



Article

Does Part of Speech Have an Influence on Cyberbullying Detection?

Jingxiu Huang , Ruofei Ding *, Yunxiang Zheng , Xiaomin Wu, Shumin Chen and Xiunan Jin

School of Information Technology in Education, South China Normal University, Guangzhou 510631, China; jimsow@m.scnu.edu.cn (J.H.); 20071206@m.scnu.edu.cn (Y.Z.); 20192821023@m.scnu.edu.cn (X.W.); 20192831014@m.scnu.edu.cn (S.C.); 20202821012@m.scnu.edu.cn (X.J.)

* Correspondence: ruofei@m.scnu.edu.cn

Abstract: With the development of the Internet, the issue of cyberbullying on social media has gained significant attention. Cyberbullying is often expressed in text. Methods of identifying such text via machine learning have been growing, most of which rely on the extraction of part-of-speech (POS) tags to improve their performance. However, the current study only arbitrarily used part-of-speech labels that it considered reasonable, without investigating whether the chosen part-of-speech labels can better enhance the effectiveness of the cyberbullying detection task. In other words, the effectiveness of different part-of-speech labels in the automatic cyberbullying detection task was not proven. This study aimed to investigate the part of speech in statements related to cyberbullying and explore how three classification models (random forest, naïve Bayes, and support vector machine) are sensitive to parts of speech in detecting cyberbullying. We also examined which part-of-speech combinations are most appropriate for the models mentioned above. The results of our experiments showed that the predictive performance of different models differs when using different part-of-speech tags as inputs. Random forest showed the best predictive performance, and naïve Bayes and support vector machine followed, respectively. Meanwhile, across the different models, the sensitivity to different part-of-speech tags was consistent, with greater sensitivity shown towards nouns, verbs, and measure words, and lower sensitivity shown towards adjectives and pronouns. We also found that the combination of different parts of speech as inputs had an influence on the predictive performance of the models. This study will help researchers to determine which combination of part-of-speech categories is appropriate to improve the accuracy of cyberbullying detection.

Keywords: machine learning; part of speech; cyberbullying detection



Citation: Huang, J.; Ding, R.; Zheng, Y.; Wu, X.; Chen, S.; Jin, X. Does Part of Speech Have an Influence on Cyberbullying Detection? *Analytics* **2024**, *3*, 1–13. <https://doi.org/10.3390/analytics3010001>

Academic Editors: Domenico Ursino and Jong-Min Kim

Received: 13 September 2023

Revised: 17 November 2023

Accepted: 28 November 2023

Published: 21 December 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As information technology becomes integrated into people's daily life, the convenience of communication may also lead to the emergence of cyberbullying [1]. Cyberbullying has become a social menace, tormenting children and young adults. Therefore, given the inherently interconnected nature of the Internet and the prevalence of cyberbullying on social networks, the development of models that can detect cyberbullying text has become increasingly urgent.

Textual features, such as keywords, document length, special characters, and part-of-speech (POS) tagging, are widely used in the development of automated cyberbullying detection models [2]. Among them, POS has been widely used as an input feature to improve the models' predictive performance, as they can explicitly display the model's syntactical bias, thus improving the performance [3]. Many studies have also used POS as an input to improve the performance of cyberbullying detection models [4–6].

Although using POS as an input feature can improve the prediction results of machine learning models for cyberbullying text, previous research still lacks an in-depth exploration of which specific part-of-speech combination can significantly improve the recognition

effect of cyberbullying text. Thus, the current study aims to investigate the influence of POS on the predictive performance of cyberbullying detection models.

We classified the collected text from Weibo into explicit bullying text, implicit bullying text, and non-bullying text based on predefined rules. Then, we conducted a sensitivity analysis of different POSs using SVM, random forests, and naive Bayes for each category separately. The research questions addressed in this study are:

- (1) Does the predictive performance of different models differ when POS is used as a feature input?
- (2) Do different models exhibit consistent sensitivity towards POS tagging?
- (3) Does the combination of POS features influence the predictive performance of the models?

2. Related Work

Research on cyberbullying-text detection focuses primarily on two aspects: algorithmic models and feature extraction for cyberbullying text. The aim is to enhance the performance of cyberbullying-text detection.

In terms of algorithmic models, the automatic detection of cyberbullying text is mainly achieved through using machine learning and deep learning. In the context of utilizing machine learning to detect cyberbullying, common algorithms include support vector machines, naive Bayes, random forests, etc. For instance, Sood employed an SVM for cyberbullying classification [7]. Nahar accomplished the task of classifying cyberbullying text by utilizing a variant of a SVM, a Fuzzy SVM, combined with information such as text and videos [8]. Hadiya utilized machine learning models, including random forest, SVM, KNN, and naive Bayes, to predict the impact of cyberbullying, respectively [9]. In the field of deep learning, models represented by Bert and CNN are commonly used for the detection of cyberbullying text. For instance, Zhou devised an attention-based B-LSTM method grounded in BERT [10]. Banerjee used a CNN with GLoVe embedding to achieve higher accuracy of cyberbullying detection [11].

On the other hand, increasing the features of cyberbullying text can also enhance the performance of cyberbullying-text detection. These features encompass keywords, part-of-speech (POS) tags, and document length, among others. Among them, POS can express the emotion of the text and, therefore, has gained many researchers' attention. Sri Nandhini proposed a cyberbullying detection system that utilized nouns, adjectives, and pronouns as data features and employed naive Bayes to classify cyberbullying text [4]. Drishya extracted POS tags such as nouns, adjectives, verbs, and pronouns, and combined naive Bayes and CNN for detecting cyberbullying in text [12]. However, Arnisha believed that distinguishing POS is an important step of understanding the meaning of a sentence, but verb and adverb tags are not very meaningful in identifying bullying features, and she fed nouns and adjectives into naive Bayes for classifying cyberbullying text with 88.6% classification accuracy [6]. And Huang classified cyberbullying text in a Twitter dataset by using textual features (expletive density and POS tagging) and social network features [13]. However, the improper selection of POS tags may lead to data redundancy, which hinders the prediction accuracy of the model. Previous research has pointed out the existence and influence of feature redundancy [14,15].

