Does BERT Pay Attention to Cyberbullying?

Fatma Elsafoury
University of the West of Scotland
School of Computing, Engineering and Physical Sciences
Paisley, United Kingdom
fatma.elsafoury@uws.ac.uk

Stamos Katsigiannis
Durham University
Department of Computer Science
Durham, United Kingdom
stamos.katsigiannis@durham.ac.uk

Steven R. Wilson
University of Edinburgh
School of Informatics
Edinburgh, United Kingdom
steven.wilson@ed.ac.uk

Naeem Ramzan
University of the West of Scotland
School of Computing, Engineering and Physical Sciences
Paisley, United Kingdom
naeem.ramzan@uws.ac.uk

ABSTRACT

Social media have brought threats like cyberbullying, which can lead to stress, anxiety, depression and in some severe cases, suicide attempts. Detecting cyberbullying can help to warn/ block bullies and provide support to victims. However, very few studies have used self-attention-based language models like BERT for cyberbullying detection and they typically only report BERT's performance without examining in depth the reasons for its performance. In this work, we examine the use of BERT for cyberbullying detection on various datasets and attempt to explain its performance by analysing its attention weights and gradient-based feature importance scores for textual and linguistic features. Our results show that attention weights do not correlate with feature importance scores and thus do not explain the model's performance. Additionally, they suggest that BERT relies on syntactical biases in the datasets to assign feature importance scores to class-related words rather than cyberbullying-related linguistic features.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Supervised learning by classification**; • **Information systems** → **Web and social media search**.

KEYWORDS

Cyberbullying, Text classification, BERT, NLP

ACM Reference Format:
Fatma Elsafoury, Stamos Katsigiannis, Steven R. Wilson, and Naeem Ramzan. 2021. Does BERT Pay Attention to Cyberbullying?. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3463029

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463029

1 INTRODUCTION

The Pew Research Centre reported in 2017 that 40% of social media users have experienced some form of cyberbullying [1, 6, 12, 15]. Cyberbullying experiences can have serious consequences for the victims, including depression, anxiety, low self-esteem and self-harm [31]. The goal of reducing these negative outcomes highlights the critical importance of research on tools for detecting, understanding and preventing cyberbullying. Cyberbullying is defined as one form or another of spreading insults using mobile or internet technology [5, 11, 20]. Over the last decade, there have been attempts to use conventional machine learning models [8, 11, 25] and deep learning models [3, 21, 26, 39] to detect cyberbullying from social media. Recent studies have used attention-based language models, like BERT, in the detection of cyberbullying [22–24, 41]. However, those studies focused mainly on enhancing the cyberbullying detection performance using BERT, without providing any analysis or insight into the model's inner-workings.

BERT [10] is a deep neural network model with an architecture based on stacked Transformer encoders [36], which each consists of multiple layers, including a multi-head self-attention mechanism. Recent studies have applied BERT on cyberbullying detection. Paul and Saha [23] used a BERT-based model on various datasets, such as Twitter (hate speech), Wikipedia Talk Pages (personal attack) and Formspring (bullying), achieving F1-scores of 0.94, 0.91 and 0.92 respectively. Despite the reported results being very good, they over-sampled the datasets before the train/test split which leads to over-fitting according to [4]. Mozafari et al. [22] proposed adding a CNN layer on top of $\text{BERT}_{\text{base}}$ for hate speech detection, achieving a maximum F1-score of 0.92. However, their proposed architecture could lead to over-fitting and a longer inference time. Although these studies show that BERT outperforms other models on the task of cyberbullying detection, none of them explain why. In recent years, there has been substantial work on the explainability of NLP and Language Models (LMs) [2, 33, 42]. With regards to attention-based models, like Transformers and BERT, [37, 38] built visualisation tools to show the attention weights in different layers between tokens in the same sentence or in two different sentences, as well as to understand the role attention weights play in pre-trained BERT [7] by analysing the behaviour of BERT's attention weights in different layers. Similarly, [19, 27] analysed the capability of BERT to capture different types of linguistic information on

