As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 8(01): 313-317, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0801054
Available online at: http://www.ijmtst.com/vol8issue01.html

Imbalanced Text Features for Toxic Comments Classification

S. Rahamat Basha, T. Bhasara Reddy

Department of Computer Science and Technology, Sri Krishnadevaraya University, India.

To Cite this Article

S. Rahamat Basha and T. Bhasara Reddy. Imbalanced Text Features for Toxic Comments Classification. *International Journal for Modern Trends in Science and Technology* 2022, *8* pp. 313-317. https://doi.org/10.46501/IJMTST0801054

Article Info

Received: 01 December 2021; Accepted: 15 January 2021; Published: 17 January 2022

ABSTRACT

Social media platforms and microblogging websites have gained accelerated popularity during the past few years. These platforms are used for expressing views and opinions about products, personalities, and events. Often during discussions and debates, fights take place on social media platforms which involves using rude, disrespectful, and hateful comments called toxic comments. The identification of toxic comments has been regarded as an essential element for social media platforms. This study introduces an ensemble approach, called regression vector voting classifier (RVVC), to identify the toxic comments on social media platforms. The ensemble merges the logistic regression and support vector classifier under soft voting criteria. Several experiments are performed on the imbalanced and balanced dataset to analyze the performance of the proposed approach. For data balance, the synthetic minority oversampling technique (SMOTE) is used on the imbalanced dataset. Furthermore, two feature extraction approaches are utilized to investigate their suitability such as term frequency-inverse document frequency (TF-IDF) and bag-of-words (BoW). The performance of the proposed approach is compared with several machine learning classifiers using accuracy, precision, recall, and F1-score. Results suggest that RVVC outperforms all other individual models when TF-IDF features are used with SMOTE balanced dataset and achieves an accuracy of 0.97

KEYWORDS: Toxic comments classification; TF-IDF; BoW; text classification; data re-sampling

1.INTRODUCTION

SOCIAL media platforms and microblogging websites have gained accelerated popularity for social communication between individuals and groups. Through these platforms, people share their thoughts, ideas, opinions and express their feelings using comments and feedback [1]. The number of internet users has been increasing gradually each year, from 2.4 billion in 2014 to 3.4 billion, 4 billion, and 4.4 billion in 2016, 2017, and June 2019, respectively [2]. As of May 2020, the number of internet users is increased to 4,648 billion [3]. Social media platforms provide a common ground for these users to share opinions and discuss ideas. However, problems arise when debates take a dirty side and fights take place on social media platforms which involves using rude, disrespectful, and hateful comments called toxic comments. Text in online comments contain many hazards such as fake news, cyberbullying, online harassment and toxicity [4]. Unfortunately, these toxic comments have become a serious issue that affects the reputation of social platforms and cause different psychological problems for users, such as depression, frustration, and even suicidal thoughts [1]. Toxic comment classification is very important to overcome the above-mentioned issues and maintain stability in online debates [5]. Toxic comments can be considered as a personalattack, online harassment, and bullying behaviors. Over the past few

years, several cases of police arrests happened where police arrested many individuals due to the abusive or negative content on personal pages [6], [7]. So a framework that can detect toxic comments and prevent publishing is of significant importance. As a result, several approaches have been introduced for the automatic detection of toxic comments using machine learning algorithms. For example, the study [8] combines machine learning and crowd-sourcing to classify the comments that are considered a personal attack. Support vector machines were also used by [9] for Cyberbullies detection. The cyberbullies are also detected in [10] using deep learning models. Despite the proposed approaches, there is a need to model more approaches to provide high accuracy for toxic comments. This study introduces an ensemble approach for toxic comments detection in imbalanced datasets and makes the following contributions

The rest of the paper is organized as follows. Section II discusses research papers from the literature which are closely related to the current study. Section III gives an overview of the machine learning algorithms adopted for the current research, as well as, the description of the dataset used for the experiment. The proposed approach is also presented in the same Section. Results are discussed in Section IV while the conclusion is given in Section V.

