applied sciences

MDPI

Article

A Study of Discriminatory Speech Classification Based on Improved Smote and SVM-RF

Chao Wu [1,†], Huijuan Hu [2], Dingju Zhu [1,2,*,†], Xilin Shan [1], Kai-Leung Yung [3] and Andrew W. H. Ip [4]

[1] School of Software, South China Normal University, Guangzhou 510631, China; 2022024225@m.scnu.edu.cn (C.W.); 2022024220@m.scnu.edu.cn (X.S.)
[2] School of Computer Science, South China Normal University, Guangzhou 510631, China; 2021023279@m.scnu.edu.cn
[3] Department of Industrial and Systems Engineering, Hong Kong Polytechnic University, Hong Kong 999077, China; kl.yung@polyu.edu.hk
[4] Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SK M4Y1M7, Canada; wh.ip@polyu.edu.hk
* Correspondence: zhudingju@m.scnu.edu.cn
† These authors contributed equally to this work.

Abstract: The rapid development of the Internet has facilitated expression, sharing, and interaction on social networks, but some speech may contain harmful discrimination. Therefore, it is crucial to classify such speech. In this paper, we collected discriminatory data from Sina Weibo and propose the improved Synthetic Minority Over-sampling Technique (SMOTE) algorithm based on Latent Dirichlet Allocation (LDA) to improve data quality and balance. And we propose a new integration method integrating Support Vector Machine (SVM) and Random Forest (RF). The experimental results demonstrate that the integrated model exhibits enhanced precision, recall, and F1 score by 6.0%, 5.4%, and 5.7%, respectively, in comparison with SVM alone. Moreover, it exhibits the best performance in comparison with other machine learning methods. Furthermore, the positive impact of improved SMOTE and this integrated method on model classification is also confirmed in ablation experiments.

Keywords: discrimination speech; latent Dirichlet allocation; support vector machine; random forest; integration method

Citation: Wu, C.; Hu, H.; Zhu, D.; Shan, X.; Yung, K.-L.; Ip, A.W.H. A Study of Discriminatory Speech Classification Based on Improved Smote and SVM-RF. *Appl. Sci.* **2024**, *14*, 6468. https://doi.org/10.3390/app14156468

Academic Editor: Douglas O'Shaughnessy

Received: 9 June 2024
Revised: 21 July 2024
Accepted: 22 July 2024
Published: 24 July 2024

1. Introduction

As the Internet continues to gain popularity and undergo rapid development, the number of Internet users continues to increase. The 53rd Statistical Report on the Development of the Internet in China [1] indicates that the number of Internet users in China reached 1.092 billion by December 2023. This represents a 24.8 million increase compared to December 2022, with the Internet penetration rate reaching 77.5%. The Internet has become an indispensable tool in the daily lives of most people, offering a multitude of conveniences in the domains of learning, work, and communication. Furthermore, it has facilitated the growth of social networks. The proliferation of online platforms and social media has led to a surge in the number of individuals sharing their daily routines, expressing their opinions, and engaging in online communication on social networks. Consequently, a considerable number of posts, tweets, comments, and replies are generated on a daily basis in the form of short videos, live broadcasts, online meetings, online classes, forums, blogs, and other online platforms. These online statements represent the opinions and emotions of online users, which may be positive, neutral, or negative. The low threshold of social media use, high degree of openness, and certain degree of anonymity facilitate the emergence of harmful and undesirable speech in online discourse, which in turn increases with the rapid growth of social media use. The United Nations Educational, Scientific and Cultural Organization defines bad speech as "speech that promotes and incites harm based

on the specific goals of a certain society or population group". Discriminatory speech is an example of bad speech, which refers to the speech that demeans, denigrates, insults, or unfairly treats an individual or group based on their race, gender, sexual orientation, religious belief, disability, age, nationality, and other characteristics [2]. Discriminatory speech may not only cause mental harm to the victims, but also cause social conflicts and undermine social stability. There are various kinds and forms of discriminatory speech, and the common forms are ridicule, sarcasm, slanders, insults, demeaning, and so on. By classifying discriminatory speech, it can detect and reduce discriminatory speech, protect the mental health, privacy, and dignity of individuals and specific groups, ensure social harmony and stability, and maintain a healthy, civilized, and good network environment.

In this paper, we gather discriminatory speech data from Sina Weibo and propose a method for classification using improved SMOTE and SVM-RF. We utilize LDA to construct a topic model for relevant data filtering and apply SMOTE to address imbalanced datasets, effectively reducing noise and imbalance issues. We compute weights based on SVM and RF classification accuracy, then aggregate weighted SVM and RF outputs to enhance precision, recall, and F1 score on the test set by 6.0%, 5.4%, and 5.7%, respectively, compared to standalone SVM. The contributions of our work are as follows:

- Produce a dataset of discriminatory speech in Chinese based on Sina Weibo. Currently, publicly available datasets of discriminatory speech are mainly based on foreign social media platforms such as Twitter, Facebook, and Youtube, while the number of datasets for Chinese is quite limited. The newly created dataset will provide valuable experience and reference for the classification of Chinese discriminatory speech.
- Propose an improved SMOTE algorithm based on LDA, which is able to obtain high-quality data more relevant to the topic and generate new data based on these high-quality data, thus reducing the interference of noisy data and also alleviating the problems caused by the imbalance of the dataset.
- Propose a new integration method to integrate SVM and RF. The method uses the correctness of the classification results of both to calculate their respective weights, and then weighs and sums the classification results of both to obtain the final classification results of the model, which improves the numerical stability and enhances the performance of the model.

