Journal of Research High School

Vol. 2024 (5) 194 – 199
ISSN 2688-3651 [online]

A Novel Approach to Detect Cyberbullying on Instagram by Using Naïve Bayes Classifier

Yuanyuan Ding[1] *

[1]Sage Hill School, Newport Coast, CA, USA
*Corresponding Author: dingyuanyuan0228@gmail.com

Advisor: Dr. Mahnaz Roshanaei, mroshana@stanford.edu

Received January 8, 2024; Revised March 4, 2023; Accepted, March 18, 2024

Abstract

Cyberbullying on social media platforms recently became a serious social issue with significant psychological impacts on individuals. This study focused on cyberbullying detection on multiple social media platforms. A related dataset was acquired from Hosseinmardi et al., and the research was conducted based on this dataset. The methodology combined both deep learning and traditional machine learning models. Experiments were conducted with various machine learning algorithms, including logistic regression, random forest, support vector machine, and Naive Bayes classifier. The DistilBERT transformer model was tested alongside these machine learning models. In the end, the Naive Bayes classifier outperformed the other models, including the transformer model, with an accuracy of 95.83%. The results indicated that while complex deep learning models like DistilBERT were often superior for natural language processing tasks, probabilistic models like the Naive Bayes classifier could sometimes yield better results. The insights from this study contributed to the broader understanding of cyberbullying detection and paved the way for future research to integrate multimodal data, explore real-time detection systems, and more.

Keywords: Cyberbullying Detection, Naive Bayes, Deep Learning, Natural Language Processing, Transformer Model, DistilBERT

1. Introduction

In the digital era, social media platforms became an essential part of people's communications and interactions. While many people shared their lives and engaged with friends in this digital community, numerous negative behaviors, such as cyberbullying, coexisted. As a new form of bullying, cyberbullying carried all the consequences that traditional bullying had, and its widespread and anonymous nature made it a concerning issue in society. Cyberbullying became a crucial issue that could lead to severe psychological consequences for victims, affect the physical and mental health of the younger generations, and cause a negative online atmosphere, among others. However, traditional methods of cyberbullying detection often relied on human oversight, which was unable to meet the extensive amounts of texts and posts on social media. Therefore, there was a pressing need to devise an efficient method for detecting cyberbullying on such platforms.

Machine learning and natural language processing, recently developed technologies, could be applied to solve this issue. Transformer models, such as DistilBERT, had shown great ability to understand the nuances of human languages and the capacity to perform text classification tasks. However, such models often required a substantial amount of labeled data and computational resources, which might not always have been available.

Previous researches had targeted on cyberbullying detection on Twitter and Instagram through machine learning models, such as the Support Vector Machines and linear regression model. However, due to the imbalanced nature of the cyberbullying datasets, in which non-cyberbullying cases outnumbered those cyberbullying ones, further research

Vol. 2024 (5) 194 – 199
ISSN 2688-3651 [online]

was needed to solve the problem and find a better model that best fitted the research objective. At the same time, previous researches did not use deep learning models, such as the transformers model, to perform the cyberbullying detection tasks, but it was necessary to employ such sophisticated language model to perform the task and analyze its results.

Understanding the importance of the problem and addressing these challenges, this paper presented a comprehensive approach to detect cyberbullying on social media platforms. After being tested to perform this text classification task, the results obtained by those models were compared. Initially, it was expected that the transformer DistilBERT model would outperform all other machine learning models and return the highest. However, a surprising result was shown. Contrary to expectations, the findings of this research revealed that a more simplistic model, the Naive Bayes classifier, exceeded the performance of the sophisticated DistilBERT transformer in this specific scenario. This finding challenged prevailing norms in the field and underscored the importance of context and data specificity in machine learning applications.

This research focused on using transformers model to detect cyberbullying on social media, finding a way to solve the data imbalance nature, and discovering the most efficient model to detect cyberbullying. The research for the first time employed the DistilBERT model in the cyberbullying detection field and compared its results with machine learning models including the Naïve Bayes Classifier, and the results of the research indicated a surprising result that a simpler model, Naïve Bayes Classifier, could outperform the more sophisticated transformers model and handle the data imbalance better.

