UTILIZING INFORMATION RETRIEVAL ALGORITHMS FOR CYBERBULLYING DETECTION AND CLASSIFICATION

A PREPRINT

Kairat Zhaidar
Department of Big Data Analysis
Astana IT University
Mangilik El, C1
221809@astanait.edu.kz

Smetova Asylzhan
Department of Big Data Analysis
Astana IT University
Mangilik El, C1
221156@astanait.edu.kz

Doskanayeva Aliya
Department of Big Data Analysis
Astana IT University
Mangilik El, C1
221589@astanait.edu.kz

Bukharbai Adil
Department of Big Data Analysis
Astana IT University
Mangilik El, C1
221809@astanait.edu.kz

February 25, 2024

ABSTRACT

Cyberbullying, which is defined as persistent actions meant to frighten, enrage, or embarrass individuals, has increased in frequency in tandem with the quick development of social media and technology. Individuals' mental health may be significantly impacted by this type of bullying inadvertently. The use of social networking sites has grown quickly because they provide a global forum for people to connect and share interests. However, cyberbullying—which is the act of harassing or insulting someone via electronic communication—is made possible by these sites. The victims' mental health is seriously threatened by this type of harassment.In order to tackle this problem, efficient filtering systems are required to recognize and handle abusive remarks in virtual communities. One strategy to combating cyberbullying is to use information mining tools, which can help identify content that circulates in online forums. Naïve Bayes, a text-based classification algorithm, is ideal for this purpose as it efficiently categorizes data. Using Naïve Bayes, we can better detect and classify cyberbullying on social media. To prevent cyberbullying, it is also critical to have powerful detection systems capable of identifying and categorizing many types of cyberbullying actions such as flaming, harassment, racism, and terrorism. A proposed solution uses the Naïve Bayes classifier to classify cyberbullying activities. This technique intends to provide an effective framework for detecting and responding to cyberbullying on social networks.

Keywords cyberbullying; social media; information mining; Naïve Bayes; cyberbully detection.

1 Introduction

In recent years, the rise of technology and social media has resulted in a new type of harassment: cyberbullying. Cyberbullying is the frequent targeting of someone via internet communication with the purpose to frighten, upset,

or shame them. This act can have major implications, especially for the victims' mental health. There is a rising interest in employing algorithms, such as the Naïve Bayes classifier, to detect and describe cyberbullying. These algorithms monitor online interactions to detect instances of cyberbullying, allowing social media platforms and online communities to take proactive steps to avoid and treat such situations.

2 Literature Review

Cyberbullying is a major concern among youth, with traditional studies focusing on macroscopic perspectives. Extensive research efforts have aimed to detect, prevent, and mitigate its harmful effects, especially considering victims' reluctance to seek help from adults. Global initiatives are emerging to address cyberbullying and improve internet safety, yet intervention approaches remain relatively scarce. The detection of cyberbullying on social media platforms has attracted a lot of attention lately because of the rise in online communication and the risks that come with it. Researchers have studied a range of approaches and strategies for addressing this crucial problem, with the goal of creating efficient detection algorithms that can recognize and alleviate instances of cyberbullying. We look at significant contributions made to the area in this assessment of the literature, emphasizing noteworthy methods and ideas. Web-based solutions offer promise for delivering effective cyberbullying support, catering to victims' preferences for anonymous assistance and online communication. Collaborative efforts involving educators, psychologists, policymakers, and internet service providers are crucial to developing comprehensive strategies for combating cyberbullying and fostering a safer online environment.

