2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings) | 979-8-3503-2234-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIBThings58340.2023.10292460

NLP-driven Content Classification Towards Fake News and Bully Detection

Darshan Gangaram Sarkale
Computer Science
Lakehead University
Thunder Bay, Canada
dsarkale@lakeheadu.ca

Vedant Jagdishbhai Gabani
Computer Science
Lakehead University
Thunder Bay, Canada
vgabani1@lakeheadu.ca

Wandong Zhang
Electri. & Computer Engg.
Western University
London, Canada
wzhan893@uwo.ca

Thangarajah Akilan
Software Engg.
Lakehead University
Thunder Bay, Canada
takilan@lakeheadu.ca

***Abstract*—Fake news and cyberbullies proliferate in today's society, and detecting them at their onset is paramount because of their anti-social elements that affect people's life and the trustworthiness of online platforms, particularly social media, like Twitter. However, it is a herculean task due to the unstructured nature of natural language and the limited availability of curated datasets to build automated solutions. Recent studies show that deep learning (DL) inspired natural language processing (NLP)-driven solutions can overcome the challenges. In this direction, this study introduces a lightweight convolutional neural network (CNN) with an attention mechanism to address the fake news and cyberbully detection problems. A thorough experimental study conducted on multiple benchmark datasets proves that the proposed model achieves extremely competitive results. It records significant improvements of 1% - 35% when compared to existing baseline models with respect to the benchmark datasets.**

***Index Terms*—Bully detection, cnn, content classification, dl, fake news, nlp**

I. Introduction

As technology and social media have grown in popularity, more people are inclined to use them for news updates and interaction with others. However, this has also led to a significant increase in social threats in the form of fast spreading of misinformation and cyberbullying. Thus, online sources' authenticity and credibility have been questioned. Fake news has the potential to mislead people and redirect their attention away from critical social, environmental, or political problems. Besides, it can also be used in cyberbullying to discredit particular individuals or groups, as shown by recent studies comparing the endorsement of fake news with bullying [1]. Even in real-world violent incidents that endanger public safety related to fake news, such as the Pizza Gate incident.

Cyberbullying is usually seen as a broad category of planned, hostile behaviors that target particular people or communities with the main objective of harming or defaming them. Researchers face the challenge of identifying fake news and cyberbullying content on social media, and try to find the link between the two. Predictions for sensitive topics are challenging since the automated systems find it difficult to distinguish between real and fake data, or between bullying and solidarity because there are not many large annotated data samples readily available for agile solution development.

This study focuses on these issues, particularly when it comes to sensitive topics, like religion, politics, and racism. It aims to develop reliable, effective, and flexible models for improving the performance of textual content classification, more specifically identifying fake news and cyberbullies. The main contribution of this work can be summarized as follows: (i) Introducing an attention-based efficient CNN for the aforesaid tasks, and (ii) Exhaustive ablation study to validate the proposed solution.

The rest of this paper is structured as follows. Section II systematically reviews the related works under four categories and discusses their limitations. Section III elaborates on the proposed methodology step-by-step. Section IV provides an in-depth analysis of the experimental results, while Section V concludes the paper with research findings and future directions.

II. Review on Related Works

This section reviews the related literature under four categories, as shown in Fig. 1.

A. Supervised Methods

Supervised learning approaches have been widely adopted for several textual content analytical problems. For instance, Granik *et al.* [2] and Bhutani *et al.* [3] used a Naïve Bayes (NB) classifier for fake news detection. While Bhutani *et al.* [3] focused on emotions and sentiment-based detection, Granik *et al.* [2] found connections between spam email and fake news. Saravanaraj *et al.* [4] proposed a general framework for detecting rumor and bullying tweets, utilizing both NB and Random Forest (RF) classifiers based on word2vec features. Additionally, they devised their model to extract demographic information about the culprits, including name, gender, and

