Exploring usability in combining data sources for detecting toxic behavior in social media

Oliver Lohaj, Ján Paralič
Technical university of Košice
Faculty of Electrical Engineering and Informatics
Department of Cybernetics and Artificial Intelligence
Vysokoškolská 4, 040 01 Košice, Slovakia
{oliver.lohaj, jan.paralic}@tuke.sk

Ruslana Buinytska
Technical university of Košice
Faculty of Electrical Engineering and Informatics
Department of Cybernetics and Artificial Intelligence
Vysokoškolská 4, 040 01 Košice, Slovakia
ruslana.buinytska@student.tuke.sk

Abstract. *This article deals with the issue of detecting toxic behavior on social networks. The main goal of the article is to find an effective approach to detecting toxic behavior using a selected machine learning method with a combination of multiple data sources. The article is divided into a theoretical and a practical part. The theoretical part provides an overview of toxic behavior in social media, existing research and various machine learning models such as CNN (Convolutional Neural Network) and BERT (Bidirectional Representation Coding from Transformers). The practical part follows the CRISP-DM methodology and focuses on data preparation and modeling. Several experiments were conducted to evaluate the performance of different models and configurations. The results showed that combining multiple data sources and advanced machine learning models can improve the accuracy of toxic behavior detection. The article concludes that the proposed approach is effective and stresses the importance of proper data preparation and comprehensive evaluation metrics.*

Keywords. BERT, CNN, social media, toxic behavior, usability

1 Introduction

The use of social media has increased a lot in the last decade (Perrin, 2021; Chou, 2009), which is directly related to the increase in the amount of time people spend online. This trend is most pronounced among young people, who spent up to twice as much time online in 2020 compared to the previous decade, averaging two to three and a half hours a day (Lei, 2024). This trend is significant because it shows that young people are increasingly oriented towards their online lives. They use social media to communicate with friends and family, get information (Mertz, 2024), but they are also exposed to negative influences such as bullying or misinformation (Parent, 2019), which can affect their view of the world. Toxic behaviour in the online environment can take many forms and often manifests itself in the form of insults, hateful comments, the spread of disinformation or cyberbullying (Beknazar-Yuzbashev, 2022). These manifestations can have a negative impact on the psychological health of individuals, can cause emotional stress and even lead to long-term psychological problems (Zsila, 2022).

In addition, toxic behavior can also lead to the division of the community, exacerbating online discussions, and creating a hostile environment on the Internet. And with each passing year, the situation in the online environment may deteriorate (Avalle, 2024). The aim of this research is to investigate how the combination of multiple data sources affects the accuracy of detection of toxic behavior in social media and to create a model using this data to increase detection efficiency.

To solve this task, we proceeded using the CRISP-DM methodology (Wirth and Hipp, 2000) and worked in Google Colab with the Python programming language. We first modified the data and then created models, which we evaluated at the end by monitoring several criteria.

2 Analysis of the current state and motivation

Before starting our own experiments, we analyzed what experiments had been done before, what models and datasets had been used, and what results had been achieved.

In an interesting study (Fan, 2021), the authors analyzed the detection of toxic behavior on social networks using deep learning. For this purpose, they chose the BERT model, specifically the BERT-base version, which they trained on the labelled dataset Kaggle Toxic Comment Classification Challenge[1]). The authors of this paper compared the performance of the BERT model with three other models, namely

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

Multilingual BERT, RoBERTa and DistilBERT. To assess the performance of the classification model, they used the AUC-ROC as a performance metric. This metric is the standard ranking method used in competitions on the Kaggle platform.

Before the model was created, preprocessing operations were used, such as removing punctuation, references, and non-English words. A pre-trained tokenizer for "bert_base_uncased" was used for tokenization. The evaluation showed that the BERT-based model performs very well (0.98561 (AUC-ROC)) as a public score and 0.98603 (AUC-ROC) as a private score) in detecting toxic content. The multilingual BERT model achieved the second-best position, followed by the DistilBERT and RoBERTa models.

