Enhancing Arabic Fake News Detection: Evaluating Data Balancing Techniques Across Multiple Machine Learning Models

Eman Aljohani
Department of Artificial Intelligence and Data Science, College of Computer Science and Engineering, Taibah University, Saudi Arabia
emmjohani@taibahu.edu.sa (corresponding author)

Received: 3 June 2024 | Revised: 20 June 2024 | Accepted: 27 June 2024

ABSTRACT

The spread of fake news has become a serious concern in the era of rapid information dissemination through social networks, especially when it comes to Arabic-language content, where automated detection systems are not as advanced as those for English-language content. This study evaluates the effectiveness of various data balancing techniques, such as class weights, random under-sampling, SMOTE, and SMOTEENN, across multiple machine learning models, namely XGBoost, Random Forest, CNN, BIGRU, BILSTM, CNN-LSTM, and CNN-BIGRU, to address the critical challenge of dataset imbalance in Arabic fake news detection. Accuracy, AUC, precision, recall, and F1-score were used to evaluate the performance of these models on balanced and imbalanced datasets. The results show that SMOTEENN greatly improves model performance, especially the F1-score, precision, and recall. In addition to advancing the larger objective of preserving information credibility on social networks, this study emphasizes the need for advanced data balancing strategies to improve Arabic fake news detection systems.

Keywords-text classification; fake news; imbalanced datasets; balancing techniques

I. INTRODUCTION

The exponential growth of social networks has resulted in a unique set of challenges when it comes to determining the authenticity of shared content. Users can freely publish content, adding to the abundance of information whose veracity is often in doubt. It is more important than ever for automated systems to identify fraudulent content quickly and accurately, keeping up with the rapid dissemination of information. In this particular context, it is especially difficult to identify rumors and fake news in Arabic-language social networks. The development of automated detection systems for Arabic has lagged behind efforts in English and other languages. This delay can be partly attributed to the inherent dialectical and linguistic complexity of Arabic, as well as the relative lack of large-scale, well-balanced datasets for machine learning model training. Furthermore, the problem of imbalanced datasets, where real articles greatly outnumber fake ones, makes the creation of efficient detection tools even more difficult. This imbalance may result in biased models that are less accurate in detecting falsehoods, but more successful in identifying true content. To improve the ability of machine learning models to detect fake news in Arabic language content on social networks, this imbalance must be addressed using a variety of data-balancing strategies. This study examines how various data balancing techniques affect the performance of different machine learning models when it comes to detecting fake news in Arabic.

Current approaches for fake news detection in Arabic mostly focus on scenarios with balanced data. This study aims to address a major gap in the current research landscape by presenting a novel approach using imbalanced classification techniques. The primary objective is to improve the robustness and accuracy of topic categorization associated with misinformation by customizing these techniques to the unique challenges of skewed class distributions commonly found in fake news datasets. The current study investigates several approaches to address the problem of unbalanced datasets in machine learning, improving the efficiency and precision of classification models. Two different datasets, an imbalanced and a balanced one, were used to train these models and thoroughly assess their performance in managing both dataset types by conducting a thorough analysis of their efficacy and adaptability in a variety of data scenarios. The primary goal is to address the critical challenge of dataset imbalance in Arabic fake news detection and to improve detection systems through the evaluation of various data balancing techniques. This study is among the first ones to thoroughly evaluate how different data balancing strategies affect several machine learning models designed especially for the identification of fake news in Arabic. The lack of studies on this subject highlights the necessity to address a critical gap in the literature by examining the influence of various data-balancing techniques.

