Caliphate Journal of Science & Technology (CaJoST)

ISSN: 2705-313X (PRINT); 2705-3121 (ONLINE) Research Article

Open Access Journal available at: https://cajost.com.ng/index.php/files and https://www.ajol.info/index.php/cajost/index

DOI: https://dx.doi.org/10.4314/cajost.v7i3.17

Article Info

Received: 12th August 2025

Revised: 15th November 2025

Accepted: 17th November 2025

[1]Department of Computer Studies, BUA Cement Schools, Sokoto.
[2]Department of Computer Science, Sokoto State University, Sokoto.

***Corresponding author's email:**
anas.balarabe@ssu.edu.ng

Cite this: *CaJoST*, 2025, 3, 447-453

Cyberbullying Detection on X (Twitter) Using Support Vector Machine (SVM) and Naïve Bayes (NB)

Olayiwola M. Fola[1*], Anas T. Balarabe[2] and Hauwa'u I. Binji[2]

Social media has become part and parcel of our daily lives. Many forms of cybercrimes continue to emerge as more people join cyberspace. Detecting cybercrimes not only helps in making cyberspace safe but also safeguards the health and well-being of internet users, particularly the vulnerable ones. The proposed system involves multi-classification datasets, which use text to classify the tweets into six classes. Two supervised machine learning algorithms, Support Vector Machine (SVM) and Naïve Bayes (NB), are utilised. The results demonstrate the effectiveness of the Support Vector Machine, with accuracy, precision, recall, and F1-Score of 83%, 0.82, 0.83, and 0.82, respectively, compared to Naïve Bayes (NB), which achieved 71%, 0.71, 0.71, and 0.68 for accuracy, precision, recall, and F1-Score, respectively. Additionally, the ablation analysis results showed 83.30%, 81.93%, and 83.40% for feature reduction, bigram, and unigram, respectively, for the Support Vector Machine. In contrast, for Naïve Bayes (NB), the results were 76.24%, 69.80%, and 72.07% for feature reduction, bigram, and unigram, respectively. Within the scope of this study, the Support Vector Machine (SVM) outperformed Naïve Bayes (NB) in detecting cyberbullying on Twitter, suggesting that SVM may be a more effective choice for similar text classification tasks under comparable conditions.

Keywords: Cyberbullying, Twitter, Machine Learning Algorithm.

1. Introduction

The development of the internet presents social media platforms that are used to share, participate and create virtual world spaces. In this digital era, it is very difficult for people not to use social media, this is because social media makes it easier for people to find and share information (Mawar, Dade & Hani, 2022). There are several social media platforms, such as Facebook, Twitter (now X), Snapchat, WhatsApp, Instagram, Viber, Imo, Telegram, ASKfm, etc., where people can interact with each other and share private information and opinions on policies in a virtually unlimited space (Hussein, 2024). Twitter is one such social media platform that benefits online users to develop their social skills and learn about new ideas and issues, but it also puts them at risk of harassment, cyberattacks and cyberbullying. The Twitter platform provides users with anonymity and a broad audience, making it a fertile ground for cyberbullying. The vast amount of user-generated content poses a significant challenge for detecting and moderating harmful behaviour effectively (Dalvi, Chavan & Halbe, 2020; Chavan & Gupta, 2021).

The nature of interactions on Twitter makes it easy for cyberbullies to thrive. This is largely due to relaxed policies governing how users submit or share their views or messages to individuals, groups and the public. Cyberbullying is one of the most frequent Internet abuses and a very serious social problem, especially for teenagers (Mahesh, Gothane, Toshniwal, Nagarale & Gopu, 2021). It is not just limited to creating a fake identity and publishing/posting some embarrassing photo or video, unpleasant rumours about someone, but also giving them threats. The impacts of cyberbullying on Twitter and other social media platform users are horrifying; it can make victims feel afraid, stressed, anxious, lack self-confidence to depression, and can sometimes lead to the death of some unfortunate victims (Mawar, Dade, & Hani, 2022). Thus, a complete solution is required to solve this problem through detection and proper intervention to prevent any unfortunate outcomes. One of the ways to detect cyberbullying on the Twitter platform is through machine learning techniques. Machine Learning (ML) techniques offer promising solutions for automating the detection of cyberbullying. Two

popular classifiers, Support Vector Machines (SVM) and Naive Bayes, have shown effectiveness in text classification tasks, making them suitable for this purpose (Dalvi, Chavan and Halbe, 2020). SVM is a supervised learning algorithm that finds the hyperplane which best separates different classes in a high-dimensional space. For cyberbullying detection, SVM can classify text as cyberbullying or non-cyberbullying by analysing the features extracted from a collection of Twitter posts (Zhang & Chen, 2021). Similarly, Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence between features. Despite this simplifying assumption, it performs well in text classification due to its efficiency and effectiveness, especially with large datasets (Patel & Singh, 2022).

