Applying SC-SMOTE model On Unlabelled Data To Classify Drug Related Webpages On Dark Web

Abstract— **The dark web has become a hub for illegal activities, including drug trafficking. Accurate classification of drug-related web-pages is crucial for law enforcement agencies to monitor and control these illegal activities. However, labelled data for dark web classification is scarce, and the available data is often imbalanced. This study aims to address these issues by applying Synthetic Minority Over-sampling Technique (SMOTE) on unlabelled data to generate synthetic samples that help labelling the data. The results show that using SMOTE on unlabelled data can significantly improve the performance of the classifier in terms of precision, recall, and F1-score. This research considers an improvement on SMOTE called Semantic-Cosine-based Synthetic Minority Over-sampling Technique (SC-SMOTE) to generate synthetic textual data and improving the accuracy of dark web classification.**

I. INTRODUCTION

The dark web is a portion of the internet that is not indexed by conventional search engines and can only be accessed via specialised software such as Tor. Drug trafficking, cybercrime, and the sale of stolen items are frequently associated with the dark web. It has become a refuge for criminal organisations and individuals to operate illegally, making it a top priority for law enforcement agencies around the world. The drug trade on the dark web has made it simpler for individuals to get narcotics without face-to-face interaction, thereby increasing the possibility of drug abuse and addiction. Moreover, drugs sold on the dark web can be laced with other dangerous substances, making them more potent and harmful. Classifying drug webpages on the dark web presents several challenges. Insufficient and skewed data from illegal and unregulated markets make it difficult to train classification models that are correct. The content on these pages can be disorganized, unstructured, and written in different languages, making it hard to extract relevant features for classification. The availability of labelled data for dark web classification is limited and often imbalanced. Labelling datasets in data mining for supervised classification algorithms is a manual process, requiring human intervention to assign the appropriate class to each data point. This process can be time consuming, costly, and prone to errors.

To address this issue, this research aims to explore the use of Synthetic Minority Over-sampling Technique (SMOTE) on textual data to help label the unlabelled data for classifying drug-related webpages on the dark web. SMOTE is a widely used oversampling technique that generates synthetic samples of the minority class to balance the data distribution. It works by creating synthetic samples of the minority class by interpolating between existing samples. This helps to balance the class distribution and improve the performance of machine learning models. SMOTE has been successfully applied to various domains, such as fraud detection, and medical diagnosis. However, SMOTE is designed to work with numerical data, and applying it to textual data is not straightforward.

To achieve this goal, the following research questions (RQ) will be addressed:

RQ1: How effective are SMOTE and its variants in handling imbalanced data for drug-related webpages on the dark web?

RQ2: How does the proposed SC-SMOTE approach compare to traditional SMOTE in improving the performance of classifiers for drug-related webpages on the dark web?

The proposed approach, Semantic Cosine SMOTE (SC-SMOTE), presents a promising solution to the challenges of generating synthetic textual samples that are both semantically relevant and contextually appropriate. This variant of the SMOTE algorithm leverages the semantic meaning of the

textual data to generate additional samples using cosine similarity. By generating additional data, SC-SMOTE aims to reduce the time and effort required for manual data labeling and improve the effectiveness of machine learning algorithms in detecting potential drug-related crimes on the dark web. The objective of this study is to evaluate the impact of using SC-SMOTE on unlabelled data on the performance of the classifier in terms of precision, recall, and F1-score. The research approach will involve conducting experiments on a dataset consisting of 20250 records and four columns that have been curated from the dark web. The study's findings are expected to contribute to the development of effective methods for monitoring and controlling illegal drug transactions on the dark web. The paper is organized as follows: a literature review on related work in the field, the proposed methodology of SC-SMOTE, an evaluation of the approach's impact on the performance of the classifier, and a discussion of the study's results and future research directions.

II. Literature Review

The dark web is a part of the internet that is hidden and inaccessible through conventional search engines. It is a hotbed of illegal activities, including drug trade, human trafficking, and cybercrime. Due to the anonymous nature of the dark web, it is difficult to monitor and regulate its activities. Therefore, the development of effective techniques for identifying and classifying drug-related web pages on the dark web is essential for law enforcement agencies. Several studies have attempted to use machine learning algorithms to classify web pages on the dark web. One such paper proposed a method that can effectively classify illegal activities on the dark web. Instead of relying on the massive dark web training set, they creatively selected laws and regulations related to each type of illegal activities to train the machine learning classifiers and achieve a good classification performance. In the areas of pornography, drugs, weapons, hackers, and counterfeit credit cards, they selected relevant legal documents from the United States Code for supervised training and conducted a classification experiment on the illegal content of the real dark web. Their approach allows researchers and the network law enforcement to check whether their dark web corpus contains such illegal activities based on the relevant laws of the illegal categories they care about in order to detect and monitor potential illegal websites in a timely manner.[4]Another proposed work attempts to identify the dark web drug marketplace with an ML-based approach. The proposed work attempts NER model-based drug page classification on the dark web data. First, the authors introduced two categories of drug-named entities in their Named Entity Recognition (NER) dataset: street names and chemical names. Next, the work presents the classification of drug-related pages based on Street and Chemical drug name entities.[5] Another research aims to give forensic investigators an efficient method for determining the themes currently being discussed in underground marketplace forums. To accomplish this, the authors studied whether it is possible to construct a classifier using semantic word representations. They desired to expedite the classifier's training process, which entails labor-intensive annotation work. [6] In another study, the authors developed an expert system to automatically classify the category and threat level of harmful content spread on the dark web by using a white box method.The KBs are the core of their expert system. [7] In terms of oversampling techniques, SMOTE is one of the most widely used methods in machine learning. It has been successfully applied in various fields, including medical diagnosis, fraud detection, and credit scoring (Chawla et al., 2002). The results of this paper indicate that the SMOTE method can increase the accuracy of classifiers for a minority class. SMOTE provides a new approach to oversampling. This paper shows that a combination of the method of oversampling the minority (abnormal) class and under-sampling the majority (normal) class can achieve better classifier performance (in ROC space) than only under-sampling the majority class. SMOTE was evaluated on many datasets with varied degrees of imbalance and varying volumes of data in the training set, offering a diverse testbed. Based on dominance in the ROC space, the combination of SMOTE and under-sampling also performs better than altering loss ratios in Ripper or by varying the class priors