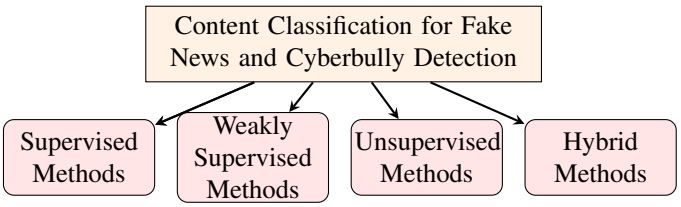


Fig. 1: An abstract categorization of content classification methods for fake news and cyberbully detection.

age. Few researchers, like Saleh *et al.* [5] and Kaliyar *et al.* [6] deployed deep CNNs to learn strong discriminatory features for increasing textual content classification results. Similarly, Dong *et al.* [7] built a dual-head CNN by taking two streams of inputs–title and content. Lately, there is a new research direction that focuses on exploiting the power of graph representation of textual data for content classification. For instance, Chandra *et al.* [8] came up with Graph Neural Network (GNN)-based model that examines the user behavior and shared content to identify fake news. On the other hand, Mahmud *et al.* [9] conducted a comparative analysis between commonly used machine learning approaches and GNNs for fake news detection. According to their finding, the GNNs outperform every other model considered in their study.

B. Weakly Supervised Methods

This strategy of building an analytical model uses supervision signals with imprecision, noise, or limited labeled data to overcome the cost involved in creating a massive amount of data samples with ground truth. For instance, Syed *et al.* [10] showed a hybrid weakly supervised model that gives the best performance with machine learning algorithms, such as Support Vector Machine (SVM), RF, and Logistic Regression. However, when deep learning approaches are applied, the model becomes complicated and is unable to provide accurate outcomes. Similarly, Wang *et al.* [11] proposed a learning framework that combines weak supervision with reinforcement learning. It subsumes three stages: (1) the unlabeled data is weekly annotated, (2) high-quality samples are selected using a reinforced selector, and (3) the selected samples are used to train a classifier for fake news detection. Xie *et al.* [12] employed a Label Noise-Resistant Mean Teaching (LNMT) approach to training their model. Hence, Raisi *et al.* [13] presented a weakly-supervised ensemble architecture, in which two deep learners co-train one another. While one learner analyses the input text's content, the other learner takes the user's social structure into account. The experiment accomplished by Abhishek [14] reveals that weakly supervised machine learning models are more efficient than state-of-the-art supervised machine learning models in tackling noise and variability in unstructured text data.

C. Unsupervised Methods

These methods perform predictions without having been exposed to labeled samples [15]–[17]. For example, Yang *et al.* [18] proposed a tri-relationship algorithm for detecting fake news. It makes use of social media communication to find user opinions, the reliability of news, and their connections. The dependencies are represented by a Bayesian network model, and an efficient version of Gibbs sampling is used for inference without labeled data. Silva *et al.* [19] implemented a novel unsupervised fake news detection framework that encodes the news data as a low-dimensional vector using a domain-agnostic technique and then passes the vectors to a noise-robust teacher-student architecture to determine the labels. The existence of a few clean labels additionally enhanced the model's performance. Likewise, Di Capua *et al.* [20] developed an unsupervised bully behavior detection model with Growing Hierarchical Self Organizing Maps (GSOM). On the other hand, Kaliyar *et al.* [21] proposed a deep learning-based unsupervised approach utilizing Bidirectional Encoder Representations from Transformers (BERT). The authors combined the parallel blocks of single-layer CNN with BERT, which helps to manage text data ambiguities.

D. Hybrid Methods

Umer *et al.* [22] combined Long Short-Term Memory (LSTM) units with CNN, and a feature dimensionality reduction technique, the Principal Component Analysis (PCA). Another work by Jain *et al.* [23] proposed a mixed learning strategy using a Naïve Bayes classifier and an SVM. Prathyusha *et al.* [24] introduced a hybrid model to identify cyberbullying text in crime investigation forums. It uses the Multiple Correlation Coefficient (MCC) to find the word correlation in the input representation for feature selection and SVM for final classification. Akhter *et al.* [25] demonstrated a robust hybrid ML model that uses the TfidfVectorizer for feature extraction followed by an Instance Hardness Threshold (IHT) for data re-sampling to overcome the overfitting and underfitting problems. Since the study only focuses on cyberbullying in Bengali, its applicability is restricted.