2. Literature Review

Machine learning was employed in cyberbullying detection tasks. Xu et al. employed Support Vector Machines (SVM) with linguistic features to detect cyberbullying content on Twitter, demonstrating promising results (Xu et al., 2012). Reynolds et al. also utilized SVM and incorporated sentiment analysis that yielded a true positive of 78.5% accuracy (Reynolds et al., 2011).

Transformer models revolutionized natural language processing, providing new and effective ways in various texts and language-related tasks. Delvin et al. presented BERT, a model that used transformers for creating contextual embeddings, significantly improved performance across a multitude of NLP tasks (Delvin et al., 2019). In the same year, Sanh et al. proposed a lighter version of BERT that aimed to maintain the performance of BERT while reducing the model size and computational requirements (Sanh et al., 2019).

However, cyberbullying detection was still a challenging task. Rosa et al. highlighted the issues of data imbalance in cyberbullying datasets that non-cyberbullying instances usually outnumbered those cyberbullying instances (Rosa et al., 2019). Van Hee et al. researched the complexity of defining and annotating cyberbullying and emphasized the need for clear guidelines and a nuanced understanding of the issue (Van Hee et al., 2018).

Specifically, previous research also focused on cyberbullying on Instagram. Hosseinmardi et al. analyzed user interactions and comments on Instagram to understand the prevalence of cyberbullying (Hosseinmardi et al., 2015). They also addressed the difference between cyberbullying and cyber aggression in their research. They scrapped raw data directly from Instagram and employed manual labeling techniques to label their data. They also kindly provided us with their datasets for subsequent research.

3. Method

3.1 Data Acquisition and Preprocessing

The dataset that contained an extensive collection of raw Instagram post data was acquired from Hosseinmardi et al. (Hosseinmardi et al. 2015). The dataset included information such as comments posted by the owner, reptiles to these comments, timestamps indicating when each comment was made, and the contents of the posts themselves (Hosseinmardi et al. 2015).

The dataset under examination was annotated with indicators of cyberbullying and cyberaggression. In alignment with the research objectives of identifying cyberbullying instances on Instagram, the dataset was restructured. This

Vol. 2024 (5) 194 – 199
ISSN 2688-3651 [online]

restructuring process entailed the selective extraction of replies and their associated cyberbullying labels. For each Instagram post, all corresponding replies were concatenated into a single sample to create a data point. The cyberbullying status of a sample was determined by analyzing the five cyberbullying labels assigned to each reply. A sample was designated as indicative of cyberbullying if it contained more than two labels suggesting such behavior. Concurrently, text features were cleaned and preprocessed, involving the removal of unnecessary information, including font details, timestamps, and common stop words.

Following the refinement and reorganization of the dataset originally compiled by Hosseinmardi et al., it was integrated as the foundation for the supervised learning of the study. Concurrently, the Sentiment140 dataset, established by Go et al. (2009), was employed for the unsupervised learning aspect of the research. A subset of two hundred data points was randomly chosen from this latter dataset. Only the textual content of these data points was extracted, and labels were randomly assigned to each. Subsequently, the data points designated for supervised and unsupervised learning were merged into a single, expansive dataset object.

The final dataset contained 478 datapoints, including 278 human-labeled datapoints. This dataset was novel because it contained both datapoints from Instagram and Twitter, allowing models to perform cyberbullying detection on multiple social media platforms. However, due to the limited number of human-labeled datapoints, the performances of the transformers model could be limited since it may have not yet reached the enough volume for the transformers model to yield its best results. However, since labeled cyberbullying dataset was rare, this research provided the best insight for using machine learning and deep learning to detect cyberbullying.

3.2 Transformers Model Selection

The subsequent step in the study was the selection of a suitable deep-learning model for the classification task at hand. Considering that the classification of cyberbullying from texts entailed the application of natural language processing (NLP), a transformer language model was deemed appropriate for performing the text classification task. The DistilBERT model was chosen for initial experimentation. This model offered advantages over other language models, such as OpenAI GPT and the Llama language models, due to its relatively smaller size and faster processing capabilities. Given the scope of the project, which was limited to a small-scale NLP task, the more compact DistilBERT model was found to be the optimal choice. With the dataset object restructured accordingly, it was possible to directly input the data points into the model for training and analysis.

