Fake News Detection Using Feature Extraction, Resampling Methods, and Deep Learning

Mirmorsal Madani[1], Homayun Motameni[2*], Hosein Mohamadi[2]

Abstract–The production of fake news were practiced even before the advent of the internet. However, with the development of the internet and traditional media giving way to social media, the growing and unstoppable process of making and spreading this kind of news have become a widespread concern. Fake news by disrupting the proper flow of information and deluding public opinion, potentially causes serious problems in society. Therefore, it is necessary to detect such news, which is associated with some challenges. These challenges may be related to various issues such as datasets, events, or audiences. Lack of sufficient information about news samples, or an imbalance are the main problems in some of these datasets, which will be addressed in this paper. In the proposed model, firstly the key features in relevant datasets will be extracted to increase information about news samples. After that, using the K-nearest neighbors, a genetic, and TomekLink algorithms as the cleaning techniques, as well as designing a Generative Adversarial network, as a technique for generating synthetic data, three novel methods in the area of hybrid resampling will be presented to balance these datasets. The presented methods cause a significant increase in the performance of the deep learning algorithms to detect fake news.

Keywords:Fake news,Feature extraction, Imbalanced classification, Resampling, Deep learning

1. Introduction

Fake news is defined as untruthful news or false information spread in society to mislead the masses. This type of deceptive information is so masterfully crafted that people believe it as reality. Unfortunately, the audience, especially the ones with insufficient literacy about social media, believes the fake news, which has a significant impact on their overall attitudes toward the subject. Thus, mistrust is created in society [1]. The idea of fake news is not something new; it had existed long before the Internet [2]. However, in the current century, the speed at which the news propagates is exceptionally high thanks to the dramatic growth and dominance of social media, and due to its rapid spread over social media and the societal threat of changing public opinion, fake news has gained massive attention [3]. The producers of this kind of news, whether deliberately or not, make the most of the situation of the society and important coming events, and engage in producing fake news. They actually attack the free and secure flow of information by producing and spreading fake news and denying the audience their right to access trustworthy news, potentially causing critical problems in society [4]. The evaluation of the truth in and validity of the news has remained a challenge since it was represented as printed papers until now that it is distributed by social media [5]. Even with the advances made in the field of artificial intelligence [6] the issue of recognizing fake news has not been completely solved yet because of various causes such as the privacy policies of the social media and permissions to access information, the necessity for freedom of the media to publish news, low media literacy among the audience, and the absence of high-quality reference datasets [7]. These datasets face issues such as a) the lack of sufficient news samples in various subjects (i.e., Fake or Real news dataset) [8]; b) The short length of news samples (i.e., Liar dataset) [9]; c) unavailability of sufficient information and needed features in news samples [10] such as headline and news source (i.e., FakeNewsNet dataset) [11]; and d) imbalance in the dataset (i.e., FakeNewsNet dataset). These issues can reduce the performance of classification and detection of fake news while implementing machine learning algorithms. Many proposals have been made to tackle these issues in various papers; however, the problem of precise detection of fake

1 Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran. Email:mt_madani@yahoo.com

2* **CorrespondingAuthor:**Department of Computer Engineering, Sari Branch, Islamic Azad University, and Sari, Iran

Email:homayun_motameni@yahoo.com

3 Department of Computer Engineering, Azadshahr Branch, Islamic Azad University, Azadshahr, Iran. Email: h.mohamadi1983@gmail.com

Received: 2022.09.24; Accepted: 2022.11.07

news still exists and researchers struggle to represent classifiers with higher performance. The policies of social media make it difficult to obtain ample data about news samples. On the other hand, in the actual world, the real news about a specific subject (i.e., only political news, only economic news, etc.) can outnumber the fake news. Therefore, there will always be a possibility of facing a lack of sufficient features for news samples and having imbalance for the datasets of this area. Thus, in this paper strategies to approach these two issues will be discussed and presented. One strategy to tackle the lack of sufficient data in datasets is to extract important features from the news body or the news title (if any) in datasets. In this paper, by implementing Natural Language Processing (NLP) and related methods, some of the key features of news samples were extracted and recognized as metadata. The datasets used in this paper, both being among reference datasets in detecting fake news, have an imbalance in their structure. Therefore, we are faced with the issue of binary imbalanced data. This problem occurs when the number of samples in one class (majority class) is significantly more than the samples in the other class (minority class). The strategies to tackle this issue can be classified into two categories [12]. The first strategy involves content-based methods, and the second strategy is user-based. The content-based methods use the hidden information in the news samples such as news body. And the user-based methods involve actions of users in the social media, such as comments about news, posts, number of likes, and etc. The content-based strategy is used in this paper. This strategy can be classified to data-level or algorithm-level methods. The data-level methods manipulate the data, including feature extraction methods, resampling methods, and etc. To decrease the rate of imbalance, the data-level methods engage in decreasing the number of samples in the majority class (under sampling methods), increasing the number of samples in the minority class (oversampling methods) [13], or a combination of both. Among the data-level methods, one can mention SMOTE [14] which is a popular oversampling method. This method involves the production of synthetic samples in the minority class, which is considered a basic approach in many oversampling methods [15, 16]. The SMOTE method has many derivative and combined versions, among which safe-level SMOTE [17], ADASYN [18], Borderline SMOTE [19], LN-SMOTE [20] are a few. Some of these methods have been implemented for the purpose of comparison to the methods represented in this paper.

Among the under sampling methods, one can mention ENN [21], TomekLink removal [22], and CNN [23]. In these methods, the decision-making about the cleaning of the dataset is done based on the target class of the samples and the target class of their nearest neighbors. Among the methods that perform the cleaning of the dataset based on the distance between the samples present in the majority class, one can name Near Miss [24] and the methods represented in [25], and [26]. The [27, 28] methods also implement clustering algorithms to replace the samples with new observations. Moreover, resampling methods have been used in combination with each other in various papers. For instance, popular methods such as SMOTEENN and SMOTETomek [29, 30] initially perform the oversampling operation on the minority class, then engage in noise cleaning by implementing under sampling techniques, which cause a rise in the performance of the detection algorithms. Koziarski et al. have also implemented a combination of cleaning and resampling methods in [31] to tackle the issues of binary imbalanced data classification, calling it CCR. Furthermore, they developed a method for multi-class imbalanced data in [32]. Moreover, the core method in [33] is worth mentioning as a combined method in the oversampling field, focusing on producing synthetic samples in borderlines between classes.

The second strategy involves algorithm-level methods that try to decrease the defects in the algorithms and increase their performance when facing imbalanced data by making alterations to learning algorithms. Among the algorithm-level methods, one can name kernel functions [34], splitting criteria in decision trees [6], and modification of the underlying loss function [35].

In this paper, three methods are represented in the hybrid resampling domain based on the data-level strategy and implementing KNN, Genetic, and TomeLinks algorithms as under sampling techniques and designing a Generative Adversarial Network as the synthetic data generator. By implementing the methods represented the performance of detecting fake news increased.

The following sections of this study are arranged as follows. A review of relevant studies is carried out in Sect. 2. In Sec. 3 the datasets used in this paper which are among the popular and reference datasets in this area are discussed. Sect. 4 is dedicated to the representation of a proposed model which itself consists of sub-sections such as text preprocessing, feature extraction, resampling, and classification and detection of fake news. Sect. 5 engages in the description the results of proposed model and compares it to other methods represented in this area. Eventually, the conclusion and further suggestions are mentioned in Sect. 6.

2. Related work

The focus of this paper is on the study and representation of a strategy to tackle the issue of insufficient data in datasets of the fake news domain and their problem of being imbalanced because these issues cause a drop in the performance of the machine learning algorithms to detect fake news. Therefore, a review of previous studies in the areas of feature extraction by Natural Language Processing methods, resampling methods to balance the imbalanced datasets, and machine learning algorithms to detect fake news was carried out. To simplify, these studies were placed in two categories which will be described as follows: Sect. 2.1 Studies related to feature extraction and the implementation of machine learning methods; Sect. 2.2 resampling methods to balance the dataset.

2.1 Feature extraction and machine learning methods

In [36], the authors detected fake news using word vector representation and text features such as stylometric features. They combined the imbalanced FakeNewsNet dataset with the McIntire dataset to form a single dataset. This allowed them to create a single balanced dataset with 49.9% real news and 50.1% fake news. Three groups of stylometric features were created during the feature extraction phase. Various classification methods such as random forest,SVM,KNN, and bagging are used for fake news detection.

The authors in [37], used n-grams for feature extraction and capsule neural networks to detect fake news. They used static and non-static word embedding models to convert text features in the Liar and FakeNewsNet datasets to numeric vectors. The accuracy metric is used to evaluate the proposed method and compare it to benchmark methods like SVM. With an accuracy of 99.8%, the best result was obtained using the ISOT dataset and the non-static capsule network. Due to the problems with the dataset and the short length of the text, the authors concentrated on metadata for the Liar dataset.

