).

[19] *M.-A. Tanase, D.-C. Cercel, and C. Chiru*, "Upb at semeval-2020 task 12: Multilingual offensive language detection on social media by fine-tuning a variety of bert-based models," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (2020), 2222–2231.

[20] *G.-A. Vlad, G.-E. Zaharia, D.-C. Cercel, C. Chiru, and S. Trausan-Matu*, "Upb at semeval-2020 task 8: Joint textual and visual modeling in a multi-task learning architecture for memotion analysis," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (2020), 1208–1214.

[21] *G.-A. Vlad, G.-E. Zaharia, D.-C. Cercel, and M. Dascalu*, "Upb@ dankmemes: Italian memes analysis-employing visual models and graph convolutional networks for meme identification and hate speech detection," in *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online. CEUR. org* (2020).

[22] *G.-A. Vlad, M.-A. Tanase, C. Onose, and D.-C. Cercel*, "Sentence-level propaganda detection in news articles with transfer learning and bert-bilstm-capsule model," in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda* (2019), 148–154.

[23] *A. Paraschiv, D.-C. Cercel, and M. Dascalu*, "Upb at semeval-2020 task 11: Propaganda detection with domain-specific trained bert," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (2020), 1853–1857.

[24] *C. Onose, C.-M. Nedelcu, D.-C. Cercel, and S. Trausan-Matu*, "A hierarchical attention network for bots and gender profiling.," in *CLEF (Working Notes)* (2019).

[25] *C. Busioc, S. Ruseti, and M. Dascalu*, "A literature review of nlp approaches to fake news detection and their applicability to romanian-language news analysis," *Revista Transilvania*, no. 10, 2020.

[26] *T. Van de Cruys*, "Two multivariate generalizations of pointwise mutual information," in *Proceedings of the Workshop on Distributional Semantics and Compositionality* (2011), 16–20.

[27] *C. Manning and H. Schutze*, *Foundations of statistical natural language processing.* MIT press, 1999.

[28] *S. Hagen, S. Peeters, E. Jokubauskaitė, and D. de Zeeuw*, "Cross-platform mentions of the QAnon conspiracy theory," *Zenodo https://doi.org/10.5281/zenodo.3758479*, Apr. 2020.

[29] *D. M. Low, L. Rumker, J. Torous, G. Cecchi, S. S. Ghosh, and T. Talkar*, "Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study," *Journal of medical Internet research*, vol. 22, no. 10, e22635, 2020.

[30] *J. Gorodkin*, "Comparing two k-category assignments by a k-category correlation coefficient," *Computational biology and chemistry*, vol. 28, no. 5-6, 367–374, 2004.

[31] *S. J. Pan and Q. Yang*, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, 1345–1359, 2009.

[32] *A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 5998–6008, 2017.

[33] *S. Sabour, N. Frosst, and G. E. Hinton*, "Dynamic routing between capsules," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017), 3859–3869.