In other article (Anand and Enswari, 2019) simple neural networks, CNNs and LSTMs are used with or without pre-trained GloVe models. The Jigsaw dataset was used to train the model (Sorensen, 2017). The metrics used to evaluate the performance of each model are classification success (accuracy), which describes the percentage of samples correctly placed in their actual class, and the value of the loss. The results show that ANN achieved a high classification success rate (98%) on trained data with minimal loss during training.

However, it achieved a high loss value on the test data and was the third worst in terms of classification success among all models. CNN achieved a relatively high success rate of classification in training (97.8%) and a low loss value (5.42%). It has a lower loss on test data than ANN, but also a lower classification success rate. LSTM achieved a lower classification success rate and a higher loss value during training compared to ANN and CNN. However, it achieved a better classification success rate and a lower loss value than previous models on test data.

The GloVe & CNN model had a lower classification success rate during training, so its loss value was higher than previous models. It performed better than other models on test data, although its loss value was similar to that of the LSTM. GloVe & LSTM had a lower classification success rate during training and its loss value was higher than that of ANN and CNN. On test data, it achieved significant results in terms of classification success and loss rate. GloVe & LSTM & CNN achieved the lowest classification success rate and the highest loss value during both training and testing. This model achieved the worst results of all.

These analyses of existing research show that the CNN and BERT models are among the best approaches for detecting toxic comments on social media (Lee, 2020). Deep-learning-based models such as CNN can capture complex patterns and structures in data, allowing for successful classification of toxic comments (Georgakopoulos, 2018).

Transformer-based models, such as BERT, on the other hand, can better understand the context of text and capture its meaning, which also leads to excellent results in detecting toxic content (Ashwin Geet d'Sa, 2020).

Based on these findings, we decided to train CNN and BERT models to detect toxic comments on social media and find out the usability of these models. These models offer a combination of advanced word processing techniques and the ability to capture complex patterns in data, which should lead to a high success rate in detecting toxic content. The goal of our work from a business perspective is to find a portable model that will be able to detect toxic behavior in text data. In the future, it could be used to create a safe and positive environment for users on social media. In terms of data mining, we expect the creation of a classification model into two classes. Data analysis makes it possible to identify relevant patterns and trends in user behavior and provides the basis for creating effective and accurate models for detecting toxic behavior on social media.

3 Datasets

As the title of the article suggests, we chose to combine multiple data sources for training our models. Our approach involved training the model on one dataset and evaluating it on a second dataset. This method allows us to assess whether a model trained on one dataset can generalize knowledge sufficiently to identify toxic content across different social networks.

Our research was based on pre-existing datasets. The first dataset is the Twitter dataset by Davidson. Tweets are collected based on keywords from the hatebase.org lexicon. The dataset contains 24783 tweets and annotations made by CrowdFlower. It is important to mention here, that the app named Twitter has been rebranded by new owner to X in the April of 2023 (Hayes, 2024).

Each tweet has been annotated by at least 3 annotators, and the consensus of annotators is 92%. The labels correspond to three classes: hate speech, offensive language and none, with the percentages of each class being 6%, 77% and 17% respectively.

Before reviewing the dataset, we reduced the number of classes to two: toxic marked as 1 and non-toxic as 0. When examining the dataset, we found that the average length of toxic comments is smaller than the average length of non-toxic comments, see. Fig. 1 and the following words are most often used:

b**ch – 11480, h*e – 4352, like – 2873, f**k – 2269, p**si – 2267, n**ga – 2019, get – 1785, a*s – 1599, s**t – 1312, u – 1307.

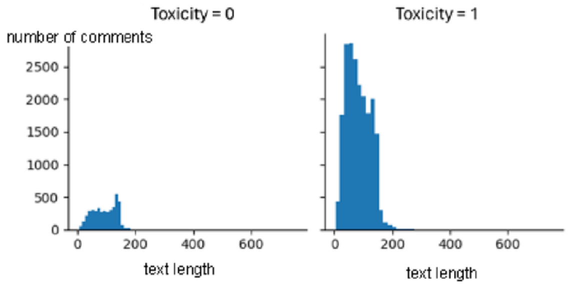

Figure 1. Comments length of different classes on Twitter dataset by Davidson

The second dataset is the Toxic Tweets Dataset, which contains tweet data from different sources that showed a significant imbalance in the number of examples between classes. This dataset is the result of a combination of various existing datasets that have been obtained from publicly available sources. The goal of combining these datasets was to achieve a balance between classes, which could lead to better results when training classification models.