The restructured dataset object enabled the direct input of data points into the deep-learning model. The initial step involved tokenizing the texts using the DistilBERT tokenizer. Subsequently, the dataset was partitioned into training and evaluation sets at a ratio of 1.8:1. The model was trained using a learning rate of 2e-5, with 32 training batches, 16 evaluation batches, and over ten epochs. This training configuration yielded an accuracy of 78% and an evaluation loss of 53.3%. As these results were suboptimal, further adjustments were made.

The training proceeded with an altered ratio of training to evaluation data, set at 2.5:1, maintaining the learning rate of 2e-5, with unchanged batch sizes and epoch count. This resulted in a slight increase in accuracy to 78.75% and a reduction in evaluation loss to 50.50%. Maintaining these parameters but extending the number of epochs in a subsequent test, the accuracy stabilized at 78.75%, while the evaluation loss marginally decreased to 49.50%.

An experiment with an increased number of epochs to 20 led to an improved accuracy of 80%, albeit with an evaluation loss that slightly increased to 53.40%. A further attempt with an extended training over 50 epochs resulted in a decrease in accuracy to 75% and a substantial increase in evaluation loss to 72.40%, indicating potential overfitting.

Consequently, the configuration from the penultimate training was selected as the standard: a ratio of 2.5:1 for training to evaluation data, a learning rate of 2e-5, 32 training batches, 16 evaluation batches, and an epoch count of 20 for the transformer model. This setup was determined to be the most effective in balancing accuracy and generalization.

3.3 Exploring Machine Learning Models

The methodology to enhance the accuracy of cyberbullying detection was broadened through the incorporation of diverse machine-learning models. Various models were trained and assessed, including logistic regression, random

Vol. 2024 (5) 194 – 199
ISSN 2688-3651 [online]

forest classifier, support vector machine, and Naive Bayes classifier. A conventional data split of 80-20 was implemented for training and testing, respectively. Moreover, K-Fold cross-validation was utilized to ensure the robustness of the model validation process, with a cross-validation (CV) parameter set to 5.

The research encompassed an extensive array of deep learning and traditional machine learning models to examine and discover the most efficacious methodologies for detecting cyberbullying on social media platforms.

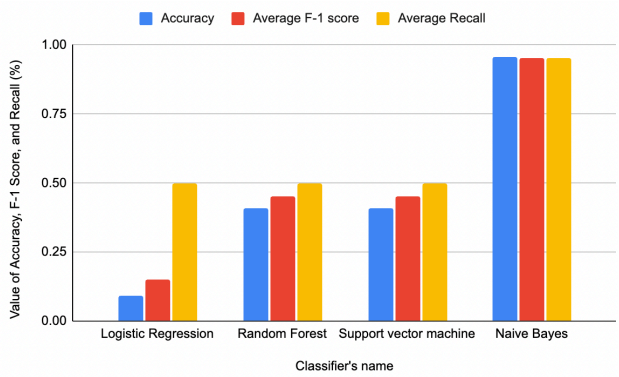

Figure 1. This figure shows the results of different machine learning models.

4. Results and discussions

As mentioned previously, the transformer model achieved an 80% accuracy with its optimal configurations. However, extensive experimentation with machine learning models yielded promising results, offering new insights into their capabilities for cyberbullying detection on social media platforms.

The performance of logistic regression was found to be relatively low, with an accuracy of only 9.00%, an average F-1 score of 15.24%, and an average recall of 50%. The linear nature of logistic regression might have limited its capacity to capture the complexity and nuances of textual data on social media, which often included slang and context-dependent language. Consequently, logistic regression was not the preferred model for this project.