In [38], the authors used machine learning techniques and extracted a set of linguistic inquiry and word count (LIWC) features to classify news articles into two categories: real and fake. They used the ISOT dataset as well as the two Kaggle datasets [39]. The authors used various ensemble techniques such as Bagging and Boosting and created two voting classifiers based on these algorithms. For evaluation purposes, they used performance metrics like accuracy, f1score, and recall. In [40], the authors presented a hybrid RNN-CNN deep learning model. The ISOT and FA-KES datasets were used to test the model. After preprocessing the news texts, they divided the dataset into two sets: training (80%) and test (20%). The Glove was used to embed the data. The proposed method was implemented on the ISOT dataset with 99% accuracy. In [41], after text preprocessing and incorporation of the text title and description, the texts were converted to a Term-document-Matrix. The TF-IDF and Information Gain techniques were then used to extract the features. For classification purposes, they used both supervised and unsupervised methods. A method was presented in [42], to increase the classification accuracy based on the most recent topics. The authors found frequently-used words using the TF-IDF technique and ranked them using the PageRank algorithm. They then used the KNN method to classify news based on the most up-to-date information in news texts.

The authors in [43] detected fake news using the N-Gram analysis and machine learning techniques. They used TF and TF-IDF to extract features after preprocessing the text. They used classical methods like KNN, as well as 5-fold cross-validation for evaluation. The specialized team gathered and prepared the used dataset, as well as the dataset in [44], and applied the proposed techniques to both. According to the results, the linear SVM and TF-IDF feature extraction techniques performed better than the other methods.

The authors in [45] attempted to detect fake news on social media. Based on users’ activities and performance, they detected and filtered misinformation-spreading websites. They attempted to detect fake news by extracting key features from the news title and news body. The logistic classifier was also used for detection and classification purposes. L. Waikhom et al. used the Liar dataset and embedding techniques suggested in [46] to detect fake news. For further simplicity, they assumed the target class was binary, in the form of either real or fake news. They used the BoW, TF-IDF, and N-gram methods for embedding and feature extraction, respectively. AdaBoost, Extra Trees, Random Forest, XGBoost, and Bagging methods were used to detect fake news. The Bagging method produced the best results, with 70% accuracy.

2.2 Resampling methods

A combined method is used in [3] to overcome the problem of imbalanced multi-class. The authors believe that the efficiency of SMOTE method and its derivatives decreases in the face of multi-class data where mutual imbalance relations are established between classes. They developed an energy-based hybrid method for clearing the data from noise, as well as finding the right area in the dataset for oversampling. The authors in [47], focused on textual data, that collected by human resources and the work experiences of different people in different occupations. The purpose of this paper was to classify the data based on different types of jobs that this job diversity had led to the production of imbalanced datasets. Authors have used different methods including numerical representation of text data, different classification algorithms, different data balancing techniques SMOTE, SMOTE-SVM, and cost-sensitive learning to analyze the costs in the numerical optimization problem. In [10], the authors presented a new fake news detection method using ensemble machine learning classifier (XGBoost) and a deep neural network model (DeepFakE) as classification methods as well as information extraction from news body and social context information. They used the BUZZFeed and FakeNewsNet datasets for evaluation. The authors in [48]investigated the effect of oversampling methods on the performance of artificial neural networks in classifying multi-class data. They applied various resampling methods such as random oversampling, random undersampling, random undersampling with ADASYN on benchmark datasets such as KDD99 and UNSW-NB15. The results of the paper included items such as increasing training time as well as increasing the detection rate of data in the minority class when using oversampling methods, reducing training time when using undersampling methods, increasing recall metrics when using resampling methods.

In [49], the authors presented a hybrid sampling method to balance the datasets. In the proposed method (HSDD), they first divided the data into several sections in each dataset. They then removed minority class samples detected as noise from the dataset and selected samples located in the borderline with minority class samples based on the results of the DD algorithm as suitable samples for oversampling operations. These selected samples were used as input to oversampling techniques to generate synthetic data.

The authors in [50] have proposed a new method using a combination of undersampling and ensemble classification methods. In the proposed method, suitable samples are identified using the results of classifiers and BinaryPSO method as an undersampling method and a new balanced database is created.In [51], the authors used a method based on a genetic algorithm (Gen-sample) for oversampling operations in the minority class. The Gen-sample Calculates the oversampling rate for each sample in the minority class based on the difficulty of that sample during learning as well as the effect of that sample on increasing the performance of the oversampling technique. The method continues to generate synthetic data as long as the classifier performance is increasing. The authors compared the performance of the method on 9 imbalanced datasets with methods such as SMOTE and ADASYN in terms of performance metrics such as accuracy.The authors in [52] focused on artificial data sampling. They developed a method called self-adaptive synthetic oversampling (SASYNO) to deal with the problem of data imbalance, which produces artificial samples close to real samples. The authors believe that oversampling methods, which randomly select samples from the minority class and produce synthetic samples at the boundary line between these samples and their neighbors, may cause the problem of overlapping between these samples and majority class samples, which can reduce the efficiency of classification algorithms. In [53], a method called MSMOTE, which is based on the SMOTE method, is presented to prevent the samples mixture. The MSMOTE selects the instances according to the distance metric and based on the random selection and nearest neighbor strategies and places them in three groups: border, safe and latent noise. The results of the method show an increase in accuracy during classification compared to the SMOTE method. A method similar to the MSMOTE method has been proposed by Sáez et al in [54]. This method uses the SMOTE method to produce artificial samples. The authors have divided the samples into four groups based on distance metrics, which include safe, borderline, rare, and outliers. In [55], the authors have reliably detected fake news so that its damage is minimized. They developed a model called Grover as a fake news generator that has the ability to detect fake news generated by itself or other neural network models. The authors believe that the best models for producing fake news can also be the best models for detecting them.

The SMOTE method is known as a basic and common method in the field of oversampling [15, 16]. This method produces synthetic samples in the minority class. First, for each sample p in the minority class, the k-nearest neighbors of p are identified, and depending on how many artificial samples are needed to balance the data, or randomly, the samples are selected. A random point is then identified along the lines that connect p and each of the r points to each other and added to the set of synthetic samples [14]. This method, unlike random oversampling, in which samples from the minority class are randomly selected and oversampled in the dataset, prevents overfitting [56]. The SMOTE method (especially in high imbalanced datasets) because it does not pay attention to the distribution of samples in the majority class and blindly produces synthetic data in the minority class, which can lead to the problem of mixing the data of the minority and majority classes [57]. In addition, if noise samples are selected, the oversampling operation will result in more production of these samples and the number of noise samples in the dataset will increase. [58]. in the borderline-SMOTE method, minority class samples are grouped into noisy, safe, and danger sets. A sample is grouped based on the number of close neighbors that have different target classes. Then in the borderline-SMOTE1 method, the production of synthetic samples is based on the samples of the danger set. The borderline-SMOTE2 method is the same as the previous method, with the difference that new synthetic samples are taken along the joining line between the samples of the danger set with their close neighbors (both in the minority class and in the majority class) [19]. This method sometimes leads to the production of synthetic samples in unsuitable areas, including noise areas.

The basic idea of the Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN) is to use a weighted distribution for different minority class samples based on their difficulty level in learning. When generating synthetic samples, this method prioritizes samples that are more difficult to learn and generates more data in those areas. This method can generate according to the number of samples of the majority class that are in the neighborhood of these samples [18].

The SMOTETomek method falls into the field of hybrid methods. This method uses SMOTE algorithm as oversampling method and the TomekLink algorithm as an undersampling method. First, the training dataset is balanced with SMOTE method. Then, using TomekLink, all pairs of nearest neighbors in the data that are next to each other in the feature space but have different target classes, are identified. Now, according to the desired strategy, the both neighbors or one of the neighbors will be removed from the database [59, 30]. This technique reduces noise in the final dataset and consequently increases performance when using machine learning algorithms.