Table 1: Cyberbullying dataset statistics

Dataset	Size	Positive samples	Avg.post length (words)	Max.post length (words)
Kaggle	7425	2578 (35%)	25.28	1419
Twitter-sex	14742	3370 (23%)	15.04	41
Twitter-rac	13349	1969 (15%)	15.05	41
WTP-agg	114649	14641 (13%)	75.45	2846
WTP-tox	157671	15221 (10%)	73.51	2320

the General Language Understanding Evaluation (GLUE) tasks. Regarding attention mechanisms and model explainability, Jain and Wallace [17] showed that contrary to the assumption that attention provides a form of explainability, attention weights do not provide meaningful explanations, with the same finding being supported by [29, 32, 35]. Inspired by this work on the analysis of BERT models, our goal is to gain a better understanding of BERT's strong performance on cyberbullying detection tasks.

In this work, we attempt to answer the following research questions: i) What is BERT's performance on different cyberbullying-related datasets? ii) What is the role that attention weights play in BERT's performance? iii) What are the features that BERT relies on for its performance? The contributions of this work can be summarised as follows: (i) We demonstrate that fine-tuning BERT with a simple single layer on top of BERT's pooled output outperforms other popular deep learning models on a range of cyberbullying-related datasets. (ii) We show that, as previously suggested [17] for some other domains, attention weights are less meaningful when it comes to explaining model performance in comparison to gradient-based feature importance scores for the task of cyberbullying detection. (iii) We provide evidence that BERT's performance may be due to reliance on syntactical biases in the datasets. The code to reproduce the experiments in this paper is shared in [13].

2 METHODOLOGY

We compared fine-tuned BERT to state-of-the-art LSTM and Bi-LSTM models on five social media cyberbullying detection datasets from different sources and with different sizes. Furthermore, to examine how fine-tuning affects attention weights, we show the difference in attention weights' patterns between BERT with and without fine-tuning. Then, to investigate the role of attention weights of fine-tuned BERT in the model's performance, we compared the mean feature importance score of individual tokens, obtained using Integrated Gradients, to their mean attention weights by computing the Pearson's linear correlation between the mean attention weights of fine-tuned BERT of all heads across the last layers (9-12) and the tokens' absolute importance score, as it has been shown that fine-tuning affects mostly BERT's last layers (9-12) [27]. Finally, we analysed the importance scores of POS tags of fine-tuned BERT to find out the features that BERT relies on to make its prediction.

2.1 Datasets

We used five cyberbullying-related datasets of varying sizes from several social media sources that contained different types of cyberbullying: (i) *Twitter-Racism*, a collection of Twitter messages containing tweets that are labelled as racist or not [39], (ii) *Twitter-Sexism*, Twitter messages containing tweets labelled as sexist or not [39], (iii) *Kaggle-Insults* [18], a dataset that contains social media comments that are labelled as insulting or not, (vi) *WTP-Toxicity*, a collection of conversations from Wikipedia Talk Pages (WTP) annotated as friendly or toxic [40], and (v) *WTP-Aggression*, conversations from WTP annotated as friendly or aggressive [40]. Information about the datasets is provided in Table 1.

2.2 Dataset pre-processing

For BERT, we followed [9]'s pre-processing steps: (1) We removed URLs, user mentions, non-ASCII characters, and the retweet abbreviation "RT" (Twitter datasets). (2) All letters were lower cased. (3) Contractions were converted to their formal format. (4) A space was added between words and punctuation marks. For the RNN models, we additionally removed punctuation and English stop words, as proposed in [3]. However, second-person pronouns like "you", "yours" and "your", and third-person pronouns like "he/she/they", "his/her/their" and "him/her/them" were not removed because we noticed in our datasets that sometimes, profane words on their own, e.g. "f**k", are not necessarily used for bullying reasons, while their combination with a pronoun, e.g. "f**k you", is used to insult someone. Then, each dataset was randomly split into a training (70%) and test (30%) set, preserving class ratios.