in Naive Bayes Classifier: approaches that could directly manage the skewed class distribution. SMOTE forces focused learning and establishes a bias towards the minority class. The author concludes that SMOTE-classifier performs better than Under-classifier, Loss Ratio, and Naive Bayes. Out of a total of 48 experiments performed, SMOTE-classifier does not perform the best only for 4 experiments [8]. The authors of this research compared oversampling techniques for the problem of multi-class topics classifications. In the work, the authors compare the standard SMOTE 14 algorithm with its two modifications (Borderline SMOTE and ADASYN) and the random oversampling strategy on the example of one of text classification tasks. The paper discusses the k-nearest neighbor algorithm, the support vector machine algorithm and three types of neural networks (feedforward network, long short-term memory (LSTM) and bidirectional LSTM). The authors combined these machine learning algorithms with different text representations and compared synthetic oversampling methods. In most cases, the use of oversampling techniques can significantly improve the quality of classification. The authors conclude that for this task, the quality of the KNN and SVM algorithms is more influenced by class imbalance than neural networks. [9] This research aims to compare the performance of seven distinct SMOTE-based algorithms depending on a variety of ters by applying each approach to distinct datasets.After testing these algorithms on various datasets, the authors have determined that different methods provide varying degrees of accuracy when applied to datasets with varying class imbalance ratios. It can be concluded that all of these algorithms are very essential in order to get good accuracy scores, though different models give the best precision score depending on the imbalance ratio of the datasets used.[10] This paper evaluates the performance of several machine learning models for the classification of toxic comments and suggests an ensemble method called regression vector voting classifier (RVVC). Extensive tests are conducted to determine the effect of an im- balanced dataset and a balanced dataset utilising random under-sampling and SMOTE over-sampling on the performance of the models. Two feature extraction approaches including TF-IDF and BoW are used to get the feature vector for models' training. The results reveal that models perform poorly on the imbalanced dataset, but the balanced dataset tends to improve classification accuracy. Besides the machine learning classifiers like SVM, RF, GBM, and LR, the proposed RVVC and RNN deep learning models perform well with the balanced dataset. The performance with an over-sampled dataset is superior to that with an under-sampled dataset because the over-sampled dataset has a larger feature set, which improves the performance of the models. The results indicate that balancing the data minimises the likelihood of model overfitting that occurs when an imbalanced dataset is utilised for training. Furthermore, TF-IDF has greater classification accuracy for toxic comments than BoW, as TF-IDF captures the significance of a word, whereas BoW merely counts the occurrence of a word. The effectiveness of the suggested ensemble technique RVVC for classifying toxic and non-toxic comments is demonstrated. The performance of RVVC is superior both with the imbalanced and balanced dataset, yet, it achieves the highest accuracy of 0.97 when used with TF-IDF features from SMOTE over-sampled dataset. When compared to cutting-edge approaches in terms of performance, RVVC performs better and works well with both small and large feature vectors [11]. In this study, the authors propose a novel adaptive learning algorithm, ADASYN, for classification tasks using imbalanced data. Based on the original data distribution, ADASYN can adaptively generate synthetic data samples for the minority class to reduce the bias introduced by the imbalanced data distribution. Furthermore, ADASYN can also autonomously shift the classifier decision boundary to be more focused on those difficult to learn examples, therefore improving learning performance. The dynamic modification of weights and an adaptive learning technique based on data distributions achieve these two goals. This method's efficacy is demonstrated by simulation results on five data sets using various evaluation metrics.[12] The authors of this study attempted to apply the KNN-Undersampling approach prior to the actual classifying process in order to deal with imbalanced data. They compared their results to those obtained

by SVM with SMOTE and normal SVM. As with the majority of machine learning algorithms, SVM can perform poorly on the minority class because it was designed to generate a model based on the overall error. The author implemented Support Vector Machine Classifier in order to differentiate between the tweets of cyber trolls and those of regular users in order to combat cyber trolling. KNNUndersampling has demonstrated somewhat superior results in balancing data when compared to SMOTE, which reached 63.83%, according to the research that was carried out on a total of 20000 data points.[13]

Figure 1 provides an overview of how the various SMOTE variants are distributed based on the domains of the datasets they were tested on. It is worth noting that all of the datasets used for testing were either numeric or categorical, and no textual datasets were utilized in any of the SMOTE variants.

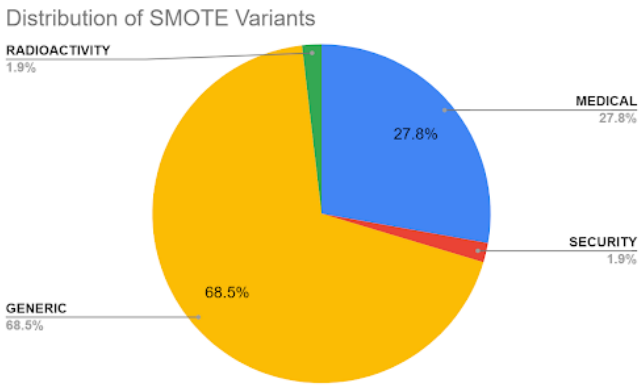


Fig. 1. Distribution of SMOTE Variants based on the domain of datasets

SMOTE-Cosine[5], proposed by F. Koto, is an oversampling technique that generates new synthetic samples by interpolating between minority class samples in the feature space using the cosine similarity metric. SMOTE-Cosine, like other SMOTE-based oversampling techniques, was originally developed for numerical data and has been applied to textual data without considering the specific characteristics of text. However, textual data differs from numerical data in several ways, such as its high-dimensional and sparse nature, making it unsuitable for direct application of oversampling techniques developed for numerical data. One of the major limitations of Cosine-SMOTE when applied to textual data is that it does not take into account the semantic relationships between words, which are crucial in determining the meaning and context of text.

III. Methodology

A. Datasets and Methods

1) Dataset: In the course of conducting a thorough investigation of the dark web, an exhaustive search was conducted to obtain a meticulously vetted dataset, consisting of 20250 records and four distinct columns, namely ”body,” ”description,” ”title,” and ”url.” Two salient features, specifically ”body” and ”url,” were extracted from the collected dataset to construct the final dataset. This final dataset comprises 20250 records and 2 columns, which include the unprocessed HTML content of the dark web pages, along with their corresponding URLs.

2) Data Preprocessing: Data preprocessing is crucial in Natural Language Processing to improve the accuracy of machine learning models. In this study, a dark web dataset was collected, consisting of 20250 records and 4 columns. To preprocess the ”body” feature, all non-letter characters such as punctuation and special characters were removed. Stopwords were also removed, which do not contribute much to the overall meaning of the document and can add noise to the features. Additionally, all letters were converted to lowercase to focus on the semantic meaning of the terms. Lemmatization was applied to group words with similar meanings into a single term, enabling the model to treat them similarly. After preprocessing, 16 records of text in the ”body” column were found to be null and were removed. The data preprocessing steps resulted in a reduction in the size of the document and the production of clean data, ready for use in machine learning models. Overall, the preprocessing steps significantly improve the accuracy of NLP models by reducing noise and making the dataset cleaner.

3) Creating word pool: To annotate the webpages as drug-related or non-drug related manually, it was necessary to have a drug-related word

dictionary. A word pool of drug related words was created manually. Figure 2 shows drug related word pool that consists of 323 words.

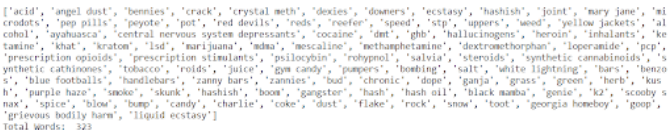


Fig. 2. Drug Word Pool

4) Converting to GloVe embeddings: GloVe stands for Global Vectors for word representation. It is an unsupervised learning algorithm developed by researchers at Stanford University aiming to generate word embeddings by aggregating global word co-occurrence matrices from a given corpus. GloVe was used to generate word embeddings from the list of words. Not all words were converted to vectors due to their absence from the GloVe training dataset and were thus discarded, resulting in a reduced word pool of 261 words.

B. Generating synthetic data

Synthetic word generation would greatly save the time and effort spent on brainstorming new words. SMOTE algorithm generates synthetic data based on existing data.