The SMOTEENN method is one of the hybrid methods in the field of resampling. This method uses SMOTE algorithm as oversampling method and the ENN algorithm as the undersampling method. First, the training dataset is balanced by SMOTE method. Then, using the ENN technique, samples of data whose target class is different from the target class of their k-nearest neighbors are identified and removed from the dataset. Because the idea of ENN is that these samples are misclassified. Of course, according to the desired strategy, all samples or only samples belonging to one of the two classes can be removed. This technique is more biased than TomekLink in removing samples. [59, 30]

In [60], the authors have used a combination of random forest and decision tree methods in the fields of oversampling and undersampling to address the problem of imbalanced datasets. Using this combined method, they have determined the number of synthetic samples required for resampling operations.The authors in [31] proposed a hybrid oversampling method called CCR for the binary class imbalanced data problem. In the cleaning phase, they used an energy-based algorithm to clean up the neighbors close to the minority class samples. Then they tried to generate synthetic data in the safe area. The authors also developed the CRC method in [32] for multi-class imbalanced data.The authors in [61] presented a cost function based on the G-mean metric for the ELM optimization problem in learning imbalanced data. In [62], the authors proposed an entropy-based hybrid method called EHSEL to deal with the problem of data imbalance. They believed that oversampling methods in high imbalanced datasets can lead to overfitting. They also stated that the use of undersampling could lead to the omission of some important samples. In the proposed model, the training set distribution is considered based on information entropy. In the undersampling operation, they then separate the important samples. According to the authors, this operation can solve the problem of overfitting.In the paper presented in [63], a method is presented based on a genetic algorithm and PCA for classifying errors in imbalanced datasets is presented. The authors used PCA for data processing and error detection. The output of the method is a list of errors in the data and the location of the errors.The authors in [64] used common resampling methods such as SMOTE, Borderline SMOTE, and their combination with SSO-PSO to balance the datasets. SSO-PSO is an intelligent undersampling technique resulting from a combination of sample subset optimization and particle swarm optimization. This technique uses a cost function to determine the optimal and balanced subsets of samples.In [65], the authors used a GAN-based method called GS-GAN to generate text. Due to the fact that GAN networks have limitations in the production of discrete data, GS-GAN uses the Gumbel-softmax distribution [66] to deal with this problem and generate text as output. The authors used a dataset containing 5,000 samples with a maximum length of 12 characters with Context-free grammar.

3Methodology

In this section, the used datasets and proposed model for fake news detection by taking into account the related operations is described. Since the focus of this study on the detection of fake news with high precession, is examining and tackling the two issues of not having enough data about news samples in datasets and the imbalance in some datasets, the proposed model in Fig. 1 has different phases which will be elaborated in the following section. Firstly, the preprocessing of news samples in datasets will be discussed in Sect. 3. 2. After that, the feature extraction phase will be explained in Sect. 3. 3, in which three operations will be carried out to increase the performance of the detection algorithms. In the next phase, the datasets are to be balanced. In this phase, three methods in the area of hybrid resampling are represented. Finally, after balancing the datasets, fake news detection will be discussed implementing the Long-short time memory (LSTM) model as a deep learning algorithm [77].

3.1Datasets

To name a few among the reference datasets in fake news detection scope, one can refer to Liar [9], ISOT fake news [29, 43], FakeNewsNet [11], and Fake or Real news [8]. The FakeNewsNet dataset which is demonstrated in Table 1 has an imbalanced structure and is therefore addressed in this study. This dataset was gathered and established at Arizona State University. Labeling was carried out based on fact-checking websites such as PolitiFact and Gossipcop. The dataset included information about the news content, social content, and spatiotemporal information. The ISOT dataset was represented by Ahmed et al. in [29, 43]. This dataset has a balanced structure and includes the two types of fake and real news. The real news is gathered from the news articles of the Reuters.com website, and the fake news from various news sources. Each news sample included text, title, and date of issue. The news samples are about diverse subjects such as Political News, World News, Government News, US News, etc. As the focus of this study is on imbalanced datasets, in the ISOT dataset only news samples about political news were selected to create an imbalanced dataset. Therefore, after the preprocessing of news samples and eliminating ones of a null value in text or title features, a dataset was formed which consisted of 11270 real news samples and 6448 fake news samples, named EISOT. These two datasets had features such as the news text, title, and date of issue. We used the news text in the FakeNewsNet dataset and news text and title features in the EISOT dataset.

Table 1. The structural information of datasets

Data Sets	#Fake news	#Real news	Length of News samples	Subject
FakeNewsNet	5755	17441	Medium/long	Social Context/ News Content/ Spatiotemporal info
EISOT	6448	11270	Medium/long	Political News

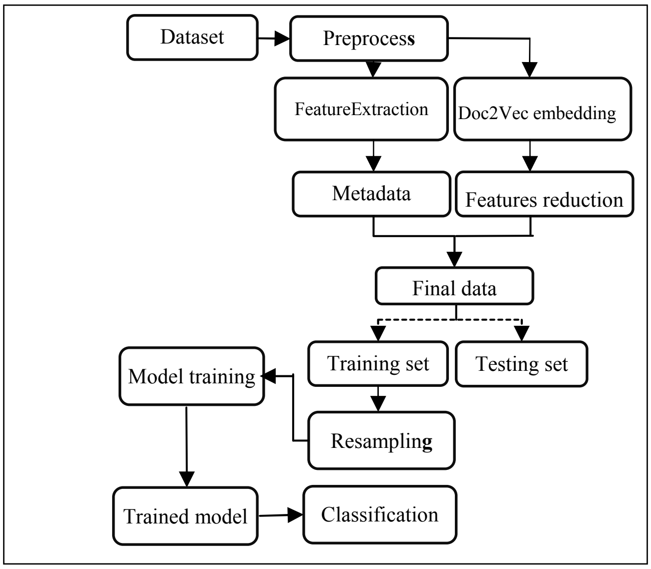


Fig. 1. The proposed model

The components of proposed model and the relevant pseudo-code are described in following.

The proposed model is summarized in the form of pseudo-code as follows

```
Input: dataset (including textual features)

for each news sample i in dataset:
  preprocess i using Beautiful Soup and NLTK
for each news sample i in dataset:
  extract key features from i and add to "Metadata"
 Use min − max normalization for each feature in "Metadata"
 Let's split dataset to training (80%) and testing (20%)
 convert training and testing to numeric vectors using doc2vec
 input = PCA.transform ($doc2vec_{raining}$)
 Training set = outer join (input ,"Metadata")
 input = PCA.transform ($doc2vec_{testing}$)
 Testing set = outer join (input ,"Metadata")
 Use resampling methods on Training set for balancing
 Use LSTM for fake news detection
```

3.2Feature Extraction (FE)

The text needs to be preprocessed before being used, a procedure is used for elimination of irrelevant data, extra characters, lemmatization, and turning the text into a usable format for machine algorithms using NLTK library of python.

Three operations were carried out in this phase after the preprocessing of the datasets. a) Due to the fact that the textual and imbalanced datasets of FakeNewsNet and EISOT only included the textual features of news body or news title and give no other key information, some important features are extracted, normalized and deemed as metadata. b) The textual features of the datasets were made into numerical vectors with fixed lengths implementing Doc2ve paragraph embedding so that they could be used in machine learning algorithms. c) In order to reduce time and memory complexities, feature-based dimensionality reduction is addressed using PCA. The related pseudo code is presented as follow.

Extraction of key Features

```
$input: Text\ (preprocessed\ Text)$
$Output: Final\ Metadata$
$Final\ metadate = [N][M\ ]$
$TS = [\ ], PS = [\ ], TW = [\ ]$
$//\boldsymbol{Surface\ features\ extraction}$
$for\ each\ news\ sample\ ns_i\ in\ Text$:
    $TS = Tokenize\ ns_i\ to\ Sentences\ using\ NLTK$
    $TW = Tokenize\ ns_i\ to\ words\ using\ NLTK$
    $PS = Part\ of\ speech\ tagging\ (POS)\ using\ NLTK$
    $SF_j = the\ number\ of\ sentences, words, characters,$
          $adjectives, adverbs,$
    $specific\ nouns, verbs, proper\ nouns, etc$
                    $Add\ SF_j\ to\ Final\ features$
               $//\boldsymbol{Polarity\ features\ extraction}$
$for\ the\ news\ body\ and\ title\ of\ the\ news\ (if\ it\ exists)$:
    $Calculate\ the\ polarity\ score\ in\ PS\ using\ SentWordNet$
$body - score = polarity\ score\ of\ news\ body$
$title - score = polarity\ score\ of\ news\ title$
$If\ body - score == title - score\ then\ matched = 1\ else\ matched = 0$
$Add\ matched\ to\ metadata$
```

3.2.1Extraction of Key Features

The Researchers have reached the conclusion that by extracting lexical, linguistic, etc. features from the text and using them as inputs in the machine learning algorithms, a more effective classification can be achieved. As an instance, one can mention the works done in [37, 67, 38, 40, 42 and 68] which will be explained in the Sect. 2. In this paper, three feature groups namely surface features, semantic features, and text polarity are extracted from the textual features present in the dataset, which will be explained in the following section. The statistical analysis of the results of this phase is presented in Sect. 4.2.

Surface Features

By implementing NLP tools and functions such as word tokenization, lemmatization, and POS tagging, features like sentence numbers, word numbers, adjective numbers, adverb numbers, proper name numbers, verb numbers, sentence length based on word numbers, etc. were extracted from news samples. After that, the features having the most dependence on target output based on Spearman's rank correlation coefficient [69] were kept and the others were removed.