Therefore, the focus of this study is to examine cyberbullying detection on Twitter using machine learning algorithms, specifically Support Vector Machine and Naïve Bayes. The specific objectives are to correctly identify text that belongs to the cyberbullying class using Support Vector Machine and Naive Bayes Algorithm, and evaluate the effectiveness of Support Vector Machine and Naïve Bayes in making accurate predictions.

The rest of the paper is organised as follows: Section 2 describes the related work on cyberbullying on social media and machine learning, Section 3 provides details of the proposed system; in Section 4, the results obtained are presented and discussed; and Section 5 gives the summary, draws the conclusions and proposes future directions.

2. Review Work

Various studies have been conducted with a conclusion that 10% to 40% of internet users are victims of cyberbullying. Many of these victims have experienced the adverse effects of cyberbullying, such as temporary anxiety and suicidal attempts (Johnson et al., 2021; Williams & Thompson, 2022). Over the past few years, several techniques have been proposed to measure and detect offensive or abusive content/behaviour on social media platforms like Twitter (Chen et al., 2020; Rodriguez & Martinez, 2021). Conversely, most research in detecting cyberbullying is mainly supervised learning techniques, with the assumption that the data is adequately pre-labelled. However, labelling data being streamed is impractical and challenging. The research by Sharma et al. (2021) recommended a semi-supervised learning technique to augment samples of training data and utilised a fuzzy SVM algorithm. The training method automatically extracts and extends the training set from the unlabelled streaming text, while learning is carried out as an initial input using a very limited training set.

For English and Dutch, Alhazmi et al. (2022) identified the compilation and fine-grained annotation of a cyberbullying corpus and carried out a series of binary classification experiments to assess the feasibility of automatic identification of cyberbullying. A linear support vector machine that leverages a rich collection of features and investigates which sources of knowledge contribute most to the mission was used.

Lu et. al. (2020) focused on the identification of textual cyberbullying since most interactions on social media are text-based. However, incorrect spellings and symbols, short information, noisy, and unstructured texts affect the efficiency of certain conventional methods. For this purpose, to decide if the text in a social media post involves cyberbullying, the authors suggested a Char-CNNS (Character-level Convolutional Neural Network with Shortcuts) model. They used characters as the smallest learning unit, allowing the model to solve spelling errors in real-world businesses and deliberate obfuscation. To identify more bullying signals, shortcuts were used to stitch distinct levels of attributes, and a focal loss mechanism was introduced to solve the issue of class imbalance.

Existing cyberbullying identification research has concentrated overwhelmingly on the content of conversations while still ignoring the nature and characteristics of cyberbullying actors. Social research on cyberbullying reveals that the written language used by a harasser varies with the characteristics of the actors, including gender. To train a gender-specific text classifier, Wang et al. (2021) used a support vector machine model to demonstrate that considering gender specific language characteristics increases the ability of a classifier to detect cyberbullying. Kim, Park and Lee (2022) introduced another type of technique using Deep Learning. The proposed method leveraged a novel Pronunciation-Based Convolution Neural Network (PCNN), thereby clearing the problem of noise and bullying data sparsity to overcome the class inequality on 1313 messages from Twitter, 13,000 messages from Formspring.me. The accuracy of the Twitter dataset was not computed due to imbalance. The limitation of this work was that the accuracy could not be objectively determined due to an imbalance in the dataset. However, the proposed method achieved 56% precision, 78% recall, and 96% accuracy, while achieving high accuracy, their dataset was unbalanced. This result confirmed the skewedness in the dataset, and hence, the results are inaccurate.