In summary, POS may have some influence on the cyberbullying prediction results of research into using POS as a text feature input to improve the effectiveness of cyberbullying-text detection. However, there is a lack of research on how to choose which part of speech is better to improve the performance of cyberbullying detection models. Therefore, this study provides a reference for improving the effectiveness of cyberbullying detection tasks by comparing the effects of different POS on random forests, support vector machines, and naive Bayes.

3. Methods

3.1. The Determination of Classification Criteria

Before collecting the data, it was necessary to perform the pre-classification and labeling of Weibo text. We divided it into three categories: explicit, implicit, and non-bullying, and they were defined as follows:

- Non-bullying text: This is unrelated to the incident and used to comfort the victim. It is divided into four main categories: comforting, mediating, judging rationally, and gaining attention.
- Implicit bullying text: This usually has complex forms of linguistic expressions, which hide bullying semantics behind instructive, persuasive, discursive, and subjective judgment as well as attributive and exaggerated statements. The cyberbully uses it to control the discourse, thus creating suppression and causing great psychological pressure on victims.
- Explicit bullying text: This generally has obvious negative tendencies and contains obvious offensive words, and is generally expressed through sarcasm, ridicule, insults, curses, threats, provocations, etc. [16,17].

3.2. Data Collection

The data for this study were obtained by crawling relevant comments from Sina Weibo. With a large user base and influence, Sina Weibo has become an important platform for public opinion in China, including politics, business, culture, and more. Although Weibo can shield offensive comments automatically, a significant number of unpleasant comments still exist. Particularly when it comes to contentious topics, a considerable number of comments can be found that contain vulgar or offensive sentences, which can be quite distasteful for readers.

We used a python program to collect text data of comments from Weibo on topics such as the economy and entertainment. Then, manual data labeling was performed. Each piece of textual data was encoded by two annotators, and the final consistency result in kappa value is 0.71. Then, the final labeling results were determined based on the majority principle.

For the purpose of sample balance, we selected 4000 pieces of data under each of the three categories, thereby making 12,000 pieces of data to form the dataset for the subsequent study.

3.3. Preprocessing of Data

After data acquisition, we removed stop words and irrelevant character sequences, such as usernames and “@” symbols. Then, we employed the widely used LTP segmentation method to segment the text [18]. The obtained results were further filtered to remove any incorrect words. In addition, we used Tencent Ailab [19] to vectorize all the words in each POS tag obtained after separation.

Based on the related work mentioned above, we observed that nouns, verbs, pronouns, and adjectives are frequently used as features in detecting cyberbullying text to enhance prediction performance. However, we consider that measure words may lead to exaggerated and untrue statements in cyberbullying text. Therefore, we used nouns (n), verbs (v), adjectives (a), measure words (m), and pronouns (r) as inputs for the next step of the training model.

3.4. Training Model

As detailed in this section, we took the word embedding vectors of all words under those five POS tags mentioned above as the input text features, and took the reduced-dimensional data as inputs to three classical machine learning models, random forest, SVM, and naive Bayes, respectively, for training and observed the results.

4. Results

4.1. The Performance Results of SVM

Table 1 shows the accuracy (acc), precision (pre), recall (re), and F1 (f1) score of the SVM processing results.

Table 1. Prediction performance of SVM.

Label	Index	POS					
		All	Noun	Verb	Adjective	Measure Word	Pronoun
All	acc	0.37	0.36	0.35	0.34	0.35	0.34
	pre	0.44	0.41	0.35	0.36	0.35	0.37
	re	0.38	0.36	0.35	0.34	0.36	0.33
	f1	0.35	0.33	0.30	0.27	0.28	0.23
Non-bullying text	pre	0.37	0.35	0.35	0.34	0.35	0.29
	re	0.5	0.41	0.51	0.69	0.77	0.12
	f1	0.40	0.37	0.42	0.45	0.48	0.17
Implicit bullying text	pre	0.35	0.36	0.35	0.4	0.35	0.35
	re	0.47	0.56	0.02	0.003	0.06	0.84
	f1	0.40	0.44	0.04	0.006	0.01	0.49
Explicit bullying text	pre	0.60	0.52	0.34	0.34	0.34	0.47
	re	0.12	0.12	0.51	0.33	0.31	0.02
	f1	0.21	0.19	0.41	0.33	0.35	0.05

The table displays the accuracy, recall, and F1 score of SVM in classifying different types of text. The model achieved an accuracy of 0.37, precision of 0.44, recall of 0.38, and F1 score of 0.35 when using word vectors from all parts of speech as inputs. Regarding the classification of parts of speech, the highest accuracy and F1 score were observed in nouns, while the lowest scores were seen in adverbs. With respect to non-bullying text, the model had a recall of 0.5 and an F1 score of 0.4, whereas the precision was merely 0.37. In the case of implicit bullying text, the accuracy and F1 score of the support vector machine (SVM) using all parts of speech as the input lay in the middle. For explicit bullying text, the precision of using all parts of speech as the input was 0.6; however, the recall and F1 score were low, at 0.12 and 0.21, respectively. The results indicate that the model suffers from low recall and a low F1 score, as evidenced by the high omission rate of cyberbullying texts. The model performed relatively well for non-bullying texts and explicit bullying texts, but not for the implicit bullying texts. Regarding the parts of speech, nouns yielded the best results, whereas adverbs gave the worst.

In Figure 1, the classification results of the SVM for different POS categories are visualized, and it is evident that the predicted labels obtained show clear spatial demarcation and are relatively clustered.

We found that when using SVM for bullying-text detection, the effect of different POS categories on the results was variable. Therefore, we used SVM for training via the traversal combination of POS, and the obtained results are shown in Figure 2.