1) SMOTE Algorithm: SMOTE is a method that creates a new sample by taking a sample from a location on a line drawn between two nearby samples in the feature space. In particular, one member of the underrepresented group is selected at random to serve as an initial example. Then, the k nearest neighbours of that instance are identified. A neighbour is picked at random, and a synthetic example is generated at a random position between the two instances in feature space.
Let $\vec{p}$ represent the randomly selected vector from the minority class. To find its nearest neighbours, euclidean distance is calculated.

$$d(\vec{p},\vec{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{1}$$

where $\vec{q}$ = neigbour of $\vec{p}$ in Euclidean n-space q_i , p_i = Euclidean vectors, starting from the origin of the space (initial point) n = n-space Let $\vec{X}$ be the synthetic vector generated using SMOTE.

$$\vec{X} = \vec{p} + [d(\vec{p},\ \vec{q}) * rand(0,1)] \tag{2}$$

2) Semantic-Cosine-based Synthetic Minority Over-sampling Technique (SC-SMOTE) Algorithm: This research paper presents a novel approach, Semantic-Cosine-based Synthetic Minority Over-sampling Technique (SC-SMOTE), to generating new words with semantic meaning by leveraging the concept of cosine similarity. The proposed method involves converting words to vectors using GloVe embeddings and then generating new vectors using SMOTE oversampling. The closest GloVe embedding is then identified and considered as the new word generated. However, the key innovation of this approach lies in how it ensures the semantic relevance of the generated words. To achieve this, cosine similarity is calculated between the newly generated vector and a relevant domain-specific term. Let $\vec{X}$ and $\vec{Y}$ be two vectors in n-dimensional space. The cosine similarity $\cos\theta$ between these two vectors is defined as:

$$\cos\theta = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \tag{3}$$

where X_i and Y_i are the ith components of $\vec{X}$ and $\vec{Y}$, respectively. This formula calculates cosine similarity between a newly created vector and a domain-specific term in the proposed technique. The newly created vector is semantically relevant and added to the dataset if its cosine similarity exceeds a threshold. This ensures that the generated words are similar to those in the dataset and not random vector combinations.

Algorithm 1 SC-SMOTE

Require: A dataset of words, D

Ensure: An updated dataset of words, D'

Convert the words in D to vectors using pre-trained GloVe embeddings. Let V_D be the resulting set of vectors.

Generate new vectors using SMOTE oversampling on V_D. Let V_{new} be the set of newly generated vectors.

Create a nearest neighbor model using V_D. Let NN be the nearest neighbor model.

for each newly generated vector, $\vec{v}new \in Vnew$ **do**

Find the closest GloVe vector to $\vec{v}new$ using NN. Let $\vec{v}closest$ be the closest GloVe vector.

Calculate the cosine similarity between $\vec{v}new$ and a relevant domain-specific term vector using the following formula:

if the cosine similarity is above a certain threshold, τ **then**

Add the closest GloVe word to D'. Let $w_{closest}$ be the closest GloVe word.

end if

end for

Output the updated dataset, D'.

Here, V_D and V_{new} are sets of GloVe vectors representing the words in the original dataset and the newly generated vectors, respectively. NN is a nearest neighbor model created using V_D. $\vec{v}closest$ and $wclosest$ are the closest GloVe vector and word to the newly generated vector, respectively, found using the nearest neighbor model. Finally, τ is a threshold value that determines whether a newly generated word is semantically relevant or not.

C. Experiment Setup

Python is used as an implementation language. Libraries such as Pandas- used for data analysis and associated manipulation of tabular data in Data Frames, Numpy - used to create embeddings of words, Sklearn - used for PCA, Spacy - cleaning of raw data, SMOTE-variants used for validating and comparing SMOTE and its variants on our data, Plotly used for visualisation of data, Beautifulsoup used for crawling web pages, gloVe used to create word embeddings are utilised in the implementation. The dataset is prepared using Microsoft Excel. Code compilation and execution are performed using a cloud service, google colab with GPU. The Microsoft Windows 11 operating system is used for all of the experiments.

D. Results and Discussion

The data collected in this section represents the outcomes of several different experiments conducted to address the problem.

1) Labelling the dataset: The dataset was to be annotated with ”drug” and ”non-drug” labels.

a) Using original word pool: A total of 8489 websites were flagged as ”non-drug” after an initial 323-word list was used to label the content. That is 42 percent of the data is labelled as non-drug. 11745 data is labelled as ”drug” and 8489 rows are labelled as ”non-drug” data.

b) Using SMOTE generated word pool: SMOTE is used to generate synthetic data and using that to label the data, only 0.7 percent of the data is labelled as non-drug. 20091 data is labelled as ”drug” data and only 143 rows are labelled as ”non-drug” data.

c) Using improved SMOTE generated word pool: The proposed Semantic-Cosine-based Synthetic Minority Over-sampling Technique (SC-SMOTE) is used to generate synthetic data and using that to label the data, and 2669 records are labelled as non-drug. 17565 data is labelled as ”drug” and 2669 data is labelled as ”non-drug”.

2) Synthetic data generation: To evaluate the performance of SMOTE and SC-SMOTE algorithms on textual data, a subset of data was used where the minority class consisted of words related to ”drugs” and the majority class consisted of ”non-drug” words. Figure 3 displays the distribution of data in the form of a scatter plot, where the red points indicate the minority class ”drugs” and the blue points represent the majority class ”non-drugs”.

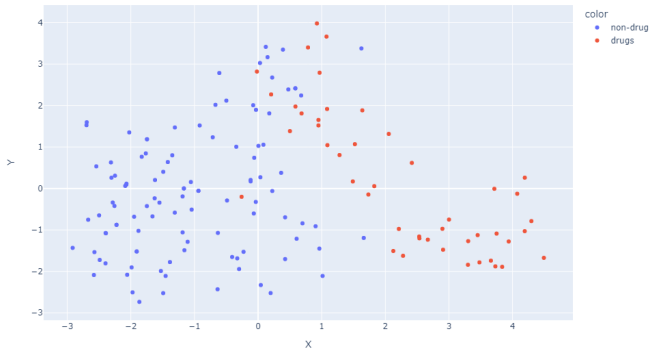


Fig. 3. Distribution of Data

a) SMOTE Algorithm: Analysis of the data that was labelled by using only the original word pool showed that a lot of sites were wrongly labelled as ”non-drug”, even though they had content directly related to illegal substances. Upon analysis, it was identified that the drug data was mislabeled as non-drug due to the limited number of keywords used for labeling. To address this issue, synthetic words were generated using SMOTE. However, this resulted in an overabundance of keywords, causing some non-drug content to be incorrectly labeled as drug-related. Figure 4 illustrates this issue, with red representing synthetic words generated using SMOTE, blue representing the original word pool, and green representing non-drug words.

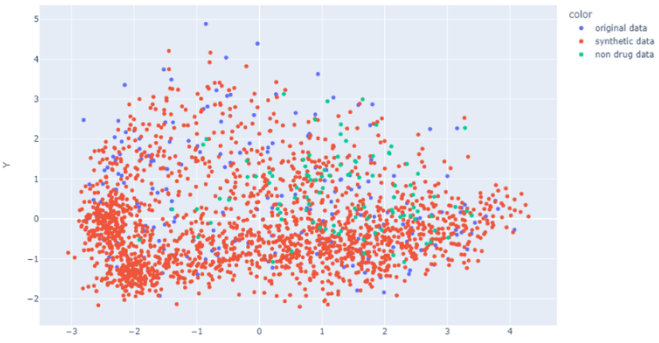


Fig. 4. Synthetic data generated using SMOTE algorithm

The synthetic data generated by the SMOTE algorithm appears to be scattered randomly throughout the feature space, as can be observed in Figure 4. It is clear from the figure that a significant amount of synthetic data generated by SMOTE is not related to the minority class ”drugs”, which could result in a decrease in classification performance on the minority class.