2.3 Deep learning models

BERT with fine-tuning was used for the task of text classification on the examined datasets, by employing $\text{BERT}_{\text{base}}$(uncased) [14]. For fine-tuning, BERT was trained for 10 epochs with a batch size of 32 and a learning rate of $2e^{-5}$, as suggested in [10]. The sequence length parameter changed across datasets depending on their maximum token length. For the Twitter-sexism and Twitter-racism datasets, a sequence length of 64 was used because it is the closest to the maximum observed sequence length in the dataset, while 128 was used for the rest because it is the maximum we could use due to available computational resources limitations. A single linear layer was added on top of the pooled output of BERT for sentence classification. We also used LSTM [16] and Bi-directional LSTM [28], with the same architecture as in [3], who used RNN models to detect cyberbullying. To this end, we first used the Keras tokeniser [34] to convert the text into numerical vectors (each integer being the index of a token in a dictionary) with a maximum length of 600 (the maximum we could use due to computational resources limitations) for the Kaggle and WTP datasets and 41 (maximum observed sequence length in the dataset) for the Twitter datasets. A trainable embedding layer was used as the first hidden layer of the LSTM and Bi-LSTM-based networks, with an input size equal to the number of unique tokens of the dataset after pre-processing and an output size of 128. The two models were then trained for 100 epochs with a batch size of 32, using the Adam optimiser and a learning rate of 0.01 which is the default of the Keras Optimiser.

3 EXPERIMENTAL RESULTS

3.1 Classification performance

The performance of the trained models on the test set is reported in Table 2. The initial training set for each model and dataset was randomly stratified-split into a training (70%) and validation (30%) set. The model was then trained using the training set, validated on

Table 2: F1-scores achieved for each dataset

Dataset	LSTM	Bi-LSTM	BERT(FT)
Kaggle	0.6420	0.653	0.768
Twitter-sex	0.6569	0.649	0.760
Twitter-rac	0.6400	0.678	0.757
WTP-agg	0.7110	0.679	0.753
WTP-tox	0.7230	0.737	0.786

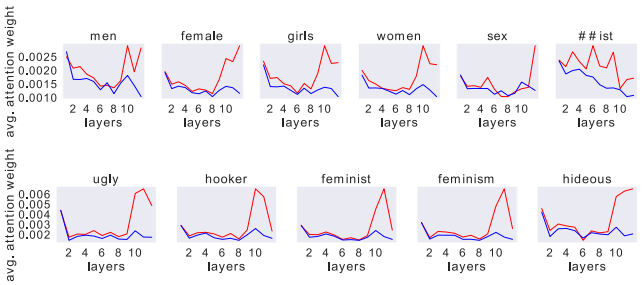

Figure 1: Mean attention weights of 12 heads per layer for fine-tuned BERT (red) and BERT without fine-tuning (blue), for the most important cyberbullying class-related tokens in the Twitter-sexism dataset according to Naive Bayes (top) and gradient-based importance scores (bottom).

the validation set and tested on the original test set. This procedure was repeated five times and the final performance of each model for each dataset was reported as the mean F1-score for the test set across the five iterations. From Table 2, it is evident that BERT with fine-tuning (FT) outperformed all the other examined models, reaching a highest F1-score of 78.6% for the WTP-Toxicity dataset. The Friedman test [43] was used to compare the F1-scores of LSTM, Bi-LSTM and BERT (FT) across the five datasets, showing that BERT (FT) performed significantly better ($p < 0.05$). We then analysed the inner-workings of BERT to get insight into the reasons behind BERT's performance, starting with BERT's attention weights.