III. Methodology

A. Model Description

Fig. 2 depicts the architecture of the proposed attention-based CNN. It subsumes an embedding layer followed by multiple 1D convolutional (Conv) layers, subsampling layers, and fully connected layers that work together to extract relevant features from the input data hierarchically. The embedding layer is the front end of the model. It takes a 700-dimensional sparse vector representation and encodes it into 64 feature maps. The 1D Conv layers transform their input sequence, $\mathbf{x}_{seq}$ through the following computation.

$$\mathbf{y} = \text{ReLU}\left(\mathbf{w} * \mathbf{x}_{seq} + b\right), \tag{1}$$

where $\mathbf{y}$ is the output feature, ReLU is the rectified linear unit activation, $\mathbf{w}$ and b are the set of trainable parameters (weights, bias) of the Conv kernel, and $*$ denotes the convolution operation. The proposed model contains three Conv layers interspaced with the subsampling layer. In this case, the subsampling operations are carried out in the proposed model using the max pooling operation. The first Conv layer uses 32 feature detectors with a kernel of size 1×5. The second Conv layer uses 16 feature detectors with a kernel of size 1×3, while the third Conv layer uses 8 feature detectors with a kernel of size 1×3. The flattened layer transforms the learned features from the third Conv layer into a 1-dimensional array. These flattened feature values are fed to the dense layer having 256 neurons, ReLU activation, and a dropout rate of 0.5. On top of this, an attention mechanism is implemented. The activation function of the output layer is customized based on the type of classification problem. For fake news detection, it is set to

