Detecting Antisocial Behavior on Social Media during COVID-19 Lockdown

Andrew Asante[1][0009−0002−9917−6147] and Petr Hajek[1][0000−0001−5579−1215]

Science and Research Centre, Faculty of Economics and Administration, University of Pardubice, Studentska 84, Pardubice 53210, Czech Republic
andrew.asante@student.upce.cz
petr.hajek@upce.cz

Abstract. The widespread availability of the internet has rendered the engagement with social media an integral component of contemporary society. Platforms such as Facebook, Twitter/X, YouTube, among others, are designed to facilitate extensive, efficient, and sustained user participation, offering both anonymity and opportunities for positive engagement. However, these platforms have also become arenas for antisocial behaviors, including disregard for others' rights, lack of empathy, trolling, and aggression, leading to significant negative psychological impacts on affected individuals. These impacts range from anxiety and emotional trauma to depression, psychological disorders, self-isolation, diminished self-esteem, and even suicidal thoughts. This study focuses on antisocial behavior (ASB) manifested in tweets from Ghana during the 21-day COVID-19 lockdown. We develop a gold-standard annotated ASB corpus from collected and pre-processed data. We then assess the performance of different baseline classifiers against three transformer models—BERT, RoBERTa, and ELECTRA—in a binary classification task designed to detect ASB. Each model demonstrated varying degrees of success; however, the RoBERTa model, upon fine-tuning, exhibited superior performance, achieving an accuracy rate of 95.59% and an F1 score of 94.99%, thereby outperforming the other models.

Keywords: Social media · Large language model · Transformer · Antisocial behavior.

1 Introduction

Societies, communities, and economies are governed by a complex interplay of written and unwritten rules, norms, and values that regulate the behavior of their members. Deviations from these accepted behavioral norms are often characterized as personality disorders or antisocial behaviors (ASB), marked by persistent disregard for others' rights, lack of empathy, engagement in trolling, and displays of aggression and hostility [24].

The rise of online social platforms, where anonymity can shield and enable misbehavior, has exacerbated these issues, creating environments conducive to the proliferation of ASB. Such online conduct can lead victims to experience a

range of psychological impacts, including anxiety, emotional trauma, depression, psychological disorders, self-isolation, low self-esteem, and even suicidal ideation [30].

Research on ASB spans various disciplines including psychology, education, and information systems [24]. Despite its multidisciplinary nature, the integration of computational linguistic analysis through natural language processing (NLP) within the realms of machine learning (ML) and deep learning (DL) has been underexplored in the context of ASB studies. This gap is particularly noticeable when considering the exacerbation of online ASB during the COVID-19 pandemic, underscoring the need for comprehensive theoretical and computational frameworks to analyze ASB across social media platforms.

The existing literature on ASB during the COVID-19 pandemic remains sparse. Some comparative analyses, such as those examining Twitter and Weibo, have identified trends related to cyberbullying [18], yet most studies focus on specific ASB manifestations without a comprehensive approach. Moreover, these studies vary significantly in terms of data collection methods, annotation techniques, and computational approaches. For instance, research on abusive language in social media often employs diverse computational strategies to address different aspects of offensive language detection [20]. Data sources for these studies typically include platforms like Twitter and Facebook, where offensive, abusive, and hateful content is prevalent [3]. Further, the literature reveals a variety of ASB forms studied in different online communities, often comparing individuals who have faced platform bans with those who have not. The predominant analytical framework in these studies involves ML techniques utilizing NLP to categorize text-based features into categories such as lexical, syntactic, linguistic, knowledge-based, and sentiment analysis [28]. Most research employs supervised and unsupervised learning algorithms, with decision trees and Naïve Bayes classifiers combined with NLP techniques being commonly used to detect cyberbullying [21]. Annotation techniques in ASB research also vary, including manual and automatic processes to identify harassment, flaming, hostility, and trolling [13]. For instance, some studies leverage traditional ML and models like BERT to analyze abusive language expressions in emoticons [27], while others use manual annotation and classification algorithms to distinguish between hateful speech and offensive language [6].

In this study, we employ DL algorithms integrated with NLP techniques to automate the detection of ASB within a corpus of tweets related to the COVID-19 pandemic. This approach aims to contribute to the improvement of social welfare and national security by providing insights into online ASB. To address the identified gaps in the literature, we formulated the following research questions:

1. How can we scrape, preprocess, and construct a labeled tweet-based dataset for monitoring ASB?
2. In what ways can DL and NLP methodologies be utilized to identify ASB within a tweet-based corpus related to the COVID-19 pandemic?