The remainder of this paper is as follows: Section 2 introduces the related work. Section 3 presents the related technology used, the data collection and pre-processing, the improved SMOTE algorithm based on LDA, and the construction and integration of the model. Section 4 explains the experimental evaluation and result, and finally Section 5 summarizes the study.

2. Related Work

This section focuses on the related work in three areas: topic modeling, data imbalance, and discriminatory speech. We first review related work on topic modeling and the application of LDA to topic modeling, followed by the related work on discriminatory speech classification and addressing data imbalance.

2.1. Topic Modeling

LDA [3] is an unsupervised generative probabilistic approach for corpus modeling. It assumes that each document can be represented as a probability distribution of latent topics and that the topic distributions of all documents share a Dirichlet prior. Each latent topic is represented as a probability distribution over words, and the word distributions of topics also share a Dirichlet prior [4]. LDA-based topic modeling methods have been applied to natural language processing, text mining, social media analysis, and other fields [5]. In order to identify hate speech in the Twitter corpus, Xiang et al. used a semi-supervised learning method, the bootstrapping technique, and linguistic rules to classify unlabeled data, and finally used LDA to classify based on statistical topic model [6]. Zhong Yi-bo et al. used the LDA-based model to extract text topics and analyze topic head words, and

completed the research of agricultural products e-commerce text classification [7]. When solving the sparsity problem of short text data, researchers have pointed out that traditional LDA may have limitations. For example, Lin et al. proposed the Dual-sparse Topic Model (Dsparse TM), which solves the sparsity problem by applying a "Spike and Slab" prior to separate the sparsity and smoothness of the distribution of document topics and topic terms [8]. The Biterm Topic Model (BTM) proposed by Cheng et al. focuses on capturing the relationship between topics by using biterm information in short texts, and introduces two online BTM algorithms to deal with short texts [9]. Sun et al. proposed a novel supervised pseudo-document-based maximum entropy discrimination latent Dirichlet allocation model (PSLDA), which combines the maximum entropy method and the pseudo-document technique to solve the sparsity of word occurrence [10]. However, LDA is a relatively simple and easy-to-interpret model based on the bag-of-words model and Dirichlet priors, which makes its parameter inference and topic explanation relatively intuitive and easy to understand. Moreover, the inference process of LDA is relatively simple, and there are mature optimization methods and tool libraries that can be used. In addition, LDA, as a model that has been widely studied and applied, may have a higher acceptability and widespread use in academia. To balance the need for complexity, interpretability, and use case, LDA might be an option [11].

2.2. Discriminatory Speech Classification

Discriminatory speech classification belongs to the text classification problem in natural language processing [12]. In recent years, some researchers have used machine learning methods to classify discriminatory speech. Burnap et al. employed an N-gram feature engineering approach to extract features from a pre-written dictionary of hate words and then fed the feature vectors into an SVM classifier, resulting in a final F-score of 0.67 [13]. Greevy et al. used a supervised machine learning approach to classify racist text datasets. To convert raw text into numeric vectors, the authors used binary feature extraction techniques and bag-of-words (BOW) feature representation techniques. This resulted in an accuracy of 87% [14]. Nobata et al. used machine learning methods to detect abusive language in online user content, and the authors adopted character-level N-gram feature representation techniques to represent features and feed them into an SVM classifier for processing, which obtained an F-measure of 77%, according to their research results [15]. Malmasi et al. used supervised classification methods to classify hate speech in social media, and the authors used character-level, word-level N-gram, and word Skip-gram methods to obtain features, which were then fed into an SVM classifier, and the experimental results showed the highest accuracy of 78% [16]. Combined with dictionary-based methods, bag-of-words methods, and N-gram methods, SVM has shown good performance in text classification tasks. Khan et al. aimed to investigate the application of machine learning techniques in sentiment analysis and conducted experiments with SVM and RF, which showed that both have considerable accuracy and are superior among machine learning algorithms [17]. Sahoo et al. used the standard hate speech dataset to compare their classification effects on three feature engineering techniques and eight machine learning algorithms [18]. In the experimental results of these studies, SVM and RF performed well, and SVM achieved the best performance, possibly because it could effectively deal with high-dimensional data by using hyperplanes and deal with nonlinear problems by using its kernel function. RF performs slightly worse than SVM, but higher than other algorithms.

2.3. Imbalanced Data

Imbalanced datasets are one of the most challenging problems in data mining and machine learning. Imbalanced data can cause a classifier to be biased toward the larger class, as the classifier will make predictions toward the majority class in order to minimize the overall classification error. The main approaches to deal with the data imbalance problem are algorithm-level and data-level approaches [19]. The algorithm-level approach is to modify the steps of the classification algorithm to achieve the goal of balanced data,

such as Liu et al. modifying the DT (Decision Tree) classifier [20] and Imam et al. modifying the SVM classifier [21]. Data-level methods rebalance the minority and majority classes by modifying the number and content of the dataset, usually through undersampling or oversampling, such as the SMOTE method. SMOTE is a commonly used oversampling method for generating new data [22], which generates new data based on the similarity between minority class data. Early et al. used filtering feature selection based on fast correlation to remove irrelevant features in the field of disease classification, and used the SMOTE method to balance the dataset to improve accuracy. The results showed that the RF classifier performed well, achieving 99.35% AUC value and 97.81% accuracy [23]. Singgalen et al. evaluated the performance of DT and SVM combined with SMOTE, respectively, in their study on the effectiveness of emotion analysis models. Indicators such as accuracy, precision, and F1 score all showed that these models could effectively distinguish positive and negative emotions, and SVM obtained 98.91% accuracy [24]. Their studies demonstrated the effectiveness of SMOTE in combination with SVM, RF, and in dealing with imbalanced dataset problems. There are also works to address data imbalances in discriminatory speech and related areas. For instance, Rathpisey and Adji used four resampling methods, including Random Oversampling (ROS), SMOTE, Adaptive Synthetic (ADASYN), and Random Undersampling (RUS) to handle inequality of class distribution in a hate speech dataset [25]. Sanya and Suadaa used the SVM model, SVM with various resampling methods, and the fine-tuned models to detect Indonesian hate speech posts in several imbalanced conditions, and the combination of SVM and SMOTE models performed the best in handling imbalanced problem based on the results [26]. Lu et al. proposed a novel dual contrastive learning (DCL) framework for hate speech detection, which alleviated the problem of data imbalance by integrating the focal loss into the dual contrastive learning framework [27].

