biCourage: ngram and syntax GCNs for Hate Speech detection

Rodrigo Wilkens[1], Dimitri Ognibene[1,2]

[1]*University of Milano-Bicocca, Italy*

[2]*University of Essex, UK*

Abstract

Hate Speech identification is a challenging task given the world knowledge required. Moreover, it is even more complex in the social media context due to language and media specificities. Despite these challenges, advances in this task may help improving collective well-being on social media. In this context, the *biCourage team* participated in the English version of Task 1 of HASOC 2021, a shared task for "Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages". Our participation in this campaign aimed to examine the suitability of Graph Convolutional Neural Networks (GCN), due to their capability to integrate flexible contextual priors, as a computationally effective solution compared to more computationally expensive and relatively data-hungry methods, such as fine-tuning. Specifically, we explored and combined two text-to-graph strategies based on different language modelling objectives, comparing them with fine-tuned Bert. We submitted the results of several deep learning architectures, comprised of different arrangements of GCNs and transformer architectures. Our team achieved the best results in both subtasks using the GCNs based architectures combining two text-to-graph strategies ranked in 21st and 20th positions in Subtasks 1A and 1B. Assessing the models' prediction, we identify complementary capabilities in the text-to-graph strategies that further research on their combination can explore. Moreover, the proposed GCN model is 3.85 times faster than fine-tuned Bert in training speed and still outperforms it by 2.3% and 5.41% on the F1 score of Subtasks 1A and 1B, respectively.

Keywords

hate speech, graph convolutional network, text-to-graph, biCourage, Bert fine-tuning

1. Introduction

The freedom to publish [1] on social media marked a new era that goes beyond just-in-time connectivity with a network of friends and like-minded people. Moreover, social media has also fostered negative social phenomena [2]. Consequently, threats, including Hate Speech (HS), have become subject of interest in different research contexts.

The automatic classification of text generated in social media, and Twitter, in particular [3], poses several significant challenges. Indeed, their informality, noisiness, and limited size lead first to a lack of features for classification, negatively affecting results, second to a lack of context and thus ambiguity, and, finally to a mismatch with models based on standard language corpora. Moreover, some works discriminate the type and target of the hate (e.g., sexism, offensive and

profane) and may have to deal with noisy labels in the ground truth, as the classification of this type of content is often subjective. However, fostering this line of research may help improving collective well-being on social media by enabling AI supported governance and moderation strategies [2].

Concerning HS detection on social media, HASOC[1] [4, 5] provides a forum and a data challenge for multilingual research on the identification of problematic content. This shared task campaign in 2021 offered two tasks targeting English and Hindi. In this work, we report the approach proposed by the *biCourage team* in the English version of Task 1. This task is divided into two subtasks, both using the same Twitter corpus but with different annotation granularity. Subtask-1A targets a binary classification for identifying hate and non hate posts. In a more fine-grained perspective, Subtask-1B also proposes the distinction between hate, profane and offensive posts.

In this paper, we report the participation of the *biCourage team* at the HASOC shared task. Our goal in this work is to explore the suitability of Graph Convolutional Neural Networks (GCN) as a computationally and data efficient solution. Specifically, we explore and combine two text-to-graph strategies with different modelling objectives, comparing them with fine-tuned Bert, which is our baseline.[2] In specific, this paper is organised as follows. We start presenting initiatives for HS classification in the machine learning and language encoding perspectives, then, in Section 3, describing the proposed GCN model, the word encoding and the text-to-graph strategies. A discussion of the performance of the different models is presented in Section 4. Finally, in Section 5, we summarize our finds.