This study evaluates the efficacy of data balancing techniques across multiple machine learning models to improve the performance of fake news detection systems for Arabic content. The objectives of the present study can be summarized as follows:

- Evaluate how well different methods, such as random under-sampling, Synthetic Minority Over-sampling Technique (SMOTE), and SMOTEENN, and class weights solve dataset imbalance in Arabic fake news detection.
- Investigate the performance of different models, including Extreme Gradient Boosting (XGBoost), Random Forest (RF), CNN, Bidirectional Gated Recurrent Unit (BiGRU), Bidirectional Long Short-Term Memory (BiLSTM), CNN-LSTM, and CNN-BiGRU, on both imbalanced and balanced datasets to identify models that effectively handle Arabic text.
- Enhance Arabic fake news detection systems to increase the accuracy of information on social media.

Previous studies on fake news detection in Arabic mostly focused on classifying the authenticity of news deploying deep learning and conventional machine learning models. In [1], two sets of features, word frequency and word embeddings, and three machine learning algorithms, Logistic Regression (LR), Support Vector Classification (SVC), and Naïve Bayes (NB), were used to classify rumor-related tweets. The best-performing classifiers were the LR classifier using count vector features and the SVC using TF-IDF, both achieving 84.03% accuracy. This method employed a top-down approach, first identifying features associated with rumors and then selecting data according to those findings. Recent studies investigated the efficacy of several feature extraction methods, such as TF-IDF, Glove, and BERT embeddings, in identifying fake news in English, with BERT performing better than the others [2]. In [3], a thorough analysis was carried out utilizing eight different machine learning models that included both traditional and deep learning techniques. XGBoost was found to be the most accurate classifier for detecting COVID-19 misinformation compared to three deep learning classifiers (CNN, RNN, and CRNN) and several traditional classifiers (RF, NB, SGD, and SVM). In [4], the BiLSTM model continuously achieved the highest accuracy rate in various datasets and training modes. In [5], CNN and BiLSTM were used to identify fake news on Twitter. In [6], a state-of-the-art deep learning approach that involved resampling in the latent space was proposed to address the problem of class imbalance in fake news detection. The process started with the deployment of a bidirectional variational autoencoder to extract rich and insightful latent representations from textual data. After that, a variety of resampling techniques, oversampling, undersampling, and hybrid sampling, were carefully applied to the latent vectors, which by nature display class bias. In [7], a combination of CNN and Bi-LSTM techniques was utilized to develop a machine-learning model to identify fake news in Arabic. This method greatly improved feature extraction, resulting in an increase in accuracy of more than 7% for both binary and multiclass classifications, highlighting how well the model worked to reduce misclassification problems in fake news in Arabic.

In [8], imbalanced classification techniques, such as SMOTE, Random Over Sample (ROS), and Random Under-Sample (RUS), were adopted to classify imbalanced Qur'anic topics. Previous studies used Keras embedding layers with pre-trained word embeddings, such as word2vec, fastText, ARBERT, and MARBERT, and CNNs or BiLSTM for false news detection. Oversampling and undersampling techniques are commonly followed to address class imbalance. Oversampling, including methods such as SMOTE, increases minority class samples, whereas undersampling reduces majority class samples. These methods are critical to improving model performance on imbalanced datasets [5]. In [9], SMOTE in conjunction with word2vec embeddings was deployed to address the problem of imbalanced data in Arabic sentiment analysis. With this method, the minority class is better represented, improving sentiment classification accuracy within the unique Arabic linguistic framework. In [10], a novel method combining SMOTE with a Complementary Neural Network (CMTNN) was presented to address the problem of imbalanced data classification. The efficacy of this approach was confirmed by contrasting its performance with well-known classification algorithms, such as Artificial Neural Networks (ANN), k-Nearest Neighbor (k-NN), and SVM, demonstrating its potential to improve predictive accuracy in situations where data are not evenly distributed.