On the other hand, the random forest classifier and the support vector machine both achieved an accuracy of 41.01%, specifically, they both had an average F-1 score of 45.1% and an average recall of 50%. The two models' aligned results revealed some potential errors in the experiment. Both models seemed to be challenged by the dataset's imbalanced nature, with non-cyberbullying instances outnumbering those cyberbullying instances. This disproportion may have led the models to exhibit a bias toward predicting the majority class, resulting in a skewed performance. However, both models handled the complexity of the data better than the logistic regression model. Random forest classifier combined the predictions from multiple machine learning algorithms to make more accurate and robust predictions. Its output was the most frequently occurring class from all the individual decision trees comprising the ensemble. This ensemble nature allowed the random forest classifier to capture a broader range of patterns and relationships within the data, enabling it to better adapt to the complexity of the textual data. The support vector machine (SVM) found the hyperplane that best separated different classes in a high-dimensional space. It could handle non-linear relationships because of the use of kernel functions. The kernel functions transformed the input space in a way that allowed SVM to find a hyperplane in the transformed space, even when the data was not linearly separable in the original space, which made SVM more capable of handling the complexities. Although both classifiers could handle complexities more efficiently than the logistic regression model, the imbalanced nature of the dataset still affected the performance of the two models, yielding a relatively low result.

On the brighter side, the Naive Bayes classifier emerged as the star performer, achieving an impressive accuracy of 95.83%, an average F-1 score of 95.30%, and an average recall score of 95.33%. The probabilistic nature of the Naive Bayes model allowed it to handle the imbalance data, as it calculated the likelihood of a post belonging to a particular class. At the same time, on the other hand, random forest and SVM's performances might have be limited by the imbalance. Naive Bayes also considered each feature of the textual data independently, reducing the risk of overfitting and ensuring the relevant features effectively contributed to the classification result. However, Random Forest, although capable of handling high-dimensional textual data, may still have been influenced by irrelevant features. When dealing with sensitive issues liked cyberbullying, the straightforward probabilistic calculations in Naive Bayes contributed to its interpretability. At the same time, Random Forest and SVM became "black boxes," struggling to track how they arrive at a particular decision. Therefore, we concluded that the Naive Bayes classifier was the best fit for our cyberbullying detection task.

Vol. 2024 (5) 194 – 199
ISSN 2688-3651 [online]

The F-1 score considered both the precision and the recall of the test to compute the score. Recall, also known as sensitivity, measured the number of true positive results divided by the number of all samples that should have been identified as positive. The F-1 score was the harmonic mean of precision and recall, providing a single metric that balanced both concerns, especially in cases where one may have been more important than the other. The F-1 score reached its best value at 1 (perfect precision and recall) and worst at 0. The F-1 score was calculated by the following formula:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

The F-1 score of the logistic regression model indicated that it struggled on capturing the complexity and nuances of textual data effectively, especially when dealing with slang and context-dependent language. As mentioned previously, the ensemble approach of the Random Forrest Classifier and the kernel function of SVM allowed it to have a higher F-1 score than the one of the logistic regression model. However, their performance was still limited by the imbalanced nature of the dataset, affecting their precision and recall balance, as reflected in their F-1 scores. The high F-1 score of the Naïve Bayes Classifier indicated that the probabilistic nature of the Naïve Bayes Classifier allowed it to have a superior ability in balancing precision and recall.

Compared to the results that our transformer model yielded, the results that Naive Bayes returned even gave a higher accuracy. Several factors that contributed to the surprising result include the volume of the dataset and stability. Our dataset, while substantial, may have yet to reach the volume that a transformer model needed to perform effectively. The Naive Bayes classifier, however, could excel in smaller datasets, therefore performing better than the transformer model in this case. While the transformer model exhibited susceptibility to the variations in data quality and distributions, the Naive Bayes classifier's stability and consistency across various data distributions allowed it to handle unpredictable domains of social media texts more effectively. However, if more human-labeled datapoints were available, the deep learning models could possibly outperform the Naïve Bayes model due to its complex and sophisticated nature. Nevertheless, in this case, due to the lack of labeled cyberbullying datapoints on the internet, the Naïve Bayes Classifier for now was the best fitted model.