2. Related Work

A critical component in a Hate Speech classifier is the language encoding. It may employ methods that are context-sensitive (e.g. BERT [6, 7, 8]) or context-independent, which may be based on language models trained on external corpus (e.g. word embeddings [9, 10] and doc2vec [11]) or they can be based on the studied corpus (e.g. TF-IDF [6, 8, 12]). In addition to language encoding, various machine learning approaches have been explored in HS classification literature. For example, Rodríguez-Sánchez et al. [6], Liu et al. [11], Canós [12], Wang and Manning [13] resorted to Support Vector Machine (SVM), Wang and Manning [13] used Naïve Bayes, Liu et al. [11] employed Random Forests and Gradient Boosted Trees, Rodríguez-Sánchez et al. [6] used Bi-LSTM, and Rodríguez-Sánchez et al. [6], Lavergne et al. [7] fine-tuned BERT models. In addition, some works, such as Liu et al. [11] employing a soft vote approach, Shushkevich and Cardiff [14] using a blended model [13], Hoffmann and Kruschwitz [8] exploring ensemble of three SVMs, combine classification models. These models, except for LSTM models, process the input as a set of features without sequential information. Word independence is usually a common assumption in machine learning architectures. However, this assumption is not held for Graph Neural Networks because nodes (i.e. words) are associated [15]. Moreover, GCNs allow explicitly specify the associations of the words.

GCN, convolutional networks that operate on graphs, can explicitly model the relationships

---

[1]https://hasocfire.github.io/hasoc/2021
[2]The models are available at https://github.com/rswilkens/biCourage.

between words by representing words as nodes and their relations as edges in the graph. Aiming to study the possible advantages of this representation, we explore the GCN as a solution for identifying Hate Speech on social networks. GCN mainly differs from other neural networks in the forward step that connects nodes by considering an adjacency matrix. In other words, while a forward step in traditional neural networks is defined as $H_{i+1} = \sigma(W_i H_i + b_i)$ in GCN it is $H_{i+1} = \sigma(W_i H_i A + b_i)$, where W_i are the weights at layer i, H_i is the feature representation at layer i, b_i is the bias at layer i, H_{i+1} is the feature representation at layer $i+1$, and A is the adjacency matrix. In a broad sense, a GCN can be seen as sequences map-reduce operations, corresponding to transformation and several aggregation operations on graphs [16]. The aggregation aims to combine multiple messages between a node and its context and reduce them into one element. The graph pooling aims to aggregate elements in a graph, reducing them into high-order graph-level representations.

Inspired by Kipf and Welling [17], Yao et al. [18] proposed TextGCN, a GCN for text document classification. TextGCN model produces a single graph where words and documents are nodes. In this graph, all documents are connected, and edges also indicate words in the same document. Yao et al. [18] also compared the performance of TextGCN and 13 other widely-used models (e.g., fastText, LSTM, CNN, and doc2vec) in different corpora, showing impressive results. However, this model is based on transductive learning. In other words, it needs all documents available during the training phase and cannot make predictions for new documents, which poses an issue for applying it in the social media context, given the speed at which new data is created. In a different perspective, Wilkens and Ognibene [19] explored graph classification using GCN models for sexism identification on social media. This work explored the MeanPool (a simple model that pools the graph after aggregations using the mean), SAGpool$_h$ [20] (a GCN that applies a self-attention mechanism to select nodes to drop) and set2set [16] (the inclusion of an LSTM in the aggregation step). The results obtained in this work point out that for the generalisation of the GCN models due to they showed similar results in social media different of the one used in the training [19].

3. Proposed model

This work explores GCN models for HS classification on social media, extending the MeanPool model used by Wilkens and Ognibene [19]. Here, we add features normalisation steps in each GCN layer, and the GCN layers are replaced by GraphSAGE layer [21], which is closely related to the GCN layer [17]. GraphSAGE aggregates information from node local neighbours, and, as this process iterates, nodes incrementally gain information from further reaches of the graph [21]. In addition to the MeanPool model that used only a mean pooling approach to compress the matrix representation into a vector that summarise the post, we propose a richer representation concatenating min, max, mean and sum pooling operations.

In order the associate the nodes, i.e. to define the graph structure, we explore two different text-to-graph strategies (ngram and parser), similarly to Wilkens and Ognibene [19]. The parser strategy, inspired by syntacticGCN [22, 23], links nodes using the dependency attachment based on parsing information, while the ngram strategy associates all words in a context window of three words. In this work, we set the edge weight as the normalised distance of words

in the document, contrarily to Wilkens and Ognibene [19] who propose the word similarity. Moreover, we use only word embeddings from a *roberta-base* model[3] trained on 58M tweets [24] as node features and truncate all sentences with more than 300 words long due to computational limitations. We use this model targeting a text encoding trained in documents close to those used in HASOC.