3. How do we evaluate the efficacy of DL algorithms in identifying ASB within tweets?

The structure of this paper is as follows. Section 2 reviews relevant empirical literature, highlighting instances of online ASB on social media platforms as manifested in textual posts or comments. Section 3 describes the corpus of documents, detailing the preprocessing steps undertaken. Section 4 presents the analysis and findings from the application of the proposed predictive models. Finally, Section 5 concludes the paper.

2 Literature Review

The dissemination of information through social media and microblogging platforms, particularly in the context of tragic events, has become a staple of contemporary communication. Scholarly research has highlighted the utilization of platforms like Twitter for the rapid sharing of critical information during disasters, ranging from typhoons and Ebola outbreaks to flooding incidents [25]. While the majority of online participants adhere to and uphold social norms, a faction engages in disruptive ASB, undermining both the cohesion of online communities and the platforms' operational efficacy [15].

The manifestation of ASB within digital communities not only fractures social bonds but also exposes individuals to a plethora of social challenges and public health risks. The motivations behind such behaviors are multifaceted; some scholars argue that these actions stem from a desire for amusement, while others suggest they are retaliatory responses to perceived grievances [24]. Irrespective of the underlying reasons, the impact of online antisocial conduct is undeniably harmful, leading to severe psychological consequences for victims.

ASB is characterized by distinct behavioral patterns that fall under the broader categories of psychopathy and sociopathy. These conditions, despite sharing common traits, are delineated by nuanced differences [12]. Such behaviors contribute to a range of societal issues, including increased violence and criminal activities, thereby imposing significant burdens on the criminal justice system, societal well-being, and healthcare infrastructure [10]. The research by Moor and Anderson [16] provides a systematic examination of the link between antisocial tendencies and the dark triad personality traits in the context of social media use. Their findings reveal a strong correlation between psychopathic behaviors and various forms of online misconduct, such as trolling, hostility, and cyber-aggression, thereby affirming the predictive capabilities of models like the one proposed by Munezero et al. [17] for identifying and forecasting antisocial conduct online.

Despite the extensive investigation into these behaviors, the literature remains incomplete, particularly concerning the shifts in online ASB during and following the COVID-19 pandemic. This gap underscores the need for ongoing research to adapt and expand our understanding of digital antisocial dynamics in the face of global crises. Table 1 summarizes key studies on ASB identified in the literature, underscoring the diversity and complexity of this area of study.

Table 1: Summary of previous studies on online antisocial behavior

Ref.	Platform	Behavior classes	Model	Result
[5]	Twitter, Facebook, WhatsApp	Hostility 3,834, Non-Hostility 4,358	SVM, DT, RF, LR	SVM achieved a higher F1-score (84.11%) compared to all other models evaluated, including LR, MLP, and RF.
[22]	Twitter (X)	Hate speech 20,620, Non-Hate speech 4,163	biLSTM and BERT	The model predicted an F1-score of 93% using biLSTM, and achieved 96% on a combined balanced hate speech dataset.
[1]	Google, Facebook, Twitter (X), Instagram	Bullying 600, Non-Bullying 600	CNN (VGG16), BiLSTM	F1-score of 88% and accuracy of 87% were achieved in predicting cyberbullying on social media platforms.
[2]	Youtube	Abusive 6,109 Non-Abusive 6,062	NB, SVM, LR, CNN, LSTM, biLSTM	Using the CNN, accuracies of 96.2% and 91.4% were achieved, outperforming the other models used in the study.
[7]	Twitter (X)	Toxic Comment 47,308 Neutral 63,828	LR,SVM, LDA, BERT	All models demonstrated good performance, with F1-scores close to or above the 90% benchmark. However, a 91.40% F1-score for BERTweet is considered less impressive in the context of this study.

3 Methodology

To explore the prevalence of ASB among web users in Ghana, we systematically collected daily tweets from Twitter/X, forming a comprehensive corpus. Twitter's relatively open access to information, facilitated by its well-structured data and the availability of various tools for programmatic access (such as Twitter APIs and free tweet-collection applications), made it an ideal platform for this research. The data collection phase was strategically timed during the partial lockdown in Ghana, from March 30 to April 20, 2020, to capture the pandemic's impact on online behavior.