The dataset contains 56745 tweets. The labels correspond to two classes: toxic behavior marked as 1 and non-toxic as 0, with the percentages of each class being 57% and 43% respectively. Regarding the length of comments of different classes, the result is the same as in the previous dataset: the average length of toxic comments is less than the average length of non-toxic comments, see Fig. 2. It also turns out that the words that are most often found in toxic messages are similar to the first set of data: b**ch – 11553, h*e – 4370, like – 4122, love – 4020, day – 3381, get – 3046, u – 2501, f*ck – 2480, p**si – 2318.

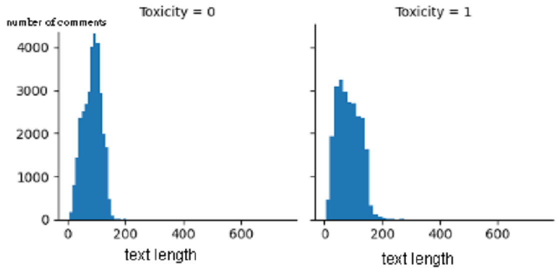

Figure 2. Comments length of different classes on Toxic Tweets Dataset

The third dataset is a dataset from Wikipedia's Talk edit pages. Wikipedia is the largest and most popular reference resource on the internet, with about 500 million unique visitors per month. These editing pages are a crucial forum for community interaction, where the contributing community discusses, debates, and communicates about changes related to a specific topic.

Overall, this dataset provides a data source for research and development of NLP models and community management tools on the online platform Wikipedia. The dataset contains 5000 messages. The labels correspond to two classes: toxic and non-toxic behavior, with the percentages of each class being 9% and 91% respectively. The graph (see Fig. 3) shows that the average length of toxic comments is less than the average length of non-toxic comments. We can see that, unlike datasets from the social network X (Twitter), the keywords in toxic comments are different: articl – 2232, page – 1874, wikipedia – 1455, edit – 1378, talk – 1322, f**ck – 1198, use - 1131, please – 1046, a*s – 990, would – 966.

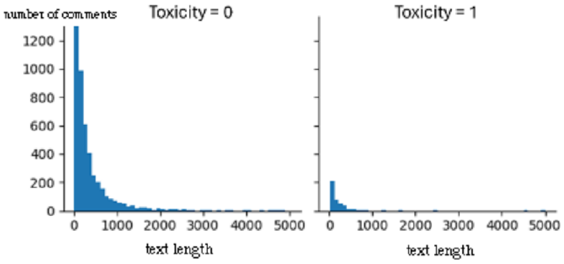

Figure 3. Comments length of different classes of Wikipedia toxicity dataset

4 Modelling

In the modelling phase, we focused on several types of models: BERT and CNN. We used two metrics to evaluate the performance of the models: accuracy and false negative rate. Accuracy provides us with an overall level of classification correctness, while false negative rates help us identify the risk of misclassification of toxic comments.

The sequence of our experiments is designed in such a way that almost every experiment depends on the results of the previous experiment. Based on each experiment we decided how to design further procedures to consider various aspects that can affect the success of detection.

The 1st experiment was designed to provide a basic view of the CNN model's performance in detecting toxic comments without the use of under sampling and using specific model parameters. As part of the experiment, we used the CNN model. We divided the Twitter dataset by Davidson into a training and test set in an 80:20 ratio without the use of under sampling. Before training, we performed tokenization and vectorization of the text data. We trained the model over 10 epochs. After training, we evaluated the performance of the model on a test set. We achieved a classification success rate: 0.934 false negative rate: 0.026.