This observation highlights the potential limitations of using SMOTE for oversampling in text data, as the generated synthetic samples may not be semantically meaningful or relevant to the minority class. Consequently, this could lead to overfitting and poor generalization performance of the resulting model on real-world data.

b) SC-SMOTE Algorithm: The proposed SC-SMOTE algorithm effectively addresses the issues of limited data and irrelevant data by generating class-related keywords. The algorithm utilizes SMOTE to generate new data, and cosine similarity is used to discard irrelevant generated words. The combination of SMOTE and cosine similarity doubled the original word pool from 323 to 616, while reducing the number of words compared to the word pool generated by SMOTE alone, thus eliminating superfluous words. Research revealed that 5768 web pages classified as ”non-drug” by the original word pool were classified as ”drug” by the SC-SMOTE word pool, indicating that the initial labeling of drug-related content was incomplete. Figure 5 illustrates synthetic data generated by the improved SMOTE algorithm. Red indicates synthetic words generated using the SC-SMOTE, blue represents the original word pool, and green represents non-drug words

Fig. 5. Synthetic data generated using SC-SMOTE algorithm

3) SMOTE Variants: Very little effort is made to adapt smote to textual data, however smote has been enhanced for numerical data, and 86 smote variations are currently available. Smote-variants package implements 86 variants of the Synthetic

Minority Oversampling Technique (SMOTE). Besides the implementations, an easy to use model selection framework is supplied to enable the rapid evaluation of oversampling techniques on unseen datasets. The majority class and minority class are required parameters for the smote function in this package. To accomplish this, the drug-related word pool was oversampled and a different pool of non-drug-related terms was used as the majority class. This was applied to all the different smote variants and analyse the results. Figures 6, 7 and 8 show graphs generated to analyse the various SMOTE variants. Data represented by 'blue' are "non-drug" related words. 'Red' represents "drug" related words and "green" depicts synthetic words.

Figure 9 plots a more comprehensive understanding of the performance of different SMOTE variants on textual datasets. The results of the experiment demonstrate that while some SMOTE variants were able to generate synthetic data, they did not guarantee the semantic relevance of the generated words to the minority class.

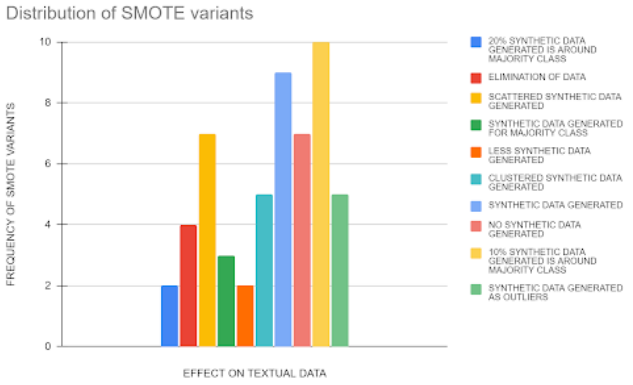
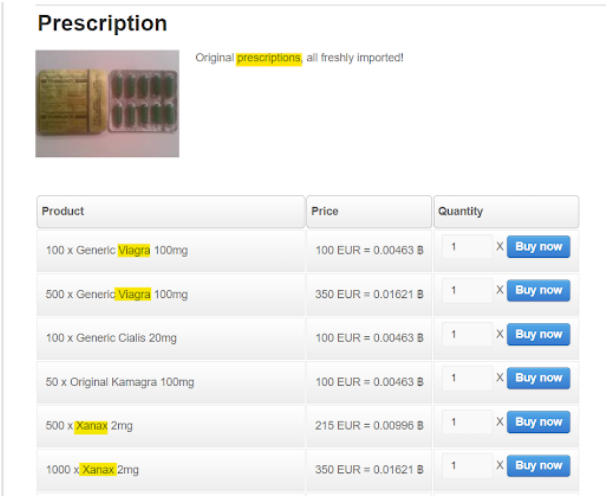
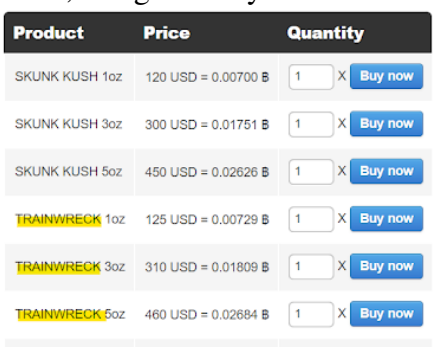


Fig. 6. Distribution of SMOTE Variants based on their effect on textual data

4) Comparison of results:



The above snippet shows a record that was mislabelled by the original word pool and was correctly labelled by using SC-SMOTE algorithm. The words that help correctly label the data were generated by the SC-SMOTE algorithm.
Words that were detected without synthetic data: 'meth', 'cannabis', 'lsd', 'cocaine', 'crystal', 'crystal meth', 'ecstasy'.
Words that were detected by SC-SMOTE: 'meth', 'cannabis', 'lsd', 'cocaine', 'crystal', 'crystal meth', 'ecstasy', 'prescription', 'prescriptions', 'xanax', 'viagra'. 'oxycodone'.



The above snippet shows a record that was mislabelled by the original word pool and was correctly labelled by using SC-SMOTE algorithm. The words that help correctly label the data were generated by the SC-SMOTE algorithm.
Words that were detected without synthetic data: 'cannabis', 'cocaine', 'crystal', 'crystal meth', 'ecstasy', 'lsd', 'meth'
Words that were detected by SC-SMOTE: 'cannabis', 'cocaine', 'crystal', 'crystal meth', 'ecstasy', 'lsd', 'meth', 'trainwreck', 'pills'.

5) Machine Learning Algorithms: The TF-IDF method is used to generate vectors, which are then used to train machine learning models such as Logistic Regression, Random Forest Classifier, and Linear Support Vector Classifier. Each of the 20234 texts in the dataset is represented by 37368 features.

Table 1 presents the precision, recall, F1-score, and accuracy of the three machine learning models trained on the three datasets. These performance metrics provide an evaluation of the models' effectiveness in predicting drug entities. Looking at the results in Table 1, it can be inferred that the Linear Support Vector Classifier (SVC) outperformed the Logistic Regression and Random Forest Classifier on all three datasets, regardless of the sampling technique used. Additionally, the use of the Synthetic Minority Over-sampling Technique (SMOTE) resulted in a significant improvement in performance metrics on all models, except for the Random Forest Classifier, which performed poorly with SMOTE. However, the use of SMOTE alone resulted in overfitting on the drug-related dataset, which is evident from the perfect precision, recall, and F1-score achieved by the Logistic Regression and Linear SVC models on this dataset. On the other hand, the use of the SC-SMOTE resulted in a more balanced improvement in performance metrics across all three datasets, without overfitting. Therefore, it can be inferred that the SC-SMOTE sampling technique is the best approach for handling imbalanced data for drug-related webpages on the dark web, especially when using the Linear SVC model. Confusion matrices for all the different algorithms on the different datasets are shown in Figure 10.

ROC curves for each dataset are shown in figure 11 and 12. ROC curve is a graphical representation of the trade-off between true positive rate (TPR) and false positive rate (FPR) for different classification thresholds. A good model should have a higher TPR and a lower FPR, resulting in a curve that is closer to the top left corner of the plot. If the ROC curve for the SMOTE dataset is at the top left corner (perfect), it indicates that the model has achieved high true positive rate and low false positive rate. This means that the model has successfully learned to distinguish between the positive and negative classes and has a high predictive power.

However, this result is not necessarily a good thing as it could also indicate that the SMOTE technique has led to overfitting of the model on the minority class samples. Overfitting occurs when the model learns to fit the training data too closely, including the noise or outliers in the data. As a result, the model performs well on the training data but poorly on the test data. On the other hand, the lower performance of the model on the no-SMOTE dataset could indicate that is negatively affecting the model's performance. This means that the model is having difficulty learning the features of the minority class and is biased towards the majority class.