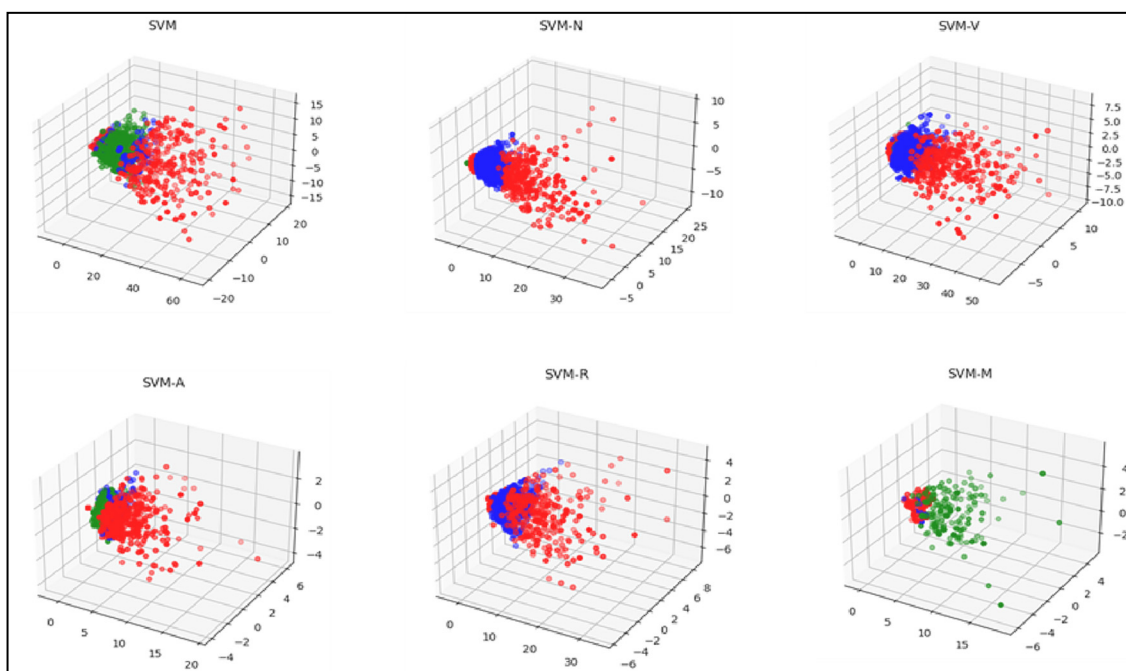


Figure 1. Prediction performance of SVM for each POS category.

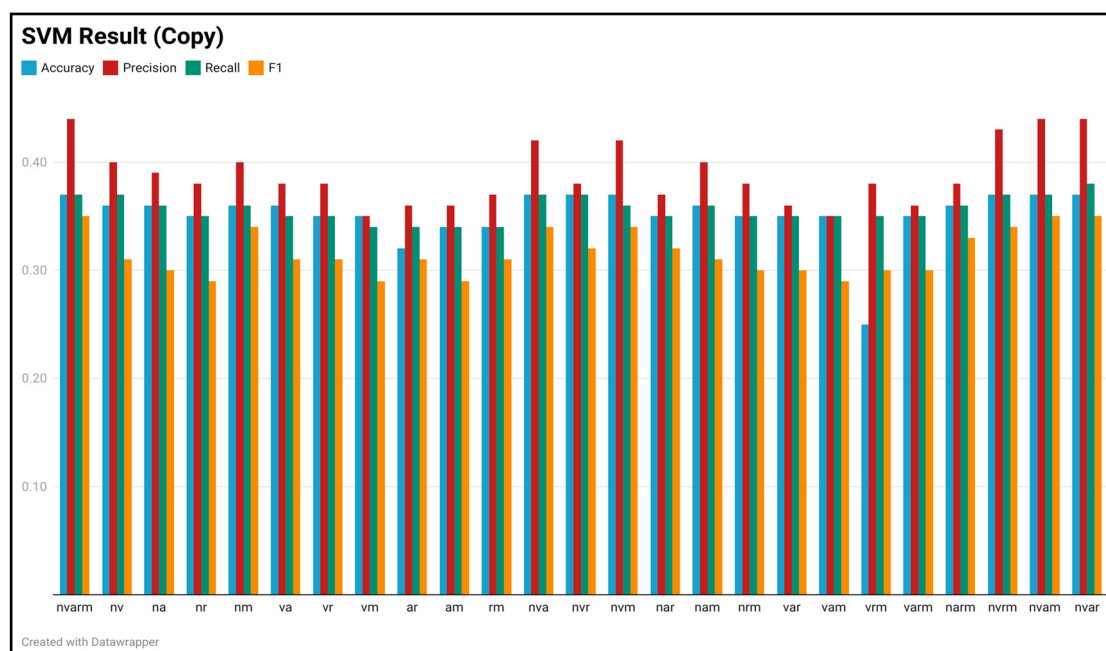


Figure 2. Prediction performance of SVM for each POS combination.

The graph shows the change in the prediction effect of the model when different combinations of POS were used as inputs. We can see that the results have high precision and low recall, which means that SVM gives high confidence but misses some correct results when determining cyberbullying texts via POS. The fit of the model did not improve significantly after the variety of POS was increased. In addition, different POS combinations were used as inputs, causing the model to predict performance differently. The best fit was achieved when nouns, verbs, adjectives, pronouns, and measure words were used as POS inputs. When *nm* (noun and pronoun), *nvar* (noun, verb, adjective, and pronoun), and *nvam* (noun, verb, adjective, and measure word) were used as inputs, the fitting results

were $F1 = 0.34$, $F1 = 0.35$, and $F1 = 0.35$, respectively, and the effect was similar to that of inputting all POS features. The worst fitting POS combination input was *vam* (verb, adjective, and measure word), with $F1 = 0.29$.

4.2. The Performance Results of Random Forest

Table 2 shows the accuracy (acc), precision (pre), recall (re), and F1 (f1) score of the random forest processing results.

Table 2. Prediction performance of random forest.