Parser and ngram text-to-graph strategies have different modelling objectives. The first one aims to connect words according to their syntactic function in the sentence, while the second strategy prioritises local context. They have also different robustness to the noise present on social media text. Therefore, it could be desirable to combine the two strategies. To carry out this combination, there are several different possible approaches. We propose a siamese network keeping the two networks independent of each other and concatenating the output of the pooling approaches. In this way, the convolutional layer can adjust the weights for each text-to-graph strategy, and the classification layers (i.e. set of dense layers) learn to separate the classes upon the combination of eight poolings (2 networks by 4 pooling operations) by 200 features characterising high level nodes. The joined architecture is presented in Figure 1, where the dark red and dark green dashed-boxes respectively indicate the ngramGCN and parserGCN, each one using a copy of the same node features and different adjacency matrices. Moreover, the blue dashed-box indicates the biCourage that uses the same components of the other two GCN except for their classification layers.

Finally, we propose Bert fine-tuning as baseline. We start this process by fine-tuning the English Bert model [25] on the binary classification version of the English corpora used in the shared task's previous editions [26, 27]. This new model is then fine-tuned using the 2021 version of the corpus. We merely continue the fine-tuning process for Subtask-1A, and, for Subtask-1B, we replace the binary classification layer with one capable of classifying four classes.

Despite the model, we start our pipeline by cleaning and tokenizing the text before training the models. The cleaning step tokenizes the text, standardizes symbols, and replaces URL and emojis by the domain and the emoji textual representation based on the post's language. The text is then annotated with dependency attachment [28].

Concerning the data preprocessing, we identified 60 posts with repeated or near-repeated content. Based on a quick analysis, we concluded that different users posted those, and they most likely come from defamation/hate campaigns. Given the substantial similarity between these posts, we randomly keep one and discard the others because they impact our internal comparison of the models. Then, we use a 90/10 split to obtain training and validation sets.

4. Results

The official evaluation system used in HASOC shared task ranks only the best run of each team, then compares only the best models. Consequently, our biCourage[4] model respectively ranked in the 21st (out of 56 teams) in Subtask-1A and 20th positions (out of 37 teams) in Subtask-1B. Tables 1 and 2 show the results of all our models.