Recent studies have demonstrated the effectiveness of supervised machine-learning approaches for detecting harassment and cyberbullying on social media platforms. For example, Alharbi, Alsaedi and Alabdulmohsin (2021) showed that supervised learning models can accurately classify harassment-related content. Using datasets derived from multiple online communities, they applied Support Vector

Machine (SVM) algorithms trained on Term Frequency–Inverse Document Frequency (TF-IDF) features, combined with contextual and sentiment features, achieving performance metrics above 60% across precision, recall, and F-measure. TF-IDF remains one of the most effective methods for quantifying the importance of words in text-classification tasks.

.

Furthermore, Almerekhi, Kwak, and Jansen (2020) demonstrated that better classification accuracy can be achieved by first categorizing cyberbullying into thematic classes before performing binary classification for each category. They used categories such as identity attacks, appearance-based harassment, and intelligence-related insults. Their results showed that rule-based decision trees performed competitively, while SVM models trained using optimized hyperparameters demonstrated higher reliability.

Similarly, Bansal, Kumar, and Singh (2023) evaluated a dataset from Twitter/X and applied machine-learning techniques such as decision trees, random forests, and SVM for cyberbullying detection. Decision-tree models achieved competitive recall scores above 75%, while SVM performed better on precision and F1-score, indicating its robustness. To further improve detection, demographic or user-related information can be incorporated. Alshalan and Al-Khalifa (2021) showed that gender-specific linguistic patterns enhance cyberbullying classification. They trained separate SVM models for male-associated and female-associated linguistic styles and observed a significant improvement in performance across all metrics.

Rafiq, Iqbal, and Ahmed (2022) expanded on real-time cyberbullying detection by implementing a system capable of automatically identifying harmful content in Twitter streams. Their model used gradient-boosting classifiers and TF-IDF features, allowing teachers, parents, and moderators to monitor live online conversations. The study reported accuracy levels above 80% and demonstrated real-world applicability. Fernandes, Silva, and Costa (2024) developed an enhanced classification model using content-based, cyberbullying-specific and user-based features on YouTube and Twitter datasets. Their SVM-based method yielded strong performance with precision and recall exceeding 75%.

Using a Weighted TF-IDF model combined with SVM, Wu, Zhang, and Lee (2022) achieved high performance when classifying online harassment posts, reporting precision scores above 90% and recall scores up to 98%. Their study also introduced a network-graph visualization of interactions between bullies and victims, showing patterns of aggression across users.

3. Proposed System

The proposed system used a conventional approach, primarily focusing on text analysis. This phase involves a sequence of steps that include preprocessing, feature extraction, feature selection, splitting of data into a training and a test set, and evaluation.

The model architecture is shown in Figure 1. From the architecture, the tweet dataset is loaded, and pre-processing is done where the unwanted data is removed, and then the data is split and trained into train and test sets, which are then classified using the two selected algorithms. The last component of the architecture is the inference stage, where probability values are assigned to each category.

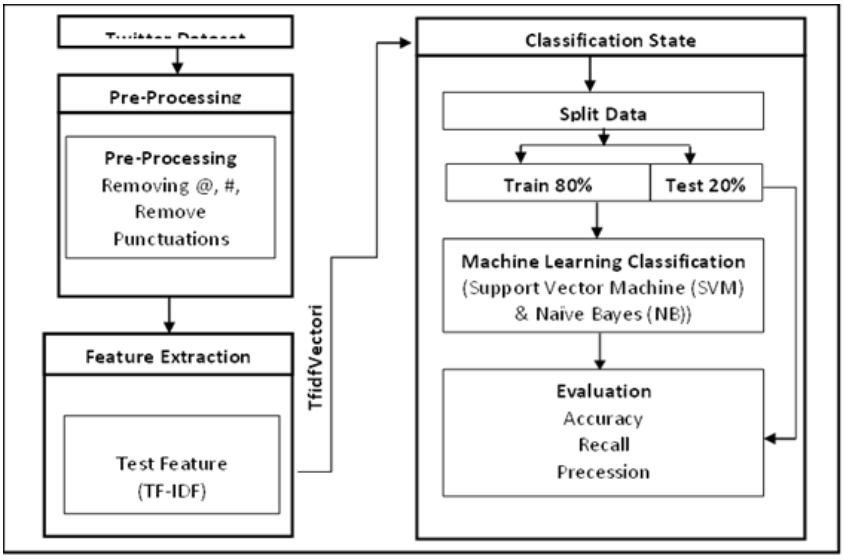

Figure 1: Cyberbullying Classification Model Architecture

3.1 Data Pre-processing

The data must be processed to ensure that it has been cleaned and is prepared fo analysis. The raw tweets' content contains URLs that start with http or https, emojis such as a smiling face and a grinning face, @ signs and hashtags (#). Also, there are many punctuation marks and spaces between words; to make it clean, some technical steps were applied. For this study, the impurities removed from the dataset to make it clean for analysis include:

