Journal of Big Data

RESEARCH Open Access

Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering

Muhammad Mujahid[1], EROL Kına[2], Furqan Rustam[3], Monica Gracia Villar[4,5,6], Eduardo Silva Alvarado[4,7,8], Isabel De La Torre Diez[9] and Imran Ashraf[10*]

*Correspondence:
imranashraf@ynu.ac.kr

[1] Artificial Intelligence and Data Analytics (AIDA) Lab, CCIS, Prince Sultan University, Riyadh 11586, Saudi Arabia
[2] Özalp Vocational School, Van Yüzüncü Yil University, Van 65100, Turkey
[3] School of Computer Science, University College Dublin, Dublin D04 V1W8, Ireland
[4] Universidad Europea del Atlantico, Isabel Torres 21, Santander 39011, Spain
[5] Universidad Internacional Iberoamericana Arecibo, Puerto Rico 00613, USA
[6] Universidade Internacional do Cuanza, Bie, Angola
[7] Universidad Internacional Iberoamericana, Campeche 24560, Mexico
[8] Universidad de La Romana, La Romana, Dominican Republic
[9] Department of Signal Theory and Communications and Telematic Engineering, University of Valladolid, Paseo de Belen 15, Valladolid 47011, Spain
[10] Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea

Abstract

The classification of imbalanced datasets is a prominent task in text mining and machine learning. The number of samples in each class is not uniformly distributed; one class contains a large number of samples while the other has a small number. Overfitting of the model occurs as a result of imbalanced datasets, resulting in poor performance. In this study, we compare different oversampling techniques like synthetic minority oversampling technique (SMOTE), support vector machine SMOTE (SVM-SMOTE), Border-line SMOTE, K-means SMOTE, and adaptive synthetic (ADASYN) oversampling to address the issue of imbalanced datasets and enhance the performance of machine learning models. Preprocessing significantly enhances the quality of input data by reducing noise, redundant data, and unnecessary data. This enables the machines to identify crucial patterns that facilitate the extraction of significant and pertinent information from the preprocessed data. This study preprocesses the data using various top-level preprocessing steps. Furthermore, two imbalanced Twitter datasets are used to compare the performance of oversampling techniques with six machine learning models including random forest (RF), SVM, K-nearest neighbor (KNN), AdaBoost (ADA), logistic regression (LR), and decision tree (DT). In addition, the bag of words (BoW) and term frequency and inverse document frequency (TF-IDF) features extraction approaches are used to extract features from the tweets. The experiments indicate that SMOTE and ADASYN perform much better than other techniques thus providing higher accuracy. Additionally, overall results show that SVM with 'linear' kernel tends to attain the highest accuracy and recall score of 99.67% and 1.00% on ADASYN oversampled datasets and 99.57% accuracy on SMOTE oversampled dataset with TF-IDF features. The SVM model using 10-fold cross-validation experiments achieved 97.40 mean accuracy with a 0.008 standard deviation. Our approach achieved 2.62% greater accuracy as compared to other current methods.

Keywords: Machine learning, Bag of words, Oversampling techniques, SMOTE, K-Means SMOTE

Introduction

Text mining and machine learning with imbalanced datasets have received an increasing amount of attention in recent years, both from a theoretical and practical standpoint and have a considerable classification challenge. The datasets are imbalanced because one or more classes have a much lower number of samples than other classes, resulting in an imbalance of class distribution. There are a variety of strategies available to deal with this issue; approaches that produce synthetic data to establish a balanced class are more adaptable than methods that manually tweak data. For example, various methods are used for text categorization [1], retrieval of information and filtration [2], and fraud detection [3].

The class imbalance has a negative impact on the prediction capabilities of classification methods. The prediction accuracy of machine learning algorithms is often used to evaluate the overall performance of the algorithms. Many of these algorithms are designed to maximize classification accuracy, which is a metric that is skewed in favor of the majority class. A classifier can obtain better classification accuracy even if it does not accurately predict a single occurrence of a minority class. According to Japkowicz and Stephen [4], the complexities of the model increase as the intensity of the class imbalance increases, and the impact of this issue on the classifiers is intensified as the quantity of the training examples decreases. Another study [5] focuses on the class imbalanced problem that is encountered when utilizing machine-based models to analyze the sentiment of tweets. The authors performed experiments on the imbalanced dataset collected from Twitter using two well-known classifiers to prove their point. Later, the synthetic minority oversampling technique (SMOTE) approach was used to ensure that the dataset was properly balanced. Their findings demonstrated that using an oversampling strategy can improve the performance of machine-based models.

Similarly, the study [6] employed SMOTE to balance the datasets and conducted research on the classification of political tweets in two different languages. It appears that their strategy makes it possible to deal with the class imbalance distribution issue by improving recognition of such minority classes while simultaneously achieving a substantial increase in total geometric mean criterion, as demonstrated by the obtained findings. Despite several works investigating the influence of individual oversampling approaches, the literature lacks a comprehensive study that compares the performance of various oversampling approaches.

Scope and importance

Oversampling is a machine learning approach that may successfully handle issues associated with class imbalance. When the number of schemas used by various classes is drastically different, there is class inconsistency. This animosity deprives minority groups of positive role models. To address this issue, oversampling procedures aim to consciously increase the minority representation in the sample. There is room for improvement in the efficiency of machine learning techniques such as neural networks, random forests, and support vector machines (SVM). Incorporating a more evenly distributed set of training data allows these algorithms to better predict outcomes from novel data and set limits for their development. To adapt to different data distributions,

quickly manage large datasets, and fine-tune algorithmic parameters, the oversampling approach employs powerful machine learning skills. Furthermore, machine learning can facilitate the development of sophisticated sampling algorithms. These algorithms can produce synthetic samples while maintaining crucial data properties.

Machine learning pipelines rely heavily on feature engineering. It includes steps to clean and prepare the raw data for use in training the models. Two widely used methods for extracting text characteristics in certain settings are BoW and TF-IDF. BoW makes use of a vector that represents the frequency of phrases in the corpus. While this method works well for sentiment analysis, document categorization, and spam detection, it doesn't take word order or other contextual aspects into account. After processing text, the Bag-of-Words (BoW) technique converts it into a high-dimensional sparse matrix. In this matrix, each row represents a document, and every column represents a word. By taking document-level word importance into account in context with the full corpus, the TF-IDF algorithm improves upon the BoW approach. To make frequent words less important and rare terms more significant in the computation, multiply TF and IDF. The analysis of textual data, including consumer comments, social media marketing, and product evaluations. Common approaches employed in this study include term frequency, BoW, and TF-IDF. Data retrieval methods such as TF-IDF and BoW compare user queries with relevant documents, using the textual content of the documents to evaluate the correctness of the documents. Businesses have begun integrating engineering methods like TF-IDF and bag-of-sampling into machine learning to make predictive models more accurate and versatile for use in real-world applications like sentiment analysis and fraud detection. They may work together to use oversampling methods. A company's value can only rise as a result of better decision-making.

Most modern companies embrace the strategy of enhancing their products. Many businesses use customer exposure as a strategic tool to assess their customers' perceptions. Remote monitoring, surveys, assessments, and questionnaires are all effective methods for collecting valuable customer feedback. These statements provide valuable feedback that companies can leverage to enhance their products and services. Consumers rely on online resources and social media platforms to make informed decisions about the most suitable products. E-commerce has become increasingly essential in our daily lives, offering a wide range of information, communication, educational, shopping, and entertainment options. Twitter is highly valuable for businesses compared to other platforms due to its ability to facilitate users sharing concise or detailed feedback on any product. Organizations and businesses often face difficulties in gathering tweets and analyzing the sentiments expressed. Automated sentiment analysis enables rapid evaluation and categorization of tweets as positive or negative.

Contributions

This paper focuses on analyzing the suitability and efficacy of different sampling approaches to solve the problem of the imbalanced dataset as well as improve the performance of machine learning models. The model's performance varies depending on whether the number of class samples is fewer or higher if the class is not properly balanced. This research makes use of two extremely imbalanced Twitter datasets, including

the EndViolence tweets dataset, and the E-commerce-related tweets dataset, which are obtained from Kaggle. So, this study makes the following contributions in this regard

1. The Impact of various sampling approaches is investigated concerning the performance of machine learning models for imbalanced datasets. For this purpose, five sampling approaches are employed including SMOTE, support vector machine-SMOTE (SVM-SMOTE), K-means SMOTE, adaptive synthetic (ADASYN) oversampling, and Border-Line SMOTE. The selection of these sampling approaches is made regarding their wide use in the existing literature and reported performances.
2. For analyzing the influence of these approaches on the performance of models, six widely used machine learning models are selected which include random forest (RF), SVM, K-nearest neighbor (KNN), AdaBoost (ADA), logistic regression (LR), decision tree (DT), and gradient boosting (GB). The performance is measured in terms of accuracy, precision, recall, and F1 score.
3. Two publicly available and highly imbalanced datasets are utilized for experiments to check the performance of oversampling techniques on machine learning models. For feature engineering, the bag of words (BoW) and term frequency-inverse document frequency (TF-IDF) are also utilized.

The following part of the paper is organized into four sections. Section "Related work" covers the relevant work on oversampling approaches. Section "Materials and proposed methodology" presents the methodology adopted to carry out experiments. In Sect. "Results and discussions", experiments and results of our study are discussed while Sect. Conclusion" provides the conclusion.

Related work

Data balancing is an important task to reduce model skewness and as such several works have made use of oversampling approaches to reduce model overfitting and increase performance [7–9]. Sarakit et al. [10] employed the SMOTE method to detect emotion in unbalanced YouTube datasets using three machine learning classifiers. The findings suggest that the proposed method improves emotion classification while also addressing the issue of an unbalanced dataset. For the categorization of toxic comments, Rupapara et al. [11] introduced an ensemble approach based on imbalanced features while SMOTE is used for data balancing. A soft voting ensemble model is used by combining the LR and support vector classifiers. Moreover, BoW and TFIDF are used for the proposed approach. Using TF-IDF features and the SMOTE method, the RV-VC model can achieve a 97% accuracy.

Flores et al. [12] used the SMOTE method for sentiment analysis to test the SVM and Naive Bayes on imbalanced datasets. Results show that preprocessing, training, and testing split effectiveness and data balancing are all prevalent aspects in achieving improved results. For sentiment analysis on the Arabic tweets related to COVID-19, the study [13] used ensemble classifiers with SMOTE. Word2vec embedding is used, as well as single and ensemble classifiers are used with and without SMOTe. The ensemble approach using word embedding and the SMOTE technique outperforms the competition. The research [14] focused on the performance of machine learning models in determining

the polarity score for extremely imbalanced datasets using word-embedding features. A number of basic classifiers and ensemble classifiers are investigated with and without SMOTE. The results show that using ensemble methods with SMOTE and embed words improved the F1 score by more than 15% on average over the baseline technique. Another research [15] classified news using the SMOTE and Borderline SMOTE techniques on imbalanced datasets to show that using SMOTE produces better results.

Along the same lines, the study [16] used oversampling approaches to overcome the problem of imbalance dataset classes in sarcasm detection from social media tweets. The authors demonstrated that SMOTE and Borderline SMOTE are efficient in classifying sarcasm sentiments. Borderline SMOTE and ADASYN oversampling approaches for multi-text classification problems are compared in the study [17].

KNN and SVM models are used in the study. Results show the supremacy of SMOTE to improve the performance of models. Another research [18] employed the SMOTE approach for sentiment and emotion classification on an imbalanced dataset. Sentiment analysis regarding spam reviews is carried out in [19] using the SMOTE approach. Similarly, the study [20] employed the machine learning techniques, and K-means SMOTE method to balance the dataset to predict school student performance and showed superior results using the balanced dataset.