The higher performance of the model on the SC-SMOTE dataset indicates that the proposed SC-SMOTE approach has effectively addressed the class imbalance problem and improved the model's performance without causing overfitting. In conclusion, while a perfect ROC curve for the SMOTE dataset might seem desirable, it is important to consider the potential impact of overfitting on the model's generalization performance. The SC-SMOTE approach appears to strike a balance between addressing the class imbalance problem and avoiding overfitting, leading to better performance on the test data.

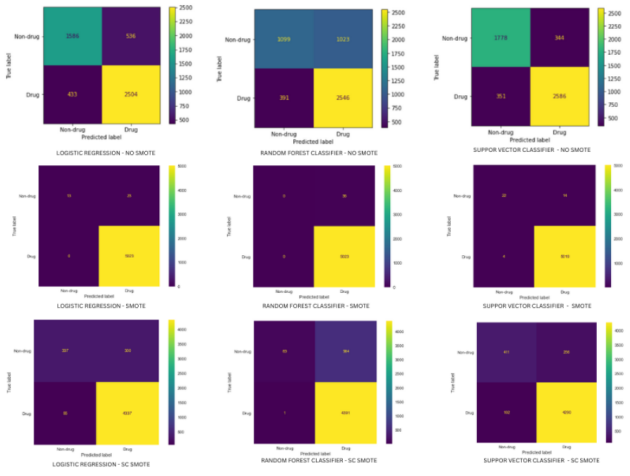


Fig. 7. Confusion Matrices for all the different algorithms on different datasets

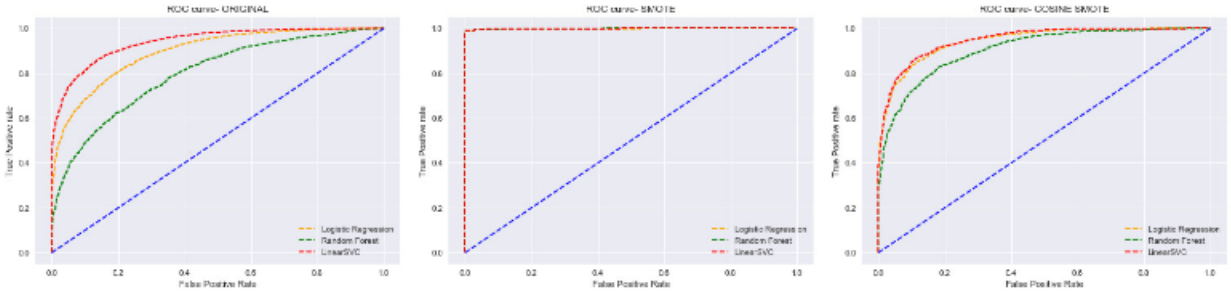


Fig. 8. ROC Curves

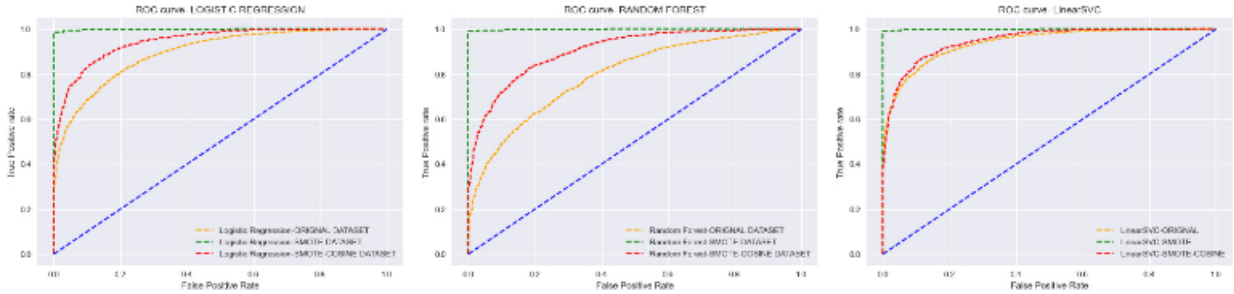


Fig. 9. ROC Curves

TABLE I

PERFORMANCE METRICS FOR MACHINE LEARNING MODELS ON THE THREE DATASETS

Dataset	**Model**	**Drug class**			**Non-drug class**			**Overall Accuracy**
		Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	
No SMOTE	Logistic Regression	0.82	0.85	0.84	0.79	0.75	0.77	0.80
	Random Forest Classifier	0.71	0.87	0.78	0.74	0.52	0.61	0.72
	Linear Support Vector Classifier	0.88	0.88	0.88	0.84	0.84	0.84	0.86
SMOTE	Logistic Regression	1.0	1.0	1.0	1.0	0.36	0.53	0.995
	Random Forest Classifier	0.99	1.0	1.0	0.0	0.0	0.0	0.992
	Linear Support Vector Classifier	1.0	1.0	1.0	0.85	0.61	0.71	0.992
SC-SMOTE	Logistic Regression	0.93	0.99	0.96	0.86	0.51	0.64	0.92
	Random Forest Classifier	0.88	1.0	0.94	0.99	0.12	0.22	0.88
	Linear Support Vector Classifier	0.94	0.98	0.96	0.80	0.62	0.70	0.93

IV. CONCLUSION

This paper proposes an improvement to the Synthetic Minority Over-sampling Technique (SMOTE) called the Semantic-Cosine-based Synthetic Minority Over-sampling Technique (SC-SMOTE). The goal of this approach is to generate new words that have semantic meaning and are contextually appropriate, which can be used to label datasets for supervised classification algorithms. The experimental results demonstrate that SC-SMOTE is successful in generating textual words that are both semantically relevant and contextually appropriate. Furthermore, the accuracy of the labelled datasets using SC-SMOTE is significantly higher than those labelled without SMOTE or with the original SMOTE.

It is worth noting that labelling datasets for supervised classification algorithms is a manual process that often requires human intervention, which can be time consuming, costly, and error prone. Therefore, the development of techniques like SC-SMOTE can greatly improve the efficiency and accuracy of this labelling process. Overall, this research paper presents a promising approach to generating new words that are both semantically relevant and contextually appropriate, and has practical applications for accurate classification of data in various fields, including dark web drug classification.