3.2 Attention weights (FT vs. NFT)

We examined the difference in attention weights' patterns between fine-tuned BERT (FT) and BERT without fine-tuning (NFT) on the Twitter-sexism dataset. To this end, we examined the attention weights of the five words with the highest probability for the cyberbullying class (according to a Multinomial Naive Bayes model) in BERT (FT) and BERT (NFT). From Figure 1(top), it is evident that the mean weights of the attention heads in the last layers of BERT (FT) (red lines) were much higher than for BERT (NFT) (blue lines), showing that the pattern of BERT (FT) in the last layers changed after fine-tuning compared to BERT (NFT). We repeated the same experiment using gradient-based importance scores [33] to get the most important words for the cyberbullying class and found a similar pattern, as shown in Figure 1(bottom). Similar results were observed for all the datasets: WTP, Kaggle and Twitter-racism.

3.3 Attention weights vs. importance scores

In the previous experiment, we demonstrated that fine-tuned BERT assigns higher attention weights to the last layers, compared to BERT without fine-tuning. This raises the following question: "*Do the attention weights of the last layers (9-12) of fine-tuned BERT explain the model's outcome?*" To answer this, we examined the correlation between gradient-based feature importance score and attention weights of fine-tuned BERT. Gradient-based feature importance scores provide a measure of the importance of individual features with known semantics [33] and have been used in previous studies for attention weights' analysis [7, 29, 32]. To compute the importance scores for all the datasets, we used the Integrated Gradients algorithm [33]. A subset of 1000 samples was randomly selected from the test set of each dataset and the absolute importance scores of all the tokens in these subsets were computed. Then, all scores were grouped by the tokens, and the mean absolute feature importance score was computed for each unique token. The same strategy was also followed for the attention weights. We computed the mean attention weight across all 12 heads per each layer, as well as the mean attention weight of the last layers (9-12), where BERT's fine-tuning is most impactful. Then, we grouped the mean attention weights by tokens and computed the mean attention weights per each token. Pearson's correlation coefficient (PCC) was used to measure the linear correlation between the mean importance scores, the mean attention weights, and the occurrences of different tokens, as shown in Table 3. Our usage of PCC was inspired by early work on attention weights by [17]. We found no linear correlation between the absolute importance score and the mean attention weights of BERT for the examined datasets ($0.056 \leq \text{PCC} \leq 0.171$), as well as between the number of occurrences of a token and the mean attention weights ($-0.101 \leq \text{PCC} \leq -0.015$) or the mean importance scores ($-0.011 \leq \text{PCC} \leq -0.002$). These results suggest that attention weights don't play a direct role in explaining BERT's performance, which is in line with previous studies [29, 32, 35].

Table 3: PCC between mean attention weights of fine-tuned BERT, mean absolute feature importance and number of occurrences per token

Dataset	No. tokens	PCC (attention vs importance)	PCC (attention vs no. occurrences)	PCC (importance vs no. occurrences)
Twitter-Sexism	3878	0.108	-0.047	-0.002
Twitter-Racism	3991	0.056	-0.015	-0.002
Kaggle-Insults	4452	0.171	-0.023	-0.004
WTP-Aggression	4457	0.125	-0.101	-0.009
WTP-Toxicity	4524	0.163	-0.076	-0.011

3.4 What does BERT learn during fine-tuning?

We used spaCy [30] to compute the absolute gradient-based importance scores of the POS tags from the examined datasets and normalised them to the range $[0, 1]$ per dataset in order to examine whether BERT learns, during fine-tuning, general cyberbullying-related features or if it relies on syntactical biases, which means the model relies only on a certain syntax to make its decision, in the dataset. Our hypothesis is that, if BERT learns cyberbullying-related features, the POS tags that receive the highest importance scores will be nouns, adjectives, adverbs, proper nouns, and pronouns and that there will be similarities in the pattern across all the datasets. On the other hand, if BERT relies on syntactical biases, the POS tags that receive the highest importance scores will be tags like punctuation, auxiliaries, determiners, and adpositions and the patterns will differ across datasets from different domains.