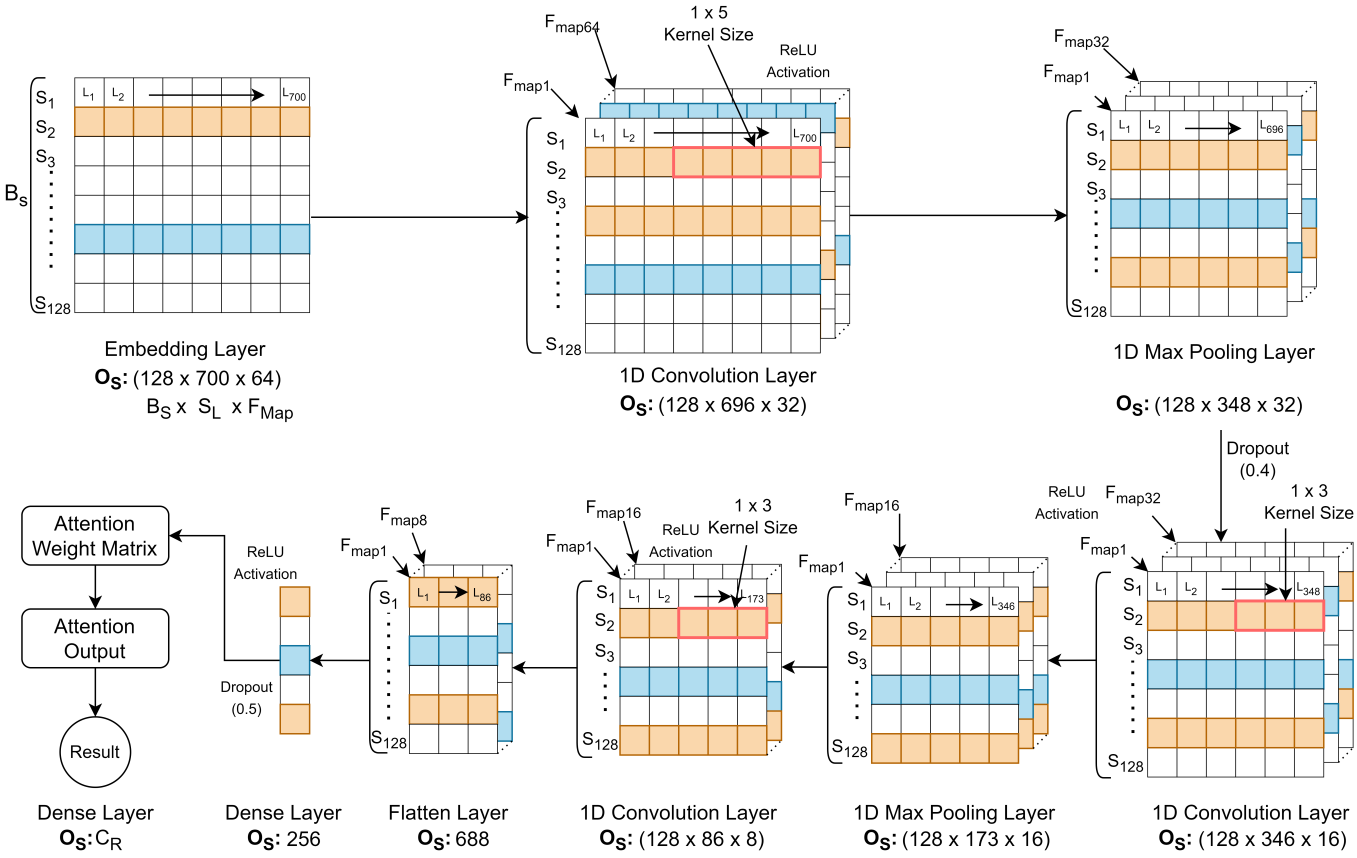


Fig. 2: A Detailed layer-wise flow diagram of the proposed CNN model. This contains several steps starting from getting input data to producing the final output classification result. B_s - Batch size, S_L - Sequence length, F_{map} - Feature map, O_s - Output shape, C_R - Classification result.

Sigmoid (cf. (2)), while for bully detection, it is set to Softmax (cf. (3)).

$$\sigma(z) = 1/(1 + exp(-z)), \tag{2}$$

where z is the computed logit of the output neuron. On the other hand, the Softmax function estimates the probability distribution over predicted output classes as in (3).

$$\sigma(\mathbf{z})_i = \frac{exp(z_i)}{\sum_{j=1}^{K} exp(z_j)}, \tag{3}$$

where k is the number of classes and $\mathbf{z}$ is the logits of the output layer.

In general, an attention mechanism calculates attention weights for each input element and applies them to produce a more focused and informative representation of the input. For a given input sequence $\mathbf{x} = (x_1, x_2, \cdots, x_n)$, the attention mechanism produces a sequence of attention weights $\mathbf{a} = (a_1, a_2, \cdots, a_n)$, where each attention weight a_i represents the importance of the i-th element in the input sequence for producing the final output as defined in (4).

$$a_i = \frac{exp(e_i)}{\sum_{j=1}^{n} exp(e_j)}, \tag{4}$$

where e_i is the energy of the i-th element, computed using a function as $e_i = f(x_i)$. Thus, the attention head generates the output z, as in (5).