The study [21] conducted sentiment analysis on tweets concerning online education using an imbalanced dataset. The study employed an e-learning dataset that was extracted using various keywords from the Twitter API. The authors employed lexicon-based and feature engineering techniques to label the tweets and extract features from them. The experiments are carried out using machine and deep learning models using SMOTE-balanced datasets. The performance of the models is better using SMOTE. The authors employ SMOTE in [22] to undertake a comparative analysis of three classifiers for sentiment analysis. The study states that SMOTE used with TF-IDF features provides the best results. Another research [23] analyzed the sentiments of people regarding e-sports education using an imbalanced Twitter dataset. The study compared the performance of Naive Bayes and SVM using SMOTE oversampling. Experimental results reveal that the Naive Bayes technique has the best accuracy score when used with the SMOTE balanced dataset.

Balaji et al. [24] employ robust machine learning methods to provide a comprehensive evaluation of numerous applications of sentiment analysis. The study commences with a comprehensive examination of machine learning methods specifically utilized for sentiment analysis. Subsequently, they provide a comprehensive examination of machine learning methodologies for sentiment analysis. Additionally, they offer a comprehensive examination of the limitations and benefits associated with employing machine learning in social media analysis. Another study [25] offers a comprehensive statistical analysis of Extensive Feature Selector (EFS), a new method for feature selection that uses probability based on classes and corpora. On four benchmark data sets, it compares EFS to nine alternative FS approaches using KNN, support vector machines, and multinomial Naive Bayes classifiers. The results show that out of the ten methods tested, EFS consistently yields the best results.

The study [26] examined how globalization tactics impacted local feature selection (LFS) techniques using feature-rich datasets. The authors used the weighted-sum,

summation, and maximum approaches to analyze globalization. Utilizing DFSS, odds ratio (OR), and chi-square (CHI) techniques, the researchers examined the effects of globalization initiatives. The approach with the highest degree of internationalization success, according to the findings, was the AVG technique. The DFSS approach outperformed RE and CHI2 in terms of MCB and MCU characteristics. The authors found the CHI2 method to be more accurate in terms of DFSS and OR techniques.

The paper [27] conducted a comprehensive examination of the DL and ML techniques used in the diagnosis of depression. They also emphasize the constraints of the current research efforts. However, this research lacks the capacity to conduct a comprehensive review of prior investigations. The authors in the paper [28] used a neural network architecture in the development of the proposed deep learning model to accurately identify the musical genre from aural inputs. The proposed approach boasted an impressive accuracy rate of 90.3%, surpassing the results of previous research in its competition. The deep model's performance consistency was evaluated by conducting K-fold cross-validation with different values of k. The study [29] provided a cutting-edge deep learning model. By deftly combining deep learning approaches with various word embedding techniques, the model conducts multi-class sentiment analysis on a dataset consisting of tweets from six prominent US airlines. To detect emotional content and insert words, the selected systems use a range of deep learning approaches. The approach begins with cleaning tweets and applying CNN's pre-processing algorithms to the raw data from DNN.

There is a lot of research on oversampling and performance analysis that uses ensemble methods or simple machine learning models. They used word embedding models, also known as "bags of words," to extract semantic information from texts. Even though these models work, it's hard for them to generalize complicated speech patterns and attitudes that depend on the situation. In machine learning, class imbalance, or the improper display of one mood class, is a major issue. Inequalities in the data may make it harder for skewed models to work well with minority classes. In an effort to address this problem, researchers have only focused on SMOTE and borderline SMOTE in the entire literature. We are still debating how to address class imbalances in this research, as these methods do not perform very well.

Text data preparation is another drawback yet frequently overlooked aspect of the analysis of tweets in the performance analysis of machine learning research. Preparing raw text for machine learning models is an essential step. Numerous studies either do not pretest or employ methods that consume a great deal of computational power without significantly enhancing performance. Ineffective preparation has an impact on the model's performance, the duration of the training process, and operational costs. Negligently normalized and extracted feature models from textual data could be problematic, thereby degrading the performance of mood classifiers.

Another drawback of existing approaches is the selection of appropriate oversampling and feature engineering techniques. The authors in studies such as [14–19] do not select appropriate sampling and feature techniques. Some studies used only one SMOTE technique to validate the performance of machine learning on balanced data. Only the single oversampling technique does not validate or provide assurance of the superiority of the model in text classification. Also, feature engineering is most important to enhance the

performance of models. Another aspect is the use of some basic machine learning models without optimization and hyper-parameter tuning for training large datasets.

In the literature, we noticed that without proper preprocessing, inadequate selection of oversampling and feature engineering techniques may lead to unsatisfactory results. The authors in the study [10] only utilized SMOTE and simple models; the performance of machine learning was not satisfactory, and hence the overall results were very poor. The study [11] also utilized the SMOTE and Ensemble ML models. An ensemble of three simple models takes a lot of time to process and provide predictions in its early stages, so only using a single SMOTE technique can ensure that the ensemble works better on SMOTE data. The study [12] also utilized the single SMOTE technique and did not perform data preprocessing techniques properly to lessen the training resources. Their study achieved poor results. The study [14] employed SMOTE to balance and enhance the dataset samples, but their study conducted experiments only with a total of 1798 samples, which is very limited and causes overfitting with ensemble methods. They used SMOTE but achieved only 78% accuracy.

From the above-discussed research works, several points can be deduced. First, the performance of machine learning tends to improve when a balanced dataset is used. Dataset balancing provided an approximately equal number of samples for model training and overcomes the problem of model skewness in the majority class. Second, SMOTe is the most widely used oversampling approach for data balancing and is predominantly used by existing studies. Third, TF-IDF is most widely used along with the SMOTE and shows better results than other feature engineering approaches. Last and foremost, despite several works investigating the influence of oversampling approaches on the performance of machine learning models, studies analyzing the comparative performance of various oversampling approaches rarely exist. Therefore, this study focuses on selecting the most commonly used oversampling approaches and performs a comparative study to fill this gap. Table 1 illustrates the summary of the state-of-the-art summary along with their limitations and contributions.

Materials and proposed methodology

This section describes the datasets, oversampling approaches, machine learning models, and evaluation parameters used in this study in detail. Figure 1 shows the flow of the methodology adopted for this study. Starting with the data acquisition, the study performs the preprocessing followed by applying various oversampling approaches to balance the dataset. The data is then split into train and test subsets for training and testing, respectively.

Description of datasets

This study utilized the EndViolence dataset that was taken from the Kaggle databases [30]. Twitter API as well as Python code are used to gather tweets. A daily search for the hashtag #EndViolence is carried out for a specified number of days in order to acquire a bigger number of tweet data. The tweets include information such as the user's name, location, description, time, followers, friends, favorites, verified status, hashtags, and source. Over 10,000 tweets have been taken, of which 8537 are neutral and 919 are positive, while 544 tweets are deemed negative. The second

Table 1 Summary of the state-of-the-art works along with their limitations and contributions

References	Techniques/sampling	Objectives	Limitations	Contributions
[10]	SMOTE, machine learning	The major objective is to address the class imbalance issues with oversampling techniques and machine learning.	The authors only utilized one sampling technique, SMOTE, and simple machine learning.	The authors employed the SMOTE method to identify emotions in imbalanced datasets using machine learning. This approach enhances emotion classification and effectively tackles the problem of dataset imbalance.
[11]	SMOTE, RV-VC	The impact of smote on uneven text features for the RVVC method classification of toxic comments	The authors only utilized one sampling technique, SMOTE, and employed an ensemble model that is time-consuming.	The authors employed the RVVC model, which enhances the performance of machine learning with a balanced dataset.
[12]	NB, SVM, SMOTE	The main aim is to perform sentiment analysis using balanced datasets with oversampling.	The entire work uses SVM in addition to NB, despite the fact that text data preprocessing techniques are not utilized in an appropriate manner.	The SMOTE technique is used to assess the efficacy of Naive Bayes and SVM algorithms on imbalanced datasets within the domain of sentiment analysis. The data demonstrate that attaining improved results depends on various shared elements.
[13]	Word2vec, Ensemble classifiers, SMOTENC	Theories concerning COVID-19 that are linked to Conspiracy and the Application of Ensemble Classifiers Performing a sentiment analysis on data taken from Arabic social networks on Twitter.	The tweets are extracted and classified into sentiments in Arabic so that only limited geographical regions can explore the concepts.	They test how well it works to add many classifiers to the model using ensemble learning techniques to make the model more accurate.
[14]	Word2vec, bagging classifiers, SMOTE	Initially, they evaluate word embedding using several classifiers on a significantly imbalanced sample. Following the implementation of SMOTE, an evaluation is conducted to assess its effects on performance.	This research focuses on the imbalanced data sensitization analysis of the short Arabic language.	This study compares the effectiveness of various classifiers in determining sentiment polarity in extremely imbalanced short-text samples. Instead of using manually created features, the study utilizes features derived from word embedding.
[16]	Machine learning, SMOTE, Borderline SMOTE	The objective of this study is to address the issue of imbalanced sarcasm detection in social media by employing oversampling of synthetic minorities.	The experiments were only tested on one dataset for sarcasm detection in social media.	Six machine learning algorithms, along with the SMOTE and Borderline SMOTE techniques, are used to do full sentiment analysis to address the imbalance concerns.
[18]	SMOTE, recursive neural networks, word embedding	The researchers' objective is to use SMOTE as an oversampling technique to generate more data for minority classes and achieve an optimal result.	They applied word embedding Composition on data that was unbalanced, but the results that they got were not beneficial.	The authors propose a technique for oversampling that relies on the composition of word embeddings, resulting in the generation of well-balanced training data with semantic significance.
[20]	Machine learning, K-Means SMOTE	The key objective is to utilize explainable AI and machine learning techniques to predict the performance of secondary-level students.	The dataset that was used is small, and the criteria for evaluation were not thoroughly investigated.	The K-Means SMOTE method was utilized to accomplish the following objectives: remove erratic data, evaluate machine learning methodologies, and determine the algorithm that would produce the maximum accuracy.

Table 1 (continued)

References	Techniques/sampling	Objectives	Limitations	Contributions
[22]	TFIDF, SMOTE, NBC, SVM, LR	Its main objective is to use social media to predict concerns about the COVID-19 vaccine.	Most users on the website are either teenagers or young adults in their twenties and thirties. People of advanced age and younger people have less involvement.	Their paper encompasses significant contributions, including a substantial number of studies, a comparison analysis, and the use of the TF-IDF and SMOTE techniques.
[23]	TFIDF, SMOTE, NB, SVM	Analyses the sentiment of an educational curriculum incorporating e-sports by employing Naive Bayes and Support Vector Machine algorithms.	The authors conducted experiments using solely SVM and NB algorithms, both with and without the use of SMOTE. However, the results obtained were somewhat poor in terms of accuracy and AUC.	An experiment comparing different algorithms was carried out to identify the one that provided the most accurate results. The strategy of data mining was utilized to examine it in comparison to the Naive Bayes method and the support vector machine.
Proposed	BoW, TDIDF, machine learning, SMOTE, SVM-SMOTE, ADASYN, K-means SMOTE, Borderline SMOTE	This paper focuses on analyzing the suitability and efficacy of different sampling approaches to solve the problem of the imbalanced dataset as well as improve the performance of machine learning models. The model's performance varies depending on whether the number of class samples is fewer or higher if the class is not properly balanced. This research makes use of two extremely imbalanced Twitter datasets.	First, using borderline SMOTE could result in an excessive number of synthetic samples, which would introduce bias into the model. Secondly, the influence of different oversampling approaches was not examined throughout the tests for deep learning.	The Impact of various sampling approaches is investigated with respect to the performance of machine learning models for imbalanced datasets. For this purpose, five sampling approaches are employed including SMOTE, support vector machine-SMOTE (SVM-SMOTE), K-means SMOTE, adaptive synthetic (ADASYN) oversampling, and Border-Line SMOTE. The selection of these sampling approaches is made regarding their wide use in the existing literature and reported performances. Kfold, BoW, and TFIDF features are used in the experiments.