REFERENCES

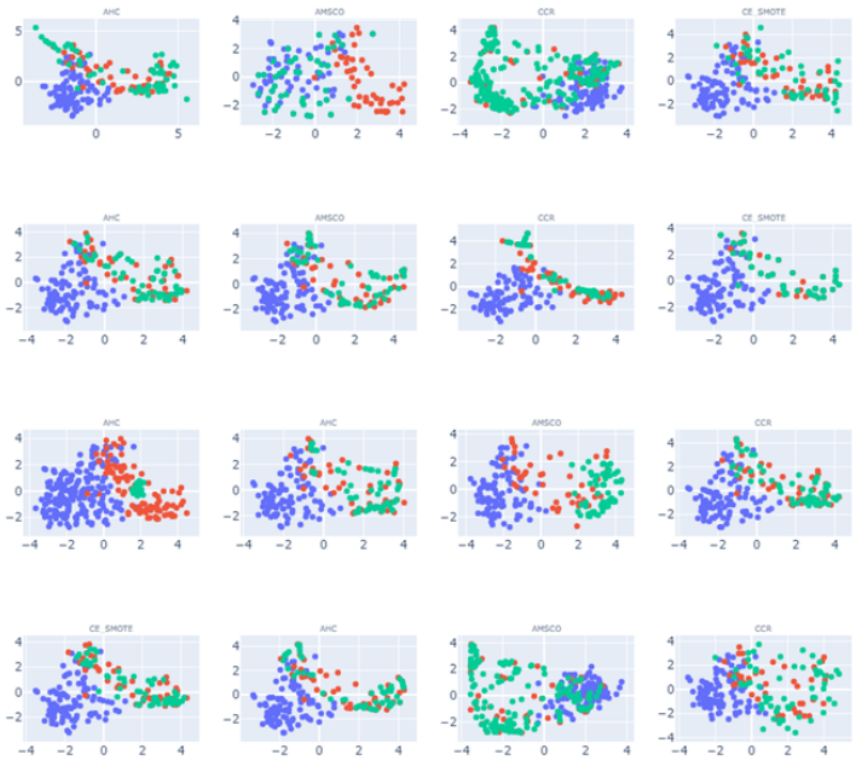


Fig. 10. SMOTE Variants

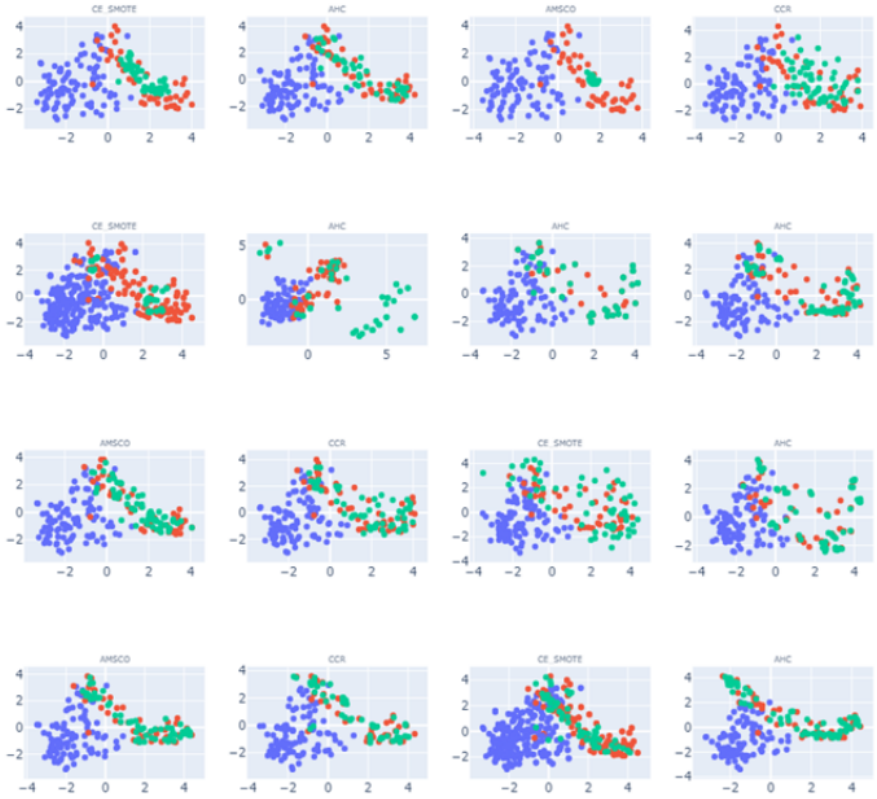


Fig. 11. SMOTE Variants

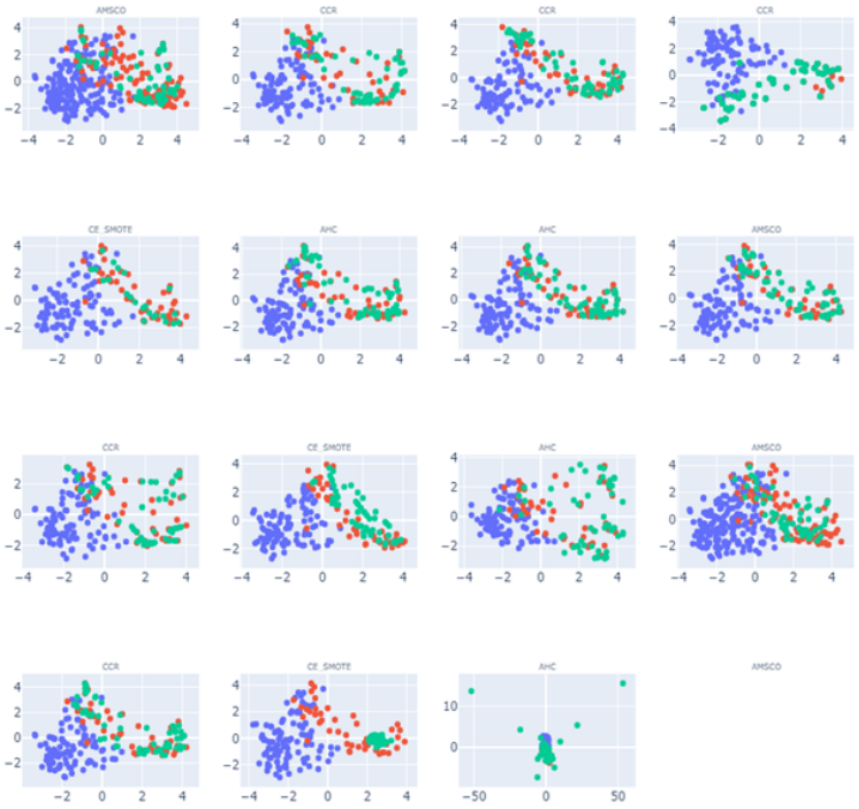


Fig. 12. SMOTE Variants

APPENDIX I
SMOTE VARIANTS

Label	Model	Data Characteristics	Data Domain	Data Collection Method	Effect in our data	Publication details
ADOMS	SMOTE+PRINCIPAL COMPONENT ANALYSIS	12 DATASETS FROM UCI -2 CLASS DATA, NUMERIC DATA	MEDICAL	UCI REPOSITORY	ELIMINATION OF DATA	THE GENERATION MECHANISM OF SYNTHETIC MINORITY CLASS EXAMPLES
AHC	SMOTE + SUPPORT VECTOR MACHINES	REFERRED A HOSPITAL DATASET AND FETCHED MEDICAL RECORDS - TEXTUAL AND NUMERICAL DATA	MEDICAL	REAL-TIME HOSPITAL PATIENTS RECORDS from The University Hospital of Geneva	SYNTHETIC DATA GENERATED FOR MAJORITY CLASS	LEARNING FROM IMBALANCED DATA IN SURVEILLANCE OF NOSOCOMIAL INFECTION