---

[3]https://huggingface.co/cardiffnlp/twitter-roberta-base
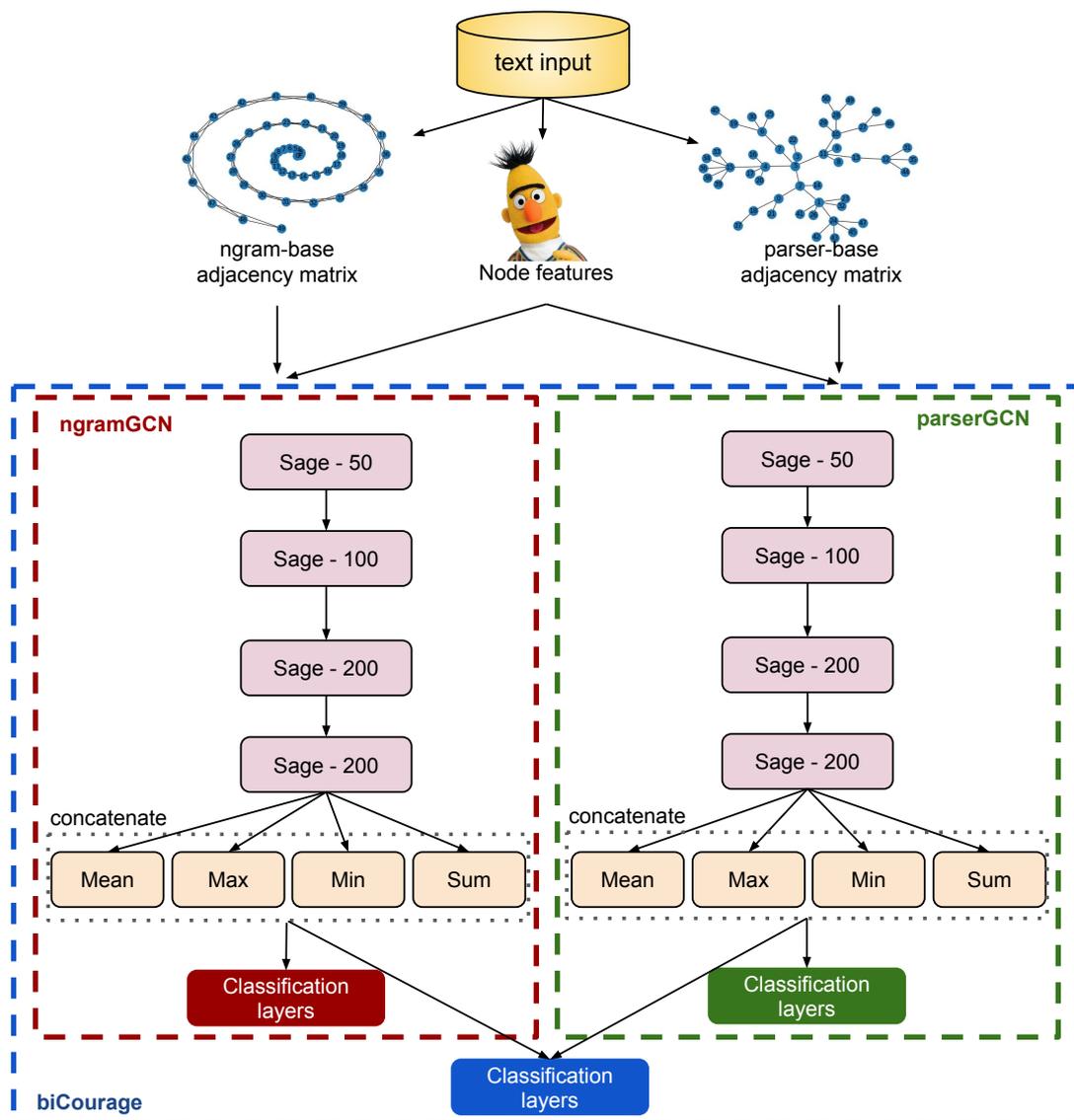[4]In the official leaderboard, biCourage is shown as 2GCNs and biGCN.

Figure 1: ngramGCN, parserGCN and biCourage architectures

In Subtask-1A, its precision and accuracy are considerably better than the other models, but the recall is below the line. On the other hand, in Subtask-1B, its precision is close to the GCNs, but its recall is considerable low, even Bert achieving the best accuracy.

Focusing on the single graph input GCN models, we observe that the ngram-based GCN consistently outperforms the syntactic one. Moreover, the performance difference between this two GCNs is bigger in the Subtask-1A. Considering that both models share the same information about the words, this suggests that the local context plays a more meaningful role in the task, at

Table 1
Evaluation of the proposed models at English Subtask-1A

Model	Macro F1	Macro Precision	Macro Recall	Accuracy
Bert	0.7666	**0.8290**	0.7518	**80.172%**
ngramGCN	0.7737	0.7901	0.7656	79.547%
parserGCN	0.7548	0.7648	0.7491	77.596%
biCourage	**0.7900**	0.7866	**0.7957**	79.938%

Table 2
Evaluation of the proposed models at English Subtask-1B

Model	Macro F1	Macro Precision	Macro Recall	Accuracy
Bert	0.5425	0.5954	0.5508	**64.715%**
ngramGCN	0.5799	0.5920	0.5885	62.920%
parserGCN	0.5639	0.5896	0.5828	61.593%
biCourage	**0.5966**	**0.6086**	**0.6110**	63.232%

least for English. Analysing the biCourage models, we note that the combination of parserGCN and ngramGCN models (i.e. biCourage) consistently improves precision in both subtasks. This indicates that ngramGCN and parserGCN models are apparently modelling different clues of Hate Speech classification.