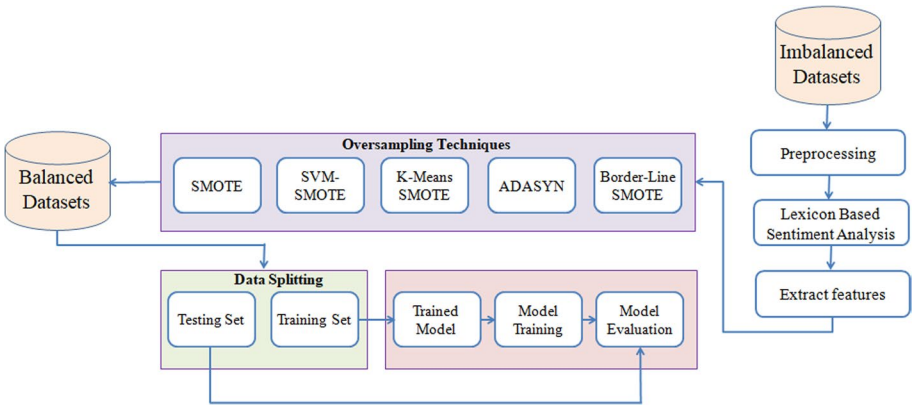

Fig. 1 Proposed work diagram

Table 2 The quantity of the Twitter dataset before and after performing sampling techniques

Sampling techniques	EndViolence dataset	Ecommerce dataset
Before applying sampling Technique	10000	14896
After applying SMOTE	25611	21870
After applying SVM_SMOTE	25611	21870
After applying K-Means SMOTE	25611	21870
After applying ADASYN	25967	20464
After applying Border-Line SMOTE	25611	21870

dataset is related to E-commerce tweets which are also taken from Kaggle. It contains tweets related to various online e-commerce websites, online business stores, etc. Both datasets have an imbalanced distribution of positive, negative, and neutral tweets and are highly imbalanced. We are dealing with multiclass text classification and imbalanced data classification problems. The authors employ imbalanced datasets as a means of addressing the challenge of multiclass text classification. Several oversampling procedures are employed to address the issue of imbalanced classification. One class has a much higher level of samples compared to the others in an imbalanced classification. The problems of minority classes are solved using sampling techniques. The quantity of the Twitter dataset before and after performing sampling techniques is presented in Table 2.

Preprocessig

Preprocessing is the process by which unstructured data is transformed into representations that are intelligible and suitable for machine learning models [31]. Preprocessing is largely utilized in order to improve the quality of input data by minimizing the amount of noise, redundant data, and unnecessary data. This allows the machines to detect the essential patterns that can be used to extract meaningful and relevant information from the preprocessed data. The data in this research is preprocessed using the steps described below in Table 3.

Table 3 Pre-processing steps followed in this study

Step	Description
Convert to lower case	In the process of preprocessing, we converted all of the text to lowercase. This conversion is critical because of the case sensitivity issue of machine learning models. The models treat the words 'happy' and 'Happy' differently if conversion is not performed on the data.
Stop words removal	The elimination of information that is irrelevant to classification tasks is a significant aspects of preprocessing. Models can not employ stop words to make decisions since they have no meaning. To improve the model's performance, unnecessary words must be removed from the data during preprocessing.
Removal of punctuation	Data preprocessing includes removing punctuation, such as "[#,$,@,""!,"" &," and "*," from the data. Nothing is changed, except that it makes it easier for the machine learning models to distinguish between distinct characters when it is removed.
Number and URL's removal	Numerical values are removed in order to enhance the data's quality. There are several challenges with feature extraction in text data since it contains many quantitative values that are of little utility in decision-making processes. In order to classify data, it is not necessary to use values that include any numeric information. Also, the uniform resource locators (URLs) are removed from the text.
Stemming and Lemmatization	As a result of stemming, machine learning models will be able to perform better. Stemming removes the 's' or 'es' at the end of the word to change it back to its root form. It is the process of turning a word into its original form, which is called lemmatization. For example, lemmatization evaluates context and translates words to their base form, while stemming just eliminates the final few letters. Stemming frequently results in wrong meaning and misspelling problems.

Feature engineering

For training the machine learning model, feature engineering is used to extract important information from the raw data. Preparing data for the models is a critical step; before building a learning model, one must first pick what features to employ [32]. In order to increase prediction accuracy, learning algorithms use features retrieved from data depending on the data's structure. The model's performance is improved as a result of selecting the appropriate feature extraction approach. For sentiment classification, the choice of feature engineering approach plays a significant role. In view of their wide use for sentiment classification, this study uses BoW and TF-IDF features.

In feature engineering, features are taken from a network intrusion dataset using the Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction methods. The utilization of the BoW method in this work was motivated by its user-friendly nature, computational efficiency, and satisfactory performance on extensive textual datasets. However, to overcome the constraints associated with BoW and enhance the efficacy of the feature engineering process, we employed the more sophisticated TFIDF methodology. These methodologies demonstrate enhanced efficiency in handling intricate textual data. The utilization of other methodologies, such as word embedding and transformers, is limited in our study due to their high computing requirements and intricate architectural design. Transformers have a substantial computational overhead, necessitating a significant amount of time for computation. The technique of feature extraction employed by transformers is characterized by a higher degree of complexity when compared to alternative methodologies. Furthermore, the alternative approaches exhibit limitations in successfully capturing contextual information and introduce biases in the data as a result of their large model size and limited training data.

Bag of words

The BoW approach is easy to comprehend and execute the approach, and it provides a great deal of customization for individual text data. It has been utilized successfully on a variety of prediction tasks including language modeling and classification of documentation. The BoW model [33] disregards the syntax and order of a text but retains the score of its words, which aids in feature generation. The BoW features of the following given sample text are represented in Table 4

Sample 1: help create multiple online platforms work online

Sample 2: help multiple internet platforms

Sample 3: collaborate across online media

Term frequency-inverse document-frequency

TFIDF scores on word frequency are used to assign weights to the most important terms in a text, even if they do not appear frequently in the content as a whole. It is an approach to make sparse features out of text data. Word frequency in a document is calculated and multiplied by the inverse document frequency, the total number of documents containing that word. TF is calculated using

$$tf = \frac{f_{xy}}{n_y} \tag{1}$$

In Eq. (1), f_{xy} is the occurrence of term x in document y and n_y is the total words in the given document. The inverse document frequency of each term is calculated by

$$idf = 1 + log\frac{N}{d_{fx}} \tag{2}$$

In Eq. (2), idf represents the inverse document frequency, N represents the number of times term x appears in document y, and d_{fx} represents the occurrence of total documents appearing in term x. Also, the log function is used to streamline their interpretation and facilitate their influence to compress the large corpus. The term frequency and inverse document frequency of each term are calculated by

$$w_{xy} = tf_{xy} \times idf_x \tag{3}$$

Table 4 BoW features for the sample text

	Help	**Create**	**Multiple**	**online**	**platform**
Sample 1	1	1	1	2	1
textbfSample 2	1	0	1	0	1
Sample 3	0	0	0	1	0
	Work	**Internet**	**Collaborat**	**Across**	**Media**
Sample 1	1	0	0	0	0
Sample 2	0	1	0	0	0
Sample 3	0	0	1	1	1

In Eq. (3), we multiply tf_{xy} and idf_x to attained the TFIDF score (w_{xy}) for term x in document y. Machine learning models do not take the data in textual form, all the data is vectorized by TFIDF and fed to the machine learning models for training [34].

Comparison between TFIDF and BoW

TF-IDF is considered to be a more advanced and efficient method for text analysis compared to BoW. This is due to its ability to take into account the significance of terms and their contextual relevance. By doing so, TF-IDF can minimize noise and enhance the overall quality of data obtained from text data. The BoW approach operates based on the frequency score of each word. One limitation of the BoW approach is its tendency to obscure the semantic meaning of individual words. For example, the semantic counterpart of "not awful" might be described as "decent" or even "pleasant." However, when employed in isolation, the phrases "not" and "awful" elicit negative emotions. There exist several techniques for reducing the dimensionality of the feature space. However, the initial phase of the BoW approach suffers from the drawback of just prioritizing terms based on their frequency counts. To address this issue, the TF-IDF strategy is presented, including a simple modification to the conventional BoW approach. TFIDF addresses the following problems faced by the Bow approach.

- TFIDF handles word importance and common words.
- The sequential nature of data is not taken into consideration by the BoW method used in text classification, potentially leading to the loss of important information and data sequence.
- The BoW method considers synonyms of terms to have discrete and distinct features. The BoW method is ineffective at capturing the data's inherent semantic information.
- The presence of a large number of datasets or data with a large number of dimensions may result in computational challenges and increased memory requirements.
- The BoW method may struggle with negations and modifiers. This method struggles to manage out-of-vocabulary terms not encountered during training.

Description of oversampling techniques

Machine-based methods give a poor performance on imbalanced classification datasets. The main reason is that machine-based methods are designed to operate on an equal number of samples for each class on classification datasets. In the case of an imbalanced dataset, the models tend to train on the majority class and show skewed performance. Data sampling techniques provide a variety of ways to handle and transform the dataset into a balanced class distribution. In oversampling, there is no loss of information from the original data since it keeps all the samples of minority and majority classes.

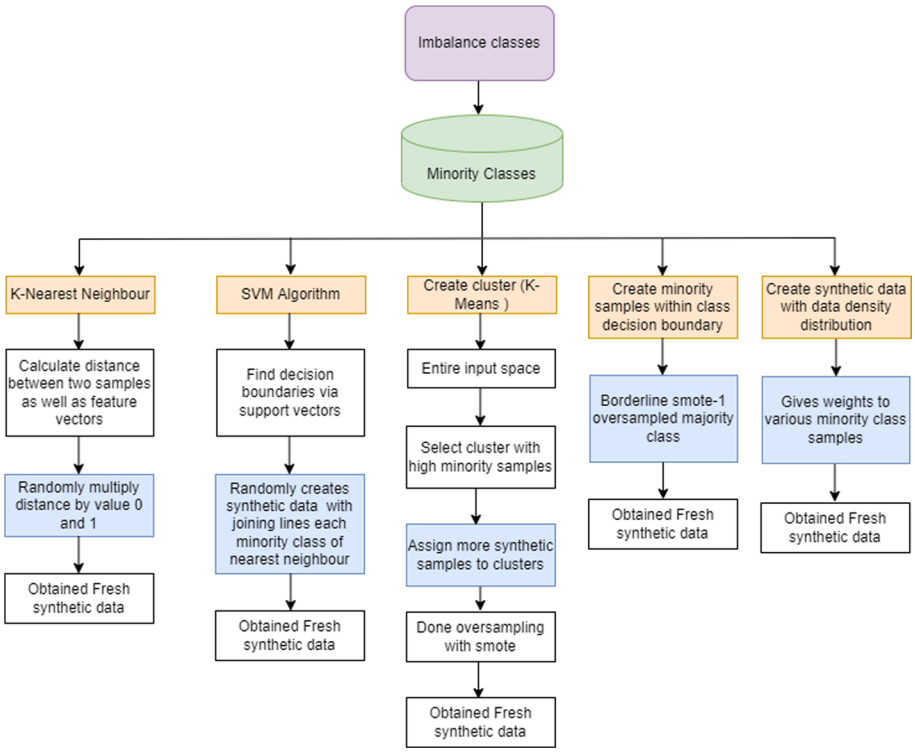

Fig. 2 Workflow diagram for oversampling techniques

On the other hand, in undersampling, the samples from the majority class are randomly removed which causes information loss. So, predominantly, oversampling approaches are recommended by the researchers. This study also adopts the concept of oversampling. The workflow of oversampling techniques is represented in Fig. 2.