For the 2nd experiment, we used the same parameters as for the first model, except that we used Twitter dataset by Davidson data with under sampling for training set. Despite the fact that we modified the training set, the results of the second model were worse than those of the first model. We achieved a lower classification success rate of 0.929 and a higher false negative rate of 0.068.

We have several explanations as to why this procedure may have negatively affected the results. First, CNN works well with local patterns and dependencies in data, but they can miss global patterns, especially if certain data classes are underrepresented. When the number of training examples is limited, this can have a negative impact on the CNN's ability to identify and generalize these patterns to a wider range of data.

Furthermore, in the case of significant under sampling, the model may be overtrained for limited patterns and specific interference elements in the training data. This phenomenon leads to a reduced ability of the model to generalize to new ones, as the model may tend to reproduce errors learned from an incomplete training set. In addition, CNNs require a relatively large amount of data to function effectively due to their depth and large number of parameters. Data reduction causes a reduction in the variability and complexity of information available for learning, which can limit the model's ability to learn and adapt to different situations or diverse data.

For the 3rd experiment, we used again the same parameters as for the first model. However, unlike the first two experiments, we used the Twitter dataset by Davidson, which was preprocessed using a tweet preprocessing script written in the Ruby programming language, which was interpreted in Python (Davidson, 2017). Nevertheless, the results of the third model were worse compared to the first model. We achieved a lower classification success rate of 0.929 and a higher false negative rate of 0.028.

In the 4th experiment, we decided to test how well the BERT model can cope with the classification of toxic comments. We took advantage of BERT's built-in tokenization feature, which is optimized for this model. For training and testing, we still use the Twitter dataset by Davidson, split 80:20 in favor of the training set. Initially, we initialized the model and tokenizer, using the "*bert-base-uncased*" model, which is one of the most common BERT models. We have specified that the number of output classes is 2, which corresponds to our binary classification task. Next, we tokenized the texts and prepared the data for training and testing. For training, we used *DataLoader* to load data in the form of batches, which we then used to update the model weights during training. To minimize losses, we used the *AdamW* optimizer with a learning coefficient $5x10^{-5}$.

The training cycle itself was implemented for 3 epochs. In each epoch, we iterated over the training data and performed back-propagation of the error and update the model weights according to the gradient. After completing the training, we tested and evaluated the model. The success rate of the model classification reached a value of 0.955, indicating that more than 95% of the cases were classified correctly. The false negative rate was 0.021, indicating that only about 2% of toxic comments were misclassified as harmless.

In the following experiments, we focused on testing the robustness of models on other datasets, which was the main goal. We wanted to see, if model trained on one kind of data is able to classify toxic comments from another platform. If yes, that would mean that the trained models are universal for wider range of use, than just using it on the same data as it was trained on.

During 5th experiment, we decided to use the model we successfully tested in the first experiment (CNN) and applied it to other Toxic Tweets Dataset. The results of this experiment gave us an important insight into how our model behaves on different data. Although we achieved a relatively good classification success rate on the different dataset (0.778), it was important to note that the false negative rate was slightly higher (0.103) compared to previous experiments. During the *"data preparation"* phase, we found that, there is a small difference, between the frequently used words in toxic comments, which may have prevented the comments from being classified correctly.

In the 6th experiment, we used the BERT model that we trained in the 4th experiment. As in the previous experiment with the CNN model, the aim of the experiment was to test the model on the Toxic Tweets Dataset, i.e. different from the one used during the training. In the previous experiment, we described how the difference between the data affected the success of the classification, now let's look at the differences at the model level. By comparing these results with the previous CNN experiment, we found that the BERT model achieved a higher classification success rate (0.882 vs. 0.778). This clearly shows that the BERT model is able to identify toxic texts better. However, the BERT model also achieved a slightly higher false negative rate (0.106 versus 0.103), which means that more toxic texts were misclassified as harmless. Although we are aware of this imperfection, the results clearly show that the BERT model surpassed the CNN model in the success rate of classifying toxic texts. This result shows that the BERT model is a better choice for toxicity detection compared to the CNN model. In this case, the difference in the false negative rate between the two models is relatively small. In case of a larger difference, we would have to consider other factors and perhaps we would prefer the CNN model, although its classification success rate is not as high as that of the BERT model.