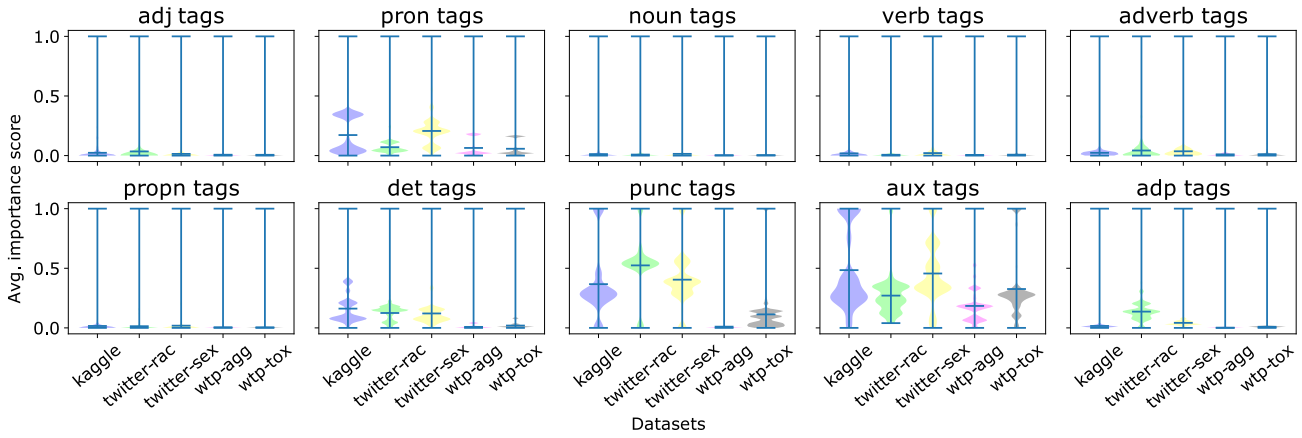

Figure 2: Mean normalised importance scores assigned by fine-tuned BERT to POS tags in the datasets.

Table 4: p-values for the Wilcoxon sign-ranked test between the mean importance scores of the datasets.

	Kaggle	Twitter-rac	Twitter-sex	WTP-agg	WTP-tox
Kaggle	-	0.845	0.556	0.001	0.001
Twitter-rac	0.845	-	0.921	0.001	0.048
Twitter-sex	0.556	0.921	-	0.001	0.001
WTP-agg	0.001	0.001	0.001	-	0.064
WTP-tox	0.001	0.048	0.001	0.064	-

The reason behind using POS tags is that they are important linguistic features that can explicitly show the model's syntactical bias. Results (Figure 2) showed that the POS tags with the highest importance scores are auxiliaries, punctuation, determiners, adpositions, and pronouns. Among these, the most informative tag for cyberbullying detection is the pronoun. The distributions of the tags in Figure 2 show similarities and differences across the datasets. A Wilcoxon sign-ranked test [43] was used to test the statistical significance of the difference between the importance scores of the POS tags across different datasets (Table 4). Results showed that a statistically significant difference could not be established between WTP-agg and WTP-tox and between Twitter-sexism and Twitter-racism ($p > 0.05$). We speculate that this happens because the domain of the datasets is the same. Similar results were found between Kaggle and Twitter-racism and between Kaggle and Twitter-sexism ($p > 0.05$). A statistically significant difference was shown between WTP-agg and Twitter-racism ($p = 0.001$), WTP-agg and Twitter-sexism ($p = 0.001$), WTP-tox and twitter-racism ($p = 0.048$), WTP-tox and Twitter-sexism ($p = 0.001$), Kaggle-insults and WTP-agg ($p = 0.001$), and Kaggle-insults and WTP-tox ($p = 0.001$). We speculate that this is because the domains of the datasets differ. The results support our hypothesis that BERT does not rely on semantic features related to cyberbullying but instead relies on syntactic biases in the datasets that may change between different domains.