$$z = \sum_{i=1}^{n} a_i x_i. \tag{5}$$

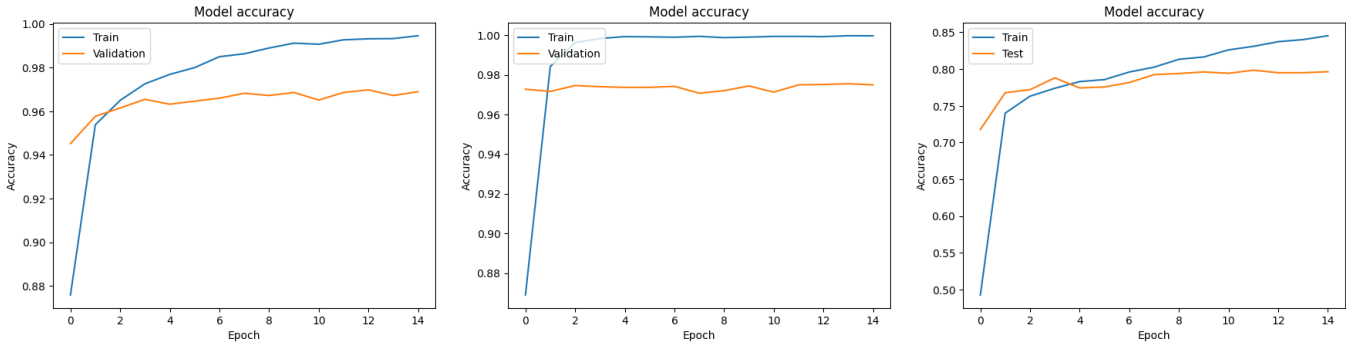
B. Model Training

To train and test the proposed model, this work uses the five benchmark datasets–four for fake news detection and one for cyberbullying detection, as summarized in Table I.

TABLE I: Dataset Description

Dataset	Main Attributes	Dimensionality
WELFake [26]	Title, Text, Label	72,095 × 4
Fake or Real News[1]	Title, Category of news	6,335 × 4
Korean Dataset[2]	Title, Content, Label	Mission 1 - 30,139 × 4 Mission 2 - 67,970 × 4
CyberBullying [27]	tweet_text, cyberbullying_type	47,692 × 2

[1]https://www.kaggle.com/datasets/jillanisofttech/fake-or-real-news
[2]https://github.com/2alive3s/Fake_news/tree/master/data



(a) WELFake (b) Korean Mission - 2 (c) CyberBullying

Fig. 3: Training progressed of the proposed content classification attention-based CNN model on two fake news detection datasets - WELFake, Korean Mission-2, and a cyberbullying dataset.

From each dataset, the model uses a mutually exclusive set of 70% of the samples for training, 10% of the samples for validation, and 20% of the samples for testing. The proposed model utilizes only the text/content of the datasets, and preprocess them with the standard Keras libraries. The preprocessing includes stop word removal, case folding, and tokenization. Hence, the model is trained using the following hyperparameters: Optimizer - Adam, batch size - 128, epochs - 15, Loss function - Categorical cross-entropy for cyberbully detection, and Binary cross-entropy for fake news detection. The training progress samples of the proposed model for three datasets are shown in Fig. 3.

IV. Experimental Study

A. Evaluation Metrics

To be consistent with baseline models, the experimental analysis uses two evaluation metrics–accuracy and area under the receiver operating characteristic curve (AU-ROC). Accuracy is the percentage of correctly classified instances out of all instances as defined in (6).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

where TP stands for true positive, which is the number of instances that are correctly classified as positive, while TN stands for true negative, which is the number of instances that are correctly classified as negative. Hence, FP and FN represent false positive–the number of instances that are incorrectly classified as positive and false negative–the number of instances that are incorrectly classified as negative, respectively. The AU-ROC measures how well the model is able to distinguish between positive and negative instances following (7).

$$\textit{AUC-ROC} = \int_{-\infty}^{\infty} ROC(\tau)\partial\tau, \tag{7}$$

where $ROC(\tau)$ represents the ROC curve at a given threshold value τ, and $d\tau$ denotes the differential element in the integral. The ROC plots the True Positive Rate (TPR) in equation (8) against the False Positive Rate (FPR) found in equation (9).

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

In (8) and (9), TP and FP represent samples that are correctly and incorrectly classified as positive, respectively. Hence, TN and FN represent samples that are correctly and incorrectly classified as negative, respectively.