In the following 7th experiment, we decided to use a BERT-based model based on the results of previous experiments. Since the Toxic Tweets Dataset contains more toxic comments, we decided to use it as a training set. We chose the Wikipedia toxicity dataset for testing. We left the parameters of the model unchanged, as we observed good results with these parameters in previous experiments. Our results show that we have achieved a classification success rate of 0.869. We also recorded a false negative rate of 0.478. Since the training dataset contained more than 55,000 comments, there is a chance that the BERT model may have been overtrained. Another reason could be that the toxic comments in Wikipedia contain different keywords than in the Toxic Tweets Dataset, which is why the model had trouble identifying new toxic patterns. In addition, differences in data annotations may have affected the results. These differences may be due to subjective annotator evaluations, inconsistent labeling standards and variability in the training process, but it showed us that training on one set and testing on other achieves pretty good results and the model can be deemed usable.

In the 8th experiment, we decided to verify our assumptions about overlearning the model from the 7th experiment. To do this, we used a Twitter dataset by Davidson containing fewer messages, with the assumption that the model would be able to better generalize dependencies across a smaller dataset. We achieved a high classification success rate of 0.947 and a lower false negative rate of 0.355. In this way, our assumptions were confirmed, and this experiment gave us valuable insight into the behavior of the model depending on the size of the training data.

We decided to create the last experiment mirroring the 8th experiment, where we used a dataset from Wikipedia for training and a Twitter dataset by Davidson for testing. This strategy is useful because in real conditions we can encounter text from different sources and different writing styles. Using multiple source datasets and testing the model on texts from different domains can provide a comprehensive view of its overall performance and ability to identify toxicity, regardless of the specific source of the text.

The results of this experiment showed that the model achieved a classification success rate of 0.822 and a false negative rate of 0.16. These results, although lower compared to previous experiments, still show the model's ability to work on new and diverse data and can be deemed usable. We have to consider that the training dataset was relatively small compared to the test dataset (5000 comments out of only 437 toxic), but at the same time it identified 17316 toxic comments, which indicates that the dataset contains a selection of high-quality and representative samples. However, the high false negative rate suggests that the model still has trouble identifying some of the toxic texts in the Twitter dataset by Davidson. This result highlights the importance of using multiple source datasets and continuously improving the model to improve its ability to identify toxicity in different contexts and text sources.

6 Evaluation and comparison

Based on the experiments carried out and the data analyzed, we can conclude that modelling toxic behavior in social media is a complex task that requires fine-tuning and understanding of the details that affect the performance of the models. In this research, we used CNN and the BERT model, which are cutting-edge methods in the field of NLP, to detect toxic behavior in datasets originating from Twitter and Wikipedia.

The results of the experiments, provided in Table 1 show that: The CNN model achieved a high level of classification success without the need for subsampling, indicating that it is capable of processing unevenly distributed data. The BERT model has shown a better classification success rate in detecting toxic comments, which confirms its robustness and adaptability to different linguistic contexts in the data.

However, in one case, it showed a higher rate of false negativity, indicating the need for further tuning. Data preprocessing was critical to model performance, with modifications such as removing special characters and transforming text data impacting the models' ability to identify toxic comments. Overfitting models is a risk to consider, especially when the model shows high accuracy on training data, but subsequently has problems with generalization on test data.

Compared to other previously published papers, this article advances the field of detecting toxic behavior on social media in several distinctive ways. Previous studies have primarily focused on evaluating the effectiveness of individual machine learning models like BERT or CNN using single data sources. For example, studies by Fan (2021) and Anand and Enswari (2019) assessed models like BERT, RoBERTa, DistilBERT, CNNs, and LSTMs individually on specific datasets, often achieving high performance metrics but within limited contexts.

This article, however, stands out by emphasizing the combination of multiple data sources for training and testing models, which significantly enhances the generalizability and robustness of the detection systems. By training models on one dataset and evaluating them on another, the research demonstrates a more comprehensive approach to understanding model performance across different social media platforms, thus addressing a key limitation in earlier studies.