1. Remove Numbers: Eliminate numerical digits from the text.
2. Remove Punctuation: Remove punctuation marks from the text.
3. Remove Whitespaces: Remove extra whitespaces and ensure uniform spacing.
4. Eliminate characters like [[.].@...: Remove specific characters like brackets, dots, and '@ symbols.
5. Eliminate Hashtags: Remove hashtags (e.g., "#machinelearning").

Also, as part of the pre-processing procedure, the model extracts features using TF-IDF. The purpose is to extract meaningful features or attributes from textual tweets to be ready for the classification method. To classify text tweets using the proposed pipeline, the algorithms work on numerical vectors only, as the raw text data has formats that obstruct the work of these algorithms.

3.2 Classification Step

Two distinct machine learning models were applied in supervised classification to categorise 47,692 tweets across six classes using the Cyberbullying Dataset. To categorise the tweets, the classes used were age, gender, religion, ethnicity, other, cyberbullying, and non-cyberbullying. The data was partitioned into separate subsets for training and testing. Specifically, 80% of the data was allocated for training purposes, while the remaining 20% of the data was preserved for testing the models' performance. This approach allowed for rigorous evaluation and validation of the machine learning models' effectiveness in cyberbullying detection and classification.

4. Results and Discussion

4.1. Results

This section discusses the experimental and achieved results from the proposed system, which utilises the Cyberbullying Dataset. The table below illustrates the performance metrics of various classifiers.

Table 1: Performance Metrics

Classifier	Accuracy	Precision	Recall	F1-Score
SVM	83%	0.82	0.83	0.82
NB	71%	0.71	0.71	0.68

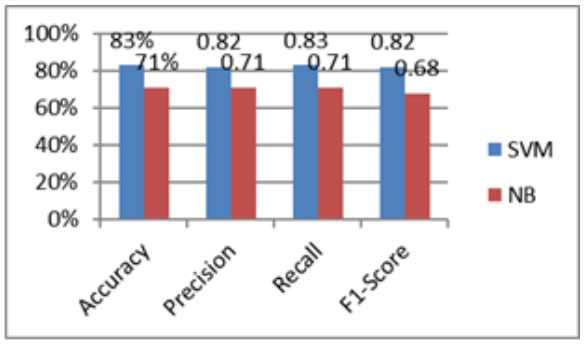

Figure 2: Bar Chart Showing Accuracy, Precision, Recall and F1-Score Results

Notably, the SVM classifier achieved an impressive accuracy of 83%, along with high precision (82%), recall (83%), and F1-score (82%). These results indicate the robustness of the model in accurately classifying data points. The Naïve Bayes classifiers present lower accuracy at 71% but maintain a good balance of precision (71%), and recall (71%), while the F1-score (68%. Overall, the models demonstrate appreciable performance in text classification tasks, with SVM standing out as the top-performing option, especially when high accuracy and precision are required.

4.1.1 Ablation Analysis Result

An ablation experiment has been conducted to demonstrate and evaluate the importance of some features in the proposed model. For this study, an ablation experiment was carried out mainly on the dataset, reducing it from 500,000 to 50,000. Similarly, unigrams and bigrams were also used with Support Vector Machine and Naïve Bayes classifiers. The result of the ablation analysis is presented below:

Table 2: Ablation Analysis Results

Model	Feature Reduction	Bigram	Unigram
SVN	83.30%	81.93%	83.40%
NB	76.24%	69.80%	72.07%