The collection process involved defining a set of case-insensitive, frequently used COVID-19-related keywords to ensure the data's relevance to the study's focus. Given the study's emphasis on the Ghanaian populace, we employed distinct hashtags and keywords in our search criteria, such as "ghCovid," "COVID-19GH," "14DaysLockDown," "GhanaCoronavirus," "fightCovid," and "ChinesevirusGh." This approach yielded a dataset[1] of 23,951 unfiltered tweets related to the COVID-19 pandemic, which underwent rigorous pre-processing.

The data pre-processing involved two main stages: cleaning and annotation, followed by tokenization and feature encoding. Data cleaning aimed to remove extraneous elements like alphanumeric characters, punctuation, misspelled words, slang, and URLs, thereby enhancing data quality and analysis accuracy. For example, a tweet reading *"So #Obour ein stupidity that has cost human lives when we go continue chop the matter for Twitter? #stayhomeGH*

[1] The original dataset contains metadata available upon request.

#COVID19Ghana. http://dlvr.it/RT908H" was refined to *"So Obour stupidity that has cost human lives when we go continue chop the matter for Twitter"* through a customized Python script. Following the cleaning process, the curated corpus was structured into sentence-like formats, ready for annotation and further analysis.

In this study, we opted for manual annotation, which involved the selection of two annotators, henceforth referred to as "Annotator A" and "Annotator B," to ensure a rigorous annotation process. The primary goal was to establish a "gold standard" corpus, identifying tweets potentially containing ASB. A key motivation for manual annotation was the absence of a publicly available, annotated dataset on tweet-based ASB during COVID-19 lockdown. Thus, our annotation effort aimed to fill this critical gap, leveraging the annotators' expertise to enrich the corpus with high-level information [4]. Table 2 showcases a selection of annotated data samples, illustrating the diverse tags applied.

Tokenization involves dissecting a textual stream into discrete units—words, phrases, or symbols—termed tokens. This process also includes normalizing uppercase words to lowercase to streamline the corpus and enhance parsing algorithm efficiency [23]. Encoding in the context of DL entails representing tokens as high-dimensional vectors. During pre-processing, this task coincides with tokenization, segmenting the text into tokens and assigning each a unique identifier based on the model's vocabulary [26]. This dual process is pivotal for preparing input representations for subsequent transformer layers, where the self-attention mechanism is applied to the input-output data, ensuring the model's effective comprehension and processing of the textual information.

Table 2: Randomly selected instances of tweets and their associated ASB tags

Sample of Data	**Annotator A**	**Annotator B**	**Gold Standard**
"Coronavirus tough decisions Akufo-Addo has made since coronavirus hit Ghana..."	No ASB	No ASB	No ASB
"Germany demands an amount of billion from China for coronavirus damages Germany. Covid is a Chinese disease....."	Lack of Empathy	No ASB	Lack of Empathy
"Avoid False Sense of Security Coronavirus Not Under Control in Ghana. Akuffo-Addo is a failure and stupid...."	Hostility	Aggressive	Aggressive
"I think Obinim can be used as the Test Lab for the Corona Vaccine to redeem his image now....."	Trolling	Hostility	Trolling

3.1 Data Description

The purpose of the descriptive data analysis conducted on the corpus was to gain insights into the data. The analysis commenced with an examination of the

value counts and the percentage distribution within the dataset. From the total pool of scraped and cleansed data, 21,234 tweet instances were selected for the annotation process. This selection excluded certain instances due to duplications or because they were numerical representations that lacked relevance to the study's focus. Such instances were removed in the early stages of pre-processing to ensure the integrity and relevance of the data analyzed.

Table 3: Frequencies of ASB classes in the annotated tweet corpus

ASB Class	**Annotator A**	**Annotator B**	**Disagreement Level**	**Final Annotation**
No ASB	19,929	19,832	97	20,033
Trolling	759	868	109	731
Aggressive	403	340	63	349
Lack of Empathy	112	145	33	92
Hostility	31	53	22	29
Total	21,234	21,234	324	21,234

Table 3 presents the distribution of annotated classes and highlights the instances of disagreement between the annotators. It also illustrates the frequency distribution of the final annotated classes. This finalization occurred after discussions between the annotators led to a consensus on the divergent interpretations of certain instances. This step was crucial to ensure the reliability of the dataset. The creation of a gold-standard annotated corpus reflects a high level of inter-annotator agreement on the identification of ASB. Inter-annotator agreement is vital in machine learning, as it enhances the accuracy and ease with which algorithms can process and learn from textual data [19]. The robustness of the dataset was further validated through the computation of a weighted Kappa score, using a Python library, which yielded a 97% agreement level. Consequently, we consolidated all classes indicative of ASB into a single category, contrasting them with the non-ASB class to form a binary distribution. This approach was strategically adopted to enhance the model's scalability and simplify both the problem definition and the interpretation of results.