This research presented the intriguing finding that a simpler probabilistic model like the Naive Bayes classifier could outperform a more advanced deep learning model in specific contexts, such as detecting cyberbullying on social media. Previous research employed the SVM and linear regression models to detect cyberbullying on a single social media platform, either Instagram or Twitter. However, since this research combined the datapoints from multiple social media platform, the Naïve Bayes Classifier in this research could perform cyberbullying detection on multiple social media platforms. Similarly, this research was the first one using the Naïve Bayes Classifier to detect cyberbullying and showing that it best fitted the task compared to other models. Despite the limited size of the dataset, which may not have been ideal for training the DistilBERT model fully, the Naive Bayes classifier proved to be more effective. As mentioned previously, the probabilistic nature of the Naive Bayes model allowed it to handle the imbalance data, as it calculated the likelihood of a post belonging to a particular class. Meanwhile, the transformers DistilBERT model could be affected by the imbalance nature of the dataset, which the non-cyberbullying cases outnumbering those cyberbullying ones. At the same time, the transformers model was also limited by the small scale of the dataset, which would be not sufficient to allow the DistilBERT model yielding its best results. However, the Naïve Bayes Classifier, due to its small scale, could handle these drawbacks of the dataset more effectively. This outcome could be particularly beneficial to researchers with restricted data access, small-scale projects, or limited computational resources. The combined labeled and balanced dataset with an additional random and potentially imbalanced unsupervised dataset mirrored the real-world distribution of cyberbullying instances on social media platforms, which was inherently skewed. The Naive Bayes model's performance in this study indicated its robustness and reliability in such conditions. These findings encouraged a reassessment within the machine learning community of the assumption that complexity correlated with superiority, advocating instead for model selection that was contextually and data-specific.

Vol. 2024 (5) 194 – 199
ISSN 2688-3651 [online]

5. Conclusion and Future Work

In summary, this research provided novel findings for the field of cyberbullying detection. A comprehensive dataset, acquired from the work of Hosseinmardi et al. (2015), was meticulously processed for analytical purposes, establishing a robust base for the investigation. Various models, including the DistilBERT transformer and machine learning models such as logistic regression, Random Forest Classifier, Support Vector Machine, and Naive Bayes classifier, were rigorously trained and evaluated.

The experimental outcomes underscored the Naive Bayes classifier's superior performance in this specific context, attaining an accuracy of 95.83% and surpassing other machine learning models, as well as the more sophisticated DistilBERT transformer. These findings emphasized the Naive Bayes model's adeptness at managing the nuances of textual data characteristic of cyberbullying instances, despite data imbalances. The superiority of the Naive Bayes model over the DistilBERT also suggested its potential for text classification tasks in scenarios where adequate and balanced training data for a transformer model were unavailable.

For future work, there were several ways to enhance this research. Firstly, it was possible to increase the size of the dataset with more balanced data points. By doing so, researchers could train transformer models well with enough data and analyze the results. Exploring other advanced transformer models, such as the Llama3, OpenAI GPT, etc., would also help researchers to have a further understanding of cyberbullying. Second, the research could also gather diverse data samples from various social media platforms so they could potentially perform cyberbullying detection on different social media platforms.

Potential applications of this research were significant. With a focus on real-world implementation, the development of a real-time detection and intervention system on social media could enable immediate identification of cyberbullying incidents. Moreover, from a psychological standpoint, user studies could be conducted to measure the tangible impact of cyberbullying detection and to obtain feedback from those affected, ensuring that the developed technologies are in harmony with user welfare.

References

Devlin, J., et al. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*. https://arxiv.org/abs/1810.04805

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(2009), 12.

Homa Hosseinmardi, et al. (2015). A comparison of common users across Instagram and ask.fm to better understand cyberbullying. In 2015 IEEE Fourth International Conference on Big Data and Cloud Computing (BdCloud) (pp. 1-8). IEEE. https://ieeexplore.ieee.org/document/7307092

Rosa, H., et al. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93, 333-345. https://doi.org/10.1016/j.chb.2018.12.021

Xu, J.-M.et al. (2012). Learning from bullying traces in social media. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 656-666). Association for Computational Linguistics.

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. In 2011 10th International Conference on Machine Learning and Applications and Workshops (Vol. 2, pp. 241-244). IEEE.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *ArXiv*. https://arxiv.org/abs/1910.01108

Van Hee, et al. (2018). Automatic detection of cyberbullying in social media text. *PLoS ONE*, 13(10), e0203794. https://doi.org/10.1371/journal.pone.0203794