B. Performance Analysis

Experiments are conducted on the model with and w/o the application of the attention mechanism, and the results are summarized in Table II. The application of the attention mechanism tends to have a positive impact on the overall performance of the model up to 1.8%. For instance, on WELFake dataset, the model w/t the attention mechanism achieves 97.1% of accuracy, while w/o the attention mechanism, it achieves

TABLE II: Comparative Analysis of the Proposed Models wrt Benchmark Models on the Respective Test Sets

Dataset	Baseline	Ours w/o Attn.	Ours w/ Attn.	Improvement (%)**
WELFake	96.0 [28]	95.9	97.1	1.14 ↑
Fake or Real News	84.3 [3]	90.9	91.1	8.07 ↑
Korean - Mission 1	52.8* [7]	71.3*	71.4*	35.22 ↑
Korean - Mission 2	72.6* [7]	97.3*	97.4*	34.16 ↑
CyberBullying	80.1 [29]	79.1	80.9	1.0 ↑

Except for the results indicated with * (Result in AU-ROC) all others are the model accuracies;
** - Ours w/t attention compared to the baseline

95.9% of accuracy. So, with the attention mechanism, there is an improvement of 1.2%. Similarly, on the Cyberbullying dataset, the model w/t the attention mechanism shows an improvement of 1.8%. When compared to the baseline models, the proposed attention-based CNN model shows significant improvements. For example, in Korean - Mission 1 and 2 datasets, the percentage of performance enhancement is $> 30\%$ compared to their respective baseline models. Hence, the proposed model records the mean average performance melioration of $> 15\%$ when compared to existing baseline solutions.

V. Conclusion

Fake news and cyberbullying have become critical issues in today's society due to the widespread use of social media and internet access. The ability to quickly spread false information or engage in harmful behavior online can have severe consequences for individuals, communities, and societies. Fake news can spread misinformation, erode trust, and incite violence, while cyberbullying can cause severe mental health issues to people. Addressing these issues through advanced technology can help us to create a healthier and more informed online environment that protects individuals from harm. Thus, in line with the current research efforts, this work introduces an attention-based deep learning model for fake news and cyberbullying detection. The experimental results demonstrate the effectiveness of the proposed model. The limitation of the proposed model is that it utilizes only text content. Hence, future work aims to improve the model to consider emoji and meme-based cyberbullying and misinformation as social media users increasingly engage with these forms of communication. Overall, leveraging deep learning and natural language processing can help us to detect and mitigate the spread of fake news and minimize the severity of cyberbullying, ultimately creating a more informed and respectful online community.

References

[1] D. Zizumbo-Colunga and M. del Pilar Fuerte-Celis, "The political psychology of lynching: Whatsapp rumors, anti-government appeals, and violence," 2020.

[2] M. Granik and V. Mesyura, "Fake news detection using naive bayes classifier," in *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*. IEEE, 2017, pp. 900–903.

[3] B. Bhutani, N. Rastogi, P. Sehgal, and A. Purwar, "Fake news detection using sentiment analysis," in *2019 twelfth international conference on contemporary computing (IC3)*. IEEE, 2019, pp. 1–5.

[4] A. Saravanaraj, J. Sheeba, and S. P. Devaneyan, "Automatic detection of cyberbullying from twitter," *International Journal of Computer Science and Information Technology & Security (IJCSITS)*, 2016.

[5] H. Saleh, A. Alharbi, and S. H. Alsamhi, "Opcnn-fake: Optimized convolutional neural network for fake news detection," *IEEE Access*, vol. 9, pp. 129 471–129 489, 2021.

[6] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "Fndnet–a deep convolutional neural network for fake news detection," *Cognitive Systems Research*, vol. 61, pp. 32–44, 2020.

[7] D.-H. Lee, Y.-R. Kim, H.-J. Kim, S.-M. Park, and Y.-J. Yang, "Fake news detection using deep learning," *Journal of Information Processing Systems*, vol. 15, no. 5, pp. 1119–1130, 2019.

[8] S. Chandra, P. Mishra, H. Yannakoudakis, M. Nimishakavi, M. Saeidi, and E. Shutova, "Graph-based modeling of online communities for fake news detection," *arXiv preprint arXiv:2008.06274*, 2020.