Synthetic minority oversampling technique

The minority data is replicated from the minority set of data in a traditional oversampling process. Although it increases the quantity of data available, it does not provide the classification model with any additional knowledge or variation. To address this problem, Chawla et al. [35] introduced the SMOTE which is the most effective and commonly used technique. For balancing the dataset, it generates fresh synthetic samples. SMOTE generates synthetic data using the KNN technique. It calculates the distance between the two sample points as well as the distance between the feature vector and its nearest neighbor. It then multiplies the distance by a value between 0 and 1 at random. At the computed distance, a new point is picked on the line segment. This procedure is repeated for each of the feature vectors that have been discovered. In terms of preventing overfitting and under-fitting, SMOTE performs better than basic sampling approaches [36].

Border-line SMOTE

Bridges of minority class points are produced due to the presence of some minority points inside the domain of majority class points. This problem in the case of SMOTE can be solved via Borderline SMOTE which is a modified version of the SMOTE. Borderline-SMOTE simply generates synthetic data within the class decision boundary and only a small number of examples along the borderline are over-sampled [37]. Most classification methods learn the borderline for every class throughout the training process in order to produce better results, which is especially crucial for classification tasks.

The SMOTE method utilizes a stochastic selection procedure to determine a sample from the set of minority samples inside the specified category. The Borderline SMOTE technique selectively selects samples for synthetic minority oversampling based on the criterion that the majority of surrounding instances belong to the majority group. To provide more clarification, it is important to note that the Borderline SMOTE is specifically designed to be applicable only to instances that are located on the outskirts of the minority class.

Advantages

- The decision boundary is defined by the boundary line, which also generates samples at the border. The task of classification has grown increasingly challenging, while performance in this area has seen significant improvement.
- The technique known as Borderline smote, as previously discussed, mitigates the risk of overfitting by producing samples in close proximity to the boundary line.
- This strategy effectively mitigates data noise by generating samples in predetermined locations with the use of the boundary line smote technique.
- Borderline SMOTE is a versatile approach that effectively improves the performance of classifiers in several areas because of its compatibility with a diverse set of machine learning techniques.

Limitations

- The utilization of computer resources in boundary line smote is greater compared to conventional smote due to its focus on borderline samples.
- The efficacy of the classification process may be influenced by the challenging endeavor of selecting and optimizing the parameters.
- The Synthetic Minority Over-sampling Technique (SMOTE) consistently generates synthetic samples, therefore effectively exploring the whole feature space. In contrast, Borderline SMOTE may not exhibit the same level of efficiency in this regard. This might provide a limitation in cases when the decision boundary is complex and encompasses many places.
- In certain regions of the feature space, the utilization of Borderline SMOTE may lead to an excessive presence of artificially generated samples, thus introducing bias into the model.

SVM-SMOTE

SVM-SMOTE is also known as borderline-SMOTE SVM. The primary distinction between SVM-SMOTE and other SMOTE variations is that rather than employing KNN to detect misclassification like in Borderline-SMOTE, SVM-SMOTE uses the SVM algorithm [38]. The SVM is used to find the decision boundary specified via support vectors, or a new sample is formed at random along the lines that connect each minority class support vector with the number of its own nearest neighbors via extrapolation, due to the density of the majority class samples near it.

Adaptive synthetic oversampling

The technique known as ADASYN [39] is based on adaptively producing minority data examples according to their densities. ADASYN is an enhanced version of SMOTE. In comparison to Borderline-SMOTE, ADASYN adopts a more unique approach. ADASYN provides synthetic data based on the data density, whereas Borderline-SMOTE seeks to synthesize data around the data decision boundary. The density of the minority class might be inversely related to the generation of synthetic data. It means that in sections of the feature set where the density of minority examples is low, more synthetic data is created, and in regions where the density is high, fewer or none are formed. The algorithm's fundamental function is to give weights to various minority class samples in terms of generating different amounts of synthetic data for each example.

K-Means SMOTE

K-Means SMOTE is a class-imbalanced data oversampling approach. It helps classification by creating minority class samples in safe and important portions of the input space. The method reduces noise while effectively resolving class imbalances. The K-means clustering technique is used to cluster all of the data and choose clusters with a high proportion of minority class samples. More synthetic samples should be assigned to clusters with sparse minority class samples [40].

Machine learning models

Machine learning is a key component of artificial intelligence because it allows machines to see patterns in large amounts of data and make future predictions based on that data. Small and large datasets can benefit from a machine's ability to enhance its performance. The Scikit-learn package is used to create supervised machine learning models. RF, DT,

Table 5 Hyperparameters settings for machine learning models

Models	Parameters setting
SVM	kernel='linear', C=3.0, random_state=200
DT	random_state=0
LR	solver="saga", multi_class="multinomial"
RF	n_estimators=300,max_depth=300
KNN	n_neighbors=5
AdaBoost	n_estimators=50, learning_rate=1

SVM, LR ADA, and KNN are utilized in this study. The hyperparameters for machine learning models are presented in Table 5.

Random forest

RF is a classification model that constructs multiple decision trees from the data. An RF [41] produces appropriate results without having adjusted hyperparameters and is one of the most used machine learning algorithms. RF is one of the most often used machine learning algorithms because of its simplicity and diversity. Using a collection of weak learners, RF creates a group of strong learners. The RF creates a set of decision trees that are independent of one another. Using the decision trees, a majority vote is obtained, which is then used to decide the classification.

Support vector machine

The SVM is a machine learning model that may be used for both classification and regression problems, and it is based on the support vectors. SVM is more suited for classification than other algorithms since it is a simpler algorithm that performs better in specific natural language processing applications [42]. The point in n-dimensional space that each data element represents is seen as a point in this technique, and the hyperplane that best distinguishes between these points is produced. The optimum hyperplane is chosen based on the highest variance between data points. This study utilizes a linear kernel for SVM since it was the most effective in our scenario.

K-nearest neighbor

KNN is one of the most straightforward supervised classification algorithms available in machine learning today. The KNN compares the new data to older examples and places it in the most comparable category. The distance between the new data and old classes is used to compute similarity. Distance estimate methods include Euclidean, Manhattan, and Cityblock. In classification issues, the KNN technique is utilized for both regression and classification. KNN is a non-parametric method and makes no inferences from the data. KNN has several settings that can be fine-tuned for accuracy. The benefit of KNN over other algorithms is that they can explain the classification result in situations where black-box models fail [43].

Logistic regression

One of the most often used machine learning algorithms LR is widely applied to classify data using a statistical function called logistic function (also known as a sigmoid function). In logistic regression, the probability ratio is directly modeled. Real values can only be between 0 and 1 for this function in order to forecast probability [44]. To ascertain the probability of an output variable, the supervised machine learning method LR is employed [45]. If the output or dependent variable is binary, it works well, but it may also be useful for multi-class data categorization. The data is categorized using the logistic function. It is possible to predict a dichotomous dependent variable using the LR regression approach. For the LR equation, the maximum-likelihood ratio (MLR) is utilized to identify which variables are statistically significant.

Decision tree

A DT is another well-known supervised machine learning technique that is used to carry out classification tasks in the context of data mining [46]. A tree-like structure is used to describe the predictive process in DT, where the implementation is performed in branch form, with the final prediction occurring at the leaf node. DT is built recursively and learned on the training dataset using the training set. The learning phase comes to an end when no more splitting can be done in the tree or when the output at the node is the same for the goal label or class (whichever comes first). DT does not need the setup of parameters or the understanding of a domain, making it an excellent tool for finding relevant and meaningful patterns in large amounts of data.

AdaBoost

To improve the overall performance of the final classifier, the AdaBoost (or Adaptive) classifier combines several weak classifiers into a single strong classifier. At each step, a powerful classifier is created by merging many classifiers together. ADA is based on the idea that one-level DTs (weak learners) are introduced sequentially to the ensemble classifier [47].

Evaluation parameters

The performance of a machine learning algorithm is determined using evaluation parameters. A machine learning model is applied to test data that the algorithms have already seen to determine how well they perform on it. Evaluation techniques analyze the model's performance and assign it a score based on its efficiency. To determine how effectively a model performs in classification tasks, it is combined with a training set of the same data. Typically, machine learning models are assessed using four parameters: accuracy, precision, recall, and F1 score. TP is the true positive rate, which refers to the true positive class that is predicted as positive. TN stands for true negative, which refers to the models' true negative predictions across all negative data. The false positive rate (FP) is the proportion of real negative forecasts that are labeled as positive. FN stands for false negative, which refers to records that belong to the positive class but are projected to be negative by the models.

The accuracy of classifiers on test data is defined as the ratio of correct predictions to the total number of predictions made by the classifiers. The highest accuracy value is 1, demonstrating that all predictions from the classifier are accurate, while the lowest accuracy score is 0. Accuracy may well be estimated using the formula

$$Accuracy = \frac{(TP + FN)}{(TP + TN + FP + FN)} \tag{4}$$

Precision, also known as positive specificity, is a measure of the proportion of properly classified cases among all correctly classified samples. A precision value of 1 indicates that every occurrence of data that has been classified as positive is positive.

$$Precision = \frac{TP}{(TP + FP)} \tag{5}$$

When comparing positive predictions, recall is a statistic that measures the proportion of accurate positive predictions produced of all possible positive predictions. In contrast to precision, which still comments upon that accurate positive prediction from all

positive predictions, recall offers an indication of the positive predictions that are not correctly predicted.

$$Recall = \frac{TP}{(TP + FN)} \tag{6}$$

Results and Discussions

In order to solve the problem of oversampling techniques on imbalanced datasets as well as to evaluate the performance of machine learning models, several experiments are carried out in this study. Two tweet datasets are used for the experiments to achieve this goal. The train-test split is in the ratio of 0.75 to 0.25 where 75% data is used for training the model and 25% for testing. To extract relevant features from the tweets, two well-known feature engineering approaches BoW and TF-IDF are adopted.