Label	Index	POS					
		All	Noun	Verb	Adjective	Measure Word	Pronoun
All	acc	0.61	0.55	0.55	0.43	0.44	0.36
	pre	0.61	0.55	0.56	0.49	0.45	0.40
	re	0.62	0.55	0.55	0.43	0.44	0.36
	f1	0.61	0.53	0.55	0.40	0.44	0.28
Non-bullying text	pre	0.58	0.48	0.48	0.40	0.37	0.34
	re	0.64	0.70	0.65	0.63	0.79	0.86
	f1	0.61	0.56	0.56	0.49	0.51	0.49
Implicit bullying text	pre	0.60	0.52	0.54	0.43	0.53	0.39
	re	0.44	0.27	0.38	0.27	0.17	0.05
	f1	0.50	0.35	0.44	0.33	0.26	0.09
Explicit bullying text	pre	0.70	0.67	0.65	0.54	0.58	0.46
	re	0.81	0.70	0.61	0.43	0.33	0.17
	f1	0.75	0.68	0.63	0.47	0.43	0.26

The random forest model achieved an accuracy of 0.61, a precision of 0.61, a recall of 0.62, and an f1 value of 0.61 when utilizing word vectors for all part-of-speech inputs. Analyzing the results from a lexical perspective shows that the model excelled in recognizing nouns and verbs, with an accuracy of around 0.55. However, it demonstrated less effectiveness in recognizing quantifiers and pronouns, with an accuracy lower than 0.4. Regarding non-bullying text, the indicators show a higher degree of stability, particularly in noun and verb recognition, with an F1 value of 0.56. For implicit bullying text, the performance of each index was more general, with the highest F1 score being approximately 0.5. For explicit bullying text, the model exhibited higher accuracy and recall, with an F1 score of 0.75 and 0.68, respectively. Considering the results in their entirety, the random forest model is less effective in detecting implicit bullying text.

The prediction results of the random forest are visualized in Figure 3. From the figure, we can see that the classification for the implicit bullying text (green points) is always the least regardless of the inputs, indicating that random forest is strict for the classification of implicit bullying, and it captures the features of the texts. The distribution of the predicted values in space shows that the distribution of the predicted results of the random forest in space is more consistent with the true results.

We then trained the random forests by traversing the combinations of POS, and the obtained results are shown in Figure 4.

From this figure, we find that the fitting effect of the random forest increased significantly after the number of the input features was increased from two to three, but the performance did not increase significantly when the POS categories were increased from three to four. Meanwhile, among all POS combinations, the POS combinations of *nva* (noun, verb, and adjective), *norm* (noun, verb, pronoun, and measure word), *nvam* (noun, verb, adjective, and measure word), *nvar* (noun, verb, adjective, and pronoun), and *nvarm* (noun, verb, adjective, pronoun, and measure word) were better fitted as inputs, with fitting effects of $F1 = 0.62$, $F1 = 0.62$, $F1 = 0.62$, and $F1 = 0.61$. The POS combinations of *am* (adjective and measure word), *rm* (pronoun and measure word) were the worst fitting, with $F1 = 0.44$ and $F1 = 0.45$, respectively.

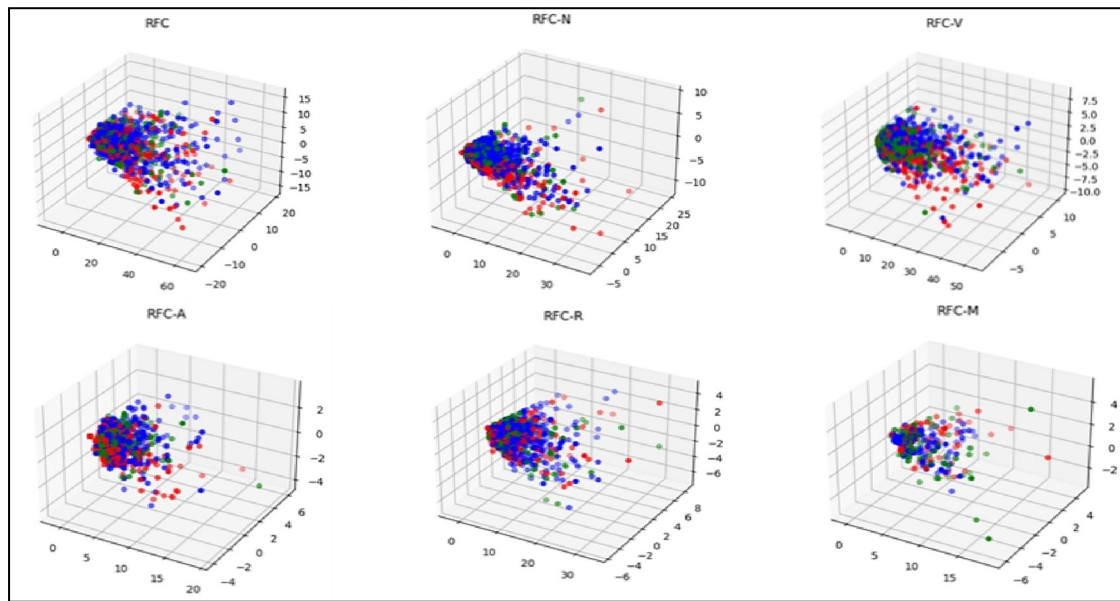


Figure 3. Prediction performance of random forest for each POS category.