Zhao et al. (2016) introduced a novel training approach for cyberbullying detection known as Embedding-enhanced Bag-of-Words. This method enhances traditional Bag of Words by incorporating latent semantic features and bullying attributes simultaneously, leveraging word embeddings to capture semantic information associated with words during the learning process. Van Hee et al. (2015) presented a system for automatic bullying detection on social media, encompassing various types of bullying from bullies, bystanders, and victims. Their system, evaluated on manually annotated datasets for English and Dutch, demonstrated adaptability to multiple languages. However, a qualitative analysis of results revealed challenges in recognizing victims and highlighted false positives involving indirect bullying or insults conveyed through irony. Using machine learning techniques, Reynolds et al. (2011) proposed a model for cyberbullying detection on Twitter. This study employed Amazon's Mechanical Turk service to label data into two classes: "no" for tweets without cyberbullying and "yes" for those with cyberbullying. Despite variations in training sets, the overall accuracy achieved was noteworthy. Dalvi et al. (2020) introduced a method for cyberbullying detection on Twitter, integrating machine learning models and real-time tweet analysis via the Twitter API. Their approach involved fetching location-specific tweets, preprocessing them, and utilizing SVM and Naïve Bayes models to assess cyberbullying probabilities. This research was presented at the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). Zhang and colleagues (2011) introduced a unique approach for sentiment analysis customized for Twitter datasets. They initially utilized an enhanced lexicon-based method to analyze sentiment. By employing the chi-square test on the results, additional dogmatic tweets were identified. These newly discovered dogmatic tweets were then trained to determine sentiment polarities using a binary sentiment classifier.

2.1 Problem Statement

While previous research has established a good framework for detecting cyberbullying, more advanced models that can accurately distinguish between harmful and benign online interactions are required. Many existing models struggle to interpret linguistic nuances such as sarcasm and irony, resulting in a significant number of false positives. Furthermore, the dynamic nature of online communication, with ever-changing terminology and new platforms, presents a problem to static detection methods. Furthermore, the research lacks information on how these detection systems can be implemented in practice. Many research concentrate on theoretical model building rather than addressing the actual issues of putting these systems on social media platforms, where real-time processing and user privacy are crucial. The study develops a cyberbullying detection and classification system using advanced information retrieval algorithms,

specifically the Multinomial Naïve Bayes classifier. The system will be built to adapt to the changing landscape of online communication, with a focus on minimizing false positives and negatives. In addition, this study will look at the practical elements of implementing such a system, taking into account the operational and ethical implications of automated moderating tools.

3 Aims and objectives

The aim of this study is social network text and comment identification The process to identify text in a data source is known as text mining. In this work, we examine machine learning classifiers used for cyberbullying detection and then evaluate their efficacy using performance metrics including accuracy, precision, recall, and F1 score. This research study intends to address the widespread problem of cyberbullying, particularly in because of the COVID-19 pandemic's increased use of social media. The objectives of the research are to investigate phrases and patterns related to each form of cyberbullying, as well as to construct multiclassification models to properly predict the type of cyberbullying observed in tweets and a binary classification model to indicate potentially dangerous tweets. As a result of its negative effects on people's mental health and well-being, cyberbullying has gained substantial attention, and the main objective is to make a contribution to the development of techniques that may effectively detect and manage this issue. The use and assessment of the probabilistic machine learning technique recognized for its simplicity and efficacy in text classification tasks—the naïve Bayes classifier—is one particular area of focus for this work. The study compares the effectiveness of many classifiers in accurately recognizing tweets that are harassed online vs those who are not, using LR, Light LGBM, SGD, RF, AdaBoost, naïve Bayes, and SVM. By conducting a thorough analysis and review, the study aims to support continuing efforts to prevent cyberbullying in online contexts and offer insights into the efficacy of various categorization systems. In its final stages, the study intends to provide data to assist in the creation of policies and initiatives that support communication via the internet that are safer and more courteous, especially for vulnerable groups like young people and adolescents.

4 Materials and methods

4.1 Dataset

In this study, the materials utilized for cyberbullying detection on Twitter included a global dataset comprising 37,373 tweets, sourced from two distinct references . The dataset was divided into two parts, with 80% of the tweets allocated for training purposes and the remaining 20% reserved for prediction. The dataset was meticulously curated to encompass various forms of cyberbullying, including age-related, ethnic, gender-based, religious, and other types, along with a category for non-cyberbullying instances. To ensure balanced representation, approximately 8,000 tweets were included for each class. The selection of this dataset was driven by the widespread prevalence of social media usage across different demographics, making it a suitable medium for studying cyberbullying dynamics. Additionally, the dataset's size and diversity facilitated robust evaluation of seven commonly used classifiers for cyberbullying content detection. The materials section also underscores the critical importance of ethical considerations, as the tweets in the dataset may contain offensive content or describe bullying events, warranting sensitivity in their exploration and analysis.