Table 6 Accuracy score of different models on balanced datasets

Models	Datasets	Features	SMOTE	SVM-SMOTE	K-Means SMOTE	ADASYN	Border-Line SMOTE
LR	EndViolence tweets	BoW	83.21	91.89	88.08	77.77	86.75
		TFIDF	98.89	98.84	98.78	98.93	98.85
	E-commerce tweets	BoW	89.52	91.20	93.94	90.12	89.08
		TFIDF	94.89	94.56	92.68	94.21	95.51
RF	EndViolence tweets	BoW	86.11	92.95	90.27	81.74	89.37
		TFIDF	99.26	99.10	98.46	99.21	99.21
	E-commerce tweets	BoW	85.55	88.86	93.59	84.01	85.42
		TFIDF	94.42	94.91	94.31	93.67	94.60
SVM	EndViolence tweets	BoW	83.32	92.30	88.34	78.06	87.25
		TFIDF	99.57	99.54	99.46	99.67	99.59
	E-commerce tweets	BoW	87.01	88.20	95.43	90.46	85.95
		TFIDF	97.64	97.43	96.41	97.55	97.60
DT	EndViolence tweets	BoW	85.49	92.94	89.53	81.28	89.00
		TFIDF	98.03	97.65	97.78	97.93	97.65
	E-commerce tweets	BoW	83.88	86.46	96.48	84.12	83.10
		TFIDF	94.29	93.70	96.45	92.13	92.81
ADA	EndViolence tweets	BoW	78.61	88.19	84.88	69.14	80.66
		TFIDF	90.98	90.64	88.97	89.18	88.83
	E-commerce tweets	BoW	63.09	68.47	85.45	62.93	66.00
		TFIDF	69.23	72.16	73.10	73.03	73.55
KNN	EndViolence tweets	BoW	84.50	92.20	83.05	74.58	83.21
		TFIDF	95.04	94.95	95.86	95.53	94.95
	E-commerce tweets	BoW	63.09	65.10	74.62	57.42	63.23
		TFIDF	71.04	68.70	76.96	61.61	71.56
GBM	EndViolence tweets	BoW	85.34	91.20	84.27	78.27	86.23
		TFIDF	96.24	95.25	96.03	96.38	95.93
	E-commerce tweets	BoW	74.19	71.24	74.19	65.35	72.49
		TFIDF	77.23	75.26	78.94	73.28	72.38

Results of ML models on oversampled datasets

Table 6 shows the accuracy results of machine learning models using different oversampling techniques with BoW and TF-IDF features. Five oversampling techniques SMOTE, SVM-SMOTE, K-Means SMOTE, ADASYN, and Border-Line SMOTE are used in this study. Of the employed machine learning models, the SVM model performs best and achieved 99.67% accuracy on the ADASYN technique, 99.57% using SMOTE, and 99.59% using Border-Line SMOTE technique using the TFIDF feature extraction approach. The SMOTE oversampling technique, which uses the KNN algorithm to artificially enhance the data, outperforms all other techniques for the EndViolence tweets dataset, which is extremely imbalanced. The SMOTE also performs well for E-commerce tweet datasets and attained a 99.26% accuracy score using TFIDF features with RF. However, when these datasets are used, the BoW features do not outperform well as compared to TF-IDF features.

Several patterns can be observed from these experiments. First, machine learning models perform better when used with TF-IDF features for both datasets. Second, predominantly, SMOTE, ADASYN, and Border-line SMOTE show better results for all the

Table 7 Precision score for positive tweets using different models with sampled datasets

Models	Datasets	Features	SMOTE	SVM-SMOTE	K-Means SMOTE	ADASYN	Border-Line SMOTE
LR	EndViolence tweets	BoW	90.06	95.45	92.76	83.18	93.13
		TFIDF	98.23	98.05	98.08	98.05	98.05
	E-commerce tweets	BoW	94.34	94.14	94.65	96.43	92.23
		TFIDF	96.07	95.34	93.76	97.23	97.34
RF	EndViolence tweets	BoW	90.76	94.23	93.02	86.06	94.77
		TFIDF	98.53	98.33	99.17	99.47	99.43
	E-commerce tweets	BoW	91.90	92.51	96.11	97.10	90.40
		TFIDF	98.03	98.32	97.23	99.42	97.85
SVM	EndViolence tweets	BoW	87.55	95.12	91.72	82.74	91.38
		TFIDF	99.07	99.19	98.06	99.26	99.03
	E-commerce tweets	BoW	92.49	90.44	96.63	97.23	82.54
		TFIDF	98.77	98.35	97.48	98.76	98.46
DT	EndViolence tweets	BoW	91.36	94.66	91.75	86.45	93.45
		TFIDF	98.12	97.29	97.23	95.76	95.56
	E-commerce tweets	BoW	90.40	88.23	96.24	92.75	84.43
		TFIDF	96.23	95.45	97.06	94.23	93.78
ADA	EndViolence tweets	BoW	92.23	96.46	90.09	78.23	91.02
		TFIDF	85.24	83.88	83.22	85.76	84.34
	E-commerce tweets	BoW	71.20	71.66	95.23	86.65	82.34
		TFIDF	64.34	73.24	62.76	78.43	80.56
KNN	EndViolence tweets	BoW	91.34	95.09	69.46	78.87	72.64
		TFIDF	97.34	95.67	91.76	95.84	88.38
	E-commerce tweets	BoW	85.45	78.35	96.45	97.09	67.97
		TFIDF	85.65	84.07	59.80	90.96	84.46
GBM	EndViolence tweets	BoW	92.98	93.13	74.38	80.73	72.64
		TFIDF	96.18	95.67	92.10	97.40	90.19
	E-commerce tweets	BoW	89.28	83.20	92.10	93.20	78.90
		TFIDF	87.34	86.21	76.36	93.91	88.28

Table 8 Recall score for positive tweets using oversampling techniques

Models	Datasets	Features	SMOTE	SVM-SMOTE	K-Means SMOTE	ADASYN	Border-Line SMOTE
LR	EndViolence tweets	BoW	71.78	85.45	73.02	64.43	72.34
		TFIDF	99.01	99.46	99.44	100.0	100.0
	E-commerce tweets	BoW	83.56	87.43	91.22	87.44	82.54
		TFIDF	89.67	91.23	91.46	84.77	91.35
RF	EndViolence tweets	BoW	78.35	87.65	78.55	71.53	79.78
		TFIDF	99.99	99.98	99.98	99.98	99.98
	E-commerce tweets	BoW	81.07	87.87	79.23	78.56	84.86
		TFIDF	94.76	87.56	91.35	81.56	87.65
SVM	EndViolence tweets	BoW	74.08	85.76	73.98	64.43	75.07
		TFIDF	99.99	99.99	100.0	99.98	100.0
	E-commerce tweets	BoW	82.27	85.44	83.23	88.76	84.55
		TFIDF	87.76	95.34	94.64	93.67	95.24
DT	EndViolence tweets	BoW	77.21	88.22	76.76	70.65	78.66
		TFIDF	97.45	97.75	98.24	98.23	97.76
	E-commerce tweets	BoW	78.04	85.24	95.05	78.45	78.05
		TFIDF	91.35	90.48	95.35	85.43	79.26
ADA	EndViolence tweets	BoW	62.45	75.07	65.35	45.36	56.98
		TFIDF	91.44	90.28	87.24	86.06	87.19
	E-commerce tweets	BoW	67.01	66.74	64.39	76.24	61.06
		TFIDF	55.17	53.39	71.42	57.49	61.29
KNN	EndViolence tweets	BoW	75.27	86.04	93.23	64.37	92.26
		TFIDF	100.0	99.98	99.98	99.99	99.99
	E-commerce tweets	BoW	52.34	65.46	32.45	28.35	62.37
		TFIDF	68.23	60.35	96.06	40.32	70.22
GBM	EndViolence tweets	BoW	81.34	88.26	95.34	70.35	95.34
		TFIDF	93.53	95.34	96.54	97.53	99.99
	E-commerce tweets	BoW	65.63	70.54	50.43	56.43	69.65
		TFIDF	73.45	74.32	97.64	69.54	80.23

machine learning models, specifically when used with TF-IDF features. Third, the performance of the models is more smooth using TF-IDF and oversampling techniques and shows more consistent results as compared to the use of BoW where the performance has large gaps for various oversampling approaches.

Table 7 shows the precision results of oversampling techniques employing machine learning models with BoW and TF-IDF features on the oversampled dataset. The precision score can also be used to evaluate oversampling techniques based on machine learning models. SMOTE, SVM-SMOTE, ADASYN, and Border-Line SMOTE techniques achieved 99.47 % with the RF model and SVM whereas RF with TF-IDF feature extraction methods achieved 98.03% with the SVM-SMOTE methodology. As for BoW features using the RF model, ADASYN and K-Means SMOTE work well and achieve 97.23% precision.

Table 8 represents the recall score for ML models on oversampling techniques with balanced datasets using TF-IDF and BoW features. The results show that RF, SVM, and LR perform superbly on the Endviolence tweets dataset using TF-IDF features with all oversampling techniques and can attain a 99.99% recall score. The KNN obtains the

Table 9 Precision score of negative tweets machine learning models using balanced dataset

Models	Datasets	Features	SMOTE	SVM-SMOTE	K-Means SMOTE	ADASYN	Border-Line SMOTE
LR	EndViolence tweets	BoW	77.34	86.48	78.44	73.25	79.05
		TFIDF	98.22	99.05	99.26	99.46	99.35
	E-commerce tweets	BoW	85.35	90.35	99.47	87.27	85.06
		TFIDF	98.35	98.06	99.23	96.24	98.37
RF	EndViolence tweets	BoW	81.24	88.43	80.64	77.34	81.12
		TFIDF	98.47	99.98	99.86	99.98	99.99
	E-commerce tweets	BoW	78.13	88.32	99.24	77.32	85.36
		TFIDF	99.34	99.18	99.34	98.16	99.46
SVM	EndViolence tweets	BoW	77.54	87.46	77.23	73.00	79.76
		TFIDF	99.87	99.99	99.06	99.99	99.05
	E-commerce tweets	BoW	80.00	83.35	98.47	85.42	79.34
		TFIDF	99.00	99.24	99.00	98.16	99.05
DT	EndViolence tweets	BoW	80.40	88.35	80.05	76.44	81.53
		TFIDF	98.32	98.28	98.34	99.23	99.45
	E-commerce tweets	BoW	78.07	83.32	98.11	82.36	82.42
		TFIDF	96.23	96.22	97.10	95.05	95.43
ADA	EndViolence tweets	BoW	71.08	81.21	73.12	60.22	69.13
		TFIDF	91.21	91.33	87.13	85.22	91.32
	E-commerce tweets	BoW	94.95	94.47	95.07	50.34	99.44
		TFIDF	93.10	92.05	97.42	89.91	86.13
KNN	EndViolence tweets	BoW	78.05	86.40	91.05	57.94	90.43
		TFIDF	89.04	90.04	97.12	90.75	98.43
	E-commerce tweets	BoW	50.32	58.10	99.36	45.97	52.04
		TFIDF	58.08	56.33	96.96	48.67	59.43
GBM	EndViolence tweets	BoW	74.54	80.54	86.44	83.22	89.44
		TFIDF	78.53	89.09	90.37	90.34	91.32
	E-commerce tweets	BoW	45.54	54.32	88.43	56.43	56.44
		TFIDF	60.43	59.62	89.53	56.89	67.83

same recall score for positive tweets using BoW and TF-IDF features on all oversampling techniques except for the K-means SMOTE where its recall score is 32.45% with BoW features and 96.06% with TF-IDF features. Results demonstrate that TFIDF features are very effective for balanced datasets while BoW features do not perform well for balanced tweets datasets in our work. Overall, the Borderline SMOTE shows superior performance than other oversampling approaches.

Table 9 represents the precision score for machine learning models on oversampled datasets with BoW and TF-IDF features. ADASYN and SVM-SMOTE attained a 99.99% precision score with TF-IDF features and ADA attained a 99.44% precision score using BoW features on the E-commerce tweets dataset.

Table 10 presents the recall score for tweets of ML models on over-sampled datasets using features BoW and TFIDF. LR, SVM, and RF these three models perform outclass using features TFIDF and increase the models' performance overall but using features BoW, these models do not perform well.