In previous studies, both SVM and RF have been shown to have advantages in text classification and sentiment analysis tasks, as well as the effectiveness of SMOTE in dealing with imbalanced dataset problems. However, the current research still faces the challenge of how to deal with data imbalance and improve classification accuracy more effectively. Further, the discussion of discrimination in Chinese datasets is missing, which is also the problem that this paper is committed to solve.

3. Data and Methods

In order to address the above issues, we propose a discriminatory speech classification method based on improved SMOTE and SVM-RF. The specific steps are as follows:

1. Collect data about discriminatory remarks on Sina Weibo and perform preprocessing, including data cleaning, text cleaning, word segmentation, and the removal of stopwords.
2. Construct an LDA-based topic model to filter out the data related to the topic.
3. Use SMOTE to synthesize new data based on existing data.
4. Classify using SVM and RF and calculate the integration weights using the correctness of the respective classification results.
5. Calculate the weighted sum of the respective classification results of SVM and RF to obtain the final classification result.
6. Perform experiments and evaluation of the model based on machine learning methods.

The overall framework is shown in Figure 1. From the framework, it can be seen that, first, the data are collected from Sina Weibo, then the data go through pre-processing and enhancement sessions. Next, the data are fed into the model for training, followed by testing and optimization of the model, and finally the optimized model classifies discriminatory speech.

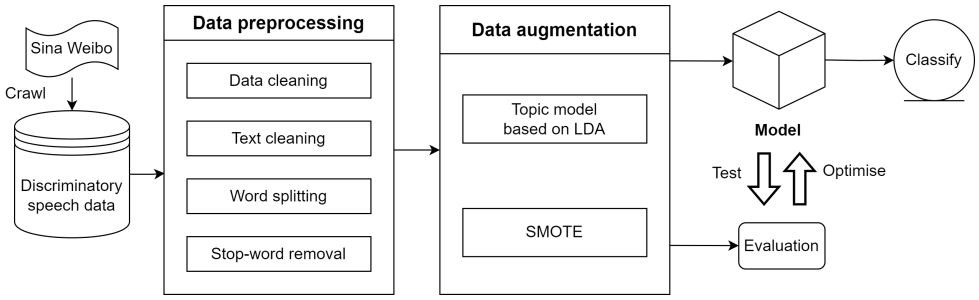

Figure 1. Overall framework.

3.1. Data

3.1.1. Data Collection

Sina Weibo is one of the largest and most widely used social media platforms in China, similar to the foreign social media, Twitter, where users can post text, pictures, videos, and other content, interact and communicate with other users, and the platform is publicly accessible. Despite the existence of content monitoring mechanisms on Weibo, there are still many users who post discriminatory remarks by some means to avoid monitoring, so in this study, we crawled discriminatory and non-discriminatory remarks on Sina Weibo as a dataset.

By analyzing the composition and characteristics of Sina Weibo's microblogs, it is found that among the three ways of accessing Sina Weibo, the mobile website m.weibo.com is the most convenient and easy way to access and analyze the data of Sina Weibo [28]. The textual information of Sina Weibo is stored in the webpage, which can be obtained through web requests. Requests 2.32.3 is an HTTP client library for Python 3.10, which supports a number of HTTP characteristics. By setting web request parameters such as cookies, headers, and keywords, the Requests library is used to simulate a browser sending requests to the Sina Weibo server to obtain microblog web page data. By analyzing the data in JSON format, it is possible to obtain the text of microblog topics, microblog content, microblog comment content, and save them locally.

We set up keywords to search for microblog topics and microblogs that contain keywords, as well as the comments of the microblogs. In order to protect users' privacy, we only collected the content of posts and comments, and other information were not collected, such as the microblog's ID, user's ID, posting time, and so on.

We collected a total of 26,542 posts and comments data, and the size of the dataset after removing duplicates and gaps is 25,726, of which 73.1% are discriminatory and 26.9% are non-discriminatory. The large difference in sample size between different labels indicates a serious sample imbalance in the collected data.

3.1.2. Data Preprocessing

Data preprocessing mainly includes data cleaning, text cleaning, word splitting, and stopwords removal.

Data cleaning is the processing of duplicate and abnormal data. Too long microblogs need to separately request the full text, which will lead to the acquisition of incomplete text; in addition, too long microblogs will have an impact on the subsequent analysis. These incomplete and too long data are deleted and processed. For duplicate microblogging content such as repeated posts by the user, only one microblog will be retained, and the rest are deleted.

Text cleaning is the processing of microblog text content. This includes removing special symbols, special characters, emoji, HTML tags, punctuation, URLs, etc. from microblog text by regular expression matching.

Word splitting is an operation performed on the cleaned plain text data, using the Jieba Chinese segmentation tool to separate words one by one, which is convenient for subsequent word frequency statistics, vectorization, and model training.

Stopwords removal is the removal of meaningless words [29], such as prepositions, auxiliaries, conjunctions, etc., by matching them using a stopwords list.