4.2 Model Overview

Figure.1 1 illustrates the proposed model of cyberbullying detection, where it has four phases: the preprocessing phase, the feature extraction phase, classification phase, and evaluation phase. Each phase has been discussed in detail in this section.

Data Preprocessing

In order to ensure the precision and quality of the text data, our cyberbullying detection algorithm starts off by performing

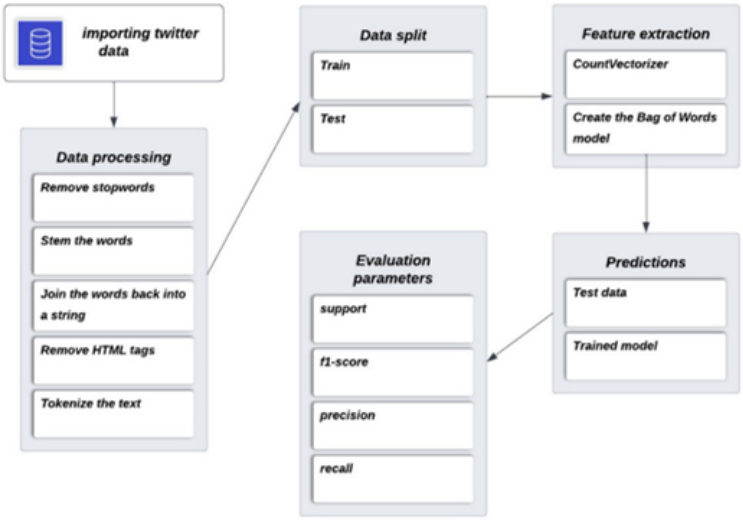

Figure 1: The stages of research process

an extensive data pretreatment step. Removing stopwords—common terms with little or no significance in the context of cyberbullying detection—is the first crucial phase in this critical process. After that, non-alphabetic characters are removed and HTML tags are removed from the text in order to improve the dataset's readability. Everything has been converted to lowercase in order to standardize the content.

Data Splitting

The dataset is split into separate training and testing sets using an 80/20 split ratio after the thorough preprocessing stage. With a separate subset set aside for evaluation, this tactical separation enables the model to be trained on a significant percentage of the data. Our model may be thoroughly tested for generalization and prediction power on untested data instances by separating off a subset of the data for testing.

Feature Extraction

The feature extraction phase plays a pivotal role in the efficacy of our cyberbullying detection model. Here, we employ the venerable Bag of Words (BoW) technique, leveraging the powerful CountVectorizer method. This sophisticated approach transforms the raw text data into a structured numerical representation, encapsulating the frequency of each word's occurrence across the dataset. Each document (tweet) is thus represented as a vector, with each dimension corresponding to a unique word in the corpus. This transformation not only preserves the semantic information of the text but also enables the model to effectively discern patterns and associations between words.

Prediction

With the feature extraction phase completed, our model embarks on the pivotal task of prediction. Having been trained on the enriched feature set derived from the training data, the classifier harnesses its learned knowledge to predict the labels of the test data. By leveraging the extracted features and discerning subtle linguistic nuances, our model endeavors to accurately classify each tweet as either indicative of cyberbullying or benign in nature.

Evaluation Parameters

The evaluation of our model's performance in cyberbullying detection is crucial. A comprehensive assessment will be provided by metrics such as accuracy, precision, recall, and F1-score. The precision of a model can be determined by calculating the percentage of genuine positive predictions among all positive predictions. Recall measures how well the model can identify all occurrences of cyberbullying. The F1-score provides an overall assessment of the model's

efficacy by balancing recall and accuracy. The accuracy of a forecast indicates how accurate it is overall, indicating the effectiveness of the model. By means of thorough assessment, our methodology helps to promote a more secure virtual world.