Finally, all oversampling techniques artificially increase the size and features of the datasets used to train machine learning models. For positive and negative tweets, the

Table 10 Recall score of negative tweets using different machine learning models

Models	Datasets	Features	SMOTE	SVM-SMOTE	K-Means SMOTE	ADASYN	Border-Line SMOTE
LR	EndViolence tweets	BoW	82.01	95.12	95.09	74.12	92.33
		TFIDF	99.99	99.99	100.0	99.98	99.99
	E-commerce tweets	BoW	95.23	94.67	92.23	95.35	96.23
		TFIDF	98.23	98.43	90.02	100.0	99.43
RF	EndViolence tweets	BoW	84.67	95.33	95.12	78.77	94.32
		TFIDF	99.99	99.98	99.97	100.0	99.99
	E-commerce tweets	BoW	94.21	94.44	92.43	95.23	92.00
		TFIDF	96.41	96.39	93.32	97.87	96.13
SVM	EndViolence tweets	BoW	81.09	94.97	95.23	75.07	90.92
		TFIDF	99.98	99.98	99.97	100.0	99.99
	E-commerce tweets	BoW	96.23	96.48	95.04	95.09	95.45
		TFIDF	99.99	99.22	96.32	98.32	99.21
DT	EndViolence tweets	BoW	85.35	94.56	95.42	79.75	93.66
		TFIDF	97.87	97.45	97.65	97.65	97.65
	E-commerce tweets	BoW	93.77	93.23	96.45	94.24	90.53
		TFIDF	95.98	95.34	95.67	95.23	94.54
ADA	EndViolence tweets	BoW	78.44	93.04	93.32	68.12	90.12
		TFIDF	84.32	83.08	82.23	85.24	80.34
	E-commerce tweets	BoW	25.00	43.45	90.34	90.34	39.34
		TFIDF	58.34	66.45	51.32	65.60	65.45
KNN	EndViolence tweets	BoW	85.48	96.34	58.67	95.43	61.95
		TFIDF	99.99	99.98	99.97	99.99	100.0
	E-commerce tweets	BoW	98.34	95.24	85.45	99.98	87.24
		TFIDF	99.99	99.20	87.15	99.98	99.99
GBM	EndViolence tweets	BoW	78.33	88.59	60.43	90.36	65.34
		TFIDF	90.87	92.89	94.36	94.89	96.10
	E-commerce tweets	BoW	88.98	92.87	88.39	95.39	90.37
		TFIDF	95.93	95.29	93.20	96.40	96.34

RF model achieves the maximum accuracy, precision, and recall score on SMOTE and ADASYN oversampled datasets, according to the results of experiments using the TF-IDF features. The only model KNN performs well with BoW and TF-IDF features with a recall score of 1.00% for the K-means SMOTE case.

In Fig. 3 accuracy of different supervised machine learning models on imbalanced data sets i.e., the EndViolence tweets dataset and E-commerce tweets dataset are represented. Experiments utilized BoW features on balanced Endviolenced tweets with six models and the graph shows that SVM-SMOTE works very well. Similarly, K-means SMOTE also performs well but in Fig. 3(b) K-means SMOTE is on top of all techniques with the highest accuracy using TF-IDF features. Figure 3(c) shows that the SVM-SMOTE performs best with some models and K-means SMOTE with RF model using BoW features. Figure 3(d) represented that SMOTE and ADASYN attained much better results with TFIDF features. Overall, SVM-SMOTE and K-means SMOTE perform well using BoW while SMOTE and ADASYN attained the highest accuracy with TF-IDF features as compared to BoW features.

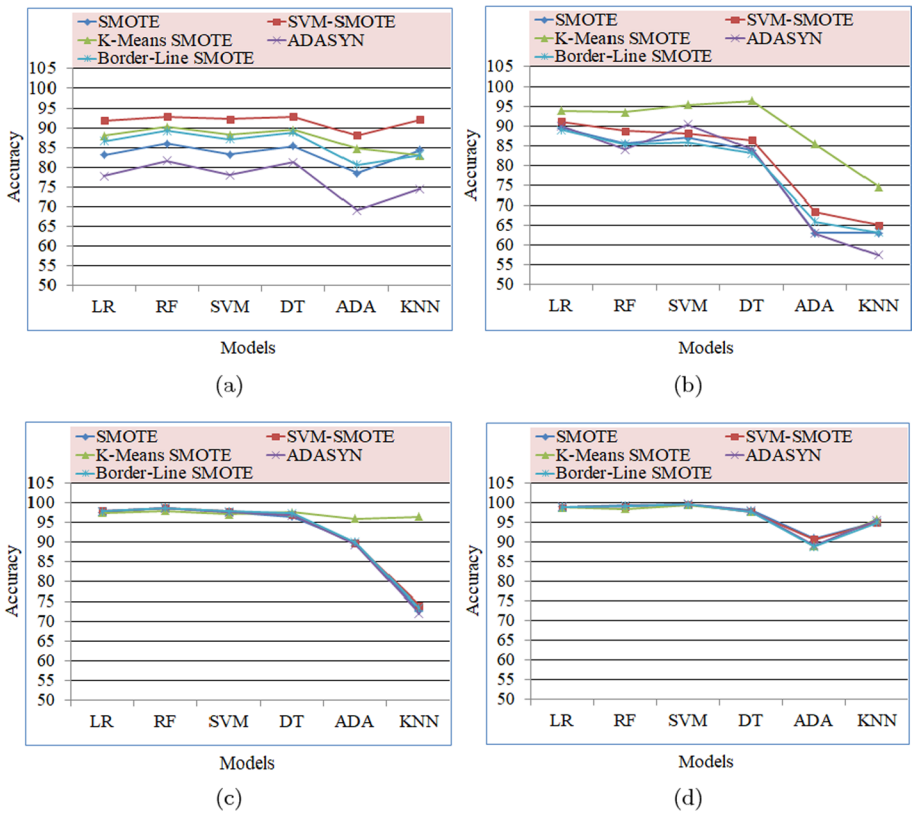

Fig. 3 Accuracy of different models on balanced datasets.**a** EndViolence tweets dataset with BoW features, **b** EndViolence tweets dataset with TFIDF features, **c** E-commerce tweets dataset with BoW features, and **d** E-commerce tweets dataset with TFIDF features

Table 11 10-fold results using models on oversampled E-Commerce dataset

Model	Oversampling-techniques					
	Features	SMOTE	SVM-SMOTE	K-Means SMOTE	ADASYN	Border-Line SMOTE
LR	BoW	0.880 ± 0.055	0.894 ± 0.060	0.925 ± 0.071	0.896 ± 0.074	0.868 ± 0.066
	TDIDF	0.944 ± 0.026	0.945 ± 0.026	0.903 ± 0.103	0.936 ± 0.020	0.948 ± 0.024
RF	BoW	0.836 ± 0.055	0.853 ± 0.065	0.931 ± 0.070	0.792 ± 0.087	0.826 ± 0.068
	TDIDF	0.941 ± 0.034	0.939 ± 0.033	0.916 ± 0.086	0.935 ± 0.020	0.934 ± 0.043
SVM	BoW	0.860 ± 0.52	0.874 ± 0.057	0.909 ± 0.058	0.897 ± 0.064	0.848 ± 0.064
	TDIDF	0.976 ± 0.010	0.974 ± 0.012	0.943 ± 0.059	0.974 ± 0.008	0.976 ± 0.013
DT	BoW	0.810 ± 0.052	0.828 ± 0.064	0.923 ± 0.049	0.767 ± 0.069	0.793 ± 0.068
	TDIDF	0.938 ± 0.033	0.930 ± 0.042	0.946 ± 0.041	0.891 ± 0.042	0.912 ± 0.053
ADA	BoW	0.634 ± 0.037	0.657 ± 0.048	0.839 ± 0.112	0.684 ± 0.114	0.591 ± 0.096
	TDIDF	0.693 ± 0.041	0.688 ± 0.036	0.823 ± 0.108	0.704 ± 0.042	0.707 ± 0.045
KNN	BoW	0.611 ± 0.057	0.610 ± 0.101	0.733 ± 0.122	0.526 ± 0.069	0.605 ± 0.064
	TDIDF	0.690 ± 0.068	0.649 ± 0.065	0.750 ± 0.126	0.605 ± 0.063	0.685 ± 0.060
GBM	BoW	0.872 ± 0.046	0.667 ± 0.102	0.787 ± 0.101	0.553 ± 0.071	0.648 ± 0.069
	TDIDF	0.710 ± 0.029	0.684 ± 0.054	0.774 ± 0.110	0.645 ± 0.046	0.656 ± 0.059

Table 12 10-fold cross-validation results using models on oversampled EndViolence tweets dataset

Model	Oversampling-techniques					
	Features	SMOTE	SVM-SMOTE	K-Means SMOTE	ADASYN	Border-Line SMOTE
LR	BoW	0.837 ± 0.008	0.906 ± 0.694	0.846 ± 0.059	0.747 ± 0.040	0.85 3 ± 0.017
	TDIDF	0.989 ± 0.010	0.988 ± 0.004	0.986 ± 0.019	0.980 ± 0.018	0.987 ± 0.014
RF	BoW	0.873 ± 0.009	0.929 ± 0.499	0.870 ± 0.051	0.772 ± 0.039	0.86 ± 0.018
	TDIDF	0.992 ± 0.008	0.990 ± 0.012	0.986 ± 0.023	0.988 ± 0.014	0.989 ± 0.018
SVM	BoW	0.837 ± 0.013	0.912 ± 0.673	0.846 ± 0.058	0.749 ± 0.030	0.860 ± 0.027
	TDIDF	0.995 ± 0.004	0.993 ± 0.007	0.994 ± 0.005	0.994 ± 0.008	0.993 ± 0.009
DT	BoW	0.868 ± 0.018	0.926 ± 0.056	0.868 ± 0.060	0.765 ± 0.033	0.883 ± 0.028
	TDIDF	0.978 ± 0.018	0.977 ± 0.021	0.980 ± 0.023	0.954 ± 0.019	0.978 ± 0.026
ADA	BoW	0.795 ± 0.009	0.856 ± 0.087	0.790 ± 0.042	0.701 ± 0.033	0.793 ± 0.013
	TDIDF	0.906 ± 0.024	0.885 ± 0.043	0.889 ± 0.0.44	0.875 ± 0.036	0.880 ± 0.027
KNN	BoW	0.842 ± 0.033	0.898 ± 0.076	0.869 ± 0.057	0.726 ± 0.048	0.876 ± 0.019
	TDIDF	0.954 ± 0.015	0.957 ± 0.018	0.754 ± 0.124	0.937 ± 0.019	0.947 ± 0.023
GBM	BoW	0.902 ± 0.012	0.901 ± 0.065	0.884 ± 0.043	0.754 ± 0.029	0.902 ± 0.012
	TDIDF	0.973 ± 0.010	0.962 ± 0.013	0.784 ± 0.131	0.956 ± 0.014	0.961 ± 0.011

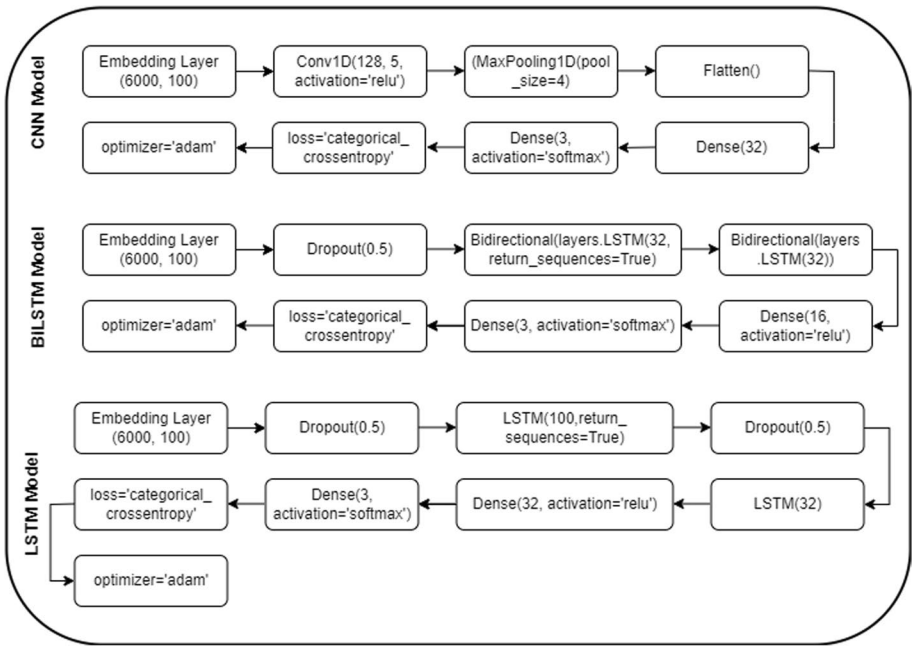

Fig. 4 Deep-learning architecture

K-fold cross-validation results on oversampled datasets

In 10-fold cross-validation, a dataset is divided into 10 parts randomly. The 90% data is used for training and the remaining 10% is used as a 'holdout-set' for testing randomly. We reserve different testing data and repeat the process 10 times at each iteration. It is a well-known and widely used method to show the efficiency and appropriateness of models. 10-fold cross-validation results are shown in Table 11 and Table 12. The experiments show that with 10-fold cross-validation, all ML models perform significantly well

while SVM attains the highest 0.973% accuracy using TF-IDF features on the SMOTE oversampled E-commerce dataset.