3.1.3. Text Length Distribution Analysis

The preprocessed data only retain text content containing Chinese characters, numbers, and English. According to Figures 2 and 3, we can see that the number of short texts is higher in discriminatory data, while the number of long texts is higher in non-discriminatory data, and most of the lengths of the texts are between 0 and 125 characters, among which the number of long texts with a length greater than 75 characters is relatively small, and the number of texts with a length greater than 125 characters is even smaller, probably because most of the data are microblog comments with fewer words on the topic of discrimination, and a small portion of them are microblogs with a larger number of words. Therefore, when dealing with the data length, we can set the data length to 75 by padding or truncation to keep the data length uniform in order to improve the efficiency of the model and avoid the problem of data distortion.

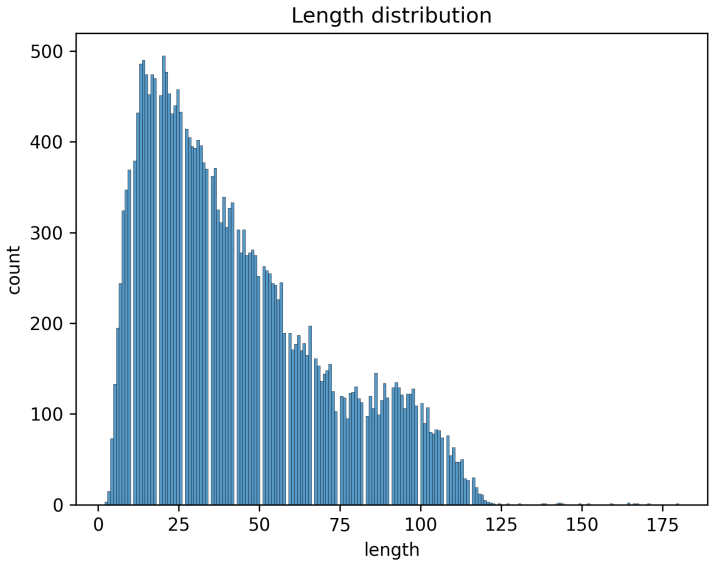

Figure 2. Length distribution.

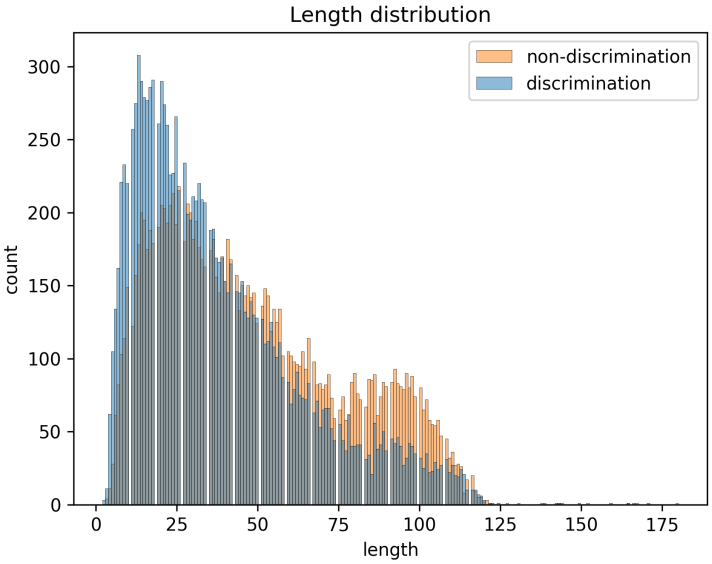

Figure 3. Length distribution of different labels.

3.2. Improved SMOTE Algorithm Based on LDA

By analyzing the above text length distribution and original data label distribution, we found that most of the microblog comments are short texts, with uneven data quality and unbalanced number of positive and negative samples. Therefore, we propose an improved SMOTE algorithm based on LDA, aiming to obtain high-quality data that are more relevant to the topic and reduce the interference caused by noisy data. At the same time, new data are generated based on the high-quality data to alleviate the problems caused by the imbalance of the dataset. Specifically, we integrate text data from both discriminatory and non-discriminatory labels, use Term Frequency-Inverse Document Frequency (TF-IDF) to extract features from the text data, and use the topic model constructed by the LDA algorithm to exploit these features to discover representative topics. Then, we filter out the high-quality data related to these topics, and finally use SMOTE to synthesize new data based on the high-quality data to balance the dataset, as shown in Algorithm 1.

Algorithm 1: Improved SMOTE based on topic model

Data: Text data D, Number of synthetic samples to generate N
Result: Synthetic samples S
1. $feature$ = TFIDF(D); // Extract feature using TF-IDF
2. $topics$ = LDA($feature$); // Find representative topic using LDA on data
3. $related\_data$ = get_related_data(D, $topics$); // Filter out data related to topics
4. S = SMOTE($related\_data$, D, N); // Employ SMOTE to oversample a specified number of samples
5. return S

Regarding the process of LDA filtering, it is as follows: We first cluster all textual information using the k-means algorithm and determine the optimal number of clusters based on the performance, and then use the optimal number of clusters as the number of topics to extract topic information for each category before filtering. As shown in Figure 4, the abscissa is the number of clusters, and the ordinate is the silhouette score, which combines the closeness within a cluster and the separation between different clusters and can measure the quality of the clustering results. A value closer to 1 means the data are better clustered, so the optimal number of clusters is 6. Then, wet use LDA to obtain the topic information of different categories and vectorize them, then calculate the similarity between the vectorized data of each category and the corresponding topic information, and retain the texts with similarity higher than 0.4. When the threshold is set low, it is easy to introduce low-quality data leading to performance degradation, and when the threshold is set too high, the diversity of samples is reduced and the robustness of the model is weakened. The results of the experiment are shown in Figure 5.