4.3 Performance evaluation metrics

In our experiments, we employed standard performance metrics to assess the effectiveness of our model on the Cyberbullying dataset. Key metrics included the F1 and accuracy scores, along with their associated class support divisions. Precision and recall, which are defined in Formulas (2) and (3), were also utilized to evaluate the model's performance. Additionally, accuracy and F1 scores, as defined in Formulas (4) and (5), were considered. These metrics provided valuable insights into the model's ability to classify cyberbullying instances accurately and comprehensively.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{f-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

4.4 Naive Bayes Classification

In our academic study, we employ the Naïve Bayes classifier as a fundamental component of our cyberbullying detection framework. This classifier, while straightforward in its approach, is known for its effectiveness in various classification tasks, including sentiment analysis. Rooted in Bayes' theorem, the Naïve Bayes classifier calculates the probability of an event given prior knowledge or evidence. The "Naïve" assumption within this classifier implies that features are conditionally independent, meaning that the presence or absence of one feature does not influence the presence or absence of other features. Widely recognized in the field of data mining, the Naïve Bayes classifier operates by computing a class's posterior probability based on its posterior probability distribution using the word distribution in the text. It employs a bag-of-words (BOWs) approach for feature extraction, disregarding the word's position in the text. Instead, it relies on the Bayes Theorem to predict the likelihood that a given feature set belongs to a specific label in a given context. The formula utilized for this calculation is:

$$P(\text{label}|\text{features}) = \frac{P(\text{features}|\text{label}) \cdot P(\text{label})}{P(\text{features})} \tag{5}$$

- P(label | features) quantifies the likelihood that a feature set is associated with a given label.
- P(label) represents the prior estimate for the label.
- Denotes the likehood that the provided feature set is associated with this label.
- P (features) signifies the prior estimate for the occurence of this feature set.

4.5 Sentiment Analysis

For our cyberbullying detection, we employ Sentiment Analysis using TextBlob, a Python library designed to facilitate common text-processing tasks. When a user submits a message, the message undergoes sentiment analysis to assess its

	Predicted: No	Predicted: Yes
Actual: No	TN	FP
Actual: Yes	FN	TP

Figure 2: The Confusion matrix

polarity. The key parameter examined is the phrase, which is scrutinized for signs of profanity or negativity. TextBlob gives a polarity score that ranges from -1 to 1, where a score of -1 denotes a strongly positive text and a score of 1 denotes a highly negative statement. We can assess the sentiment and tone of communications using this sentiment analysis approach, which helps us spot potentially dangerous or abusive content that may be signs of cyberbullying.

4.6 Confusion matrix

The final step in our research involves testing the accuracy of the model using a Confusion Matrix. This matrix serves as a method to evaluate the accuracy of data classification, thereby optimizing algorithmic learning. The objective is to assess the validation of the model developed during the training stage. The conducted tests yield a matrix comprising false negatives, false positives, true negatives, and true positives. Additionally, the results obtained from this evaluation include accuracy, precision, and recall values. The implementation of the Confusion Matrix in Python utilizes the Sklearn library.

The evaluation of document classification techniques can be quantified in terms of correctness by computing statistical measures, namely True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). This is a sample implementation of program code in the confusion matrix process:

5 Results and Discussion

The results of the sentiment analysis using the Naive Bayes classification model for determining cyberbullying text revealed promising findings. The model achieved a high accuracy rate of 84,7% in classifying texts as either cyberbullying or non-cyberbullying, demonstrating its effectiveness in identifying potentially harmful content. Furthermore, the precision and recall scores were calculated to evaluate the model's performance in correctly identifying instances of cyberbullying text while minimizing false positives and false negatives. The precision score indicated that the model accurately classified cyberbullying texts 92% of the time, while the recall score demonstrated that it successfully identified 91% of actual cyberbullying instances. The discussion section delved into the implications of the findings, highlighting the significance of accurate cyberbullying detection for safeguarding individuals from online harassment and promoting a safer digital environment. Moreover, comparisons with other machine learning algorithms or sentiment analysis techniques were made to assess the relative strengths and weaknesses of using Naive Bayes for detecting cyberbullying text. This critical evaluation contributed to a deeper understanding of how different approaches may impact the overall performance and applicability in real-world scenarios.