3.2 Deep Learning Models

Text classification by a language model represents an endeavor to harness the capabilities of NLP for the automatic assignment of labels to textual instances. Traditional ML algorithms such as Support Vector Machines (SVM), Naive Bayes (NB), Logistic Regression (LR), and others have been employed to construct predictive models, often complemented by various feature engineering techniques to optimize performance [29]. However, the exponential growth of digital data has increasingly challenged the efficacy and accuracy of these conventional methods.

In response, DL models have emerged, capable of learning intricate representations or features across multiple layers from large datasets, thereby achieving

state-of-the-art results in text classification tasks [11]. The advent of transformer-based models within the DL sphere has further revolutionized NLP research, owing to their exceptional ability to capture long-range dependencies within text. The foundational architecture of the standard transformer model is built around several key components: the self-attention mechanism, encoder-decoder structure, positional encoding, multi-head attention, and feed-forward neural networks [26]. The architecture's self-attention and multi-head attention components are designed to effectively capture and process the input sequence of data, enabling the model to discern complex dependencies and contextual relationships among tokens. The operational principle of the self-attention mechanism is outlined as follows:

$$\text{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{d_k}\right)V, \tag{1}$$

where Q are queries, K are keys, V are values, and d_k is the dimensionality of key vectors. Below is the formula for the multiple attention mechanism:

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(head_1, \ldots, head_h)\mathbf{W}^O \tag{2}$$

$$head_i = Attention(\mathbf{QW}_i^Q, \mathbf{KW}_i^K, \mathbf{VW}_i^V),$$

where h is the number of heads, $Concat$ is a concatenation operator (.), and $\mathbf{QW}_i^Q$, $\mathbf{KW}_i^K$,$\mathbf{VW}_i^V$, $\mathbf{W}^O$ are parameter metrics.

In our study, we have employed three advanced transformer-based models for this downstream task, namely: Bidirectional Encoder Representations from Transformers (BERT) [9], Robustly optimized BERT Pretraining Approach (RoBERTa) [14], and Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) [8]. BERT, since its introduction in 2018, has set a benchmark in the NLP domain by showcasing how text can be effectively represented within its semantic context, capturing real-world information through its deep feature extraction capabilities. RoBERTa serves as an enhanced version of BERT, tailored for NLP-specific objectives with modifications aimed at improving the original model's efficiency. ELECTRA introduces an innovative pretraining approach. It leverages a sample generator network to identify and replace tokens, employing a discriminator model that efficiently predicts and reflects the semantic context of the data it represents.

Figure 1 illustrates a workflow diagram summarizing the key stages in our data pre-processing and feature extraction process, which facilitates the detection of ASB in tweet-based messages. This pipeline begins with data collection from the Twitter platform, followed by feature extraction using the transformer models, and culminates in the prediction of model performance based on critical evaluation metrics.

In this study, we adopted a train-validation-test split methodology for model training and evaluation, partitioning the annotated corpus into ratios of 70:20:10, respectively. We fine-tuned three prominent language models—BERT (bert-base-uncased), ELECTRA (electra-small-discriminator), and RoBERTa (roberta-base)—employing a consistent procedure that entailed initializing each model's