Currently, multiple datasets are applied in the investigation of fake news in Arabic. The Arabic Fake News Dataset (AFND) [17] is very important in the field of detecting news credibility, especially with regard to social media platforms. In [15], an extensive Twitter dataset was created to analyze COVID-19 disinformation in both Arabic and English. ARACOVID19-MFH [14] combines ten different subtasks with predefined and mutually exclusive labels into a single unified task. In [16], a dataset of Arabic tweets related to COVID-19 was manually annotated into 13 different classes, such as misinformation (labeled "rumor"), different COVID-related topics like cure, virus info, and governmental measures, and situational information classes such as advice and support. This includes some of the most popular content that was retweeted during the early stages of the COVID-19 pandemic. In [1], a sample of 2,000 tweets was manually annotated to identify misinformation. In [3], an extensive Arabic Twitter dataset related to the COVID-19 pandemic was created. The tweets in this dataset were manually classified into two categories, false and true. In [18], an Arabic dataset of misinformation related to COVID-19, which consisted of approximately 6.7K tweets annotated with multiple classes and labels, was presented.

In [19], a manually annotated corpus was used as a baseline for the proposed automated system to identify fake news in Arabic text. Utilizing LR with distinct features, this study obtained an impressive 93.3% F1-score on the automatically annotated dataset and an F1-score of 87.8% on the manually annotated dataset. In [20], LR was employed to classify WhatsApp messages as appropriate or inappropriate, achieving an accuracy of 81% in identifying potentially harmful content. The effectiveness of LR in identifying messages as safe or harmful shows the potential and difficulties of similar techniques in Arabic fake news detection, where data imbalance is still a major obstacle. This comparison highlights how sophisticated solutions are required to increase classification accuracy in a variety of text analysis applications.

In the context of Arabic fake news detection, a significant weakness is identified in the application of imbalanced data handling techniques. Although several approaches have been proposed for other languages and domains, these sophisticated techniques have not adequately addressed the particular challenges of Arabic Natural Language Processing (NLP), including its rich morphology and dialectical variations. This study aims to close this gap by improving detection accuracy and model generalizability through the adaptation of imbalanced data techniques, specifically designed for Arabic fake news datasets. Table I illustrates an overview of recent datasets implemented in studies of Arabic-language fake news detection on social media platforms.

TABLE I. OVERVIEW OF RECENT DATASETS USED IN FAKE NEWS DETECTION IN ARABIC

Dataset Name	Size	Classification Type	Balance	Models Used	Handling Imbalance
Arabic Rumor-Non-Rumor [11]	10,000 tweets	Binary (Rumor, Non-Rumor)	Balanced	Transformer models (GigaBERT, Roberta-Base, Arabert, etc.)	-
ArCOV-19 and ArCOV19-Rumors [12, 13]	9,414 tweets	Binary (False, True)	Balanced	Bi-GCN, RNN+CNN, AraBERT, MARBERT	-
AraCOVID19-MFH [14]	10,800 tweets	Multi-class (10 labels)	Imbalanced	Transformer models	No
COVID-19 Tweets [15]	4,966 tweets	Binary and Fine-grained (10 classes)	Imbalanced	Transformer models	No
ArCorona: Analyzing Arabic Tweets [16]	8,000 tweets	Binary and Fine-grained (13 classes)	Imbalanced	SVM, AraBERT	No
COVID-19 Tweets [1]	2,000 tweets	Three classes (False, True, Unrelated)	Imbalanced	LR, SVC, NB	No
Large Arabic Twitter Dataset on COVID-19 [3]	8,786 tweets	Binary (Misleading, Not-Misleading)	Imbalanced	Traditional (RF, XGB, NB, SGD, SVC) and deep learning (CNN, RNN, CRNN)	AUCPRLoss loss function to optimize for AUC
Arabic Fake News Dataset (AFND) [17]	606,912 news articles	Three classes (Credible, Not-Credible, Undecided)	Imbalanced	CNN, LSTM	No

II. DATASETS AND DATA IMBALANCE ANALYSIS

Effective detection of fake news hinges on the quality and nature of the datasets upon which detection models are trained and tested. In machine learning, handling imbalanced data is essential, as uneven class distributions can lead to biased predictions and compromise the generalizability of algorithms. This section explores two distinct datasets: The AFND, characterized by its imbalanced class distribution, and the ArCOV-19 dataset, which is balanced.