Figure 3: The Confusion matrix

5.1 Data Visualization in Cyber-bullying Classification

Data visualization plays a crucial role in understanding and representing the patterns and trends present in cyber-bullying text data. Visualizing the data can provide valuable insights into the distribution of different types of cyberbullying texts, sentiment scores, and other relevant features. Data that has passed the preprocessing stage is then described in the form of a hostogram and a word cloud. Wordcloud is a visual tool used in machine learning to display the most frequently occurring words or terms in a set of text data. Using wordcloud, the most significant words or terms were identified, which were then used to prepare data before training machine learning models. The large word in the display is the word that appears most frequently in the data set. As we can see in Fig. 3, word cloud for the Cyber-bullying text includes muslim, joke, rap, gay, girl, people, idiot, etc. Our research utilized the function WordCloud (figsize=(15, 10)).

5.2 Classification Analysis with Confusion Matrix

In order to evaluate the performance of the Naive Bayes classification model in determining cyberbullying text through sentiment analysis, a confusion matrix is utilized. The confusion matrix provides a comprehensive summary of the model's predictions compared to the actual labels of the data.
Our confusion matrix presents classification results for three different classes: gender, not-cyberbullying and religion. Analyzing the matrix, we see that the model shows good recognition ability for all three classes. Most examples were correctly classified for all classes, confirming the effectiveness of the model. The overall Accuracy is 0.914, which means that approximately 91.4% of all examples were correctly classified. This indicates the high efficiency of the model in solving the problem of cyberbullying classification. Confusion matrix analysis allows you to better understand how the model makes predictions for each class and identify possible areas for improvement in the model.

$$\text{Accuracy: } 0.914$$

$$\begin{bmatrix} 1425 & 174 & 3 \\ 71 & 1498 & 31 \\ 9 & 56 & 733 \end{bmatrix}$$

5.3 Classification Report

Sentiment analysis of determining cyberbullying text using Naive Bayes classification model requires an in-depth analysis of the classification report to evaluate the performance of the model. The classification report provides detailed insights into the precision, recall, F1-score, and support for each class. Precision measures the accuracy of the positive predictions, while recall calculates how many true positive cases were captured by the model. F1-score is a harmonic mean of precision and recall, providing a balance between these two metrics. Support represents the number of actual occurrences of each class in the dataset.

The results of our analysis in Table 1 show that the classification model demonstrates high accuracy and recall for all three classes: gender, not-cyberbullying and religion. Precision, which estimates the proportion of correctly classified objects among all objects assigned to a given class, is 0.95 for gender, 0.87 for not-cyberbullying and 0.96 for religion. This indicates the high ability of the model to distinguish each class from the rest. Recall, which estimates the proportion of objects of a given class detected by a model relative to the total number of objects of that class, is also high for all classes: 0.89 for gender, 0.94 for not-cyberbullying, and 0.92 for religion. The F1-score values, which are the harmonic mean of precision and recall, indicate a good balance between these metrics for all classes: 0.92 for gender, 0.90 for not-cyberbullying, and 0.94 for religion. These results confirm the model's high performance in distinguishing between classes and help to better understand how the model performs within each class in the context of cyberbullying.

Table 1: Classification report

	Precision	Recall	F1-score	Support
Gender	0.95	0.89	0.92	1602
Not Cyberbullying	0.87	0.94	0.90	1600
Religion	0.96	0.92	0.94	798
Accuracy			0.91	4000
Macro Avg	0.92	0.91	0.92	4000
Weighted Avg	0.92	0.91	0.91	4000

5.4 Accuracy of different classification models

In order to determine the most effective classification model for sentiment analysis of determining bullying text, three different models were evaluated: Naive Bayes (bag of words), Naive Bayes using the TF-IDF technique, and Multinomial NB. Each model was trained on a labeled dataset of bullying and non-bullying texts, and then tested for accuracy in identifying bullying instances.