We further inspected the POS tags with the highest importance scores, like auxiliaries, determiners and punctuation across the different datasets. For **determiners** and **punctuation**, Kaggle, Twitter-sexism and Twitter-racism datasets, which have the highest scores for determiners and punctuation, contain less noise compared to WTP-agg and WTP-tox. Noise here denotes that determiners or punctuation are mixed with other nouns and/or symbols e.g. "anti-white.the". In contrast, **auxiliaries**, received the highest importance scores across all the datasets, since the detected auxiliaries did not have any noise in any of the datasets. We speculate that the noise is the cause of the low importance scores in WTP datasets. We also speculate that the domain of the dataset contributes to the amount of noise that can exist in the dataset. For example, Twitter does not allow long text, which means that even if mistakes and noise exist, the occurrences of noise is limited compared to a platform like Wikipedia Talk Pages where there is no text limit, thus allowing more space for noise. This provides additional evidence that the domain of the dataset affects its syntactical composition and in turn affects BERT's performance and potentially limits its generalisability due to BERT learning syntactical biases instead of cyberbullying-related linguistic features.

4 CONCLUSION

In this work, we conducted a series of experiments on five datasets to analyse the performance of BERT on the task of cyberbullying detection. Results showed that BERT outperforms other commonly used deep learning models on multiple cyberbullying-related datasets. In addition, even though the patterns of the attention weights of fine-tuned BERT are different from those of BERT without fine-tuning, results showed that attention weights are not meaningful when it comes to the model's prediction, and that BERT captures syntactical biases in the datasets. This might lead to some limitations in BERT's generalisability. Thus, results showed that attention weights do not explain the performance of fine-tuned BERT and that its success is due to the reliance on syntactical biases in the datasets. Our findings indicate that our understanding of cyberbullying detection using pre-trained models like BERT can be improved by using gradient-based feature importance methods, which can assist in revealing some of the biases in the model or the dataset, thus helping towards fair detection of cyberbullying. We also expect that fine-tuning BERT on datasets with diverse syntactical structures will help to improve generalisation, so that BERT does not rely on specific syntactic biases found in some datasets.

REFERENCES

[1] Ghada M Abaido. 2020. Cyberbullying on social media platforms among university students in the United Arab Emirates. *International Journal of Adolescence and Youth* 25, 1 (2020), 407–420.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[3] Sweta Agrawal and Amit Awekar. 2018. Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms. In *Advances in Information Retrieval*, Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (Eds.). Springer International Publishing, Cham, 141–153.

[4] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international ACM SIGIR Conference on Research and Development in Information Retrieval*. 45–54.

[5] Bill Belsey. 2005. Cyberbullying: An emerging threat to the "always on" generation. *Recuperado el* 5, 5 (2005), 2010.

[6] Tommy KH Chan, Christy MK Cheung, and Randy YM Wong. 2019. Cyberbullying on social networking sites: the crime opportunity and affordance perspectives. *Journal of Management Information Systems* 36, 2 (2019), 574–609.

[7] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 276–286.

[8] Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. 2014. Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies. In *Advances in Artificial Intelligence*, Marina Sokolova and Peter van Beek (Eds.). Springer International Publishing, Cham, 275–281.

[9] Huong Dang, Kahyun Lee, Sam Henry, and Ozlem Uzuner. 2020. Ensemble BERT for Classifying Medication-mentioning Tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. 37–41.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.

[11] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *5th International AAAI Conference on Weblogs and Social Media*.

[12] Maeve Duggan. 2017. Online harassment 2017. Pew Research Center, https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/.

[13] Fatma Elsafoury, Stamos Katsigiannis, Steven R. Wilson, and Naeem Ramzan. 2021. https://github.com/efatmae/Does-BERT-pay-attention-to-cyberbullying-/tree/main/Does-BERT-pay-attention-to-cyberbullying-.