Moreover, while earlier works often used standard datasets like the Kaggle Toxic Comment Classification Challenge and Jigsaw datasets, this article introduces and experiments with diverse datasets including Twitter datasets by Davidson, Toxic Tweets Dataset, and Wikipedia's Talk edit pages. This diversity in data sources provides a broader perspective on the models' capabilities and limitations, offering more generalizable and practical insights into their effectiveness and moreover, their usability.

Furthermore, the methodological rigor of following the CRISP-DM methodology ensures a structured and replicable approach to data preparation, modeling, and evaluation. This contrasts with some previously mentioned studies that might not have provided as detailed a methodological framework. The article's use of comprehensive evaluation metrics, including accuracy and false negative rate, adds depth to the analysis, offering a more nuanced understanding of model performance compared to simpler metrics like accuracy alone used in some earlier research.

This article not only confirms the effectiveness of advanced models like BERT and CNN but also highlights the critical importance of combining multiple data sources and rigorous data preparation. It provides a more robust framework for detecting toxic behavior on social media, pushing the boundaries of current research and offering practical solutions for real-world applications.

Table 1. Comparison of experiment results

Experiment number	Model	Dataset for training	Dataset for testing	Classification success rate	False negative rate
1	CNN	Twitter by Davidson (unsampled, self-processed)	Twitter by Davidson	0.934	0.026
2	CNN	Twitter by Davidson (under sampling, self-processed)	Twitter by Davidson	0.929	0.068
3	CNN	Twitter by Davidson (no subsampling, preprocessing: Ruby script)	Twitter by Davidson	0.929	0.028
4	BERT	Twitter by Davidson	Twitter by Davidson	0.955	0.021
5	CNN	Twitter by Davidson (unsampled, self-processed)	Toxic Tweets Dataset	0.778	0.103
6	BERT	Twitter by Davidson	Toxic Tweets Dataset	0.882	0.106
7	BERT	Toxic Tweets Dataset	Wikipedia dataset	0.869	0.478
8	BERT	Twitter by Davidson (100%)	Wikipedia dataset	0.947	0.355
9	BERT	Wikipedia dataset	Twitter by Davidson (100%)	0.822	0.16

7 Conclusion

Based on our experiments, we can conclude that combining multiple sources can be very effective to find a model that would work best, as it includes aspects such as model generalization, which shows how the model has learned to recognize toxic behavior in different contexts and situations, and whether it can work effectively with new, unknown data. Also, the combination of training and testing on different data points to the robustness of the model, whether the model can work with different communication styles, language variations and forms of toxic behavior, which makes it a more reliable tool for detection in different contexts. The use of a combination of multiple data sources achieves better efficiency of the model in detecting toxic behavior in different contexts and situations, increasing its practical applicability and reliability.

But before using testing datasets different from the training dataset, a sufficient level of classification success and a false negative rate on the underlying data must be achieved. We have found that this is affected by data preprocessing, not every preprocessing technique may be suitable for the selected dataset, so it is better to test several variants. Other important thing to consider is to not always considering emojis, hyperlinks and similar details. This leads to the lower efficiency of preprocessing, and sometimes it can only create unnecessary noise.

In addition, we realized that a balanced dataset is not always the best solution, during subsampling we can miss valuable data based on which the model could learn to better classify comments. It is also very important to choose the right model that is suitable for the specific task, in our experiments we used CNN and BERT and found that BERT captures context and global dependencies better. An equally important factor is what metrics will be used to evaluate the effectiveness of the model. For toxic behavior detection models, it is important to use other metrics such as false negative rates in addition to classification success.

Despite its strengths, this research has several limitations that highlight areas for improvement. The small training dataset compared to the test dataset suggests the potential for enhanced performance with larger data. While the model's high false negative rate points to areas needing refinement, it also shows capability in identifying much toxic content. The risk of overfitting calls for improved training methods to boost generalization. Data preprocessing impacts performance, indicating the need for varied techniques. Annotation subjectivity and labeling inconsistencies suggest a need for standardized approaches. Using CNN and BERT models offers insights into their strengths, with BERT excelling in context and dependency capture. Variations in model effectiveness across datasets underscore the importance of considering linguistic and contextual diversity to improve robustness and applicability.