The reason why LDA helps to identify discrimination by filtering the dataset is that, first, it is able to capture the topic information in the documents through a probability distribution, thus identifying the main topics and concepts in the documents. Secondly, keeping high-quality documents with high similarity to the topic helps to filter out off-topic or irrelevant documents and reduce noisy data.

Table 1 shows the distribution of topic words obtained by constructing topic models for discriminatory and non-discriminatory texts using LDA separately, where the weights indicate the probability of selecting each word under that topic. The distribution is obtained by counting all the words in the document labeled as discriminatory or non-discriminatory. It can be seen that the data cover topics such as region, race, and gender, and that the topics labeled in both categories are related to “discrimination” and “black”.

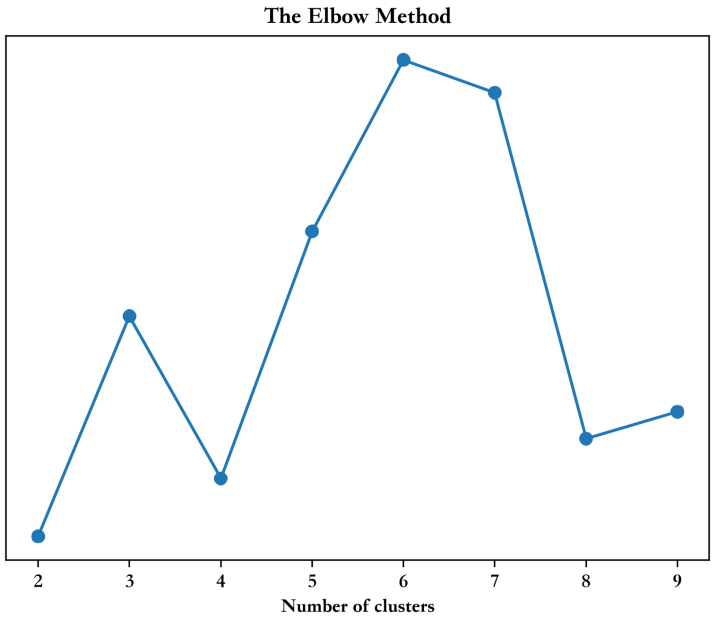

Figure 4. Clustering quality with different number of clusters.

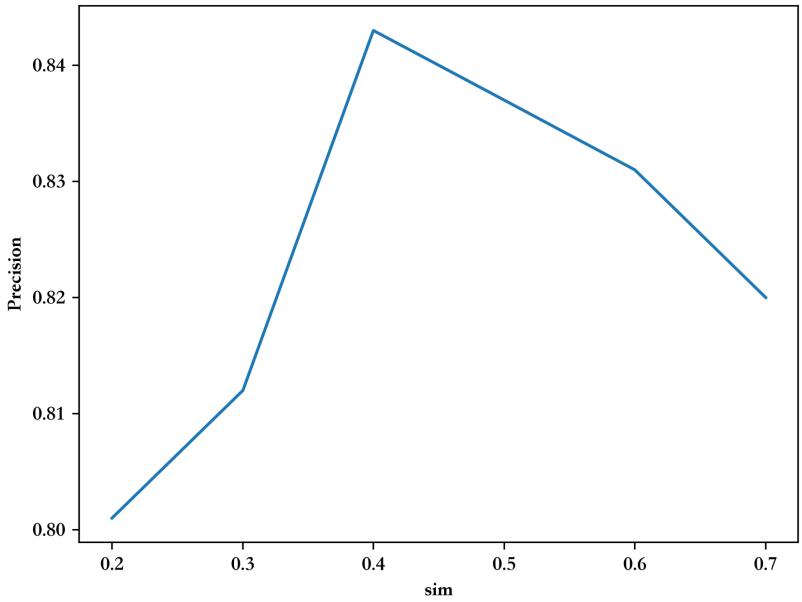

Figure 5. Precision at different similarity thresholds (sim).

Table 1. Distribution of topic words.

Label	Topic Word
Non-discrimination	0.011 × "Disgusting" + 0.010 × "Dongbei" + 0.010 × "Black" + 0.009 × "Discrimination" + 0.009 × "Henanese" + 0.008 × "Shanghai" + 0.008 × "Henan" + 0.007 × "Region"
Discrimination	0.041 × "Black" + 0.027 × "Discrimination" + 0.018 × "China" + 0.009 × "Female" + 0.008 × "Male" + 0.008 × "USA" + 0.008 × "None" + 0.007 × "White"

As shown in Table 2, the size of the original dataset is 25,726, the size of the dataset after LDA-based topic model filtering is 16,726, and the size of the dataset after SMOTE oversampling is 29,850. The discriminatory data accounted for 47.6% of the total data, the

non-discriminatory data accounted for 52.4% of the total data, and the number of samples with different labels was balanced after the processing.

Table 2. Number of data.

	Original Data	After Filtering by Topic Model	After Oversampling by SMOTE
Number	25,726	16,726	29,850

3.3. Model Construction and Integration

Figure 6 illustrates the tasks of model construction and optimization, including dataset division, model construction, and model evaluation and optimization. These are described as follows:

1. Split dataset: the preprocessed dataset is divided into training and test sets, which are used for model training and model evaluation, respectively.
2. Training and integration: the training data are applied to SVM and RF classifiers, respectively, to obtain the model classification results and calculate the correct rate. The correct rate is used to calculate the integrated weights of the model as shown in Equations (1) and (2), and finally the model classification results are weighted and summed to obtain the final classification results.
3. Evaluation and optimization: the accuracy, recall, and F1 score of the model are calculated to evaluate and optimize the model.