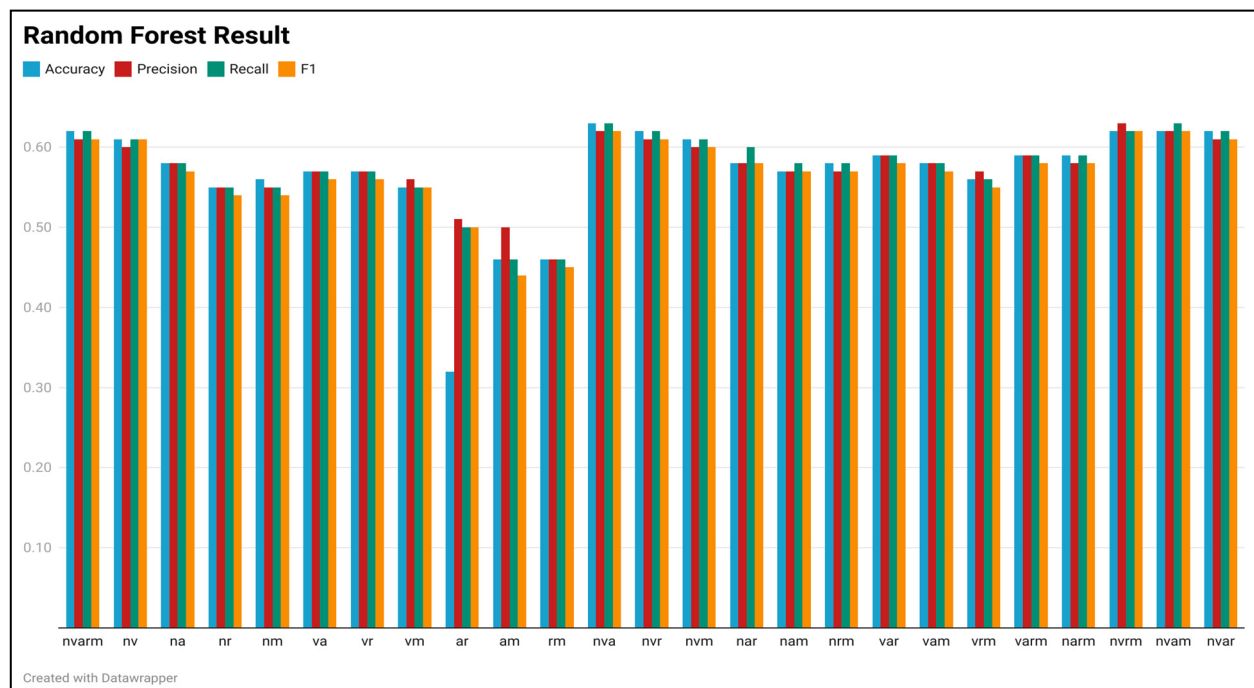


Figure 4. Prediction performance of random forest for each POS combination.

4.3. The Performance results of Naive Bayes

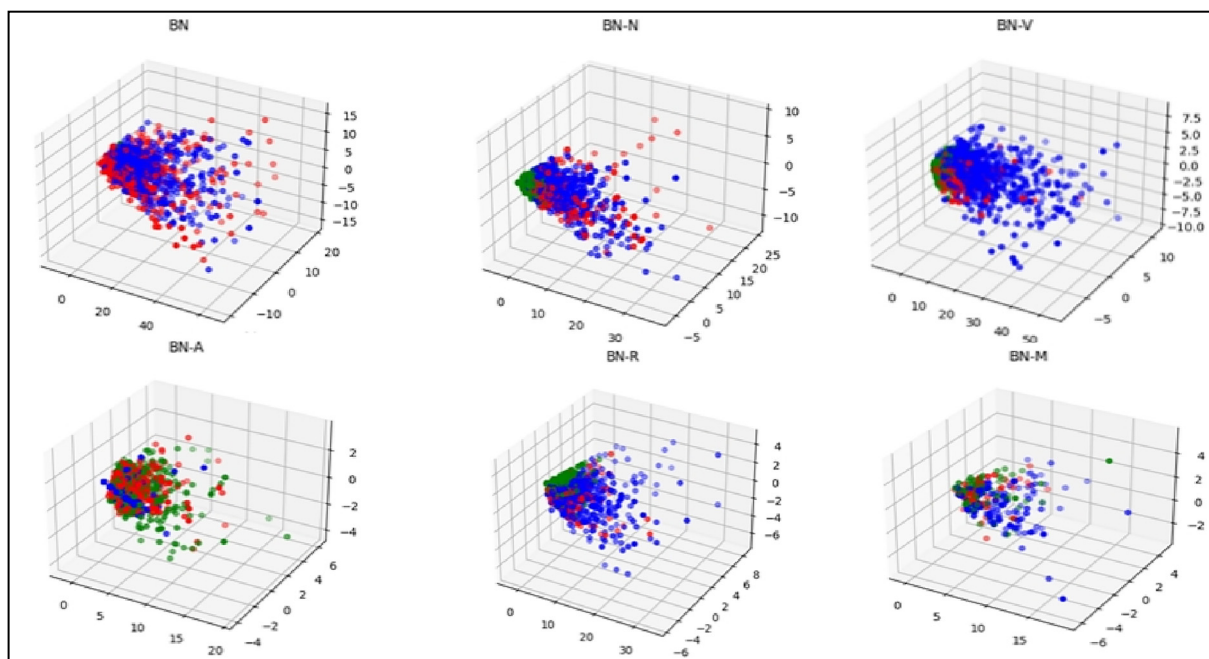
Table 3 shows the accuracy (acc), precision (pre), recall (re), and F1 (f1) score of the naive Bayes processing results. Overall, the measures of naive Bayes's accuracy and F1 value were approximately 0.3–0.4. Regarding non-bullying text classification, the precision was high, but recall was low. This suggests that the model tends to exclude non-bullying texts.

Table 3. Prediction performance of naive Bayes.

Label	Index	POS					
		All	Noun	Verb	Adjective	Measure Word	Pronoun
All	acc	0.35	0.39	0.37	0.32	0.37	0.35
	pre	0.35	0.42	0.40	0.32	0.40	0.37
	re	0.36	0.39	0.37	0.32	0.36	0.35
	f1	0.35	0.34	0.32	0.29	0.31	0.27
Non-bullying text	pre	0.30	0.41	0.42	0.32	0.40	0.34
	re	0.22	0.10	0.10	0.14	0.11	0.83
	f1	0.26	0.17	0.16	0.19	0.17	0.48
Implicit bullying text	pre	0.36	0.36	0.35	0.33	0.35	0.35
	re	0.44	0.76	0.74	0.22	0.76	0.04
	f1	0.40	0.49	0.48	0.26	0.48	0.07
Explicit bullying text	pre	0.39	0.50	0.42	0.32	0.43	0.42
	re	0.40	0.30	0.27	0.61	0.22	0.20
	f1	0.40	0.37	0.33	0.42	0.29	0.26

For implicit bullying text, the model had high recall but average precision, leading to false alarms. However, the model performed best when recognizing explicit bullying text. Regarding parts of speech, it was more accurate for nouns and verbs but less effective for quantifiers and pronouns. In summary, the model's ability to detect bullying text is limited, especially in recognizing implicit bullying.