A. The Arabic Fake News Dataset (AFND)

AFND is a large annotated dataset collected from 134 public news websites in 19 Arab countries and comprising 606,912 articles [17]. News credibility on social media is a critical problem that AFND attempts to solve. The class imbalance in the dataset presents a unique opportunity to explore approaches for dealing with imbalanced data issues. The dataset comprises 167,233 'not credible', 207,310 'credible', and 232,369 'undecided' news articles, as evidenced in Figure 1. Its diversity and size are noteworthy. The class distribution reveals a moderate imbalance: 38.29% of the dataset is classified as undecided, 34.16% as credible, and 27.55% as not credible. This distribution presents a distinct difficulty for machine learning algorithms, highlighting the necessity of strategic management to reduce bias in favor of the undecided category, which marginally predominates over the others. To ensure that predictive models derived from this dataset can generalize well across all categories of news credibility, this scenario offers a valuable opportunity to investigate and improve techniques meant to address dataset imbalance.

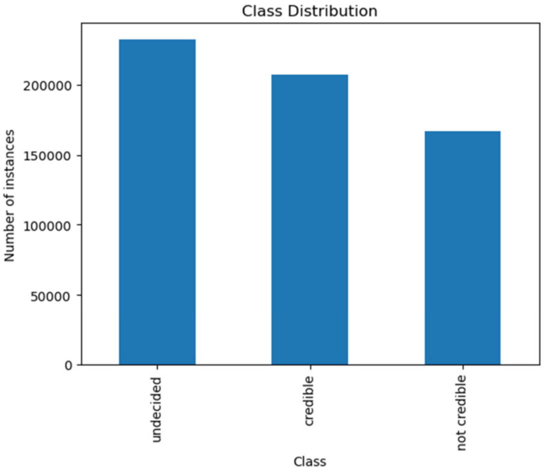


Fig. 1. Class distribution in AFND.

Given the complexity of the Arabic language, this dataset poses a special challenge, as its use sheds light on how well the models process and categorize Arabic textual data. In addition to class imbalance, there is significant variation in the amount of words per article. However, its size and diversity make it especially valuable for training deep learning models in Arabic fake news detection, as it provides a rich environment for experimenting with and improving different NLP methods for Arabic language processing, noise reduction, and error handling.

Data marked unspecified were removed because their ambiguous nature could compromise the precision and clarity needed for accurate classification, introducing noise or bias into the results. This reduced the total size of the dataset and altered the ratio of credible to non-credible categories, as shown in Figure 2. Maintaining this balance is essential to prevent classification models from being intrinsically biased in favor of one class. This study focused on classifying fake news by combining the news article text and title. This combination enables a more thorough analysis because texts provide a more comprehensive story while titles may only provide a limited amount of information. Through the integration of features derived from both the title and the text, the model gains an enhanced ability to detect patterns and indicators linked to fake information. Advanced NLP techniques are required to process and analyze complex and varied data efficiently. Integrating these two essential components is substantial to improving the classification model's dependability and accuracy and guaranteeing a more reliable detection of fake news in the dataset.

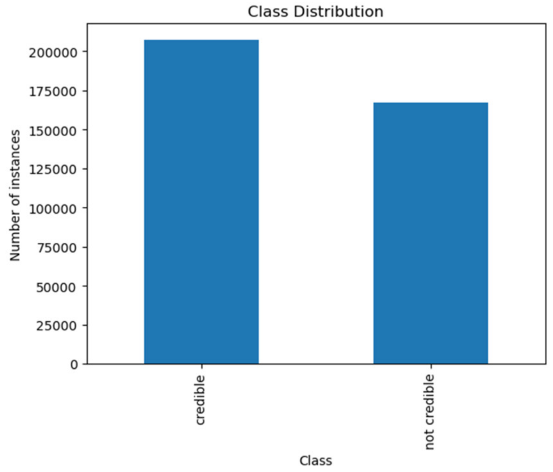


Fig. 2. Class distribution of AFND after removing the unspecified class.