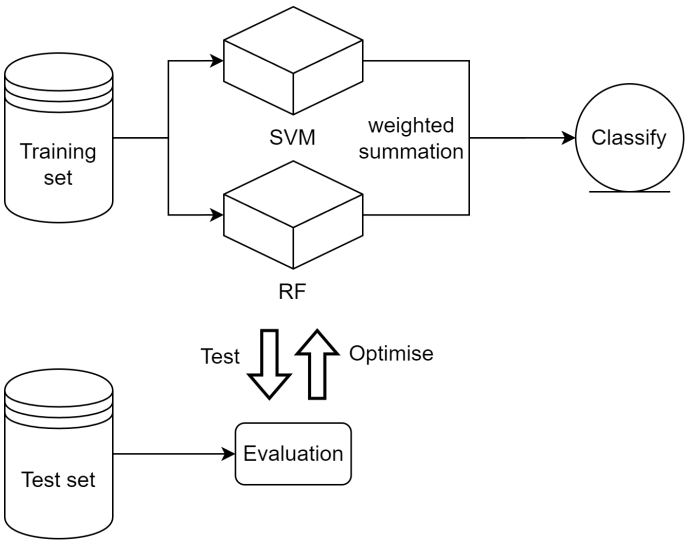

Figure 6. Model construction and optimization.

The formula for calculating the weights is as follows:

Firstly, the index value of each element in the correct rate Z of the classification results of SVM and RF is calculated, and in order to improve the numerical stability, the maximum value in the vector Z is subtracted first, and then the exponential operation is performed to avoid the overflow of the exponential value, as in Equation (1)

$$exp(z_i) = e^{z_i - max(z)}, i = 1, 2, ..., n \quad (1)$$

Then the exponential values are normalized as in Equation (2)

$$softmax(z_i) = \frac{exp(z_i)}{\sum_1^n exp(z_i)}, i = 1, 2, ..., n \quad (2)$$

where n is the number of label category.

Specifically, the integration method does not use the stacking method. SVM and RF are used to obtain the probabilities of discrimination and non-discrimination predicted by a sample, respectively, and then the final probabilities are obtained by weighted summation. If the probability of discrimination is large, the text will be judged as discriminatory speech, where the weights are obtained by Equations (1) and (2).

The metrics we used to evaluate our model are common classification task metrics (e.g., precision, recall). For model optimization, we mainly used the Grid Search cross-validation (Grid Search CV) method to optimize the performance of SVM and RF. Our optimization goal is to adjust the parameters of SVM and RF to achieve better performance. For the parameters of SVM, we set two kinds of kernels: linear and radial basis function (rbf), and three values of regularization parameter C were set: 0.1, 1, and 10. For RF, we set the number of trees (n_estimators) to 100 and 200, and the maximum depth of the tree (max_depth) to none, 10, and 20, with none representing no depth limit. We used 5-fold cross validation, performed a parameter search via Grid Search CV, and recorded the parameter search time, model fitting time, and the best parameter combination determined based on the mean test scores in cross validation, as shown in Figures 7 and 8. Through optimization, SVM and RF can achieve the best prediction performance with specific parameter settings: for SVM, using (C = 1) and a linear kernel; for RF, utilizing (n_estimators = 100) and no limit on max_depth. These settings not only enhance prediction accuracy but also improve the models' generalization ability and stability.

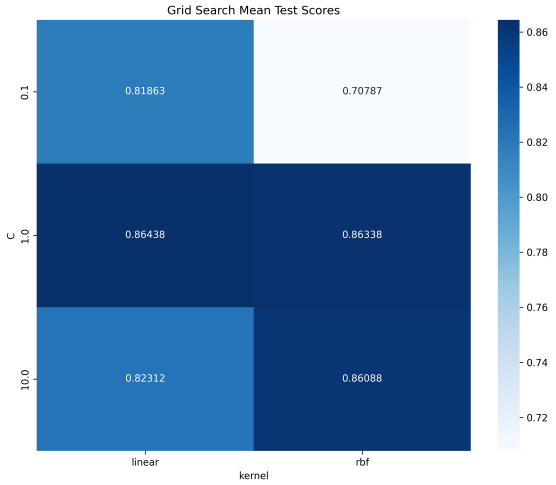

Figure 7. Grid search mean test scores of SVM.

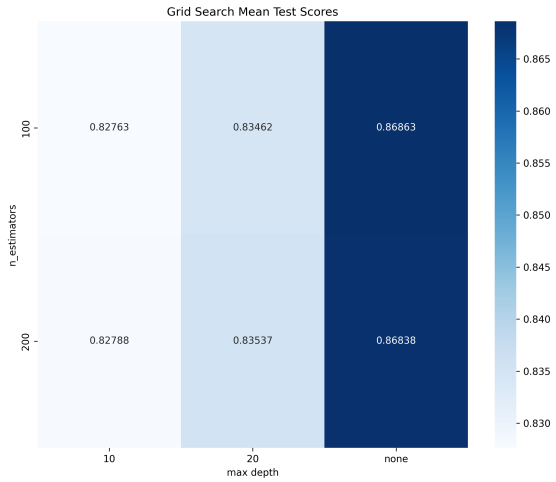

Figure 8. Grid search mean test scores of RF.

4. Experimental Results and Discussion

Figure 9 illustrates the receiver operating characteristic (ROC) curve of our model, depicting the variation of the true positive rate (TPR) and false positive rate (FPR). The blue dotted diagonal indicates that positive and negative cases are predicted with equal probability (i.e., TPR equals FPR), and if the model only makes random guesses, the ROC curve will be close to the diagonal. The curve shows that the model's performance is superior across the entire range of thresholds, with the area under the curve (AUC) value of 0.92, indicating its robustness and accuracy in discriminatory speech classification. These results reinforce the potential of the model for discriminatory speech classification.