- **Text Polarity features**

Sentiment analysis is used to examine the attitude of the author of a news sample. The attitude of a writer could be positive, negative, or neutral. Sentiment analysis is done in the three levels of sentence level, document level, and aspect level [70]. At the document level, the goal is to detect the sentiment from the whole text. In the detection of fake news, Sentiment analysis can be beneficial for two reasons: a) the producers of fake news do all they can to make a piece of news viral (including the choice of attractive pictures, using a suitable and attractive format, and writing for the title and the body) [71]. Therefore, as the positive or negative attitude in the title or body of a news sample can affect the appeal of that news and consequently the audience response dramatically, the detection of the polarity of news samples is tackled in this paper. b) Sentiment analysis can lead to the structuring of the text which is an important trait in implementing machine learning algorithms. Numerous methods and tools have been introduced for text sentiment analysis. In this paper, the SentWordNet function from the Python NLTK library based on WordNet was implemented for the sentiment analysis of news samples [72]. In the FakeNewsNet dataset, sentiment analysis was done on the titles and in the EISOT dataset, the sentiment analysis was carried out on the text and title separately. Since the title of a news story has a bigger share in the initial attraction of the audience than the text, and in real news, there must be a consistency between the sentiments in the body and title of a news story; thus, to do the sentiment analysis on the news samples in the EISOT dataset including the text and the title for each news sample, after detecting the news text polarity and the news title polarity, the consistency between the two is also examined. So eventually, one feature in the FakeNewsNet dataset and three features in the EISOT dataset are extracted and added to the metadata.

3.2.2 Paragraph embedding

In the following section, after a brief explanation about embedding methods, the reason for implementing Doc2vec paragraph embedding is elaborated. There are various methods in NLP for mapping words, phrases, or documents onto numerical vectors. Among these, one can mention Bag of Word (BoW) and Bag of n-grams, word vector methods like Word2Vec, FastText, and Glove, also paragraph vector methods like Doc2Vec and BERT. Bagging methods pay no heed to word meanings and the space between them, and in the BoW method, the order of the words is removed. The Bag of n-grams method is not appropriate for data with high dimensions. The Word2Vec method cannot reproduce words not present in the training dataset. In the FastText method which is an extension of the Word2Vec method [73], words are broken into n-grams which facilitates the reproduction of words that were even rarely repeated in the dataset. In this study, each news sample in the dataset is deemed as a paragraph including several sentences. Therefore, in order to represent each news sample, the

Doc2Vec embedding was implemented [74, 36] which is based on the Word2Vec embedding. Thus, Not only does this method has similar reproductions with similar mapping locations in the vector space for words with the same meaning as the Word2Vec method, but the order of the words and their situation in the paragraph is also restored. Implementing the Doc2Vec method, each news sample from the dataset can be turned into a numerical vector with a fixed length of 400.

3.2.3Feature-Based Dimensionality Reduction

Using high-dimensional data can lead to an increase in time and memory complexities. Therefore, implementing dimension reduction techniques can improve the processing procedure and memory usage [75]. In this paper, we have focused on reducing the dimensions in the columns (features). Among the methods for dimension reduction, one can mention CF-DF, DF, TF-IDF and PCA. To reduce the vocabulary size and categorize the training documents, the ranking technique is used in DF. Each class is then given its own set of sub-vocabularies. The terms are ranked according to the frequency with which they appear in documents. The class frequency is introduced in CF as a new quantity. Training documents are classified similarly to DF to determine the frequency of a term. The class frequency value represents the number of nodes to which each term in the vocabulary is applied. Therefore, the class frequency of low-frequency terms is lower. The terms in CF-DF are chosen in two phases: the terms with frequencies less than the threshold value t are selected in the first phase. Then, in the second phase, the DF techniques are used to select the terms and develop the reduced feature set. The TF-IDF method uses Term Frequency (TF) and Inverse Document Frequency (IDF) to determine the importance of terms. $\mathrm{IDFi} = \log \mathrm{N/n}$, where N is the total number of documents, and n is the number of documents containing the ith terms. Thus, the term with a higher IDF value is one that appears in fewer documents. Finally, the $\mathrm{TF} * \mathrm{IDF}$ product will be used to assess and determine the significance of a particular term. PCA is a statistical method for dimensionality reduction, which uses a linear transformation. It can be applied to a large number of variables in a dataset as a technique for extracting features or essential variables (as the principal component), i.e., the directions that maximize data variance. PCA aims to extract low-dimensional features from a high dimensional set to record more data with fewer variables. In order to reduce the time and memory complexity and overcome the overfitting issue, PCA was implemented in this study. The output of this phase is data with 4 features (dimensions).

3.3Proposed Resampling Methods

Three methods in the field of hybrid resampling are represented, namely GANENN, GANGen, and GANTomek by implementing a GAN network and the KNN, Genetic, and TomekLinks algorithms respectively. These methods include the two phases of generation and cleaning. Firstly, in the Generation phase and using the minority class samples and the GAN network (as per generation phase) the dataset is balanced. Then, in the cleaning phase, for the GANENN method by using the KNN algorithms, for the GANTomek method by using the TomekLinks algorithm, and for the GANGen method by using the genetic algorithm, bad samples (including outliers, noisy, etc.) are identified in both minority and majority classes. After that, as per the adopted strategy, these samples are removed from the final dataset. The strategy can choose to remove bad samples from one or both classes. In this paper, bad samples are eliminated from both classes.

- **Generation Phase**

Nowadays, using Generative Adversarial Networks (GANs) in the learning field has caused a great revolution especially in the fields of image processing, signal processing, and computer vision. The architecture of these networks is suitable for producing continuous data (like image). Therefore, implementing them in this field has been a focal point for computer science experts [76, 78]. These networks consist of deep learning algorithms and adversarial approaches; therefore, their performance is different from the performance of popular deep learning models. Numerous types of these networks have been developed in the field of image in recent years such as DC GANs [79], Capsule GANs [80] and Fictitious GANs [81]. However, the development of these networks in the field of texts has been limited due to the data being discrete [82]. These networks include two modules in a simplified model. The first module is a neural network named generator (G) which is trained to produce synthetic samples, and the second module is a neural network named discriminator (D) which is trained to detect fake samples from the real ones. In fact, the training of the GANs begins as a competition or a minmax game between the two modules. The goal of G is to generate synthetic samples with high similarities to the original ones to beat module D in detecting the samples. On the other hand, D is always training with original samples and updating its parameters to do its best at detecting synthetic samples and categorizing with the maximum accuracy. Therefore, module G tries to minimize the object function, while D tries to maximize the same function. The minmax function is defined as follows.

$$min_G max_D V(G,D) = \mathrm{E}_{x\sim\, Pdata(x)}[\log(D(X))] + E_{z\sim\, Pz(z)}[log(1 - D(G(z)))] \quad (1)$$

Where:
$V(G,D)$ *is loss function to be mininmized*
$D(x)$: *is the discriminator's estimated probabilityof a real sample x being real*
Ex is the mathematical expectancy over all samples from the real data set X
Ez is the mathematical expectancy over all random generator inputs
$P(z)$: *distribution of generator*, z: *sample from* $p(z)$
$G\ (z)$: *generate fake data given the noise vector z*
Discriminator: *maximize* $\log D(x) + \log 1 - D(G(z))$
Generator: *minimize* $\log(1 - D(G(z))$

In this paper, GANs were used to produce synthetic samples and actually balance the EISOT and FakeNewsNet datasets. The reason for this implementation is their ability to produce synthetic samples with high similarities to the original ones. The scale of similarity between the synthetic data and the original data can be calculated by statistical analysis which will be represented in the results section.In order to design the related GAN network, the two G and D modules were used. The input for module G is random samples (z) and its output is fake samples(x~). There will be two types of input in the input entrances of module D one of which is real samples(x) with the label 1 (the news samples in the minority class); and the other is fake samples(x~) with the label zero (module G output). In each epoch, with the input of each batch of real samples, D parameters are updated to minimize the loss function. Then, G is trained to make better samples in terms of similarity to real samples. Eventually, module D learns the distribution of the output probability based on the input data. This module will actually learn the $p\ (\frac{y}{x})$ probability for the input x and output y. Module D is an MLP neural network and has an input layer, 3 hidden layers with ReLU activation, and an output layer with one neuron with sigmoidal activation. Module G has two hidden layers with ReLU activation and one linear activation with (size of input layer) neurons in output. As the issue is a binary classification one, the BCE loss function is implemented.