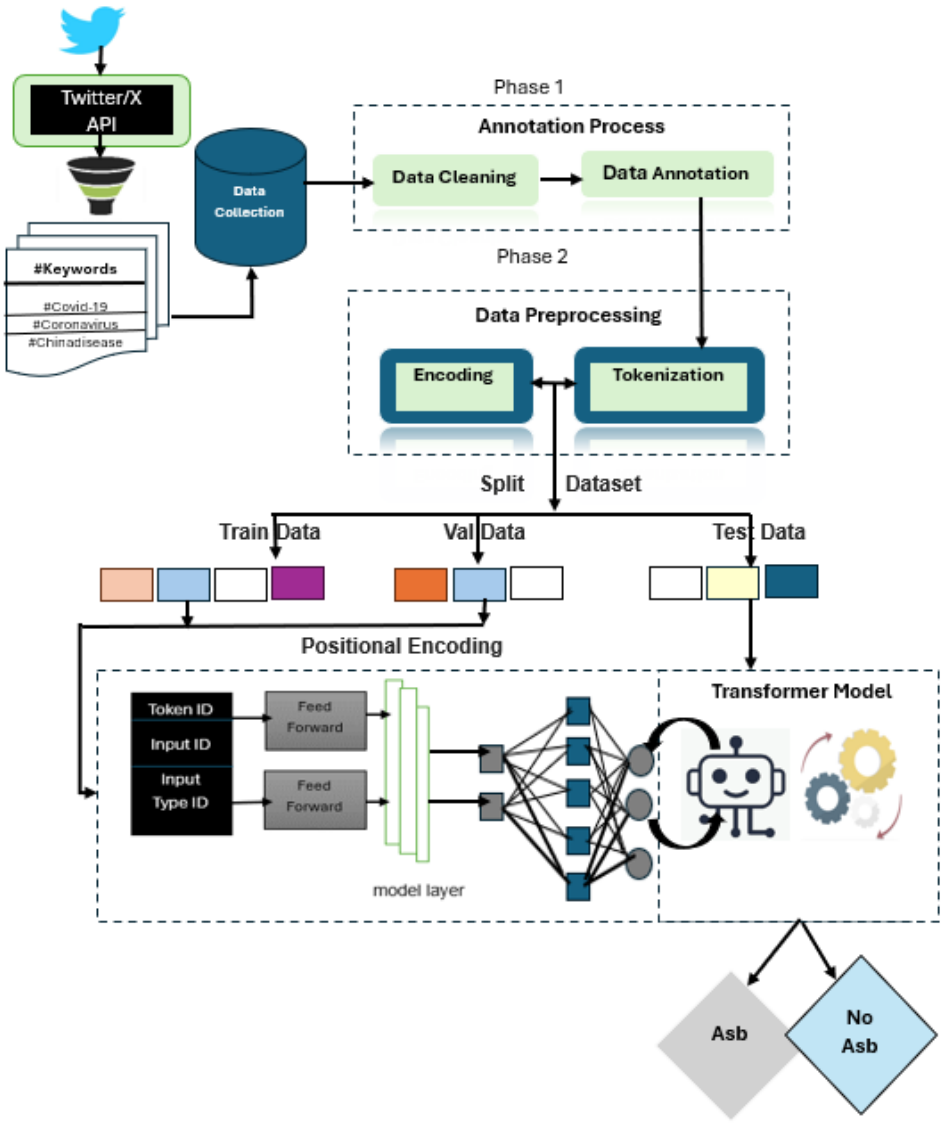

Fig. 1: The workflow for detecting ASB

specific tokenizer, fine-tuning on the training dataset, and assessing performance on the validation set. This process necessitated the careful adjustment of hyperparameters, including learning rate, batch size, and the number of training epochs, to attain optimal model performance. During the training phase, we utilized the validation dataset to fine-tune the hyperparameters, thereby mitigating the risk of model overfitting. Our findings indicated that a configuration of three training epochs, learning rate of 5e-5, and a batch size of 32 yielded the best results according to a comprehensive grid search. The test set, constituting 10% of the data and consisting of entirely unseen samples, was deployed to evaluate the final model performance.

For benchmarking purposes, we trained baseline models—LR, SVM, and NB—using the approach outlined in [2], with features vectorized through Term Frequency-Inverse Document Frequency (TF-IDF).

4 Results

This section delineates the evaluation methodology and performance metrics employed to assess the selected DL models' efficacy in identifying ASB within the COVID-19 pandemic corpus. To address the issue of class imbalance inherent in the dataset, we implemented the Synthetic Minority Oversampling Technique (SMOTE) as a remedial strategy. The evaluation framework encompassed a variety of metrics, including the standard accuracy score and supplementary metrics such as the F1 score, precision, and recall. Table 4 consolidates the performance outcomes of the three advanced DL classifiers, subsequent to the application of a train-test-validation split methodology for model training and evaluation.

Furthermore, the robustness and discriminative ability of each model were rigorously verified through the analysis of confusion matrices and Area Under the Curve (AUC) scores, providing a comprehensive overview of the models' predictive capabilities and reliability in the context of ASB detection during the pandemic.