Aiming to measure the differences between the GCN models, we examine the Cohen kappa agreement in the test set. In this way, we identified that the three models present a good agreement. The parserGCN and ngramGCN models present an agreement of 0.6296 in Subtask-1A and 0.6738 in Subtask-1B. The biCourage presents an agreement of 0.7176 with ngramGCN (Subtask-1A) and 0.6979 (Subtask-1B), but on the other hand, the biCourage agrees with the parserGCN 0.6213 for Subtask-1A and 0.7025 in Subtask-1B. These results point out that ngram- and parser-based text-to-graph approaches provide different information to the model and it can be successfully mixed throw the biCourage model. Moreover, the agreement differences in Subtask 1A and 1B between biCourage, and ngramGCN and parserGCN imply that the biCourage model can be capable of prioritising part of the network, which is a desirable feature for siamese networks. Furthermore, our results propose that syntactic information might be more relevant for recognising the subtype of HS (i.e. hate speech, offensive and profane content) than for distinguishing between hate and non-hate posts.

The results obtained from GCN, particularly the biCourage, are remarkable, especially in Subtask-1B. First, the biCourage architecture is computationally cheaper to train; i.e. biCourage trains 3.85 times faster (CPU time spent according to the OS) than the Bert fine-tuned using the same data and similar machines. Second, in terms of the F1 score it outperforms the fine-tuned Bert by 2.3% and 5.41% for Subtasks 1A and 1B respectively. However, at this point, we cannot identify the source of this performance. In other words, we are comparing models based on different language encoding approach. Therefore, we prepare a final experiment. In this, we train the biCourage model using the Bert fine-tuned on the 2019 and 2020 versions of the HASOC shared task. This analysis allows us directly compare the results from the biCourage model

Table 3
Study of the impact of language encoding in the biCourage model at English Subtask-1B

Model	Macro F1	Macro Precision	Macro Recall	Accuracy
Bert	0.5425	0.5954	0.5508	64.715%
biCourage (RoBERTa)	0.5966	0.6086	**0.6110**	63.232%
biCourage (Bert)	**0.5998**	**0.6169**	0.6001	**66.198%**

with the Bert fine-tuning approach. As shown in Table 3, the biCourage trained using Bert embeddings achieves results similar to the biCourage submitted (biCourage RoBERTa in Table 3) and better than Bert's fine-tuning. This point out that the difference between the fine-tuned Bert and the biCourage in Table 2 comes mainly from the GCN architecture. Moreover, these results reveal that the proposed model can take advantage of more suitable language encoding models.

5. Conclusion

Threats detection on social media is a complex task due to language and media specificities. Their detection may help improving Collective Well-Being on the network by enabling targeted artificial intelligence social media governance strategies [2]. Aiming for it, we (*biCourage team*) participated in the English version of Task 1 at HASOC 2021, a shared task for "Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages."

Inspired by Kipf and Welling [17], Yao et al. [18] and building on our previous work Wilkens and Ognibene [19], our approach uses Graph Convolutional Networks to associate words in a post, exploring different word association approaches (ngram and parser). We applied the Bert model fine-tuned on external HS corpora and then fine-tuned on the shared task corpus as a baseline for this approach. Among our models, our biCourage model achieved the best result in both subtasks. In Subtask-1A, the biCourage is ranked in the 21st position and 20th in Subtask-1B. Looking close at the results, we identified that the biCourage model might be capable of prioritising part of the network depending on the task, which is beneficial for networks that combine different inputs. Aiming to check the performance of biCourage in Subtask-1B, we train a new biCourage that may be directly compared to the fine-tuned Bert. This analysis indicated that the performance improvement, compared to Bert fine-tuned, may be attributed to the GCN model, and the language encoding approaches had a small impact.

The main contribution of this work goes beyond the rank in the shared task. We propose a new model capable of outperforming Bert's fine-tuning process and is 3.85 times faster to train. However, our results are limited to the shared task dataset. Thus the generalisation to other languages and tasks is still open.

Acknowledgments

References

[1] R. Baeza-Yates, B. Ribeiro-Neto (Eds.), Modern Information Retrieval, 2nd ed., Addison-Wesley, 2010.

[2] D. Ognibene, D. Taibi, U. Kruschwitz, R. S. Wilkens, D. Hernandez-Leo, E. Theophilou, L. Scifo, R. A. Lobo, F. Lomonaco, S. Eimler, et al., Challenging social media threats using collective well-being aware recommendation algorithms and an educational virtual companion, arXiv preprint arXiv:2102.04211 (2021).

[3] M. Michelson, S. A. Macskassy, Discovering users' topics of interest on twitter: a first look, in: Proceedings of the fourth workshop on Analytics for noisy unstructured text data, 2010, pp. 73–80.