Figure 5 shows the distribution of the prediction results in space for each POS for naive Bayes. In naive Bayes, when adjectives were the input feature, the prediction results were not the same as those for other single POSs. The distribution of predicted values in space for adjectives, nouns, and measure words tends to be more dispersive.

**Figure 5.** Prediction performance of naive Bayes for each POS category.

We tried to use various POS combinations as input features and tested the sensitivity of naive Bayes to different POS combinations. The results are illustrated in Figure 6.

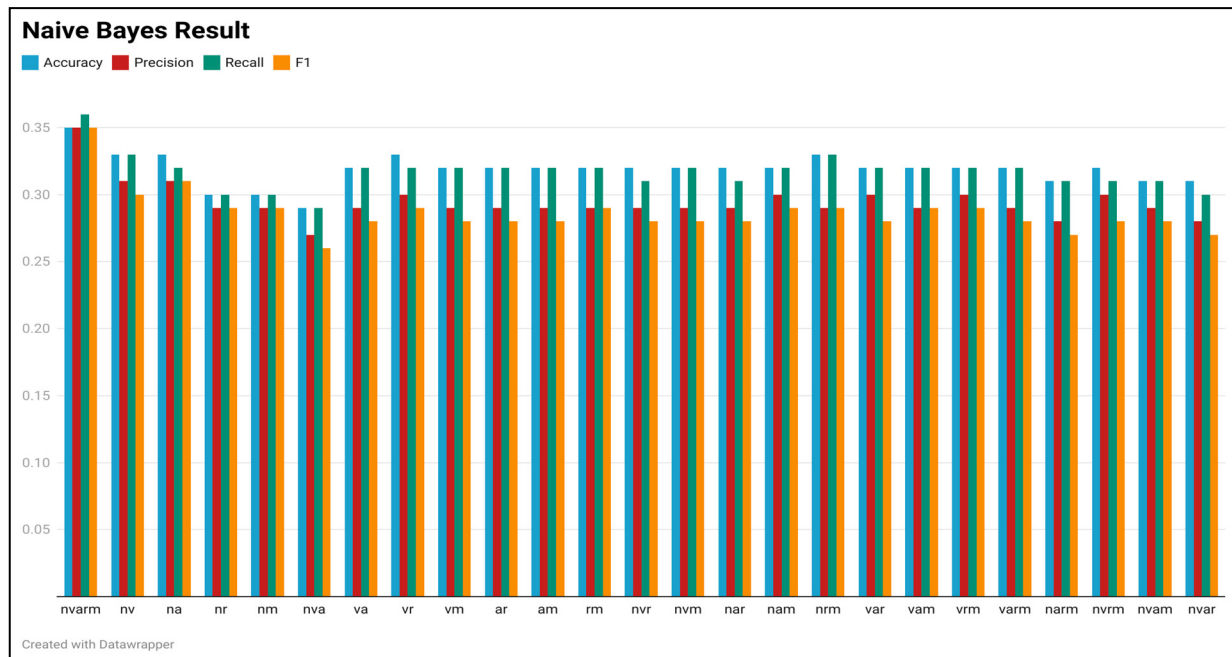


Figure 6. Prediction performance of naive Bayes for each POS combination.

We found that all POSs as inputs worked best. We speculate that this may be a problem caused by the fact that naive Bayes is not sensitive to POS. Compared to the other POS combinations, only the fits of *nvarm* (noun, verb, adjective, pronoun, and measure word) and *na* (noun and adjective) as inputs exceeded 0.3. The best POS combination fitting result was for *nvarm* (noun, verb, adjective, pronoun, and measure word) (F1 = 0.35); the worst was for *nvar* (noun, verb, adjective, and pronoun) (F1 = 0.27) and *nvm* (noun, verb, pronoun, and measure word) (F1 = 0.27).

In summary, from a POS perspective, the POSs to which naive Bayes is sensitive are nouns, verbs, and pronouns, in that order, and it is less sensitive to measure words and adjectives. In terms of different models, the sensitivity of the three models to POS in order from highest to lowest is: random forest, SVM, and Parsimonious Bayes. Meanwhile, we found that the combination of POS may affect the prediction effect of the model. For example, the POS combinations within the best SVM model are *nm* (noun and measure word), *nvar* (noun, verb, adjective, and pronoun), *nvam* (noun, verb, adjective, and measure word). And in the best random forest model, the POS combinations are *nva* (noun, verb, and adjective), *nvm* (noun, verb, pronoun, and measure word), *nvam* (noun, verb, adjective, and measure word), *nvar* (noun, verb, adjective, and pronoun), and *nvarm* (noun, verb, adjective, pronoun, and measure word). In naive Bayes, the combination with the best fit is *nvarm* (noun, verb, adjective, pronoun, and measure word).

5. Discussion

5.1. The Predictive Performance of Different Models Varies When POS Is Used as a Feature Input

We verified that when POS is used as a feature input, the predictive performance of different models varies. The data above demonstrate that, when predicting bullying texts using POS as a feature, random forest achieves the best performance ($F1_{\text{mean}} = 0.46$), followed by naive Bayes ($F1_{\text{mean}} = 0.32$), while SVM achieves the worst performance ($F1_{\text{mean}} = 0.29$). This indicates that there are differences in performance among machine learning models in predicting cyberbullying. Many studies have also had similar findings in the prediction of cyberbullying texts. Tarek discovered the same ranking of performance

in detecting Arabic online harassment text [20]. Hadiya conducted a comparison between random forest, SVM, KNN, naive Bayes, and MLP and found that, regardless of accuracy, precision, or F1 score, the predictive performance of random forest was significantly higher than that of naive Bayes [9]. Our research supports these findings. However, there are also some conclusions about the performance of SVM and naive Bayes that are contrary to our study, which are worthy of investigation. Karthik collected data by scraping comments from videos posted on YouTube and then annotated the data, ultimately finding that the predictive performance of SVM exceeded that of naive Bayes [21]. In summary, random forest is the most stable of the above models. We speculate that this may be due to the robustness of random forest, which can combine multiple decision trees, making it relatively insensitive to noise or outliers in the training data. In contrast, SVM and naive Bayes may encounter some issues under such circumstances.