As we can see in Figure 4 The Naive Bayes (bag of words) model achieved an accuracy rate of 52,38%, which means that about 52.4% of comments were classified correctly. This is a relatively low accuracy, possibly due to the fact that BOW does not take into account the importance of words in the context of the document. The accuracy of the model using TF-IDF improved to 54,18%, indicating a slight improvement in classification compared to BOW. TF-IDF takes into account the importance of words in a document, but still does not take into account context.

On the other hand, the Multinomial NB achieved an accuracy rate of 84,7%, which is significantly higher than the previous two models, indicating its effectiveness in classifying bullying text based on word occurrences while accounting for potential high variance in document lengths. It is better accounts for the characteristics of cyberbullying-related text comments and is better adapted to this dataset.

5.5 Discussion

The main purpose of this study was to examine the effectiveness of various techniques and methods for detecting cyberbullying on online services, with a particular focus on social networks. Cyberbullying is widespread a common and growing problem that affects people of all ages, especially teenagers and young adults. The negative consequences

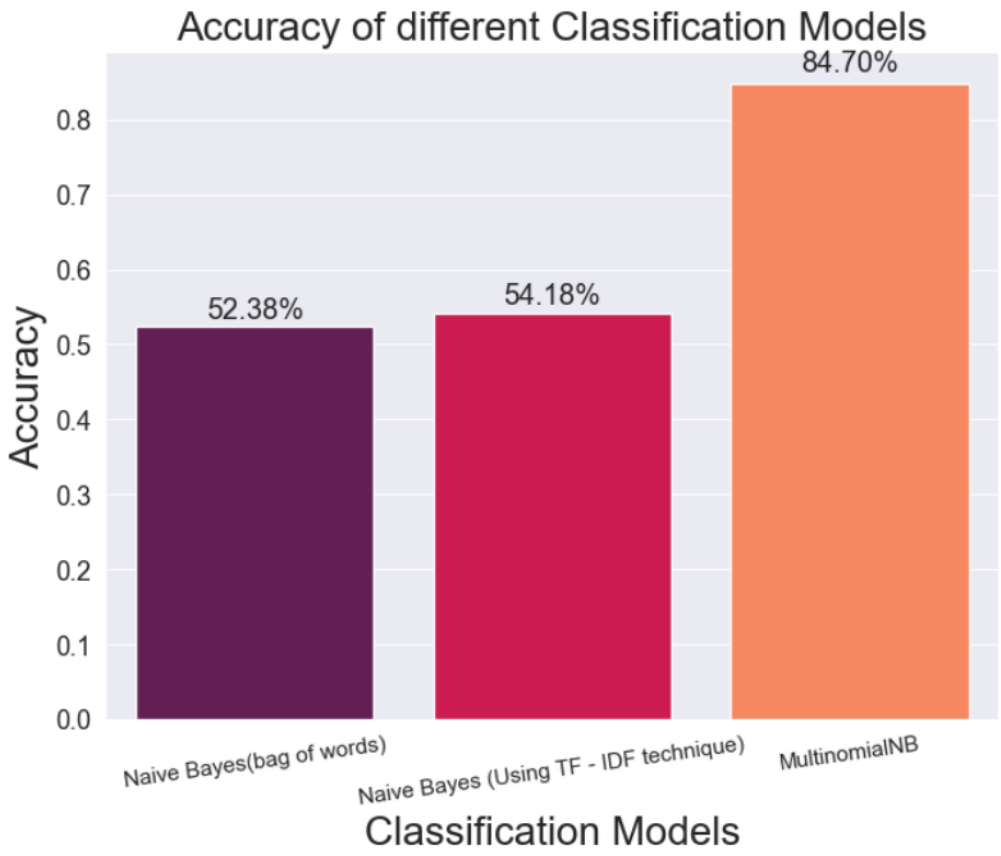

Figure 4: Accuracy of different classification models

of cyberbullying can be severe, including emotional distress, depression, and even suicidal thoughts/actions. As online communication continues to be an integral part of people's lives, it is critical to identify and combat cyberbullying early and effectively to mitigate its harmful effects.