[14] Google Research. 2020. BERT. https://github.com/google-research/bert. Accessed: 2020-09-28.

[15] Leslie Haddon and Sonia Livingstone. 2017. Risks, opportunities, and risky opportunities: How children make sense of the online environment. In *Cognitive Development in Digital Contexts*. Elsevier, 275–302.

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[17] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *NAACL-HLT (1)*. Association for Computational Linguistics, 3543–3556.

[18] Kaggle. 2012. Detecting Insults in Social Commentary. https://www.kaggle.com/c/detecting-insults-in-social-commentary/data. Accessed: 2020-09-28.

[19] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. In *EMNLP/IJCNLP (1)*. Association for Computational Linguistics, 4364–4373.

[20] Robin M Kowalski, Gary W Giumetti, Amber N Schroeder, and Heather H Reese. 2012. Cyber bullying among college students: Evidence from multiple domains of college life. In *Misbehavior online in higher education*. Emerald Grp. Pub. Ltd.

[21] Akshi Kumar, Shashwat Nayak, and Navya Chandra. 2019. Empirical Analysis of Supervised Machine Learning Techniques for Cyberbullying Detection. In *International Conference on Innovative Computing and Communications*, Siddhartha Bhattacharyya, Aboul Ella Hassanien, Deepak Gupta, Ashish Khanna, and Indrajit Pan (Eds.). Springer Singapore, Singapore, 223–230.

[22] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A BERT-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*. Springer, 928–940.

[23] Sayanta Paul and Sriparna Saha. 2020. CyberBERT: BERT for cyberbullying identification. *Multimedia Systems* (2020).

[24] John Pavlopoulos, Nithum Thain, Lucas Dixon, and Ion Androutsopoulos. 2019. Convai at semeval-2019 task 6: Offensive language identification and categorization with perspective and bert. In *Proceedings of the 13th international Workshop on Semantic Evaluation*. 571–576.

[25] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful What You Share in Six Seconds: Detecting Cyberbullying Instances in Vine. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (Paris, France) *(ASONAM '15)*. ACM, 617–622.

[26] E. Raisi and B. Huang. 2017. Cyberbullying Detection with Weakly Supervised Machine Learning. In *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 409–416.

[27] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A Primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics* 8 (2021), 842–866.

[28] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[29] Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2931–2951.

[30] Spacy. 2021. Linguistic features by Spacy. https://spacy.io/usage/linguistic-features. Accessed: 2021-03-02.

[31] Fabio Sticca, Sabrina Ruggieri, Françoise Alsaker, and Sonja Perren. 2013. Longitudinal risk factors for cyberbullying in adolescence. *Journal of community & applied social psychology* 23, 1 (2013), 52–67.

[32] Xiaobing Sun and Wei Lu. 2020. Understanding Attention for Text Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3418–3428.

[33] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. 3319–3328.

[34] Tensorflow.org. 2020. Text tokenization utility class. https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer. Accessed: 2020-09-28.

[35] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention Interpretability Across NLP Tasks. *arXiv.org* arXiv:1909.11218 (2019).

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems (NIPS 2017)*. 5998–6008.

[37] Jesse Vig. 2019. A Multiscale Visualization of Attention in the Transformer Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Florence, Italy, 37–42.

[38] Jesse Vig and Yonatan Belinkov. 2019. Analyzing the Structure of Attention in a Transformer Language Model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Florence, Italy, 63–76.

[39] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93.

[40] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, 1391–1399.

[41] Jaideep Yadav, Devesh Kumar, and Dheeraj Chauhan. 2020. Cyberbullying Detection using Pre-Trained BERT Model. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 1096–1100.

[42] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. 2019. "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. In *International Conference on Machine Learning, AI for Social Good Workshop* (Long Beach, CA, USA).

[43] Donald W Zimmerman and Bruno D Zumbo. 1993. Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks. *The Journal of Experimental Education* 62, 1 (1993), 75–86.