This work provides a contribution to understanding the dynamics of toxic behavior detection in social media using multiple data sources and modelling. The combination of different approaches to data pre-processing and the application of different models can lead to improved detection of antisocial behavior. The experiments showed the importance of the adaptability of models to new datasets and their ability to generalize learned patterns of behavior. We also found out that BERT and CNN models enhance usability in detecting toxic behavior in social media. BERT effectively captures language context, while CNN identifies patterns efficiently. Together, they improve accuracy and adaptability across various datasets and can be deemed of high usability in future research.

In the context of social media, where it is important to quickly and accurately identify toxic comments, our findings show that finding the optimal model and proper data pre-processing is crucial. For practical applications, it is important to balance detection accuracy and the model's ability to adapt to constantly changing behavior patterns in the online space.

Acknowledgments

This research was funded by the Slovak Research and Development Agency under the contracts No. APVV-22-0414 and APVV-20-0232 and by the Scientific Grant Agency of the Ministry of Education, Research, Development and Youth of the Slovak Republic and the Slovak Academy of Sciences under grants No. 1/0685/21 and 1/0259/24.

References

Anand, Mukul, and R. Eswari. "Classification of abusive comments in social media using deep learning." *2019 3rd international conference on computing methodologies and communication (ICCMC)*. IEEE, 2019.

Avalle, Michele, et al. "Persistent interaction patterns across social media platforms and over time." *Nature* 628.8008 (2024): 582-589.

Beknazar-Yuzbashev, George, et al. "Toxic content and user engagement on social media: Evidence from a field experiment." *Available at SSRN 4307346* (2022).

Chou, Wen-ying Sylvia, et al. "Social media use in the United States: implications for health communication." *Journal of medical Internet research* 11.4 (2009): e1249.

d'Sa, Ashwin Geet, Irina Illina, and Dominique Fohr. "Bert and fasttext embeddings for automatic detection of toxic speech." *2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies"(OCTA)*. IEEE, 2020.

Hayes, David. C. 2024. "X." The editors of *Encyclopaedia Britannica. Available* online: https://www.britannica.com/money/Twitter

Davidson, T., Warmsley, D., Macy, M., & Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM '17)* (pp. 512-515). Montreal, Canada. GitHub repository, https://github.com/t-davidson/hate-speech-and-offensive-language

Fan, Hong, et al. "Social media toxicity classification using deep learning: real-world application UK brexit." *Electronics* 10.11 (2021): 1332.

Georgakopoulos, Spiros V., et al. "Convolutional neural networks for toxic comment classification." *Proceedings of the 10th hellenic conference on artificial intelligence*. 2018.

Lee, Jieh-Sheng, and Jieh Hsiang. "Patent classification by fine-tuning BERT language model." World Patent Information 61 (2020): 101965.

Lei, Xiaojing, et al. "The relationship between social media use and psychosocial outcomes in older adults: a systematic review." *International Psychogeriatrics* (2024): 1-33.

Mertz, Breanne A., et al. "# SocialMediaWellness: Exploring a research agenda and conceptualization for healthy social media consumption." *Journal of Consumer Behaviour* 23.2 (2024): 321-335.

Parent, Mike C., Teresa D. Gobble, and Aaron Rochlen. "Social media behavior, toxic masculinity, and depression." *Psychology of Men & Masculinities* 20.3 (2019): 277.

Perrin, Andrew. "Social media usage." *Pew research center* 125 (2015): 52-68.

Sorensen, C. J., et al. *Toxic Comment Classification Challenge*. Kaggle, 2017. Available online: https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge

Wirth, Rüdiger, and Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining." *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Vol. 1. 2000.

Zsila, Ágnes, et al. "Toxic behaviors in online multiplayer games: Prevalence, perception, risk factors of victimization, and psychological consequences." *Aggressive Behavior* 48.3 (2022): 356-364.