Table 4: Results of the fine-tuned DL and baseline ML models

Model	**Acc. %**	**Prec. %**	**Recall %**	**F1-score %**	**AUC %**
NB	84.00	94.00	84.00	88.00	90.00
LR	88.00	94.00	88.00	91.00	88.00
SVM	87.00	94.00	87.00	90.00	85.00
ELECTRA	95.17	94.29	95.17	93.80	91.25
ROBERTA	**95.59**	94.88	**95.59**	**94.99**	**93.60**
BERT	95.00	**94.94**	95.00	94.96	88.57

Testing on an unseen dataset revealed that RoBERTa achieved the highest accuracy score at 95.59%, followed closely by ELECTRA with 95.17%, and BERT with a slightly lower score of 95.00%, though all scores surpassed the baseline

NB, LR, and SVM models. The analysis of the AUC values presented in Table 4 demonstrates that all classifiers performed well on both ASB classes. Efforts were made to minimize the impact of false negatives across all three proposed models, enhancing the models' suitability for future predictive tasks.

5 Conclusion

Social media platforms are designed to foster constructive interactions among users. However, there is an increasing trend of these platforms being exploited by individuals engaging in misconduct, often using offensive, hateful, or abusive language. This negative shift can be attributed to the inherent mobility and anonymity offered by these online environments [30]. Our study specifically investigated ASB in tweet-based discourse during the COVID-19 pandemic in Ghana, utilizing a rigorously annotated gold-standard corpus.

The analysis yielded two key findings: firstly, the proportion of tweets classified as exhibiting ASB, based on predefined categories, and secondly, the proportion deemed non-ASB. It was observed that the cloak of anonymity on social media emboldened some individuals to engage in various forms of social misconduct, including aggression, hostility, trolling, and a lack of empathy during Ghana's lockdown period. These behaviors were identified and classified in accordance with established literature within the ASB domain [18]. Further examination revealed distinct categories of ASB within the corpus, each with varying prevalence. Notably, trolling emerged as the most common form of ASB, reflecting the contentious nature of online discourse during the lockdown. While trolling is a relatively new phenomenon in Ghana's cyberspace, its impact is rapidly growing, mirroring its detrimental effects observed globally. The consequences of online ASB, regardless of intent, are invariably harmful, often leading to emotional distress, anxiety, depression, and, in extreme cases, suicide among the affected individuals [30].

This study employed three state-of-the-art transformer-based models to classify ASB within the corpus, achieving high accuracy and performance metrics. This suggests the presence of ASB in tweets during Ghana's partial lockdown, highlighting the social implications of the pandemic.

However, a limitation of this study is the narrow temporal scope of the dataset, confined to the three-week lockdown period from March 30 to April 20, 2020, presenting a potential problem of overfitting. Additionally, this study utilized a manual annotation process, which can be subject to annotator bias, and focused solely on a dataset from one platform, Twitter/X. Therefore, future research could expand on these findings by extending the dataset's timeframe and incorporating data from other social media platforms such as Facebook and Instagram. Furthermore, enriching the pre-processing phase with metadata such as user IDs, timestamps, and follower counts could offer deeper insights into ASB dynamics. Exploring other advanced models and fine-tuning hyperparameters may further enhance the detection and understanding of ASB in online environments.

Acknowledgments. This paper was supported by the grant No. SGS 2024 017 provided by the Faculty of Economics and Administration, University of Pardubice.