- **Cleaning Phase**

Numerous studies have been carried out by researchers on imbalanced datasets to detect various samples in the minority or majority classes and their impact on the performance of the classified algorithms. For instance, the authors in [53] categorize these samples based on the metric of distance, random strategies, and the nearest neighbor into the three groups of safe, border, and latent noise instances. Furthermore, the authors in [54, 56] studies categorize the samples in the minority class into four groups of safe, borderline, rare, and outlier. In this paper, the samples in the FakeNewsNet and EISOT datasets are separated into the two groups of good samples and bad samples. Finally, based on the adopted strategy, the bad samples are removed from the final dataset. The strategy can choose to remove bad samples from one or both classes. In this paper, bad samples are eliminated from both classes. In the GANENN method, by implementing the ENN algorithm, in the GANGen algorithm by using the genetic algorithm and based on the specificity metric, and in the GANTomek algorithm by using the TomekLink algorithm these groups will be identified.

KNN Samples Selection

In this method, the metric that defines a news sample as being good is having the same target class label as its nearest neighbors. Firstly, the KNN algorithm (for K=3) is used to select the nearest neighbors for each news sample. Then, the samples which have the same target class as their 3 nearest neighbors are selected as good samples and the rest of samples in related dataset are put into bad samples and are removed from dataset.

GANKNN algorithm

```
We denote the minority class by L, #minority class by nl,
the majority class by H, #majority class by nH, the #final features by y
#generation by n, #samples per population by s, #features by m,
the imbalanced training set by T and
the balanced training set by T' and the cleaned training set by T" and
threshold by t and Tomek Links by TL
Generation phase (by GAN)
Input: Input: T, x (samples in L), n (#epochs), k (#steps) Output: T'
 − Training GAN:
  for n:
    for k, x:
      (data for training the D)
      Sample batch_size of noise samples (z1 … zy) from pg(z)
Sample batch_size of real samples (x1 … xy) from Pdata(x)
      calculate the loss function and the gradients
      update the D weights by calling optimizer
    (data for training the G)
    Sample batch_size of noise samples (z1 … zy) from pg(z)
Calculate the loss function and the gradients
    Update the D weights by calling optimizer
− Generate samples:
  Generate Synthetic_samples((nH − nL) × m) using G
  T' = concat (T and Synthetic samples)
Cleaning phase (by ENN):
Input: T′ and K
Output = T"
   for each sample s in T′:
     compute its K nearest neighbors in T′
      (call this set Ns )
      If s in L and |H ∩ Ns| == K: : add s to bad samples
      If s in H and |L ∩ Ns| == K: : add s to badsamples
T" = T' – bad samples (based on desired strategy)
```

- **Genetic Samples Selection**

The genetic algorithm is an evolutional method based on repetition for the limited and unlimited enhancement issues, benefitting from Darwin's natural selection principles to find an optimal formula for prediction. The genes represent the quantity and variable of the problem and are placed as strings inside the chromosomes as the main variables of our problem. Population is the concept of a group of chromosomes related to one generation and is one other significant concept in this algorithm. In nature, the combination of better chromosomes yields a better generation. Similarly, the genetic algorithm tends to tackle problems in the same way [84]. This algorithm begins its overall process with an initial population of random samples. Each sample in the population represents a potential solution to the designated problem. The samples evolve through successive repetitions called generations, and each generation is evaluated by fitness function metric. In this algorithm, the accuracy metric is implemented as we are searching for samples which can be chosen as good samples in order to reach maximum performance in the detection of fake news samples when implementing the machine learning algorithms and designate the best parents. Then, the crossover and mutation operations are implemented to detect the population for the new generation based on the best parents and offspring. This process will be repeated according to the number of generations. At the end of the algorithm, good samples will be identified which are the best news samples in the training dataset. Now, the rest of the samples are put into the bad samples and are removed from dataset.

GANGen algorithm

```
Generation phase (by GAN)
as same as Algorlthm1
− Cleaning phase (by Genetic):
Input: T′, n, s Output = T"
Initialize population (s × m)
for n:
  //measuring the fitness of samples in the population
  for s:
     calculate the predict of samples (%accuracy)
   Select the best parents(#s/2) from samples for mating pool
  //Generating next generation using crossover
  for s/2:
      crossover parents −> child(s/2 × m)
      adding some changes to the child using mutation
  creating the new population(s × m)based on the
   parents(s/2 × m) and child(s/2 × m)
  update current population
Gen_output = best samples
T″ = best samples
```

- **TomekLink Samples Selection**

In this method, using TomekLink algorithm, all pairs of nearest neighbors in the data that are next to each other in the feature space but have different target classes, are identified. Now, according to the desired strategy, the both neighbors or one of the neighbors will be removed from the database.

GANTomek algorithm

```
Generation phase (by GAN)
as same as Algorlthm1

− Cleaning phase (by Tomek Links):
Input: T′, t Output = bad samples
Temp = []
for each sample s in T′:
  If s not in temp:
     Add s to temp
     NN = Nearest neighbor (s)
     Add NN to temp
     If s in H and |NN ∩ L|: add pair (s , NN)to badsamples
     If s in L and |NN ∩ H|: add pair (s , NN)to badsamples

T″ = T′ – bad samples
```

4. Experimental results

4.1 Performance parameters

Regarding [85], the performance evaluation metrics can be categorized into three groups. a) Ranking metrics such as AUC used for evaluating classifications; b) Threshold metrics such as accuracy to measure the classification prediction; c) Probability metrics like LogLoss which are suitable the evaluate the reliability of classifier. The popular metrics in the evaluation of classifications are some Threshold metrics such as accuracy, recall, and accuracy. It should be noted that the results of evaluation metrics are pivotal for decision-making. Therefore, choosing appropriate metrics is very important. The authors in [86] have concluded that the sole usage of accuracy metric in the evaluation of classifications in imbalanced datasets could lead to misguided decisions. By comparing the F1Score and accuracy metrics, the authors in [87] have shown that in the binary imbalanced data classification problem, these two metrics can yield very optimistic results, leading to wrong decisions. The authors in [88] have represented the Balanced Accuracy metric for the evaluation of classifications. They introduce this metric as an objective and suitable one for imbalanced datasets. The authors in [89] claim that for imbalanced datasets, the classification error rate in the minority class is more than the majority class, and there should be more focus on the evaluation of classifications in the minority class. Therefore, as the datasets in this paper are imbalanced, for the evaluation of models the sensitivity, accuracy, precision, specificity, and AUC were implemented.

Confusion matrix is a two-dimensional matrix of actual class values and predicted class values. It summarizes the classification performance of a classifier [91]. In the Confusion Matrix, True Positives (TP) denote the correctly-predicted positive values, True Negatives (TN) denote the correctly-predicted negative values, False Positives (FP) denote the incorrectly-predicted positive values, and False Negatives (FN) denote the incorrectly predicted negative values. Sensitivity (Sen) is the ratio of correctly predicted positive observations to the all observations in actual class. Accuracy (Acc) is the ratio of correctly predicted observations to the all observations. Specificity (Spec) is the ratio of correctly predicted negative to the all observations in actual class. Precision (Prec) is the ratio of correctly predicted positive observations to the all predicted positive observations.

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \qquad (2)$$

$$Prec = \frac{TP}{TP + FP} \qquad (3)$$

$$Spec = \frac{TN}{TN + FP} \qquad (4)$$

$$Sen = \frac{TP}{TP + FN} \qquad (5)$$

Since the issue of detecting fake news is a binary classification one, using the ROC [90] curve can show the scale of separation the model brings or, in other words, the performance of the model in differentiating between the two classes. Therefore, the AUC metric which actually represents the area below the ROC curve is a suitable metric for evaluating the performance of the model, especially when using imbalanced datasets.

4.2 Results

In this section, the performance of the represented models in detecting fake news is evaluated. After the operations of preprocessing, the datasets are splitted into the two parts of training set (%80) and testing set (%20). Then, %20 of the training set is deemed as validation data. The results of implementing the proposed model on the used datasets are presented in Table 2. The results are shown in two sections, before performing feature extraction and resampling operations and also after performing these two operations. According to the results in Table 2, which presented for FakeNewsNet dataset, the performance of the model has increased (on average) %17.76, after performing the feature extraction and hybrid resampling operations.

Figure 2, illustrate the impact (on average) of each of the operations (feature extraction and the proposed hybrid resampling methods) on the performance of LSTM from the viewpoint of AUC. As can be seen, after feature extraction, AUC has increased in EISOT and FakeNewsNet datasets by %1.5 and %2 respectively, which is not satisfactory. This could be because of the imbalanced structure of these datasets, and also the difficulty thereof. The impact of hybrid resampling methods on the performance metrics especially from the viewpoint of AUC is significant. For instance, after balancing the FakeNewsNet dataset using the hybrid resampling methods, the AUC rose by %16.26.