CCR	SMOTE+TOMEK LINKS SMOTE+EDITED NEAREST NEIGHBOR RULE	(42 DATASETS FROM KEEL ONLY 2 CLASS DATASETS COMPOSED SOLELY OF NUMERICAL DATA WERE USED)	GENERIC	KEEL REPOSI-TORY	20% SYNTHETIC DATA GENERATED IS AROUND MAJORITY CLASS	CCR: A COMBINED CLEANING AND RESAMPLING ALGORITHM FOR IMBALANCED DATA CLASSI-FICATION
CE SMOTE	CE-SMOTE ALGO-RITHM	10 IMBALANCED DATASET FROM UCI - NUMERICAL AND NOMINAL CATEGORICAL DATA	MEDICAL	UCI REPOS-ITORY	10% SYNTHETIC DATA GENERATED IS AROUND MAJORITY CLASS	A NEW OVER-SAMPLING METHOD BASED ON CLUSTER ENSEMBLES
cluster SMOTE	RIPPER + SMOTE	2 DATASETS USED: PACKETS AND DESTINATION - TREATING AS A 2 CLASS PROBLEM	GENERIC	PACKETS AND DESTI-NATION	20% SYNTHETIC DATA GENERATED IS AROUND MAJORITY CLASS	COMBATING IMBALANCE IN NETWORK INTRUSION DATASETS
DE over-sampling	DEC-SVM (Classifi-cation with DE and clustering hybrid re-sampling)	10 UCI datasets - Numerical	Medical	UCI repository	NO SYN-THETIC DATA GENERATED	A NOVEL DIFFERENTIAL EVOLUTION-CLUSTERING HYBRID RESAMPLING ALGORITHM ON IMBALANCED DATASETS
DEAGO	DEnoising Autoencoder-based Generative Oversampling (DEAGO)	GENERIC	Radioactive	-	CLUSTERED SYNTHETIC DATA GENERATED	SYNTHETIC OVERSAM-PLING FOR ADVANCED RADIOACTIVE THREAT DETECTION
DSMOTE	k-nearest neighbor, naïve Bayes and support vector machine	11 datasets (5 datasets from IUMS and 6 datasets from UCI), 3 classifiers and 4 resampling methods.	GENERIC	UCI , IUMS	SYNTHETIC DATA GENERATED AS OUTLIERS	DIVERSITY AND SEPARABLE METRICS IN OVER-SAMPLING TECHNIQUE FOR IMBALANCED DATA CLASSI-FICATION

DSRBF	static smote radial basis function (SSRBF) method, and the dynamic smote radial basis function (DSRBF) method.	12 datasets taken from the UCI repository	Medical	UCI	10% SYNTHETIC DATA GENERATED IS AROUND MAJORITY CLASS	A DYNAMIC OVER-SAMPLING PROCEDURE BASED ON SENSITIVITY FOR MULTI-CLASS PROBLEMS
Edge Det SMOTE	SVM +Smote	SVM is used to 9 different datasets with 5-fold cross validation	GENERIC	-	SYNTHETIC DATA GENERATED	WEIGHT DECISION ALGORITHM FOR OVER-SAMPLING TECHNIQUE ON CLASS-IMBALANCED LEARNING
G SMOTE	SVM + GP(Gaussian process classifier) +Smote (SMOTE+EnClassifier)	4 different imbalanced datasets	GENERIC	SYNTHETIC DATA GENERATED AS OUTLIERS		HANDLING IMBALANCED DATASETS BY PARTIALLY GUIDED HYBRID SAMPLING FOR PATTERN RECOGNITION
GASMOTE	N-HyperGraph Algorithm	18 UCI datasets which are downloaded from the machine learning data repository	GENERIC	UCI	SYNTHETIC DATA GENERATED FOR MAJORITY CLASS	A NOVEL ALGORITHM FOR IMBALANCE DATA CLASSIFICATION BASED ON GENETIC ALGORITHM IMPROVED SMOTE
Gaussian SMOTE	Gaussian probability distribution + SMOTE algorithm	The seeds dataset consists of seven attributes and three classes. Each class is made of seventy samples	GENERIC	-	10% SYNTHETIC DATA GENERATED IS AROUND MAJORITY CLASS	GAUSSIAN-BASED SMOTE ALGORITHM FOR SOLVING SKEWED CLASS DISTRIBUTIONS
polynom fit SMOTE star	one-against-all SVM	handwriting samples of an Arabic letter source document written by 60 persons	GENERIC	-	SYNTHETIC DATA GENERATED	NEW OVER-SAMPLING APPROACHES BASED ON POLYNOMIAL FITTING FOR IMBALANCED DATA SETS

polynom fit SMOTE bus	one-against-all SVM	handwriting samples of an Arabic letter source document written by 60 persons.	GENERIC	-	SYNTHETIC DATA GENERATED	NEW OVER-SAMPLING APPROACHES BASED ON POLYNOMIAL FITTING FOR IMBALANCED DATA SETS
polynom fit SMOTE poly	one-against-all SVM	handwriting samples of an Arabic letter source document written by 60 persons	GENERIC	-	CLUSTERED SYNTHETIC DATA GENERATED	NEW OVER-SAMPLING APPROACHES BASED ON POLYNOMIAL FITTING FOR IMBALANCED DATA SETS
polynom fit SMOTE mesh	one-against-all SVM	handwriting samples of an Arabic letter source document written by 60 persons.	GENERIC	-	SYNTHETIC DATA GENERATED	NEW OVER-SAMPLING APPROACHES BASED ON POLYNOMIAL FITTING FOR IMBALANCED DATA SETS
Gazzah	SMOTE and TL	Twelve datasets downloaded from KEEL, Numerical dataset	Medical	KEEL Datasets	NO SYNTHETIC DATA GENERATED	A HYBRID SAMPLING METHOD FOR IMBALANCED DATA
ISMOTE	Smote and TL	Twelve datasets downloaded from KEEL, Numerical dataset	Medical	KEEL Datasets	NO SYNTHETIC DATA GENERATED	A NEW COMBINATION SAMPLING METHOD FOR IMBALANCED DATA
LLE SMOTE	Bayesian + Knn + SVM	three datasets are collected from several chest x-ray image databases in automatic computerized detection of pulmonary	Medical	-	SCATTERED SYNTHETIC DATA GENERATED	CLASSIFICATION OF IMBALANCED DATA BY USING THE SMOTE ALGORITHM AND LOCALLY LINEAR EMBEDDING

LVQ SMOTE	SVM (Support Vector Machine) , Logistic Tree , Neural Network, Naive Bayes , Random Forest,and OLVQ3. SVM was implemented using a package called LIBSVM	Eight imbalanced benchmark datasets	Medical	UCI Machine Learning Repository	SCATTERED SYNTHETIC DATA GENERATED	LVQ-SMOTE – LEARNING VECTOR QUANTIZATION BASED SYNTHETIC MINORITY OVER–SAMPLING TECHNIQUE FOR BIOMEDICAL DATA
NT SMOTE	LSVM	-	GENERIC	-	SYNTHETIC DATA GENERATED AS OUTLIERS	NEIGHBORHOOD TRIANGULAR SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE FOR IMBALANCED PREDICTION ON SMALL SAMPLES OF CHINESE TOURISM AND HOSPITALITY FIRMS
ProWSyn	Proximity Weighted Synthetic Oversampling Technique	-	GENERIC	-	SYNTHETIC DATA GENERATED	PROWSYN: PROXIMITY WEIGHTED SYNTHETIC OVER-SAMPLING TECHNIQUE FOR IMBALANCED DATA SET LEARNING
ROSE	A nonparametric decision tree and a logit model have been chosen as classification models	50 balanced ROSE samples	GENERIC	-	SYNTHETIC DATA GENERATED AS OUTLIERS	TRAINING AND ASSESSING CLASSIFICATION RULES WITH IMBALANCED DATA