5.2. Across Different Models, the Sensitivity of POS Is Consistent

In this study, nouns ($F1_{svm} = 0.33$, $F1_{RandomForest} = 0.53$, and $F1_{NaiveBayes} = 0.34$) and verbs ($F1_{svm} = 0.30$, $F1_{RandomForest} = 0.55$, and $F1_{NaiveBayes} = 0.32$) were the most predictive of all POSs, closely followed by measure words ($F1_{svm} = 0.28$, $F1_{RandomForest} = 0.44$, and $F1_{NaiveBayes} = 0.31$). This finding is consistent with those of other researchers [6,22,23]. However, Arnisha holds the opposite view, stating that verb tags have little significance in identifying bullying texts [23]. For adjectives ($F1_{svm} = 0.27$, $F1_{RandomForest} = 0.40$, and $F1_{NaiveBayes} = 0.29$), it was discovered that it had only slightly greater fitting effectiveness than using pronouns. It is suggested that among the three models above, the sensitivity of the model to adjectives is not good, probably because the word formation features of Chinese are more complex than those of English. We speculate that the possible reason for this result is that adjectives, as an expression of emotion, can have an ironic meaning.

For the pronoun ($F1_{svm} = 0.23$, $F1_{RandomForest} = 0.28$, and $F1_{NaiveBayes} = 0.27$) result, we infer that the reason for its low performance may be due to different definitions of cyberbullying or differences between Chinese and other languages. Pronouns are also an important aspect in recognizing cyberbullying text, as noted in Fatma's study. Fatma noticed that profane words, such as "f**k," may not necessarily be used for bullying. However, when combined with pronouns, such as in "f**k you," it becomes a statement of bullying [3]. Gauri also performed feature extraction in preprocessing, including nouns and pronouns, and determined and recorded the frequency of words in the text, which effectively detected bullying text [24]. An alternative explanation could be that this paper fails to differentiate between first-person pronouns and second-person pronouns. The impact of first-person pronouns on cyberbullying language differs from that of second-person pronouns, such as "I felt awful in race" and "You/She felt awful in race"; while the former is not a cyberbullying instance, the latter is. Since this paper does not distinguish between first and second person, this is one of the possible reasons for the low results for pronouns.

Meanwhile, we also found some interesting things, despite the varying levels of sensitivity towards parts of speech among the different models. Compared with adjectives, it is apparent that measure words consistently yield better fitting effectiveness as inputs for SVM, random forest, and naive Bayes models ($F1_{svm} = 0.28$, $F1_{RandomForest} = 0.44$, and $F1_{NaiveBayes} = 0.31$). Though different researchers have employed POS for text detection features in handling cyberbullying, few of them specifically emphasize the use of measure words as text features for the automatic detection of cyberbullying.

5.3. Using Different Combinations of POSs as Features Has an Influence on the Predictive Performance of the Models

POS as a text feature has garnered increasing attention among researchers in the field. This study aimed to compare the sensitivity of three classic machine learning models (naive Bayes, random forest, and SVM) to POS and discovered that different combinations of POSs may also impact the predictive performance of machine learning.

In this study, we found that more effective POS combinations for SVM prediction are *nm* (noun and measure word), *nvar* (noun, verb, adjective, and pronoun), and *nvam* (noun, verb, adjective, and measure word). For random forest, the more effective combinations are *nva* (noun, verb, and adjective), *nvrm* (noun, verb, pronoun, and measure word), *nvam* (noun, verb, adjective, and measure word), *nvar* (noun, verb, adjective, and pronoun), and *nvarm* (noun, verb, adjective, pronoun, and measure word). As for naive Bayes, the more effective combination is *nvarm* (noun, verb, adjective, pronoun, and measure word).

Therefore, when it comes to selecting which POS to use as an input for the features of cyberbullying texts, it is necessary to consider the model's sensitivity to POS. Although many researchers have used SVM, random forest, and naive Bayes with POS features to detect cyberbullying texts, they have ignored the effect of lexical combinations on the detection effectiveness of cyberbullying texts [4,6,25–27].

We also observed in different models that nouns, verbs, adjectives, and pronouns frequently appear as part of the optimal POS combinations that enhance prediction results. This may be because nouns, verbs, adjectives, and pronouns can reflect the features of cyberbullying text. Belal also stated that nouns, pronouns, and adjectives are considered the main features of content, while adverbs and verbs contribute additional information [28].

6. Conclusions and Limitation

The purpose of this study was to investigate the effect of part of speech (POS) as a text feature on the performance of cyberbullying detection models. We found that random forest had the best performance in predicting bullying texts through POS, followed by SVM, while naive Bayes had the worst performance. Among all POSs, nouns and verbs had the best predictive performance, followed by pronouns, while adjectives and measure words had the worst performance. The best POS combinations for SVM were *nm* (noun and measure word), *nvar* (noun, verb, adjective, and pronoun), and *nvam* (noun, verb, adjective, and measure word). For random forest, they were *nva* (noun, verb, and adjective), *nvrm* (noun, verb, pronoun, and measure word), *nvam* (noun, verb, adjective, and measure word), *nvar* (noun, verb, adjective, and pronoun), and *nvarm* (noun, verb, adjective, pronoun, and measure word). For naive Bayes, the best was *nvarm* (noun, verb, adjective, pronoun, and measure word).

It was proven that using POSs as text features for cyberbullying can have an impact on the predictive performance of the model. The predictive performance of the model may not be enhanced by simply inputting a fixed combination of POSs. Instead, it is necessary to take the sensitivity of different models to different POSs into account in order to better optimize their predictive performance.

However, this study has certain limitations. Specifically, only nouns, verbs, adjectives, quantifiers, and pronouns were selected as POS features, and the influence of other POS features on the models' effectiveness was not analyzed. We could take other POSs, such as adverbs and prepositions, into account and investigate their impact on the models' predictive performance in cyberbullying texts in future research. Additionally, other machine learning models' sensitivity to POS could be explored in the future.