II. LITERATURE REVIEW

Toxic comments on social media platforms have been a source of a great stir between individuals and groups. A toxic comment is not only verbal violence but includes the comment that is rude, disrespectful, negative online behavior, or other similar attitudes that make someone leave a discussion. Therefore, the toxic comments identification on social platforms is an important task that can help to maintain its interruption and hatred-free operations. Consequently, a large variety of toxic comment approaches have been proposed. Three characteristics concerning toxic classification are evaluated: classification, feature dimension reduction, and feature importance. The authors use a deep learning-based toxic comments classification approach in [11] for the imbalanced toxic dataset. The performance evaluation is carried out on Kaggle Wikipedia's talk page edits dataset which contains 159,571 records of toxic comments. The proposed approach makes a multi-class classification including toxic, threat, severe toxic, obscene, insult, and identity hate. Convolutional neural network (CNN), bidirectional long short- term memory (LSTM), bidirectional gated recurrent unit (GRU), and the ensemble of the three models are used for classification. Results indicate that the ensemble approach gives the highest classification with an F1 score of 0.828 for toxic/non-toxic and 0.872 for toxicity types. The study [12] proposed a method to classify the online toxic comments using logistic regression and neural network models. Online toxic comments classification dataset is taken from Kaggle and logistic regression (LR), CNN, LSTM, and CNN+LSTM (2 layers of LSTM and 4 layers of CNN) are used. All models perform good but CNN+LSTM achieves 0.982 accuracy which is the highest among all the classifiers. In the same vein, the study [13] perform classification for online toxic comments using support vector machine (SVM), naive Bayes (NB), K-nearest neighbor (KNN), linear discriminant analysis (LDA), and CNN. The classification is conducted on KaggleWikiperida comments for toxic and non-toxic comments. CNN model achieves accuracy higher than 90% accuracy while the machine learning classifier obtains accuracy between 65% to 85%. Due to the reported high accuracy of deep learning approaches, several researchers focus on using deep CNN and LSTM architectures for classification. For example, deep neural network architectures are used for toxic comments classification in [14]. The study uses NB, LSTM, and RNN to identify toxic comments. For this purpose, a toxic comment classification challenge dataset comprising 159,000 comments is used. LSTM performs best with 67% true positive rate which is 20% higher than the NB model. On the other hand, LSTM achieved a 73% F1 score, 81% precision score, and 66% recall. Similarly, hybrid deep learning approaches are adopted in [15] for the same task. For this purpose, the Jigsaw toxic comments classification dataset is used. The hybrid deep learning achieved 98% accuracy and 80% F1 score. Another study [16] created their dataset taking comments from Facebook pages and labeled them with six categories: toxic, severe toxic, obscene, threat, insult, identity hate. Different machine learning and deep learning algorithms are applied for Bangla toxic comments

classification. SVM, Gaussian NB, Multinomial NB, Multi-Label k Nearest Neighbor (MLKNN), and Backpropagation for Multi-Label Neighbor (BP-MLL) are used to classify comments. BPMLL outperforms both machine learning and deep learning algorithms used for experiments. The study [17] proposed a methodology for the classification of toxic comments and depth error analysis. The study uses two datasets including the Wikipedia talk pages and a Twitter dataset, containing six classes of toxic comments.

III. MATERIALS AND METHODS

This study uses different techniques, methods, and tools for the classification of toxic and non-toxic comments. Also, various preprocessing steps, data re-sampling methods, features extraction techniques, and supervised machine learning models are adopted for the said task. A. DATA DESCRIPTION This study aims at the automatic classification of toxic and non-toxic comments from social media platforms. Various machine learning models are utilized for this purpose to evaluate their strength for the said task. For evaluation, the selected models are trained and tested with binary class datasets. Traditionally, toxic comments are grouped under several classes such as hate, toxic, threat, severe toxic, obscene, insult and non-toxic, etc. We follow a different approach by grouping the comments under two classes, toxic and non-toxic. The original dataset which is taken from Kaggle [18], is a multi-label dataset and contains labels such as toxic, severe toxic, obscene, threat, insult, and identityhate. The non-toxic comments belong to one class, while from the other comments only those comments are selected that have toxic labels. It means that the comments that label severe toxic, obscene, threat, insult, and identity hate are not selected. For example, Table 1 shows that 'comment 2' is only toxic and 'comment 3' is non-toxic. For our experiment, both 'comment 2' and 'comment 1' are selected under toxic and no-toxic classes, but 'comment 1' and 'comment 4' are not selected.