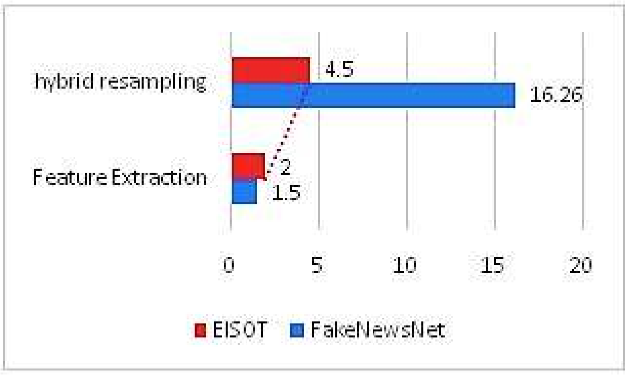


Fig.2. The effectiveness of the proposed model on performance (AUC %)

Table 2. Results of the hybrid resampling methods

dataset		method	Acc	Prec	Sen	Spec	AUC
Fake News Net	before feature extraction & hybrid resampling		80.2	83.5	92	43.5	68
	after feature extraction & hybrid resampling	GANGen	84	83.5	82.5	84	83.3
		GANENN	97.5	98.7	96.5	91.2	94
		GANTomek	83	84.5	79	80.8	80
EISOT	before feature extraction & hybrid resampling		88.3	91.2	90.5	84.5	87.3
	after feature extraction & hybrid resampling	GANGen	91	90.5	92.5	90.2	91.3
		GANENN	99.3	99.5	99.2	98.2	98.6
		GANTomek	92.5	90	93	90	91.5

Fig.3 and Fig.4, illustrate the ability of the SMOTE and GAN (presented in this paper) in two datasets used from the point of view of synthetic data generation. The figures are drawn based on the statistical descriptors including mean, Standard Deviation and Variance. The results show that the synthetic data generated by the GAN network are more similar to the original data than the SMOTE method.Comparing the performance of GANENN, GANGen, and GANTomek methods in hybrid resampling domain with other methods in the two datasets studied in this paper are represented in Table 3. The method whose implementation results were better in terms of performance metrics than other methods is bolded in the table. In FakeNewsNet and EISOT datasets, the average performance (AUC viewpoint) in GANENN, GANGen, and GANTomek methods is approximately %10.3 and %1.8 higher than SMOTEENN and SMOTETomek methods respectively.

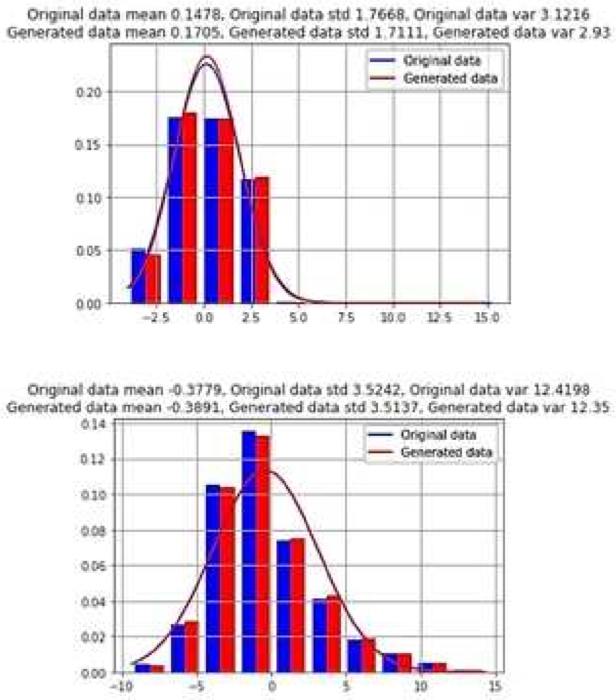
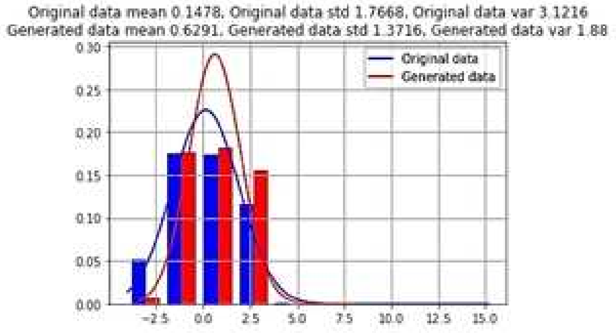


Fig .3. Statistical analysis of synthetic data generated byGANinFakeNewsNet (a), EISOT (b)



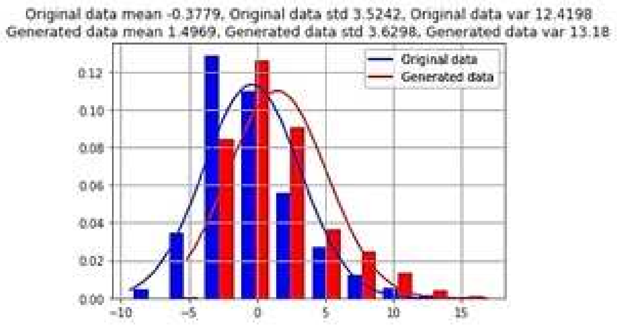


Fig .4. Statistical analysis of synthetic data generated by SMOTE in FakeNewsNet (a), EISOT (b)

Table. 3. Comparison of hybrid resampling methods

Dataset	method	Acc	Prec	Sen	AUC
FakeNewsNet	SMOTETomek	75	76	74	71
	SMOTEENN	87.5	87	83	80
	GANENN	**97.5**	**98.7**	**96.5**	**94**
	GANTomek	83	84.5	79	80
	GANGen	84	83.5	82.5	83.3
EISOT	SMOTETomek	89.5	88.5	91	88
	SMOTEENN	98	97.5	98.5	96
	GANENN	**99.3**	**99.5**	**99.2**	**98.6**
	GANTomek	92.5	90	93	91.5
	GANGen	91	90.5	92.5	91.3

5. Conclusion

This study focused on the detection of fake news that has been growing in recent years. One of the problems in this area is the absence of appropriate reference datasets. These datasets face issues such as unavailability of sufficient information about news samples and imbalanced structure. These issues can reduce the performance of fake news detection while implementing machine learning algorithms. Many proposals have been made to tackle these issues in various papers; however, the problem of precise detection of fake news still exists and researchers struggle to represent identifiers with higher performance. Thus, in this paper strategies to approach these two issues were discussed and presented. One strategy to tackle the lack of sufficient information in datasets is to extract important features from the news body or the news title (if any) in datasets. In this paper, by implementing Natural Language Processing and related tools and methods, some of the key features of news samples were extracted and recognized as metadata. The strategies to tackle the imbalanced structure of datasets can be classified into two categories including data-level and algorithm level methods. In this paper, three hybrid resampling methods based on the data-level strategy were represented. By implementing these methods, the performance of detecting fake news increased. In future work, the authors of this paper plan to overcome other problems of datasets in this field, such as the short length of news samples using algorithm-level methods.

References

[1] Desuky A.S, Hussain S (2021) an Improved Hybrid Approach for Handling Class Imbalance Problem. Arab J SciEng 46, 3853–3864(2021). https://doi.org/10.1007/s13369-021-05347-7

[2] ChenY, Conory N, Rubin.V (2015) News in an Online World: The Need for an Automatic Crap Detector ASIST '15: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community November 2015 Article No.: 81 Pages 1–4

[3]Shrestha, A., Spezzano, F. Characterizing and predicting fake news spreaders in social networks. Int J Data Sci Anal (2021). https://doi.org/10.1007/s41060-021-00291-z

[4] Zhang X, Ghorbani AA (2019) An overview of online fake news: Characterization, detection, and discussion, Information Processing & Management, Volume 57, Issue 2,2020,102025,ISSN:0306 4573,https://doi.org/10.1016/j.ipm.2019.03.004 (https://www.sciencedirect.com/science/article/pii/S0306457318306794)

[5] Figueira Á, Oliveira L (2017) the current state of fake news: challenges and opportunities. Procedia Computer Science, Volume 121, 2017, Pages 817-825, ISSN 1877-0509, https: //doi.org/10.1016/j.procs.2017.11.106. (https://www.sciencedirect.com/science/article/pii/S1877050917323086)

[6] Fenglian Li, Xueying Zhang, Xiqian Zhang, Chunlei Du, Yue Xu, Yu-Chu Tian (2018) Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets, Information Sciences, Volume 422, 2018, Pages 242-256, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2017.09.013. (https://www.sciencedirect.com/science/article/pii/S0020025517304784)

[7] Zhou X, Jain A, Phoha VV, Zafarani R (2019) Fake News Early Detection: A Theory-driven Model. arXiv preprint arXiv: 1904.11679

[8] McIntire G (2018) Fake and Real News Dataset. [Online], Available: https://github.com/GeorgeMcIntire/fake_real_news dataset, July 10, 2018

[9] Wang WY (2017) Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (p. 422426)

[10] Kaliyar R.K, Goswami A, Narang P (2021) DeepFakE: improving fake news detection using tensor decomposition-based deep neural network. J Supercomputing 77, 1015–1037. https://doi.org/10.1007/s11227-020-03294-y

[11] Shu K, Mahudeswaran D, Wang SH, Lee D, Liu H (2018) FakeNewsNet: A Data Repository with News Content, Social Context and Spatial temporal Information for Studying Fake News on Social Media [Online], Available: https://arxiv.org/abs/1809.01286, December 15, 2018

[12] Stefanowski J. (2016) Dealing with Data Difficulty Factors While Learning from Imbalanced Data. In: Matwin S., Mielniczuk J. (eds) Challenges in Computational Statistics and Data Mining. Studies in Computational Intelligence, vol 605. Springer, Cham. https://doi.org/10.1007/978-3-319-18781-5_17

[13] Michał K, Potential (2021) Anchoring for imbalanced data classification, Pattern Recognition, Volume 120, 2021, 108114, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2021.108114.