Similarly, 10-fold cross-validation results for the EndViolence dataset are given in Table 12. The highest accuracy of 0.995 is obtained using the SVM with SMOTE oversampled dataset while working with TF-IDF features. The performance of machine learning models is better and consistent with small variations when TF-IDF features are used for oversampling approaches.

Results of deep learning models on balanced datasets

In this section, the results of deep learning models on E-commerce and EndViolence datasets are represented. Figure 4 provides the details of the deep learning model implementation. These models are utilized with the 'categorical-crossentropy' loss function, for three classes. The 'softmax activation function is used while the 'Adam' optimizer is utilized for optimization. All models are fitted with 100 epochs and the batch size is set to 64.

Table 13 Results of deep learning models on E-commerce and EndViolence tweets dataset

Models	Dataset	Accuracy	Class	Precision	Recall
CNN	EndViolence tweets	96.11	Positive	84.12	80.34
			Negative	79.86	75.45
			Neutral	98.04	99.09
	E-commerce tweets	93.38	Positive	94.52	95.92
			Negative	87.23	78.23
			Neutral	95.41	96.04
LSTM	EndViolence tweets	95.95	Positive	77.33	89.19
			Negative	73.64	67.04
			Neutral	99.98	98.23
	E-commerce tweets	93.80	Positive	93.04	95.04
			Negative	88.21	82.42
			Neutral	96.23	96.23
BiLSTM	EndViolence tweets	95.65	Positive	79.32	85.21
			Negative	77.27	72.12
			Neutral	99.23	99.22
	E-commerce tweets	92.82	Positive	91.19	96.54
			Negative	86.23	82.05
			Neutral	96.04	92.23
GRU	EndViolence tweets	94.16	Positive	82.14	83.18
			Negative	75.10	67.98
			Neutral	89.09	90.02
	E-commerce tweets	90.98	Positive	90.19	92.12
			Negative	83.10	80.90
			Neutral	91.02	89.94
RNN	EndViolence tweets	90.91	Positive	74.34	82.10
			Negative	73.29	67.38
			Neutral	88.39	87.34
	E-commerce tweets	83.22	Positive	84.30	87.35
			Negative	79.37	79.96
			Neutral	83.20	85.29

Table 14 Statistical T test results

Models	Statistical t test	P-value	Hypothesis
SVM-ADASYN vs. CNN	50.346	0.003	Rejected
SVM-ADASYN vs. LSTM	13.422	0.005	Rejected
SVM-ADASYN vs. BiLSTM	13.245	0.005	Rejected
SVM-ADASYN vs. LR	5.188	0.035	Rejected
SVM-ADASYN vs. RF	4.561	0.044	Rejected
SVM-ADASYN vs. DT	19.580	0.002	Rejected
SVM-ADASYN vs. ADA	25.109	0.001	Rejected
SVM-ADASYN vs. KNN	78.637	0.000	Rejected

Table 15 Performance comparison of the latest approach introduced to other current methods

References	Authors	Models	Oversampling technique	Accuracy	Year
[21]	Mujahid et al.	SVM	SMOTE	95.45	2021
[17]	Glazkova et al.	SVM	Border-Line SMOTE	73.67	2020
[11]	Rupapara et al.	RVVC	SMOTE	97.05	2021
[12]	Floeres et al.	SVM	SMOTE	83.16	2018
[13]	Hashedi et al.	Voting	SMOTENC	78.17	2022
[22]	Liu et al.	SVM	SMOTE	87.24	2021
[28]	Hasib et al.	BMNET-5	-	90.32	2023
[48]	Alhudhaif Adi	RF	ADASYN	91.72	2021
[49]	Gonzalez et al.	RF	SMOTE	89.00	2022
[50]	Mahmud et al.	KNN	SMOTE	93.47	2023
[51]	Aditya et al.	LSTM	SMOTE	89.42	2023
[52]	Lavanya and Sasikala	SVM	ADASYN	92.23	2024
This study		SVM	ADASYN	99.67	2024
			SMOTE	99.57	
			Border-Line SMOTE	99.59	
		RF	ADASYN	99.21	
			SMOTE	99.26	
			Border-Line SMOTE	99.21	

Table 13 shows the results of CNN, LSTM, and BiLSTM models to evaluate the performance of two datasets. The CNN model shows better performance as compared to the other two models with an accuracy score of 96.11%. The BiLSTM model does not perform well and is able to attain an accuracy of 95.65% on the EndViolence tweets dataset and 92.82% on the E-commerce tweets dataset.

Statistical T-test comparison

To determine the superiority of our approach over others, we conduct statistical tests to evaluate its performance. We achieve this by comparing the performance metrics of our models with those of other approaches. We used the null hypothesis approach (H_0), which effectively differentiates between two model performances and ensures no evidence of a significant difference. To achieve this, we first used the primary performance measures and their corresponding predictions on the same datasets. Next,

perform a statistical t-test on these metrics while keeping the alpha level at 0.05. If the P-values are lower than the alpha level, the test will reject the hypothesis. Conversely, if the P-values are higher than the alpha level, the test will not reject the hypothesis. Table 14 presents the statistical T-test results. According to the statistical T-test, our approach is superior to other methods as there is a significant difference between them.

Performance comparison with existing studies

Table 15 demonstrates the comparison of different studies with the current study in terms of accuracy. Mujahid et al. [21] used different ML-based models to analyze the public sentiments towards education during COVID-19 using an oversampled dataset and achieved an accuracy of 95.45% with the SMOTE oversampling technique. Rupapara et al. [11] used the RVVC model for toxic tweets classification on highly imbalanced data with the SMOTE technique. The RF model attained 91.72% accuracy using ADASYN by Alhudhaif Adi [48]. Another study [49] employed RF with SMOTE and attained 89.00% accuracy. Mahmud et al.[50] developed a KNN classifier by employing the SMOTE to enhance accuracy and address class imbalances in the data. As a result, they achieved an accuracy of 93.47%. In order to enhance results, Aditya et al[51] utilized a deep learning LSTM model that consisted of many layers. Nevertheless, their methodology led to significant computing expenses and unsatisfactory results in the implementation of the SMOTE technology. Lavanya and Sasikala [52] utilized the ADASYN technique to train an SVM model. Their study achieved an accuracy of 92.23%. Based on the aforementioned studies, several conclusions can be inferred. Utilizing a balanced dataset has a tendency to improve the effectiveness of machine learning. Dataset balancing resolves the problem of model skewness in the majority class by evenly distributing a similar number of samples for model training. Furthermore, the majority of completed research has utilized SMOTE, the prevailing technique for data balancing through oversampling. In this study, the SVM model achieved 99.67% accuracy with ADASYN and the RF model obtained 99.26% accuracy with SMOTE oversampled tweets dataset. The Border-Line SMOTE oversampling technique also performs excellently with an accuracy of 99.59 % and 99.21% with the SVM and RF model, respectively using TFIDF features.

Discussion

The research examines the use of machine and deep learning to address the problem of oversampling strategies on unbalanced datasets, as well as to assess the efficacy of machine learning models. This research involves conducting several experiments. The studies use two Twitter datasets to accomplish the above goal. We divide the data into a train-test split with a ratio of 0.75 to 0.25, allocating 75% of the data for model training and reserving 25% for testing. In order to identify significant features from the tweets, two well-recognized methods of feature engineering, BoW and TF-IDF, are used. The proposed approach effectively examines the issue of oversampling in imbalanced classes.

The results indicate that machine learning models use various oversampling techniques that concurrently include BoW and TF-IDF features. The study utilizes five different oversampling approaches: SMOTE, SVM-SMOTE, K-Means SMOTE, ADASYN,

and Border-Line SMOTE. With an accuracy of 99.67% with the ADASYN method, 99.57% with SMOTE, and 99.59% with Border-Line SMOTE using the TFIDF feature extraction strategy, the SVM model does better than other machine learning models that are currently being used. The SMOTE oversampling strategy, when applied to the EndViolence tweets dataset, outperforms all other techniques in terms of performance because it uses the KNN algorithm to artificially augment the data. This dataset is notable for its significant imbalance. Using TFIDF features and RF, the SMOTE algorithm achieves exceptional performance on Twitter datasets pertaining to e-commerce, achieving an accuracy score of 99.26%. However, when used with these datasets, the BoW features do not outperform the TF-IDF features. As a result of these experiments, multiple patterns have emerged. Utilizing TF-IDF features for both datasets enhances the performance of machine learning models. For machine learning models, SMOTE, ADASYN, and border-line SMOTE are the most effective, especially when combined with TF-IDF features. Furthermore, the use of TF-IDF and oversampling approaches results in a more consistent and seamless performance of the models. This case differs from the use of BoW, where there are significant variations in model performance for different oversampling algorithms.

We also check the performance of machine learning with the oversampling technique using 10-fold cross-dataset experiments. Using 10-fold cross-validation, the studies demonstrate that all machine learning models perform very well, except for the support vector machine, which achieves the best accuracy of 0.973% by using TF-IDF features on the SMOTE oversampled e-commerce dataset. In addition, deep learning models were also tested to check the performance of oversampling techniques. Gated recurrent unit (GRU) attained the lowest performance, while CNN attained 96.11% accuracy. To assess the superiority of our technique compared to the alternatives, we conduct performance analyses utilizing statistical tests. To accomplish this, we assess the performance indicators of our models in comparison to those of other competing approaches.

This research showcases a significant improvement in average performance when compared to previous models. The proposed model showcased superior performance in terms of performance metrics and oversampling strategies when compared to other models. The SVM demonstrated the highest level of performance, with other existing alternative models following strongly behind. The proposed framework demonstrated an accuracy of 99.67%, further highlighting its exceptional performance.

Conclusion

A prominent issue in machine learning is the class imbalance which leads to model overfit and affects the performance of machine learning models. Machine learning models tend to over-train on the majority class and the skewed distribution of class samples reduces their performance. For imbalanced datasets, a high degree of accuracy can be achieved only by forecasting the majority class, but the minority class is missed, which is frequently the objective of developing the model in the first place. Although both undersampling and oversampling approaches can be used in this regard, the oversampling approach is preferred as there is no loss of information in it. As a means of resolving the issue of data imbalance, this study employs oversampling techniques on two extremely imbalanced Twitter datasets using BoW and TF-IDF features. Experiments involve six

machine learning and three deep learning models in this regard. Results show that balancing the data seems to lessen the possibility of model overfitting, which also happens when an imbalanced dataset is used for training. Several inferences can be made from the results. First, the RF model performed well overall on SMOTE and ADASYN oversampled data, achieving an accuracy of 99.67% and a recall score of 100.00% when used with the TF-IDF feature. Second, TF-IDF tends to show better performance as compared to BoW features when oversampled data is used for experiments. Third, despite better performance using all oversampling approaches, SMOTE, ADASYN, and Borderline SMOTE yield better results for the most part.