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Ahmed, M.T., Akter, N., Rahman, M., Das, D., AZM, T., Rashed, G.: Multimodal cyberbulling meme detection from social media using deep learning approach. International Journal of Computer Science and Information Technology (IJCSIT) **15**, 27–37 (2023)
2. Akhter, M.P., Jiangbin, Z., Naqvi, I.R., AbdelMajeed, M., Zia, T.: Abusive language detection from social media comments using conventional machine learning and deep learning approaches. Multimedia Systems **28**(6), 1925–1940 (2022)
3. Awal, M.R., Cao, R., Mitrovic, S., Lee, R.K.W.: On analyzing antisocial behaviors amid covid-19 pandemic. arXiv preprint arXiv:2007.10712 (2020)
4. Bharadiya, J.: A comprehensive survey of deep learning techniques natural language processing. European Journal of Technology **7**(1), 58–66 (2023)
5. Bhardwaj, M., Akhtar, M.S., Ekbal, A., Das, A., Chakraborty, T.: Hostility detection dataset in hindi. arXiv preprint arXiv:2011.03588 (2020)
6. Bohra, A., Vijay, D., Singh, V., Akhtar, S.S., Shrivastava, M.: A dataset of hindi-english code-mixed social media text for hate speech detection. In: Proceedings of the 2nd Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. pp. 36–41 (2018)
7. Bonetti, A., Martínez-Sober, M., Torres, J.C., Vega, J.M., Pellerin, S., Vila-Francés, J.: Comparison between machine learning and deep learning approaches for the detection of toxic comments on social networks. Applied Sciences **13**(10), 6038 (2023)
8. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Gibbon, S., Khalifa, N.R., Cheung, N.H., Völlm, B.A., McCarthy, L.: Psychological interventions for antisocial personality disorder. Cochrane Database of Systematic Reviews (9), CD007668 (2020)
11. Hajek, P., Barushka, A., Munk, M.: Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. Neural Computing and Applications **32**(23), 17259–17274 (2020)
12. Howard, R., Duggan, C.: Antisocial Personality: Theory, Research, Treatment. Cambridge University Press (2022)
13. Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T.: Aggression-annotated corpus of hindi-english code-mixed data. arXiv preprint arXiv:1803.09402 (2018)
14. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

15. Machová, K., Kolesár, D.: Recognition of antisocial behavior in online discussions. In: Information Systems Architecture and Technology: Proceedings of 40th Anniversary International Conference on Information Systems Architecture and Technology–ISAT 2019: Part II. pp. 253–262. Springer (2020)
16. Moor, L., Anderson, J.R.: A systematic literature review of the relationship between dark personality traits and antisocial online behaviours. Personality and Individual Differences **144**, 40–55 (2019)
17. Munezero, M., Montero, C.S., Kakkonen, T., Sutinen, E., Mozgovoy, M., Klyuev, V.: Automatic detection of antisocial behaviour in texts. Informatica **38**(1), 3–10 (2014)
18. Obaid, M.H., Guirguis, S.K., Elkaffas, S.M.: Cyberbullying detection and severity determination model. IEEE Access **11**, 97391–97399 (2023)
19. Parekh, M., Patel, Y.: Labeling news article's subject using uncertainty based active learning. In: International Summit Smart City 360°, pp. 200–208. Springer (2020)
20. Rizwan, H., Shakeel, M.H., Karim, A.: Hate-speech and offensive language detection in roman urdu. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2512–2522 (2020)
21. Salawu, S., He, Y., Lumsden, J.: Bullstop: A mobile app for cyberbullying prevention. In: Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations. pp. 70–74 (2020)
22. Saleh, H., Alhothali, A., Moria, K.: Detection of hate speech using bert and hate speech word embedding with deep model. Applied Artificial Intelligence **37**(1), 2166719 (2023)
23. Schmidt, C.W., Reddy, V., Zhang, H., Alameddine, A., Uzan, O., Pinter, Y., Tanner, C.: Tokenization is more than compression. arXiv preprint arXiv:2402.18376 (2024)
24. Singh, R., Subramani, S., Du, J., Zhang, Y., Wang, H., Miao, Y., Ahmed, K.: Antisocial behavior identification from twitter feeds using traditional machine learning algorithms and deep learning. EAI Endorsed Transactions on Scalable Information Systems **10**(4), 1–17 (2023)
25. Stephenson, J., Vaganay, M., Coon, D., Cameron, R., Hewitt, N.: The role of facebook and twitter as organisational communication platforms in relation to flood events in northern ireland. Journal of Flood Risk Management **11**(3), 339–350 (2018)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I.: Attention is all you need. arxiv [cs. cl]. 2017 (2023)
27. Wiegand, M., Ruppenhofer, J.: Exploiting emojis for abusive language detection. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 369–380 (2021)
28. Wiegand, M., Ruppenhofer, J., Schmidt, A., Greenberg, C.: Inducing a lexicon of abusive words–a feature-based approach. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1046–1056 (2018)
29. Yilmaz, S.F., Kaynak, E.B., Koç, A., Dibeklioğlu, H., Kozat, S.S.: Multi-label sentiment analysis on 100 languages with dynamic weighting for label imbalance. IEEE Transactions on Neural Networks and Learning Systems **34**(1), 331–343 (2021)
30. Zinovyeva, E., Härdle, W.K., Lessmann, S.: Antisocial online behavior detection using deep learning. Decision Support Systems **138**, 113362 (2020)