[14] Chawla N.V, Bowyer K. W, Hall L. O, Kegelmeyer W. P (2002) SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[15] Maria P, Pedro Antonio G, Peter T, Cesar H (2016) Oversampling the minority class in the feature space, IEEE Trans. Neural Netw. Learning Syst. 27 (9) 1947–1961.

[16] Bellinger, C, Drummond, C, Japkowicz, N (2018). Manifold-based synthetic oversampling with manifold conformance estimation. Mach Learn 107, 605–637.https://doi.org/10.1007/s10994-017-5670-4

[17] Bunkhumpornpat C., Sinapiromsaran K., Lursinsap C. (2009) Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem. In: Theeramunkong T., Kijsirikul B., Cercone N., Ho TB. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2009. Lecture Notes in Computer Science, vol 5476. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-01307-2_43

[18] He, Haibo & Bai, Yang, Garcia, Edwardo, Li, Shutao. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. Proceedings of the International Joint Conference on Neural Networks. 1322 - 1328. 10.1109/IJCNN.2008.4633969.

[19] Han H, Wang WY, Mao BH (2005) Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang DS, Zhang XP, Huang GB. (eds) Advances in Intelligent Computing. ICIC 2005.

Lecture Notes in Computer Science, vol 3644. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11538059_91

[20] Maciejewski, Tomasz, Stefanowski, Jerzy. (2011). Local neighbourhood extension of SMOTE for mining imbalanced data. Proceeding of the IEEE symposium on computational intelligence and data mining. 104-111. 10.1109/CIDM.2011.5949434.

[21] Wilson D.L (1972) Asymptotic properties of nearest neighbor rules using edited data IEEE Trans. Syst. Man. Cybern., 2 (3) (1972), pp. 408-421

[22] Two Modifications of CNN," in IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-6, no. 11, pp. 769-772, Nov. 1976, doi: 10.1109/TSMC.1976.4309452.

[23] Hart P (2006) The condensed nearest neighbor rule (corresp.). IEEE Trans. Inf. Theor., 14(3):515{516,

[24] Interject M, Zhang (2003) knn approach to unbalanced data distributions: a case study involving information extraction. In Proceedings of workshop on learning from imbalanced datasets, 2003.

[25] Drasko F, Srdjan S, Slobodan J, Silvana P, Misko S, Distance based resampling of imbalanced classes: With an application example of speech quality assessment, Engineering Applications of Artificial Intelligence, Volume 64, 2017, Pages 440-461, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2017.07.001.

[26] Peng M, Zhang Q, Xing X, Gui T, Huang X, Jiang Y.-G, Ding K., Chen Z (2019). Trainable Undersampling for Class-Imbalance Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 4707-4714. https://doi.org/10.1609/aaai.v33i01.33014707

[27] Lin W, Chih-Fong T, Ya-Han H, Jing-Shang J (2017) Clustering-based undersampling in class-imbalanced data." Inf. Sci. 409 (2017): 17-26.

[28] Show-Jane Y, Yue-Shi L (2009) Cluster-based under-sampling approaches for imbalanced data distributions, Expert Systems with Applications, Volume 36, Issue 3, Part 1, 2009, Pages 5718-5727, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2008.06.108. (https://www.sciencedirect.com/science/article/pii/S0957417408003527)

[29] Ahmed H, Traore I, Saad S (2018) Detecting opinion spams and fake news using text classification", Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.

[30] Batista, Gustavo & Prati, Ronaldo & Monard, Maria-Carolina. (2004). A Study of the Behavior of Several Methods for Balancing machine Learning Training Data. SIGKDD Explorations. 6. 20-29. 10.1145/1007730.1007735.

[31] Koziarski, Michał, Wożniak, Michał (2017) CCR: A combined cleaning and resampling algorithm for imbalanced data classification" International Journal of Applied Mathematics and Computer Science, vol.27, no.4, 2017, pp.727-736. https://doi.org/10.1515/amcs-2017-0050

[32] Michał K, Michał W, Bartosz K (2020) Combined Cleaning and Resampling algorithm for multi-class imbalanced data with label noise, Knowledge-Based Systems, Volume 204, 2020, 106223, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2020.106223. (https://www.sciencedirect.com/science/article/pii/S0950705120304330)

[33] Bunkhumpornpat C, Sinapiromsaran K (2015). CORE: Core-based synthetic minority over-sampling and borderline majority under-sampling technique, International Journal of Data Mining and Bioinformatics 12(1): 44–58.

[34] Mathew, Josey, Pang, Chee & Luo, Ming, Leong, Weng. (2017). Classification of Imbalanced Data by Oversampling in Kernel Space of Support Vector Machines. IEEE Transactions on Neural Networks and Learning Systems. PP. 1-12. 10.1109/TNNLS.2017.2751612.

[35] Khan SH, Hayat M, Bennamoun M, Sohel FA, Togneri R (2017) Cost-Sensitive Learning of Deep Feature Representations from Imbalanced Data. IEEE Trans Neural Netw Learn Syst. 2018 Aug; 29(8):3573-3587. doi: 10.1109/TNNLS.2017.2732482. Epub 2017 Aug 17. PMID: 28829320.

[36] Reddy H et al (2020) Text-mining-based Fake News Detection Using Ensemble Methods", International Journal of Automation and Computing, DOI: 10.1007/s11633-019-1216-5 (H. Reddy, 2020)

[37] Goldani MH, Momtazi S, Safabakhsh R (2021) Detecting fake news with capsule neural networks. Applied Soft Computing, Volume 101, 106991, ISSN 1568 4946, https://doi.org/10.1016/j.asoc.2020.106991. (https://www.sciencedirect.com/science/article/pii/S1568494620309303)

[38] Iftikhar A, Muhammad Y, Suhail Y, Muhammad OA (2020) Fake News Detection Using Machine Learning Ensemble Methods. Complexity, vol. 2020, Article ID 8885861, 11 pages. https://doi.org/10.1155/2020/8885861

[39] Kaggle (2018) Fake News Detection. Kaggle, San Francisco, CA, USA, https://www.kaggle.com/jruvika/fake-news-detection

[40] Nasir JA, Khan OS, Varlamis I (2020) Fake news detection: A hybrid CNN-RNN based deep learning approach. Elsevier, International Journal of Information Management Data Insights, https://doi.org/10.1016/j.jjimei.2020.100007

[41] Goseva K et al (2020) Identification of Security related Bug Reports via Text Mining using Supervised and Unsupervised Classification, https://ntrs.nasa.gov/search.jsp?R=20180004739 2020-02 02T17:46:02+00:00Z

[42] Yukari O, Ichiro K (2013) Text Classification based on the Latent Topics of Important Sentences extracted by the PageRank Algorithm", Proceedings of the ACL Student Research Workshop, pages 46–51, Sofia, Bulgaria, August 4-9 2013. Association for Computational Linguistics

[43] Ahmed H, Traore I, Saad S. (2017) "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138).