The research objectives behind this study, the most effective methods are as follows: what are the detection of cyberbullying in text data and create the most suitable model to improve the accuracy and speed of detection. To answer this question, the study examined different approaches to cyberbullying detection, including natural language processing techniques such as tokenization, stop word removal, vectorization, and the use of machine learning algorithms for text classification. Analysis of studies showed that among the various methods used to detect cyberbullying, the Multinomial NB machine learning model based on Bayes' theorem gives the best results. It works by counting the frequency of words or tokens in each document and applies a multinomial distribution to estimate the probability of a document belonging to each category. Such models can identify and use subtle linguistic signals that indicate cyberbullying by identifying relationships between words and phrases in the text. Comparing the results of different methods sheds light on their strengths and weaknesses in detecting cyberbullying content. The results and conclusions obtained open new avenues for future research and the development of even more accurate tools and strategies for detecting cyberbullying.

6 Conclusion

The study is based on the cyberbullying_tweets dataset, which comprises various text comments. We conducted a series of experiments to find the model with the highest accuracy and the greatest generality. The following classifers are used in this work: Unsupervised Learning (VADER and TextBlob) and Supervised Learning (NB including the BOW,

TF-IDF and Multinomial NB). As a result of the study, the following results were obtained:
1) A dataset was collected, pre-processed and manually classified. Our results showed that the model has good accuracy, recall and F1-score for all three classes: gender, not_cyberbullying and religion. This confirms the effectiveness of our approach to classifying cyberbullying.
2) Next was a comparison of the accuracy of Naive Bayes models using various text representation techniques. We found that the MultinomialNB model achieved the highest accuracy of 0.847, which indicates its good performance.
This approach may be useful in identifying and addressing cyberbullying, helping to create a safer online environment. However, further research is needed to improve the accuracy and robustness of the model, especially regarding the subtleties and nuances of language that can affect the classification process. Overall, this study lays the foundation for future advances in the use of sentiment analysis and Naive Bayes classification to cyberbullying.

References

1. Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. IJCSNS Int. J. Comput. Sci. Netw. Secur, 19(3), 62.
2. Cyberbullying detection through sentiment analysis. (2020, December 1). IEEE Conference Publication | IEEE Xplore.
3. Dalvi, Fahim Sajjad, Hassan Durrani, Nadir Belinkov, Yonatan. (2020). Analyzing Redundancy in Pretrained Transformer Models. 4908-4926.
4. Dewi, C., Chen, R., Christanto, H. J., Cauteruccio, F. (2023). Multinomial naïve Bayes classifier for sentiment analysis of internet movie database. Vietnam Journal of Computer Science, 10(04), 485–498. https://doi.org/10.1142/s2196888823500100
5. Langos, C. (2012). Cyberbullying: The challenge to define. Cyberpsychology, Behavior, and Social Networking, 15(6), 285–289. https://doi.org/10.1089/cyber.2011.0588
6. Mangaonkar, Amrita Hayrapetian, Allenoush Raje, Rajeev. (2015). Collaborative detection of cyberbullying behavior in Twitter data. 611-616.
7. Maria Navin, J. R., Pankaja, R. (2016). Performance analysis of text classification algorithms using confusion matrix. International Journal of Engineering and Technical Research (IJETR), 6(4), 75-8.
8. Sendra, A. L., Carrasco, A., Martín, A. J., De Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023
9. Slonje, R., Smith, P. K., Frisén, A. (2013). The nature of cyberbullying, and strategies for prevention. Computers in Human Behavior, 29(1), 26–32. https://doi.org/10.1016/j.chb.2012.05.024
10. Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W. and Hoste, V., 2015. Automatic detection and prevention of cyberbullying. In International Conference on Human and Social Analytics (HUSO 2015) (pp. 13-18). IARIA.
11. Whittaker, E., Kowalski, R. M. (2014). Cyberbullying via social media. Journal of School Violence, 14(1), 11–29. https://doi.org/10.1080/15388220.2014.949377
12. Zhao, R., Zhou, A. and Mao, K., 2016, January. Automatic detection of cyber-bullying on social networks based on bullying features. In Proceedings of the 17th international conference on distributed computing and networking (pp. 1-6).
13. Zhang, L., Ghosh, R., Dekhil, M., Hsu, M. and Liu, B., 2011. Combining lexicon-based and learning-based.