PREPROCESSING STEPS

Pre-processing techniques are applied to clean the data which helps to improve the learning efficiency of machine learning models . For this purpose, the following steps are executed in the given sequence. Tokenization: is a process of dividing a text into smaller units called 'tokens'. A token can be a number, word, or any type of symbol that contains all the important information about the data without conceding its security. Punctuation removal: involves removing the punctuation from comments using natural language processing techniques. Punctuations are the symbols that are utilized in sentences/comments to make the sentence clear and readable for humans. However, it creates problems in the learning process of machine learning algorithms and needs to be removed to improve their learning process. Some common punctuation marks are mostly used such that colon, question marks, comma, semicolon, full-stop/period, etc.Number removal: is also a part of preprocessing which helps to improve the performance of the machine learning algorithms. Numbers are unnecessary and do not contribute to the learning of text analysis approaches. Removing the numbers increases the efficiency of models and decreases the complexity of the data. Stemming: is an important part of preprocessing because it increases the performance by clarifying affixes from sentences/comments and converting the comments into the original form. Stemming is the process of transforming a word into its root form. For example, different words have the same meaning such as: 'plays', 'playing', 'played' are modified forms of 'play'. Stemming is implemented using the Porter stemmer algorithms [33]. Spelling correction: is the process of correcting the misspelled words. In this phase, the spelling checker is used to check the misspelled words and replace them with the correct word. Python library 'pyspellchecker' provides the necessary features to check the misspelled words and is used for the experiments [34]. Stopwords removal: Stopwords are those English words that do not add any meaning to a sentence. So these can be removed by stopwords removal without affecting the meaning of a sentence. The removal of stop-words increases the model's performances and decreases the complexity of input features [35].

FEATURE ENGINEERING

Feature engineering aims at discovering useful data features or constructing features from original features to train machine learning algorithms effectively. The study concludes that feature engineering can improve the efficiency of machine learning algorithms. 'Garbage out' is a corporateproverb used in machine learning

which implies that senseless data used as the input, yields meaningless output. In contrast, more information-driven data will yield favorable results. Hence, feature engineering can derive useful features from raw data which helps to improve the reliability and accurateness of learning algorithms. In the proposed methodology, two feature engineering methods are used including the bag of words and term frequency-inverse document frequency.

BAG-OF-WORDS

The bag of words (BoW) technique is used to extract features from the text data. The boW is easy to implement and understand besides being the simplest method to extract features from the text data. The boW is very suitable and useful for language modeling and text classification. The 'CountVectorizer' library is used to implement BoW. CountVectorizer calculates the occurrence of words and constructs a spare database matrix of words [38]. The boW is a pool of words or features, where every feature is categorized as a label that signifies the occurrences of the categorized feature.

IV. PROPOSED METHODOLOGY

Ensemble learning is widely used to attain high accuracy for classification tasks. The combination of various models can perform well as compared to individual models. Owing to the high accuracy of ensemble models, this study leverage an ensemble model to perform toxic comments classification. Our experiments indicate the good performance from LR and SVC, so to further improve the performance, this study combines these models. The proposed approach is called regression vector voting classifier (RVVC) and combines these models using soft voting criteria as shown in Figure 2. The soft voting criteria ensure that the class with a high predicted probability by two classifiers will be considered as the final prediction.

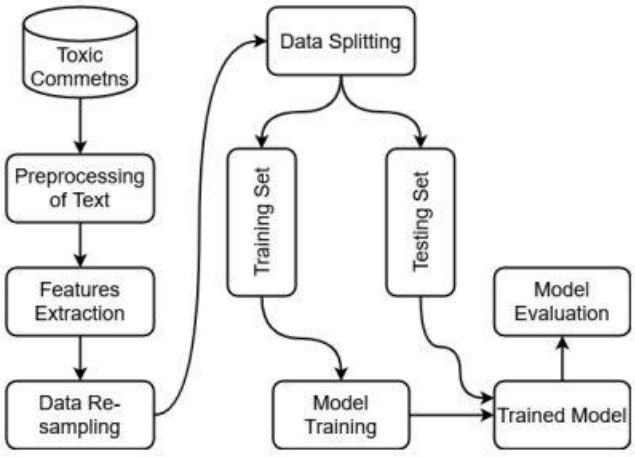

FIGURE 1: The flow of the proposed methodology.