[44] Horne B.D, Adali S (2017) This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: the 2nd International Workshop on News and Public Opinion at ICWSM

[45] Aldwairi M, Alwahedi A (2018) Detecting Fake News in Social Media Networks" ScienceDirect, Procedia Computer Science 141 (2018) 215- 222

[46] Waikhom L, Goswami, RS (2019) Fake News Detection Using Machine Learning. Proceedings of International Conference on Advancements in Computing & Management (ICACM) Available at SSRN: https://ssrn.com/abstract=3462938 or http://dx.doi.org/10.2139/ssrn.3462938 les. In Proceedings of the Eighth International Joint Conference on Natural Language Processing Short Papers pp. 252{256)

[47] Padurariu C, Breaban M (2019) Dealing with Data Imbalance in Text Classification. Procedia Computer Science. 159. 736-745. 10.1016/j.procs.2019.09.229

[48] Bagui S, Li K (2021) Resampling imbalanced data for network intrusion detection datasets. J Big Data 8, 6 (2021). https://doi.org/10.1186/s40537-020-00390-x

[49] Liping C, Jiabao J, Yong Z (2021), HSDP: A Hybrid Sampling Method for Imbalanced Big Data Based on Data Partition, Complexity, vol. 2021, Article ID 6877284, 9 pages, 2021. https://doi.org/10.1155/2021/6877284

[50] Li J, Wu Y, Fong S et al (2021) a binary PSO-based ensemble under-sampling model for rebalancing imbalanced training data. *J Supercomputing...* https://doi.org/10.1007/s11227-021-04177-6

[51] Vishwa K, Wenhao Z, Arash N, Ramin R (2019), GenSample: A Genetic Algorithm for Oversampling in Imbalanced Datasets, arXiv,abs/1910.10806

[52] Gu Xiaowei, Angelov P, Soares E (2019) A Self-Adaptive Synthetic Over-Sampling Technique for Imbalanced Classification

[53] Hu S.G, Liang Y.F, Ma L.T, He Y (2009) MSMOTE: Improving Classification Performance When Training Data is Imbalanced. In Proceedings of the 2009 Second International Workshop on Computer Science and Engineering, WCSE '09, Washington, DC, USA, 28–30 October 2009; Volume 2, pp. 13–17.

[54] Sáez J.A, Krawczyk B, Wo ´zniak M (2016) Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. Pattern Recognit. 2016, 57, 164–178

[55] Zellers, Rowan H, Ari R, Hannah B, Yonatan F, Ali R, Franziska C, Yejin. (2019). Defending Against Neural Fake News.

[56] Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2012) A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 2012, 42, 463–484.

[57] Fernández A, García, S, Herrera F (2011) Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution. In Hybrid Artificial Intelligent Systems: Proceedings of the HAIS 2011 6th International Conference, Wroclaw, Poland, 23–25 May 2011; Corchado, E.; Kurzy ´nski, M., Wo ´zniak, M., Eds.; Springer: Berlin/Heidelberg, Germmany, 2011; Part I; pp. 1–10.

[58] Pattaramon V, Eyad E (2019)Neighbourhood-based undersampling approach for handling imbalanced and overlapped data, Information Sciences, Volume 509, 2020, Pages 47-70, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2019.08.062. (https://www.sciencedirect.com/science/article/pii/S0020025519308114)

[59] Batista, Gustavo & Bazzan, Ana & Monard, Maria-Carolina. (2003). Balancing Training Data for Automated Annotation of Keywords: a Case Study.the Proc. Of Workshop on Bioinformatics. 10-18.

[60] El-Shafeiy E, Abohany A (2020) Medical imbalanced data classification based on random forests. In: Joint European-US Workshop on Applications of Invariance in Computer Vision (pp. 81–91). Springer, Cham

[61] i J, Kim H (2020) G-mean based extreme learning machie for imbalance learning. Dig. Signal Process. 98, 10267 (2020)

[62] Dongdong L, Ziqiu C, Bolu W, Zhe W, Hai Y, Wenli D (2021) Entropy-based hybrid sampling ensemble learning for imbalanced data. Int J IntelSyst. 2021; 36: 3039–3067. https://doi.org/10.1002/int.22388

[63] Babu M. Pushpa S (2020). Genetic Algorithm-Based PCA Classification for Imbalanced Dataset. 10.1007/978-981-15-2780-7_59

[64] Susan S, Amitesh (2020). Hybrid of Intelligent Minority Oversampling and PSO-Based Intelligent Majority Undersampling for Learning from Imbalanced Datasets. 10.1007/978-3-030-16660-1_74

[65] Kusner M, Hernández-Lobato, J (2016). GANS for Sequences of Discrete Elements with the Gumbel-softmax Distribution

[66] Jang E, Gu S, Poole B (2017) Categorical reparameterization with Gumbel-Soft- max, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings

[67] YounusKhan J et al (2021) A benchmark study of machine learning models for online fake news detection.Elsevier, Machine Learning with Applications Journal, https://doi.org/10.1016/j.mlwa.2021.100032

[68] Reis JCS, Correia A, Murai F, Veloso A, Benevenuto F (2019) Supervised Learning for Fake News Detection. in IEEE Intelligent Systems, vol. 34, no. 2, pp. 76-81, March-April 2019, doi: 10.1109/MIS.2019.2899143

[69] Spearman C (1987) The proof and measurement of association between two things, Am. J. Psychol. 15 (1904) 72–101

[70] Nandwani P, Verma R (2021) A review on sentiment analysis and emotion detection from text. Soc. Netw. Anal. Min. 11, 81.https://doi.org/10.1007/s13278-021-00776-6

[71] Baptista, João, Gradim, Anabela (2020) Understanding Fake News Consumption: A Review. Social Sciences. 9. 10.3390/socsci9100185.

[72] Baccianella S, Esali A, Sebastiani F (2010) SentiWordNet 3.0, An enhanced Lexical resource for sentiment analysis and opinion mining in:7th international conference on language resources and evaluation (LREC), pp 200-2204

[73] Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with sub word information, Transactions of the association for computational linguistics, vol.5, pp.135-146, 2017, Distributed under a CC-BY 4.0 license

[74] Le Q, Mikolov T (2014) Distributed Representations of Sentences and Documents. Proceedings of the 31 st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s)

[75] Chetana V, Kolisetty Soma S, Amogh K (2020). A Short Survey of Dimensionality Reduction Techniques. 10.1201/9781003043980-2.

[76] Tian L, Wang Z, Liu W et al (2021) An improved generative adversarial network with modified loss function for crack detection in electromagnetic nondestructive testing. Complex Intell. Syst. https://doi.org/10.1007/s40747-021-00477-9

[77] Sepp H, Jurgen S (1997) Long short-term memory. Neural computation", 9(8):1735–1780

[78] Yang P, Paul D.Y, Juanita F, Bing B. Z, Zili Z, Albert Y. Z (2014) Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications." IEEE transactions on cybernetics44, no. 3: 445-455

[79] Radford A, Metz L, and Chintala S, "Addressing the Classification with Imbalanced Data with deep convolutional generative adversarial networks," arXiv preprint arXiv: 1511.06434, 2015.

[80] Ayush J, Wael A, Yue W, Premkumar N, "Capsulegan: Generative adversarial capsule network,"in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 0–0.

[81] Ge H, Xia Y, Chen X, Berry R, Wu Y (2018) Fictitious GAN: Training GANs with Historical Models. In: Ferrari V., Hebert M., Sminchisescu C., Weiss Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11205. Springer, Cham. https://doi.org/10.1007/978-3-030-01246-5_8

[82] Iqbal, T., Qureshi, S., The Survey: Text Generation Models in Deep Learning., Journal of King Saud University Computer and Information Sciences (2020), doi: https://doi.org/10.1016/j.jksuci.

[83] Napierala K., Stefanowski J (2016) Types of minority class examples and their influence on learning classifiers from imbalanced data. J Intell Inf Syst 46, 563–597. https://doi.org/10.1007/s10844-015-0368-1

[84] Vallada E, Ruiz R (2011). A genetic algorithm for the unrelated parallel machine scheduling problem with sequence dependent setup times. European Journal of Operational Research. 211. 612-622. 10.1016/j.ejor.2011.01.011.

[85] Ferri C, Hernández-Orallo J, Modroiu R (2009) An experimental comparison of performance measures for classification, Pattern Recognition Letters, Volume 30, Issue 1, 2009, Pages 27-38, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2008.08.010

[86] Haibo H, Yunqian M (2013). Imbalanced Learning: Foundations, Algorithms, and Applications 10.1002/9781118646106.

[87] Davide C, Giuseppe J (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* 21. 10.1186/s12864-019-6413-7.

[88] García V, Mollineda R.A, Sánchez J.S (2009) Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions. In: Araujo H., Mendonça A.M., Pinho A.J., Torres M.I. (eds) Pattern Recognition and Image Analysis. IbPRIA 2009. Lecture Notes in Computer Science, vol 5524. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-02172-5_57

[89] Branco P, Torgo L, Ribeiro R (2015) A survey of predictive modelling under imbalanced distributions. ACM Comput Surv (CSUR). https://doi.org/10.1145/2907070

[90] Andrew P. B (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition, Volume 30, Issue 7, 1997, Pages 1145-1159, ISSN 0031-3203,

[91] Ting K.M (2011) Confusion Matrix. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_157