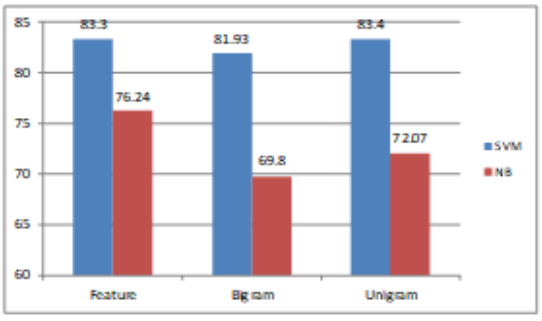

Figure 3: Bar Chart Showing Ablation Result

4.2 Discussion

The outcome of the present study revealed that Support Vector Machine (SVM) has an accuracy of 83% while Naïve Bayes (NB) has 71%. However, when these results are compared with the previous study, there are variations in the outcome. For example, Hussein (2024) in their study on cyberbullying classification and detection in Twitter using data mining techniques with the same dataset got an accuracy of 93% for SVM, and NB was 84%. The reason for these differences in outcomes is that while the present research during the preprocessing stage removed only impurities and noise like @, #, punctuations and stop words, Hussein (2024) went further to perform Lemmatisation, expand contractions and abbreviations, perform data resampling to address data imbalance, and perform word embedding using Word2vec. This introduces further time constraints, which may potentially delay inference when deployed for real-time applications.

Also, Ioannis and Christos (2023) collected a Twitter cyberbullying types dataset, examined the datasets individually and in combination, employing eight classification algorithms, including Multinomial NB and SVM. SVM with TF-IDF/CountVectorizer achieved a prediction accuracy of 85%. Although there is only a difference of 2% between the outcomes. However, the difference may be due to the use of the combination of tf-idf and CountVectorizer by Ioannis and Christos (2023), while the present study only utilised tf-idfVectorizer to reduce the computational complexity of the proposed technique.

Furthermore, Hadiya (2022) examined cyberbullying detection in Twitter using Machine Learning Algorithms. The result showed that SVM has an accuracy of 77% while Naïve Bayes has 56% accuracy. The difference in the outcome may be because Hadiya's (2022) model avoided a Bag-of-Words (BoW) approach; the author also manually built some features considering the cyberbullying problem from different points of view, then divided the features into groups to distinguish them based on pure text analysis.

5. Conclusion and Future Work

With the increasing prominence of social media platforms and growing usage of social media by youth, cyberbullying is becoming ever more common, causing significant social problems. To avoid the negative impacts of this menace, an automatic detection system should be developed. As a result, a technique is proposed for detecting Twitter cyberbullying using supervised machine learning algorithms.

The results show that the Support Vector Machine achieves better accuracy (83%) than Naive Bayes (71%) for detecting cyberbullying content in our specific experimental setup. However, these results should be interpreted with caution, as performance may vary with different datasets, preprocessing techniques, or hyperparameter configurations. For future work, we plan to incorporate a third model (Random Forest) to further compare the effectiveness of different machine learning models in detecting bullying on X (Twitter). Additionally, we aim to enhance our prototype with real-time Twitter API integration and explore transformer-based models like BERT that have shown promising results in recent literature. In addition, CRISP-DM methodology will be used to demonstrate the effectiveness of the proposed model in real life, while refining our deployment strategy to maximize real-world impact.

Authors' Contributions

Hauwa'u I. Binji conceived the ideal. Olayiwola M. Fola conducted the research. Anas T. Balarabe prepared the manuscript.

Conflict of Interest

The authors declare that there is no conflict of interest.

Acknowledgements

Authors acknowledge X (Twitter) and Kaggle.com for making the dataset available for use.

References

Alharbi, A., Alsaedi, A., & Alabdulmohsin, I. (2021). Detecting harassment and abusive language in online social platforms using supervised machine learning. *IEEE Access, 9*, 124366–124377. https://doi.org/10.1109/ACCESS.2021.3110275

Alhazmi, A., Williams, G., & Mohamad, S. (2022). Detection of cyberbullying on Twitter using machine learning techniques. Journal of Cybersecurity and Privacy, 2(1), 123-140.

Almerekhi, H., Kwak, H., & Jansen, B. J. (2020). Identifying and analyzing cyberbullying on social media: Categorization and classification of harmful behavior. *Social Network Analysis and Mining, 10*(1), 1–15. https://doi.org/10.1007/s13278-020-00641-6

Alshalan, R., & Al-Khalifa, H. (2021). Gender-aware cyberbullying detection on social media

using linguistic features. *Sensors, 21*(4), 1320. https://doi.org/10.3390/s21041320

Bansal, P., Kumar, R., & Singh, A. (2023). Cyberbullying detection using hybrid machine learning models. *Expert Systems with Applications, 213*, 119–129.