[4] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, A. K. Jaiswal, Overview of the HASOC subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021. URL: http://ceur-ws.org/.

[5] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, M. Zampieri, Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages and Conversational Hate Speech, in: FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, 13th-17th December 2021, ACM, 2021.

[6] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576.

[7] E. Lavergne, R. Saini, G. Kovács, K. Murphy, Thenorth@ haspeede 2: Bert-based language model fine-tuning for italian hate speech detection, in: 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2020, volume 2765, CEUR-WS, 2020.

[8] J. Hoffmann, U. Kruschwitz, Ur nlp@ haspeede 2 at evalita 2020: Towards robust hate speech detection with contextual embeddings, in: 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2020, 2020.

[9] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

[10] G. Gambino, R. Pirrone, Chilab@ haspeede 2: Enhancing hate speech detection with part-of-speech tagging, in: 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2020, 2020.

[11] H. Liu, F. Chiroma, M. Cocea, Identification and classification of misogynous tweets using multi-classifier fusion, in: Evaluation of Human Language Technologies for Iberian

Languages: IberEval 2018 (IberEval@SEPLN), CEUR Workshop Proceedings, 2018, pp. 268–273.

[12] J. S. Canós, Misogyny identification through svm at ibereval 2018., in: Evaluation of Human Language Technologies for Iberian Languages: IberEval 2018 (IberEval@SEPLN), 2018, pp. 229–233.

[13] S. I. Wang, C. D. Manning, Baselines and bigrams: Simple, good sentiment and topic classification, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2012, pp. 90–94.

[14] E. Shushkevich, J. Cardiff, Classifying misogynistic tweets using a blended model: The ami shared task in ibereval 2018., in: Evaluation of Human Language Technologies for Iberian Languages: IberEval 2018 (IberEval@SEPLN), 2018, pp. 255–259.

[15] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S. Y. Philip, A comprehensive survey on graph neural networks, IEEE transactions on neural networks and learning systems (2020).

[16] J. Hu, S. Qian, Q. Fang, Y. Wang, Q. Zhao, H. Zhang, C. Xu, Efficient graph deep learning in tensorflow with tf_geometric, arXiv preprint arXiv:2101.11552 (2021).

[17] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907 (2016).

[18] L. Yao, C. Mao, Y. Luo, Graph convolutional networks for text classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 7370–7377.

[19] R. Wilkens, D. Ognibene, Mb-courage @ exist: Gcn classification for sexism identification in social networks, EXIST: sEXism Identification in Social neTworks – First Shared Task at IberLEF 2021 (2021).

[20] J. Lee, I. Lee, J. Kang, Self-attention graph pooling, in: International Conference on Machine Learning, PMLR, 2019, pp. 3734–3743.

[21] W. L. Hamilton, R. Ying, J. Leskovec, Inductive representation learning on large graphs, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 1025–1035.

[22] D. Marcheggiani, I. Titov, Encoding sentences with graph convolutional networks for semantic role labeling, arXiv preprint arXiv:1703.04826 (2017).

[23] J. Bastings, I. Titov, W. Aziz, D. Marcheggiani, K. Sima'an, Graph convolutional encoders for syntax-aware neural machine translation, arXiv preprint arXiv:1704.04675 (2017).

[24] F. Barbieri, J. Camacho-Collados, L. Neves, L. Espinosa-Anke, Tweeteval: Unified benchmark and comparative evaluation for tweet classification, arXiv preprint arXiv:2010.12421 (2020).

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

[26] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in indo-european languages, volume abs/2108.05927, 2021. URL: https://arxiv.org/abs/2108.05927. arXiv:2108.05927.

[27] T. Mandl, S. Modha, A. K. M, B. R. Chakravarthi, Overview of the HASOC track at FIRE 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: P. Majumder, M. Mitra, S. Gangopadhyay, P. Mehta (Eds.), FIRE 2020: Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, ACM, 2020, pp.

29–32. URL: https://doi.org/10.1145/3441501.3441517. doi:10.1145/3441501.3441517.

[28] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A python natural language processing toolkit for many human languages, arXiv preprint arXiv:2003.07082 (2020).