SDSMOTE	Smote + SDSmote + C4.5 + adaboost and bagging	four public data-sets Glass,bloody wine data-sets are from the UCI Machine Learning Repository, JM1 is from NASA standard data-sets	GENERIC	UCI, NASA	LESS SYNTHETIC DATA GENERATED	AN IMPROVED SMOTE IMBALANCED DATA CLASSIFICATION METHOD BASED ON SUPPORT DEGREE
Selected SMOTE	SMOTE-Out, SMOTE-Cosine, and SelectedSMOTE	Eighteen different datasets from UCI.	Medical	UCI	SYNTHETIC DATA GENERATED	SMOTE-OUT, SMOTE-COSINE, AND SELECTED-SMOTE: AN ENHANCEMENT STRATEGY TO HANDLE IMBALANCE IN DATA LEVEL
SMMO	M-Smote and Smote	-	GENERIC	UCI	NO SYNTHETIC DATA GENERATED	SELECTING MINORITY EXAMPLES FROM MISCLASSIFIED DATA FOR OVER-SAMPLING.
SMOBD	Adaptive Oversampling Technique Based on samples Density, Sigmoid Function Smoothing, Linear Interpolation Smoothing, SVM with different error costs and ASMOBD-CS	9 UCI datasets	GENERIC	UCI	NO SYNTHETIC DATA GENERATED	APPLYING OVER-SAMPLING TECHNIQUE BASED ON DATA DENSITY AND COST-SENSITIVE SVM TO IMBALANCED LEARNING
SMOTE Cosine	SMOTE-Out(creating synthetic example in outside area of dash line), SMOTE-Cosine, and SelectedSMOTE	eighteen different datasets	GENERIC	datasets from UCI	10% SYNTHETIC DATA GENERATED IS AROUND MAJORITY CLASS	SMOTE-OUT, SMOTE-COSINE, AND SELECTED-SMOTE: AN ENHANCEMENT STRATEGY TO HANDLE IMBALANCE IN DATA LEVEL

SMOTE D	deterministic version of SMOTE	66 datasets	GENERIC	repository KEEL	SYNTHETIC DATA GENERATED	SMOTE-D A DETERMINISTIC VERSION OF SMOTE
SMOTE ENN	Smote + Tomek, Smote + Enn	13 UCI Data sets	GENERIC	UCI Repository	NO SYNTHETIC DATA GENERATED	A STUDY OF THE BEHAVIOR OF SEVERAL METHODS FOR BALANCING MACHINE LEARNING TRAINING DATA
SMOTE FRST 2T	Smote + Fuzzy rough set theory	Cleaning/reducing the training data using a double threshold for eliminating original majority data on the one hand, and synthetic minority data on the other hand. Nominal data	Security	-	10% SYNTHETIC DATA GENERATED IS AROUND MAJORITY CLASS	FUZZY-ROUGH IMBALANCED LEARNING FOR THE DIAGNOSIS OF HIGH VOLTAGE CIRCUIT BREAKER MAINTENANCE: THE SMOTE-FRST-2T ALGORITHM
SMOTE IPF	Smote + iterative partitioning filter + ENN + Tomek links	Real world datasets Multi-class datasets are modified to obtain two-class imbalanced problems	Medical	KEEL Repository	10% SYNTHETIC DATA GENERATED IS AROUND MAJORITY CLASS	SMOTE–IPF: ADDRESSING THE NOISY AND BORDERLINE EXAMPLES PROBLEM IN IMBALANCED CLASSIFICATION BY A RE-SAMPLING METHOD WITH FILTERING
SMOTE OUT	SMOTE-Out(creating synthetic example in outside area of dash line), SMOTE-Cosine, and SelectedSMOTE	eighteen different datasets	GENERIC	datasets from UCI	ELIMINATION OF DATA	SMOTE-OUT, SMOTE-COSINE, AND SELECTED-SMOTE: AN ENHANCEMENT STRATEGY TO HANDLE IMBALANCE IN DATA LEVEL

SMOTE PSO	PSO+SVM	18 data-sets	GENERIC	KEEL data-set repository	SCATTERED SYNTHETIC DATA GENERATED	PSO-BASED METHOD FOR SVM CLASSIFICATION ON SKEWED DATA SETS
SMOTE PSOBAT	Bat-inspired algorithm (BAT) and particle swarm optimization algorithm (PSO)	30 highly imbalanced datasets which are spanned across 10 application areas	GENERIC	data from Ding's paper (Ding, Zejin. ""Diversified ensemble classifiers for highly imbalanced data learning and their application in bioinformatics."" (2011))	LESS SYNTHETIC DATA GENERATED	OPTIMIZING SMOTE BY METAHEURISTICS WITH NEURAL NETWORK AND DECISION TREE
SMOTE RSB	Rough set theory+SMOTE	Multiclass 44 data-sets	GENERIC	the UCI repository	10% SYNTHETIC DATA GENERATED IS AROUND MAJORITY CLASS	SMOTE-RSB*: A HYBRID PREPROCESSING APPROACH BASED ON OVERSAMPLING AND UNDERSAMPLING FOR HIGH IMBALANCED DATA-SETS USING SMOTE AND ROUGH SETS THEORY
SMOTE Tomek-Links	k-NN algorithm uses the Heterogeneous Value Difference Metric (HVDM) distance function. Ten different methods of under and over-sampling .(Tomek links,CNN,One-sided selection,CNN + Tomek links,Neighborhood Cleaning Rule,Smote + ENN,Smote + Tomek Links,Smote)	thirteen UCI data sets.	GENERIC	UCI	SCATTERED SYNTHETIC DATA GENERATED	A STUDY OF THE BEHAVIOR OF SEVERAL METHODS FOR BALANCING MACHINE LEARNING TRAINING DATA

SMOTE	SMOTE algorithm	nine different datasets	GENERIC	-	SCATTERED SYNTHETIC DATA GENERATED	SMOTE: SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE
SN SMOTE	NCN+GG+RNG+ SMOTE	39 datasets	GENERIC	KEEL Data Set Repository	SCATTERED SYNTHETIC DATA GENERATED	SURROUNDING NEIGHBORHOOD BASED SMOTE FOR LEARNING FROM IMBALANCED DATA SETS
SOMO	-	-	GENERIC	-	SCATTERED SYNTHETIC DATA GENERATED	SELF-ORGANIZING MAP OVER-SAMPLING (SOMO) FOR IMBALANCED DATA SET LEARNING
SSO	Neural network ensemble + smote	10 real world datasets	GENERIC	KEEL dataset repository	10% SYNTHETIC DATA GENERATED IS AROUND MAJORITY CLASS	STOCHASTIC SENSITIVITY OVER-SAMPLING TECHNIQUE FOR IMBALANCED DATA
SUNDO	Smote + Support vector machine + Decision tree	Two real world datasets coming from the metal industry	Medical	-	NO SYNTHETIC DATA GENERATED	NOVEL RESAMPLING METHOD FOR THE CLASSIFICATION OF IMBALANCED DATASETS FOR INDUSTRIAL AND OTHER REAL-WORLD PROBLEMS
Supervised SMOTE	SOS Algorithm.	Two benchmark datasets were chosen to evaluate the efficacy of the proposed SOS algorithm.	Medical	-	10% SYNTHETIC DATA GENERATED IS AROUND MAJORITY CLASS	A NEW SUPERVISED OVER-SAMPLING ALGORITHM WITH APPLICATION TO PROTEIN-NUCLEOTIDE BINDING RESIDUE PREDICTION

SVM balance	Smote + Support vector machine.	Data taken from January to November 2018 from Dr. Cipto Mangunkusumo Hospital (RSCM). Numeric data.	Medical		10% SYNTHETIC DATA GENERATED IS AROUND MAJORITY CLASS	PREPROCESSING UNBALANCED DATA USING SUPPORT VECTOR MACHINE
SYMPROD	Smote + Friedman test + Holm-Benferroni + Random Forest + Logistic Regression + Support vector machine.	10 imbalanced dataset from UCI repository. 4 artificial datasets.	GENERIC		CLUSTERED SYNTHETIC DATA GENERATED	A SYNTHETIC MINORITY BASED ON PROBABILISTIC DISTRIBUTION (SYMPROD) OVERSAMPLING FOR IMBALANCED DATASETS
V SYNTH	-	-	GENERIC	-	CLUSTERED SYNTHETIC DATA GENERATED	USING VORONOI DIAGRAMS TO IMPROVE CLASSIFICATION PERFORMANCES WHEN MODELING IMBALANCED DATASETS