EVALUATION METRICS

We evaluate the performance of machine learning models in terms of accuracy, precision, recall, and F1 score. 1) Accuracy Accuracy indicates the ratio of correct predictions to the total predictions from the classifiers on test data. The maximum accuracy score is 1 indicating that all predictions from the classifier are correct while the minimum accuracy score can be 0. Accuracy can be calculated as Accuracy = Number of correct predictions T otal number of predictions , (14) Another form to calculated accuracy is using

TABLE 1: Performance results of all models on oversampled data using BoW features

Classifier	Accuracy	Precision	Recall	F1 score
RF	0.92	0.94	0.78	0.83
SVC	0.92	0.87	0.87	0.87
KNN	0.89	0.86	0.74	0.78
DT	0.91	0.84	0.85	0.85
LR	0.94	0.91	0.87	0.89
RVVC	0.93	0.91	0.85	0.88

V. CONCLUSIONS

This study analyzes the performance of various machine learning models to perform toxic comments classification and proposes an ensemble approached called RVVC. The influence of an imbalanced dataset

and balanced dataset using random under-sampling and SMOTE over-sampling on the performance of the models is analyzed through extensive experiments. Two feature extraction approaches including TF-IDF and BoW are used to get the feature vector for models' training. Results indicate that models perform poorly on the imbalanced dataset while the balanced dataset tends to increase the classification accuracy. Besides the machine learning classifiers like SVM, RF, GBM, and LR, the proposed RVVC and RNN deep learning models perform well with the balanced dataset. The performance with an oversampled dataset is better than the under-sampled dataset as the feature set is large when the data is over-sampled which elevates the performance of the models. Results suggest that balancing the data reduces the chances of models over-fitting which happens if the imbalanced dataset is used for training. Moreover, TF-IDF shows better classification accuracy for toxic comments than BoW as TF-IDF records the importance of a word contrary to BoW which simply counts the occurrence of a word.

REFERENCES

1. Elias Aboujaoude, Matthew W Savage, VladanStarcevic, and Wael O Salame. Cyberbullying: Review of an old problem gone viral. Journal of adolescent health, 57(1):10–18, 2015.
2. How Much Data is Created on the Internet Each Day? https://blog. microfocus.com/how-much-data-is-created-on-the-internet-each -day/. Accessed: 2020-06-06.
3. World Internet Users and 2020 Population Stats. https://www. internetworldstats.com/stats.htm. Accessed: 2020-06-06.
4. Maeve Duggan. Online harassment. Pew Research Center, 2014.
5. PinkeshBadjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 759–760, 2017.
6. Man jailed for 35 years in Thailand for insulting monarchy on Facebook. https://www.theguardian.com/world/2017/jun/09/ man-jailed-for-35-years-in-thailand-for-insulting-monarchy-on-f acebook. Accessed: 2020-06-06.
7. Mississippi teacher fired after racist Facebook post; black parent responds. https://www.clarionledger.com/story/news/2017/09/20/ mississippi-teacher-fired-after-racist-facebook-post/684264001/. Accessed: 2020-06-06.
8. Ellery Wulczyn, NithumThain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In Proceedings of the 26th International Conference on World Wide Web, pages 1391–1399, 2017.
9. Michal Ptaszynski, JuusoKalevi Kristian Eronen, and Fumito Masui. Learning deep on cyberbullying is always better than brute force. In LaCATODA@ IJCAI, pages 3–10, 2017.
10. Sweta Agrawal and Amit Awekar. Deep learning for detecting cyberbullying across multiple social media platforms. In European Conference on Information Retrieval, pages 141–153. Springer, 2018.
11. Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. Imbalanced toxic comments classification using data augmentation and deep learning. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 875–878. IEEE, 2018.
12. Mujahed A Saif, Alexander N Medvedev, Maxim A Medvedev, and TodorkaAtanasova. Classification of online toxic comments using the logistic regression and neural networks models. In AIP Conference Proceedings, volume 2048, page 060011. AIP Publishing LLC, 2018.
13. Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, and Vassilis P Plagianakos. Convolutional neural networks for toxic comment classification. In Proceedings of the 10th Hellenic Conference on Artificial Intelligence, pages 1–6, 2018.
14. Sara Zaheri, Jeff Leath, and David Stroud. Toxic comment classification. SMU Data Science Review, 3(1):13, 2020.
15. RohitBeniwal and ArchnaMaurya. Toxic comment classification using hybrid deep learning model. In Sustainable Communication Networks and Application, pages 461–473. Springer, 2021.
16. ANM Jubaer, Abu Sayem, and MdAshikur Rahman. Bangla toxic comment classification (machine learning and deep learning approach). In 2019 8th international conference system modeling and advancement in research trends (SMART), pages 62–66. IEEE, 2019.
17. Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. arXiv preprint arXiv:1809.07572, 2018.
18. Hafiz Hassaan Saeed, KhurramShahzad, and Faisal Kamiran. Overlapping toxic sentiment classification using deep neural architectures. In 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pages 1361–1366. IEEE, 2018.