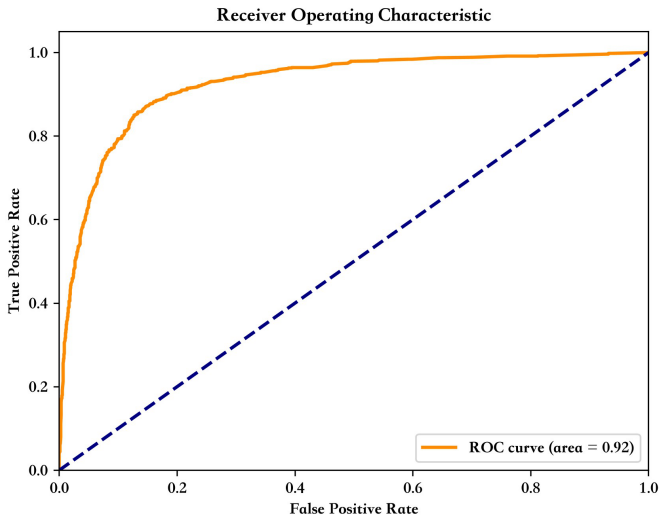

Figure 9. ROC curve.

The improved SMOTE and SVM-RF based models are compared with other base class methods (SVM, RF, DT, KNN (K-Nearest Neighbor), GaussianNB (Gaussian Naive Bayes), and XGB (eXtreme Gradient Boosting)) on the Sina Weibo dataset that we produced. As can be seen in Table 3, the KNN method has the second highest precision, but the recall and F1 scores are low and the overall performance is poor. The performance of the improved SMOTE and SVM-RF based model achieved the highest precision among all models at 90.3%, which is 6.0%, 3.0%, 11.2%, 1.6%, 37.4%, and 11.0% higher than that of SVM, RF, DT, KNN, GaussianNB, and XGB, respectively. In addition, the improved SMOTE and SVM-RF based model also achieved the highest precision in terms of recall and F1 score.

Table 3. Comparison experiment.

Model	Precision	Recall	F1
SVM	0.843	0.862	0.852
RF	0.873	0.834	0.853
DT	0.791	0.764	0.777
KNN	0.887	0.543	0.674
GaussianNB	0.529	0.733	0.615
XGB	0.793	0.884	0.836
Improved SMOTE + SVM + RF	0.903	0.916	0.909

In order to verify the impact of improved SMOTE and the integration method on model classification, ablation experiments were carried out, as shown in Table 4. It is easy to find that in the presence of improved SMOTE, the precision, recall, and F1 score of SVM and RF are increased to different degrees, indicating that the unbalanced dataset will have an impact on the model, and the model performance is increased after balancing the dataset with improved SMOTE. The performance metrics of the model based on both

improved SMOTE and SVM-RF are higher than the case without improved SMOTE and a single model; thus, it can be seen that the integrated model of weighted summation of the respective classification results of SVM and RF after normalization of the computed weights by subtracting the maximum value improves the performance of the model.

Table 4. Ablation experiment.

Model	Precision	Recall	F1
SVM	0.843	0.862	0.852
Improved SMOTE + SVM	0.886	0.893	0.889
RF	0.873	0.834	0.853
Improved SMOTE + RF	0.881	0.868	0.859
Improved SMOTE + SVM + RF	0.903	0.916	0.909

5. Conclusions

With the rapid development of the Internet, a large number of online remarks have been generated. The discussed web discourse represents the views and feelings of web users. However, there are some utterances that contain discriminatory meanings. In view of the potential harm of discriminatory speech, it is crucial to classify and detect discriminatory speech instances. Therefore, this paper collects and processes the data of discriminatory speech in Sina Weibo and proposes an improved SMOTE algorithm based on LDA, which filters out the high-quality data related to the topic and oversamples the imbalanced dataset to synthesize new data. Finally, the weights were calculated by the correct rates of the classification results of SVM and RF, and the final result of the model was obtained by weighted summation of the classification results of SVM and RF. The experimental results show that the proposed method improves the precision, recall, and F1 value by 6.0%, 5.4%, and 5.7%, respectively, compared with SVM, and also shows the best performance compared with other machine learning methods. In addition, the ablation experiment results show that the improved SMOTE and the integrated method can effectively improve the model. In the future, we will continue to optimize our model (e.g., model parameter tuning, feature engineering) to improve its overall performance. In addition, we also consider integrating other models (such as DT and XGB) and optimize the ensemble method, such as using the stacking method to synthesize the prediction results of each model to improve the generalization ability and performance of the model. In the face of imbalanced or sparse data, we will adopt the improved SMOTE method to solve these problems. Combined with the advantages of neural networks in dealing with large-scale and complex data, these combinations will likely improve the performance and robustness of the model. Our future research directions include exploring how different types of neural networks can be integrated and improving the experimental design to delve into how these deep learning models perform on our specific tasks and datasets. We use these models as baselines and also combine them with the improved SMOTE as well as the integrated model, respectively, to comprehensively evaluate and compare their performance.

Author Contributions: Conceptualization, C.W. and D.Z.; methodology, C.W. and H.H.; software, X.S.; validation, C.W. and H.H.; formal analysis, C.W. and H.H.; investigation, C.W.; resources, X.S. and D.Z.; data curation, C.W.; writing—original draft preparation, C.W.; writing—review and editing, C.W.; visualization, K.-L.Y.; supervision, A.W.H.I.; project administration, D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. The dataset presented in this article is not available because it is part of an ongoing study.

Acknowledgments: This research is in part supported by the Research Centre of Deep Space Explorations, the Hong Kong Polytechnic University.