Acknowledgements
The authors are thankful for the support of Artificial Intelligence & Data Analytics Lab (AIDA) CCIS Prince Sultan University, Riyadh, 11586, Saudi Arabia. The authors would also like to thank Prince Sultan University, Riyadh Saudi Arabia for the support of APC of this publication.

Author contributions
MM conceived the idea, performed data analysis and wrote the original draft. EK conceived the idea, performed data curation and wrote the original draft. FR performed data curation, formal analysis, and designed methodology. MGV did project administration, dealt with software and performed visualization. ESA acquired the funding for research, and performed visualization and initial investigation. IdlTD dealt with software, carried out project administration and performed validation. IA supervised the study, performed validation and review and edit the manuscript. All authors read and approved the final manuscript.

Funding
This research is funded by the European University of Atlantic.

Availability of data and materials
The data can be requested from the corresponding authors.

Code availability
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interest.

Received: 27 February 2024 Accepted: 31 May 2024
Published online: 17 June 2024

References

1. Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced data. ACM Sigkdd Explor Newsl. 2004;6(1):80–9.
2. Lewis DD, Catlett J. Heterogeneous uncertainty sampling for supervised learning. In: Cohen WW, Hirsh H, editors. Machine learning proceedings 1994. New Brunswick: Elsevier; 1994. p. 148–56.
3. Mohammed RA, Wong K-W, Shiratuddin MF, Wang X. Scalable machine learning techniques for highly imbalanced credit card fraud detection: a comparative study. In: Geng X, Kang BH, editors. Pacific Rim international conference on artificial intelligence. Nanjing: Springer; 2018. p. 237–46.
4. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intelligent data analysis. 2002;6(5):429–49.
5. Ghosh K, Banerjee A, Chatterjee S, Sen S. Imbalanced twitter sentiment analysis using minority oversampling. In: Ghosh K, editor. 2019 IEEE 10th international conference on awareness science and technology (iCAST). Morioka: IEEE; 2019.
6. Ah-Pine J, Soriano-Morales E-P. A study of synthetic oversampling for twitter imbalanced sentiment analysis. In: Ah-Pine J, editor. Workshop on interactions between data mining and natural language processing (DMNLP 2016). Riva del Garda: DMNLP; 2016.
7. Aljedaani W, Rustam F, Ludi S, Ouni A, Mkaouer MW. Learning sentiment analysis for accessibility user reviews. In: Aljedaani W, editor. 2021 36th IEEE/ACM International conference on automated software engineering workshops (ASEW). Melbourne: IEEE; 2021. p. 239–46.

8. Hasib KM, Azam S, Karim A, Al Marouf A, Shamrat FJM, Montaha S, Yeo KC, Jonkman M, Alhajj R, Rokne JG. Mcnn-lstm: combining CNN and LSTM to classify multi-class text in imbalanced news data. IEEE Access. 2023. https://doi.org/10.1109/ACCESS.2023.3309697.
9. Hasib KM, Towhid NA, Faruk KO, Al Mahmud J, Mridha M. Strategies for enhancing the performance of news article classification in bangla: handling imbalance and interpretation. Eng Appl Artif Intell. 2023;125: 106688.
10. Sarakit P, Theeramunkong T, Haruechaiyasak C. Improving emotion classification in imbalanced youtube dataset using smote algorithm. In: Sarakit P, editor. 2015 2nd International conference on advanced informatics: concepts, theory and applications (ICAICTA). Chonburi: IEEE; 2015. p. 1–5.
11. Rupapara V, Rustam F, Shahzad HF, Mehmood A, Ashraf I, Choi GS. Impact of smote on imbalanced text features for toxic comments classification using rvvc model. IEEE Access. 2021;9:78621–34.
12. Flores AC, Icoy RI, Peña CF, Gorro KD. An evaluation of SVM and naive bayes with smote on sentiment analysis data set. In: Flores AC, editor. 2018 International conference on engineering, applied sciences, and technology (ICEAST). Phuket: IEEE; 2018. p. 1–4.
13. Al-Hashedi A, Al-Fuhaidi B, Mohsen AM, Ali Y, Gamal Al-Kaf HA, Al-Sorori W, Maqtary N. Ensemble classifiers for Arabic sentiment analysis of social network (twitter data) towards COVID-19-related conspiracy theories. Appl Comput Intell Soft Comput. 2022. https://doi.org/10.1155/2022/6614730.
14. Al-Azani S, El-Alfy E-SM. Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text. Proc Comput Sci. 2017;109:359–66.
15. Rivera G, Florencia R, García V, Ruiz A, Sánchez-Solís JP. News classification for identifying traffic incident points in a spanish-speaking country: a real-world case study of class imbalance learning. Appl Sci. 2020;10(18):6253.
16. Banerjee A, Bhattacharjee M, Ghosh K, Chatterjee S. Synthetic minority oversampling in addressing imbalanced sarcasm detection in social media. Multimed Tools Appl. 2020;79(47):35995–6031.
17. Glazkova A. A comparison of synthetic oversampling methods for multi-class text classification. arXiv preprint. 2020. arXiv:2008.04636.
18. Xu R, Chen T, Xia Y, Lu Q, Liu B, Wang X. Word embedding composition for data imbalances in sentiment and emotion classification. Cogn Comput. 2015;7(2):226–40.
19. Saumya S, Singh JP. Detection of spam reviews: a sentiment analysis approach. CSI Trans ICT. 2018;6(2):137–48.
20. Hasib KM, Rahman F, Hasnat R, Alam MGR. A machine learning and explainable AI approach for predicting secondary school student performance. In: Hasib KM, editor. 2022 IEEE 12th annual computing and communication workshop and conference (CCWC). Las Vegas: IEEE; 2022. p. 399–405.
21. Mujahid M, Lee E, Rustam F, Washington PB, Ullah S, Reshi AA, Ashraf I. Sentiment analysis and topic modeling on tweets about online education during COVID-19. Appl Sci. 2021;11(18):8438.
22. Liu J, Lu S, Lu C. Exploring and monitoring the reasons for hesitation with COVID-19 vaccine based on social-platform text and classification algorithms. Healthcare. 2021;9:1353.
23. Ardianto R, Rivanie T, Alkhalifi Y, Nugraha FS, Gata W. Sentiment analysis on e-sports for education curriculum using naive bayes and support vector machine. Jurnal Ilmu Komputer dan Informasi. 2020;13(2):109–22.
24. Balaji T, Annavarapu CSR, Bablani A. Machine learning algorithms for social media analysis: a survey. Comput Sci Rev. 2021;40: 100395.
25. Parlak B, Uysal AK. A novel filter feature selection method for text classification: extensive feature selector. J Inform Sci. 2023;49(1):59–78.
26. Parlak B, Uysal AK. The effects of globalisation techniques on feature selection for text classification. J Inform Sci. 2021;47(6):727–39.
27. Hasib KM, Islam MR, Sakib S, Akbar MA, Razzak I, Alam MS. Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey. IEEE Trans Comput Soc Syst. 2023. https://doi.org/10.1109/TCSS.2023.3263128.
28. Hasib KM, Tanzim A, Shin J, Faruk KO, Al Mahmud J, Mridha MF. Bmnet-5: a novel approach of neural network to classify the genre of bengali music based on audio features. IEEE Access. 2022;10:108545–63.
29. Hasib KM, Habib MA, Towhid NA, Showrov MIH. A novel deep learning based sentiment analysis of twitter data for us airline service. In: Hasib KM, editor. 2021 International conference on information and communication technology for sustainable development (ICICT4SD). Dhaka: IEEE; 2021.
30. Kaggle: ENDviolence Tweets. 2021. https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction/metadata. Accessed 22 Feb 2024.
31. Vijayarani S, Ilamathi MJ, Nithya M, et al. Preprocessing techniques for text mining-an overview. Int J Comput Sci Commun Netw. 2015;5(1):7–16.
32. Scott S, Matwin S. Citeseer. Feature engineering for text classification. 1999;99:379–88.
33. Zhang Y, Jin R, Zhou Z-H. Understanding bag-of-words model: a statistical framework. Int J Mach Learn Cybern. 2010;1(1):43–52.
34. Cong Y, Chan Y-B, Ragan MA. A novel alignment-free method for detection of lateral genetic transfer based on tf-idf. Sci Rep. 2016;6(1):1–13.
35. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
36. Li Y, Guo H, Zhang Q, Gu M, Yang J. Imbalanced text sentiment classification using universal and domain-specific knowledge. Knowl Based Syst. 2018;160:1–15.
37. Han H, Wang W-Y, Mao B-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: Huang DS, editor. International conference on intelligent computing. Cham: Springer; 2005. p. 878–87.
38. Tang Y, Zhang Y.-Q, Chawla N.V, Krasser S. Svms modeling for highly imbalanced classification. IEEE Trans Syst Man Cybern Part B (Cybernetics). 2008;39(1):281–8.
39. He H, Bai Y, Garcia EA, Li S. Adasyn: adaptive synthetic sampling approach for imbalanced learning. In: He H, editor. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). Hong Kong: IEEE; 2008. p. 1322–8.

40. Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. Inform Sci. 2018;465:1–20.
41. Fauzi MA. Random forest approach for sentiment analysis in Indonesian. Indonesian J Elect Eng Comput Sci. 2018;12(1):46–50.
42. Yuan R, Li Z, Guan X, Xu L. An SVM-based machine learning method for accurate internet traffic classification. Inform Syst Front. 2010;12(2):149–56.
43. Chen Y, Hu X, Fan W, Shen L, Zhang Z, Liu X, Du J, Li H, Chen Y, Li H. Fast density peak clustering for large scale data based on KNN. Knowl Based Syst. 2020;187: 104824.
44. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform. 2002;35(5–6):352–9.
45. Ramadhan W, Novianty SA, Setianingsih SC. Sentiment analysis using multinomial logistic regression. In: Ramadhan W, editor. 2017 International conference control electronics, renewable energy and communications (ICCREC). Yogyakarta: IEEE; 2017. p. 46–9.
46. Sharma H, Kumar S. A survey on decision tree algorithms of classification in data mining. Int J Sci Res (IJSR). 2016;5(4):2094–7.
47. Chen S, Shen B, Wang X, Yoo S-J. A strong machine learning classifier and decision stumps based hybrid adaboost classification algorithm for cognitive radios. Sensors. 2019;19(23):5077.
48. Alhudhaif A. A novel multi-class imbalanced eeg signals classification based on the adaptive synthetic sampling (adasyn) approach. PeerJ Comput Sci. 2021;7:523.
49. Rodríguez-González A, Tuñas JM, Prieto Santamaría L, Fernández Peces-Barba D, Menasalvas Ruiz E, Jaramillo A, Cotarelo M, Conejo Fernández AJ, Arce A, Gil A. Identifying polarity in tweets from an imbalanced dataset about diseases and vaccines using a meta-model based on machine learning techniques. Appl Sci. 2020;10(24):9019.
50. Mahmud F.G, Hermanto T.I, Nugroho I.M. Implementation of k-nearest neighbor algorithm with smote for hotel reviews sentiment analysis. Sinkron. 2023;8(2):595–602.
51. Aditya K, Wicaksono GW, Heryawan HAS, Aditya CSK. Sentiment analysis of the 2024 presidential candidates using smote and long short term memory. J Inform. 2023;8(2):279–86.
52. Lavanya P, Sasikala E. Enhanced performance of drug review classification from social networks by improved adasyn training and natural language processing techniques. In: Hemanth DJ, editor. Computational intelligence methods for sentiment analysis in natural language processing applications. Amsterdam: Elsevier; 2024. p. 111–27.

Publisher's Note