Author Contributions: Conceptualization, J.H. and R.D.; methodology, R.D.; resources, Y.Z. and J.H.; writing—original draft preparation, R.D. and X.W.; writing—review and editing, S.C. and X.J.; visualization, R.D.; supervision, Y.Z.; project administration, J.H.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under grant no. 62207015 and Humanities and the Social Sciences Youth Foundation of the Chinese Ministry of Education under grant no. 22YJC880021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments, which greatly helped to improve this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Garaigordobil, M. Prevalencia y consecuencias del cyberbullying: Una revisión. *Psychol. Psychol. Ther.* **2011**, *11*, 233–254.
2. Sood, S.O.; Antin, J.; Churchill, E. Using crowdsourcing to improve profanity detection. In Proceedings of the 2012 AAAI Spring Symposium of the Conference, Stanford, CA, USA, 26–28 March 2012.
3. Elsafoury, F.; Katsigiannis, S.; Wilson, S.R.; Ramzan, N. Does BERT pay attention to cyberbullying? In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval of the Conference, Virtual, 11–15 July 2021.
4. Nandhini, B.S.; Sheeba, J.I. Cyberbullying detection and classification using information retrieval algorithm. In Proceedings of the International Conference on Advanced Research in Computer Science Engineering and Technology of the Conference, Unnao, India, 6–7 March 2015.
5. Rakib, T.B.A.; Soon, L.-K. Using the reddit corpus for cyberbully detection. In Proceedings of the 10th International Scientific Conferences on Research and Applications in the Field of Intelligent Information and Database Systems of the Conference, Dong Hoi City, Vietnam, 19–21 March 2018.
6. Akhter, A.; Uzzal, K.A.; Polash, M.A. Cyber bullying detection and classification using multinomial Naïve Bayes and fuzzy logic. *Int. J. Math. Sci. Comput.* **2019**, *5*, 1–12. [[CrossRef](#)]
7. Sood, S.O.; Churchill, E.F.; Antin, J. Automatic identification of personal insults on social news sites. *Am. Soc. Inf. Sci. Technol.* **2012**, *63*, 270–285. [[CrossRef](#)]
8. Nahar, V.; Al-Maskari, S.; Li, X.; Pang, C. Semi-supervised learning for cyberbullying detection in social networks. In Proceedings of the 25th Australasian Database of the Conference, Brisbane, QLD, Australia, 14–16 July 2014.
9. Hadiya, E.M. Cyber Bullying Detection in Twitter using Machine Learning Algorithms. *Adv. Eng. Manag.* **2022**, *4*, 1172–1184.
10. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics of the Conference, Berlin, Germany, 7–12 August 2016.
11. Banerjee, V.; Telavane, J.; Gaikwad, P.; Vartak, P. Detection of cyberbullying using deep neural network. In Proceedings of the 5th International Conference on Advanced Computing and Communication Systems of the Conference, Coimbatore, India, 15–16 March 2019.
12. Drishya, S.V.; Saranya, S.; Sheeba, J.I.; Devaneyan, S.P. Cyberbully image and text detection using convolutional neural networks. *Fuzzy Syst.* **2019**, *11*, 25–30.
13. Huang, Q.; Singh, V.K.; Atrey, P.K. Cyber bullying detection using social and textual analysis. In Proceedings of the 3rd International Workshop on Socially-Aware Multimedia of the Conference, Orlando, FL, USA, 7 November 2014.
14. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [[CrossRef](#)]
15. Koller, D.; Sahami, M. Toward optimal feature selection. In Proceedings of the International Conference on Machine Learning of the Conference, Bari, Italy, 3–6 July 1996.
16. Li, W. The language of bullying: Social issues on Chinese websites. *Aggress. Violent Behav.* **2020**, *53*, 101453. [[CrossRef](#)]
17. Caselli, T.; Basile, V.; Mitrović, J.; Kartoziya, I.; Granitzer, M. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In Proceedings of the Language Resources and Evaluation of the Conference, Marseille, France, 11–16 May 2020.
18. Che, W.; Feng, Y.; Qin, L.; Liu, T. N-LTP: An open-source neural language technology platform for Chinese. In Proceedings of the Empirical Methods in Natural Language Processing of the Conference, Punta Cana, Dominican Republic, 7–11 November 2021.
19. Song, Y.; Shi, S.; Li, J.; Zhang, H. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies of the Conference, New Orleans, LA, USA, 1–6 June 2018.
20. Kanan, T.; Aldaaja, A.; Hawashin, B. Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in Arabic social media contents. *Internet Technol.* **2020**, *21*, 1409–1421.
21. Dinakar, K.; Reichart, R.; Lieberman, H. Modeling the detection of textual cyberbullying. In Proceedings of the International AAAI Conference on Web and Social Media of the Conference, Barcelona, Spain, 21 July 2011.
22. Yuan, P.; Liu, W. The Study of Cyber-bullying from the Perspective of Critical Discourse Analysis: A Case Study of Tik Tok Comment Area Language. *Lit. Art Stud.* **2023**, *13*, 82–88.
23. Pascucci, A.; Masucci, V.; Monti, J. Computational stylometry and machine learning for gender and age detection in cyberbullying texts. In Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos of the Conference, Cambridge, UK, 3–6 September 2019.

24. Rao, G.; Goyal, M.; Wali, D.; Yadav, S. Cyber-Bullying Detection Using Machine Learning and Naïve Bayes and N-Gram Model. *Innov. Res. Technol.* **2021**, *8*, 648–651.
25. Nurrahmi, H.; Nurjanah, D. Indonesian twitter cyberbullying detection using text classification and user credibility. In Proceedings of the 1st International Conference on Information and Communications Technology of the Conference, Yogyakarta, Indonesia, 6–7 March 2018.
26. Fortuna, P.; Ferreira, J.; Pires, L.; Routar, G.; Nunes, S. Merging datasets for aggressive text identification. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying of the Conference, Santa Fe, NM, USA, 25 August 2018.
27. Yu, L.; Liu, H. Efficient feature selection via analysis of relevance and redundancy. *Mach. Learn. Res.* **2004**, *5*, 1205–1224.
28. Murshed, B.A.H.; Mallappa, S.; Saif, M.A.N.; Al-Ariki, H.D.E. DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform. *IEEE Access* **2022**, *10*, 25857–25871. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.