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

References

1. The 53rd Statistical Report on the Development of Internet in China released by China Internet Network Information Center. *J. Natl. Libr. China* **2024**, *33*, 104.
2. Xu,Y.; Liao, X. Discriminatory speech discrimination by fusing bidirectional gated recurrent Unit and convolutional neural network. *J. Wuhan Univ. (Sci. Ed.)* **2020**, *66*, 111–116. [CrossRef]
3. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
4. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L. *Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey*; Springer: New York, NY, USA, 2019.
5. Xue, Y.; Kambhampati, C.; Cheng, Y.; Mishra, N.; Wulandhari, N.; Deutz, P. A LDA-Based Social Media Data Mining Framework for Plastic Circular Economy. *Int. J. Comput. Intell. Syst.* **2024**, *17*, 8. [CrossRef]
6. Xiang, B.; Zhou, L. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Baltimore, MD, USA, 22–27 June 2014; pp. 434–439.
7. Zhong, Y.; Nong, J.; Du, Y. Text classification analysis of agricultural products e-commerce reviews based on LDA topic model. *Gansu Agric.* **2023**, *12*, 64–68. [CrossRef]
8. Lin, T.; Tian, W.; Mei, Q.; Cheng, H. The dual-sparse topic model: Mining focused topics and focused terms in short text. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Republic of Korea, 7–11 April 2014; pp. 539–550.
9. Cheng, X.; Yan, X.; Lan, Y.; Guo, J. Btm: Topic modeling over short texts. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 2928–2941. [CrossRef]
10. Sun, M.; Zhao, X.; Lin, J.; Jing, J.; Wang, D.; Jia, G. PSLDA: A novel supervised pseudo document-based topic model for short texts. *Front. Comput. Sci.* **2022**, *16*, 166350. [CrossRef]
11. Kapočiūtė-Dzikienė, J.; Ungulaitis, A. Towards Media Monitoring: Detecting Known and Emerging Topics through Multilingual and Crosslingual Text Classification. *Appl. Sci.* **2024**, *14*, 4320. [CrossRef]
12. Yan, S.; Wang, J.; Zhu, S.; Cui, Y.; Tao, Z. Research on Internet Sensitive Speech Recognition based on character and word features. *Comput. Eng. Appl.* **2023**, *59*, 129–138.
13. Burnap, P.; Williams, M.L. Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Sci.* **2016**, *5*, 1–15. [CrossRef] [PubMed]
14. Greevy, E.; Smeaton, A.F. Classifying racist texts using a support vector machine. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington, DC, USA, 14–18 July 2004; pp. 468–469.
15. Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; Chang, Y. Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web, Montreal, QC, Canada, 11–15 April 2016; pp. 145–153.
16. Malmasi, S.; Zampieri, M. Detecting hate speech in social media. *arXiv* **2017**, arXiv:1712.06427.
17. Khan, T.A.; Sadiq, R.; Shahid, Z.; Alam, M.M.; Su'ud, M.B.M. Sentiment Analysis using Support Vector Machine and Random Forest. *J. Inform. Web Eng.* **2024**, *3*, 67–75. [CrossRef]
18. Sahoo, S.; Subudhi, A.; Dash, M.; Sabut, S. Automatic classification of cardiac arrhythmias based on hybrid features and decision tree algorithm. *Int. J. Autom. Comput.* **2020**, *17*, 551–561. [CrossRef]
19. Elreedy, D.; Atiya, A.F. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Inf. Sci.* **2019**, *505*, 32–64. [CrossRef]
20. Liu, W.; Chawla, S.; Cieslak, D.A.; Chawla, N.V. A robust decision tree algorithm for imbalanced data sets. In Proceedings of the 2010 SIAM International Conference on Data Mining, SIAM, Columbus, OH, USA, 29 April–1 May 2010; pp. 766–777.
21. Imam, T.; Ting, K.M.; Kamruzzaman, J. z-SVM: An SVM for improved classification of imbalanced data. In Proceedings of the AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, 4–8 December 2006; Proceedings 19; Springer: Berlin/Heidelberg, Germany, 2006; pp. 264–273.
22. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
23. Kishor, A.; Chakraborty, C. Early and accurate prediction of diabetics based on FCBF feature selection and SMOTE. *Int. J. Syst. Assur. Eng. Manag.* **2021**. [CrossRef]
24. Singgalen, Y.A. Comparative Analysis of DT and SVM Model Performance with SMOTE in Sentiment Classification. *KLIK Kaji. Ilm. Inform. Dan Komput.* **2024**, *4*, 2485–2494.
25. Rathpisey, H.; Adji, T.B. Handling Imbalance Issue in Hate Speech Classification using Sampling-based Methods. In Proceedings of the 2019 5th International Conference on Science in Information Technology (ICSITech), Jogjakarta, Indonesia, 23–24 October 2019; pp. 193–198. [CrossRef]

26. Sanya, A.D.; Suadaa, L.H. Handling Imbalanced Dataset on Hate Speech Detection in Indonesian Online News Comments. In Proceedings of the 2022 10th International Conference on Information and Communication Technology (ICoICT), Virtual, 2–3 August 2022; pp. 380–385. [CrossRef]
27. Lu, J.; Lin, H.; Zhang, X.; Li, Z.; Zhang, T.; Zong, L.; Ma, F.; Xu, B. Hate Speech Detection via Dual Contrastive Learning. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 2787–2795. [CrossRef]
28. Jiang, A.; Yang, X.; Liu, Y.; Zubiaga, A. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Soc. Netw. Media* **2022**, *27*, 100182. [CrossRef]
29. Degife, W.A.; Lin, B.S. A Multi-Aspect Informed GRU: A Hybrid Model of Flight Fare Forecasting with Sentiment Analysis. *Appl. Sci.* **2024**, *14*, 4221. [CrossRef]