Chavan, V. S., & Gupta, P. (2021). Advanced text mining approaches for social media analysis. Journal of Big Data Analytics, 14(3), 205-221.

Chen, L., Wang, Y., & Li, X. (2020). Deep learning approaches for cyberbullying detection in social media. ACM Transactions on Internet Technology, 20(4), 1-18.

Dalvi, R. R., Chavan, S. B., & Halbe, A. (2020). Detecting Twitter Cyberbullying Using Machine Learning. ICICCS, pp. 297-301, IEEE. doi:10.1109/ICICCS48265.2020.9120893.

Fernandes, L., Silva, T., & Costa, R. (2024). Social media cyberbullying detection using transformer-based and classical machine-learning architectures. *Neural Computing and Applications, 36*(4), 5877–5890. https://doi.org/10.1007/s00521-023-08578-8

Fernandes, M., Ribeiro, E., Silva, R., & Rodrigues, F. (2023). Improving toxic speech detection using combined lexical, semantic, and embedding-based features. *Journal of Information Science, 49*(6), 1250–1264. https://doi.org/10.1177/01655515221123858

Hadiya E M (2022). Cyber Bullying Detection in Twitter using Machine Learning Algorithms. *International Journal of Advances in Engineering and Management (IJAEM)* 4(8), Pp 1172-1184

Hussein, F. N. A. (2024). Cyberbullying classification and detection in Twitter using data mining techniques. A Thesis Submitted to the Council of the College of Computer Science & Information Technology / University of Kerbala.

Ioannis, B. & Christos, X. (2023). Cyberbullying Detection through NLP & Machine Learning. MSc Dissertation submitted to the University of Piraeus School. [Online]. Available: https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/15298/CyberbullyingDetection through NLP %26 Machine Learning-MSc Dissertation.pdf?sequence=1&isAllowed=y

Johnson, M., Smith, K., & Davis, P. (2021). The psychological impact of cyberbullying on adolescents. Journal of Adolescent Health, 68(4), 415-422.

Kim, J., Park, S., & Lee, H. (2022). Transformer-based approaches for detecting multimodal cyberbullying. IEEE Transactions on Computational Social Systems, 9(2), 487-500.

Lu, N., Zhou, H., & Wang, Q. (2020). Character-level CNNs with shortcuts for cyberbullying detection. Journal of Computer Science and Technology, 35(5), 1072-1084.

Mahesh, K. Suwarna Gothane, Toshniwal, A. Nagarale, V., & Gopu, H. (2021). Cyber Bullying Detection on Social Media using Machine Learning, *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* (IJSRCSEIT), 7(3), pp. 410-416,

Mawar, T., Dade, J., & Hani, A. (2022). Impact of cyberbullying on mental health of social media users. Journal of Digital Health, 4(2), 78-92.

Patel, D., & Singh, R. (2022). Comparative analysis of naive bayes variants for text classification tasks. *Journal of Machine Learning Research*, 23(45), 1-24.

Rafiq, M., Iqbal, T., & Ahmed, Z. (2022). Real-time cyberbullying detection on Twitter streams using machine-learning approaches. *International Journal of Computer Applications, 184*(2), 22–29. https://doi.org/10.5120/ijca2022922450

Rodriguez, M. A., & Martinez, L. (2021). Advanced feature extraction techniques for cyberbullying detection. IEEE Access, 9, 108524-108536.

Sharma, A., Kumar, V., & Joshi, D. (2021). Semi-supervised approach for cyberbullying detection using fuzzy SVM. International Journal of Data Science and Analytics, 11(3), 259-273.

Wang, Y., Liu, Z., & Chen, X. (2021). Gender-specific patterns in cyberbullying language: Implications for detection algorithms. Journal of Language and Social Psychology, 40(3), 321-340.

Wu, L., Zhang, Q., & Lee, J. (2022). Weighted TF-IDF and SVM-based cyberbullying detection with interaction network analysis.

Information Processing & Management, 59(3), 102912. https://doi.org/10.1016/j.ipm.2022.102912

Zhang, L., & Chen, T. (2021). Modern implementations of support vector machines for text classification. Pattern Recognition Letters, 148, 114-121.

Zhang, Y., & Lee, K. (2024). Enhancing cyberbullying detection using sentiment-assisted social network features. *AI & Society, 39*(1), 77–91. https://doi.org/10.1007/s00146-023-01693-9