[9] F. B. Mahmud, M. M. S. Rayhan, M. H. Shuvo, I. Sadia, and M. K. Morol, "A comparative analysis of graph neural networks and commonly used machine learning algorithms on fake news detection," in *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*. IEEE, 2022, pp. 97–102.

[10] L. Syed, A. Alsaeedi, L. A. Alhuri, and H. R. Aljohani, "Hybrid weakly supervised learning with deep learning technique for detection of fake news from cyber propaganda," *Array*, p. 100309, 2023.

[11] Y. Wang, W. Yang, F. Ma, J. Xu, B. Zhong, Q. Deng, and J. Gao, "Weak supervision for fake news detection via reinforcement learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 516–523.

[12] J. Xie, J. Liu, and Z.-J. Zha, "Label noise-resistant mean teaching for weakly supervised fake news detection," *arXiv preprint arXiv:2206.12260*, 2022.

[13] E. Raisi and B. Huang, "Weakly supervised cyberbullying detection using co-trained ensembles of embedding models," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 479–486.

[14] A. Abhishek, "Cyberbullying detection using weakly supervised and fully supervised learning," 2022.

[15] T. Akilan, D. Shah, N. Patel, and R. Mehta, "Fast detection of duplicate bug reports using lda-based topic modeling and classification," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 1622–1629.

[16] T. Akilan, A. Thiagarajan, B. Venkatesan, S. Thirumeni, and S. G. Chandrasekaran, "Quantifying the impact of complementary visual and textual cues under image captioning," in *2020 IEEE International Conference on Systems, Man, and Cybernetics*, 2020, pp. 389–394.

[17] B. D. Patel, H. B. Patel, M. A. Khanvilkar, N. R. Patel, and T. Akilan, "Es2isl: An advancement in speech to sign language translation using 3d avatar animator," in *2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2020, pp. 1–5.

[18] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, and H. Liu, "Unsupervised fake news detection on social media: A generative approach," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 5644–5651.

[19] A. Silva, L. Luo, S. Karunasekera, and C. Leckie, "Unsupervised domain-agnostic fake news detection using multi-modal weak signals," *arXiv preprint arXiv:2305.11349*, 2023.

[20] M. Di Capua, E. Di Nardo, and A. Petrosino, "Unsupervised cyber bullying detection in social networks," in *2016 23rd International conference on pattern recognition (ICPR)*. IEEE, 2016, pp. 432–437.

[21] R. K. Kaliyar, A. Goswami, and P. Narang, "Fakebert: Fake news detection in social media with a bert-based deep learning approach," *Multimedia tools and applications*, vol. 80, pp. 11 765–11 788, 2021.

[22] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On, "Fake news stance detection using deep learning architecture (cnn-lstm)," *IEEE Access*, vol. 8, pp. 156 695–156 706, 2020.

[23] A. Jain, A. Shakya, H. Khatter, and A. K. Gupta, "A smart system for fake news detection using machine learning," in *2019 International conference on issues and challenges in intelligent computing techniques (ICICT)*, vol. 1. IEEE, 2019, pp. 1–4.

[24] T. Prathyusha, R. Hemavathy, and J. Sheeba, "Cyberbully detection using hybrid techniques," in *International Conference on Telecommunication, Power Analysis and Computing Techniques (ICTPACT 2017). IEEE*, 2017, pp. 1–6.

[25] A. Akhter, U. K. Acharjee, M. A. Talukder, M. M. Islam, and M. A. Uddin, "A robust hybrid machine learning model for bengali cyber bullying detection in social media," *Natural Language Processing Journal*, p. 100027, 2023.

[26] P. K. Verma, P. Agrawal, and R. Prodan, "WELFake dataset for fake news detection in text data," Feb. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4561253

[27] J. Wang, K. Fu, and C.-T. Lu, "Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 1699–1708.

[28] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "Welfake: Word embedding over linguistic features for fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 881–893, 2021.

[29] X. Guo, U. Anjum, and J. Zhan, "Cyberbully detection using bert with augmented texts," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 1246–1253.