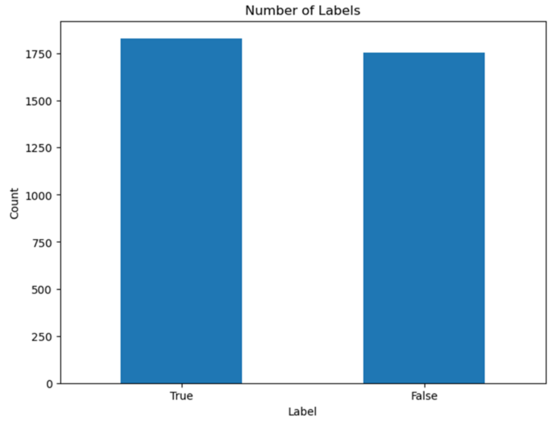


Fig. 3. Distribution analysis of the ArCOV19-Rumors dataset.

B. *ArCOV-19-Rumors Dataset*

The ArCOV19-Rumors dataset [13] was specially assembled to detect false information about the COVID-19 pandemic in Arabic. This dataset contains 9.4K tweets that are relevant to the 138 verified claims that make up this collection. The dataset covers a wide range of topics affected by the pandemic, including health, social, political, sports, entertainment, and religious claims, and facilitates both claim-level and tweet-level verification tasks. The dataset is a valuable resource for the development and training of verification systems due to its balanced distribution of true and false tweets. This dataset includes retweets and replies, increasing its usefulness for thorough disinformation analysis. Figure 3 presents the number of rows for the two labels, where 1831 occurrences are marked True and 1753 are marked False, corresponding to approximately 51.09% and 48.91%, respectively.

III. METHODOLOGY

This study is based on the effective and powerful text representation and classification tool FastText, which allows both models to be improved by offering high-quality word embeddings. The aim is to examine how different approaches handle unbalanced data, a crucial factor in enhancing the performance of machine learning models. Class weight adjustment modifies the model's loss function to increase the importance of minority classes and promote a balanced learning process. Undersampling tries to bring the majority class prevalence closer to that of the minority class, though it may cause information loss. This study also investigates oversampling techniques, such as SMOTE and SMOTEENN, to enhance the representation of underrepresented classes.

A. *Class Weight Adjustment*

Class weight adjustment gives different weights in different classes within a dataset depending on their frequencies. The specific method's fundamental idea is to give higher weights to the minority classes and lower weights to the majority classes The goal is to improve a model's predictive performance, particularly its sensitivity and accuracy in detecting instances of minority classes, by incorporating class weights into the training process. Given a dataset D with a multi-class target variable, let Y represent the set of target labels. Each label y_i corresponds to a unique class C.

Let $C = \{c_{1,}c_2, \ldots\ldots, c_C\}$ be the set of all unique classes in Y, where $f(c_j)$ is the frequency of class c_j in the dataset. After that, class weights are calculated as:

$$W(c_j) = \frac{N}{C\times f(c_j)}$$

where N is the total number of samples in D. The objective function is then used to include these weights in the training process.

$$min \sum_{i=1}^{N} w\,(y_i).L(\theta; x_i, y_i)$$

B. *Undersampling*

Undersampling involves deleting some instances of the majority class to reduce its size. This method equalizes the class distribution to facilitate algorithm learning. However, undersampling may cause a loss of potentially significant data from the majority class [21]. Random undersampling removes instances from the majority classes at random, disregarding their position in the feature space or their influence.

C. Oversampling, SMOTE, and SMOTEENN

Oversampling is an essential data preprocessing method for dealing with unbalanced datasets. In many domains, imbalanced datasets are prevalent. The minority class, which is frequently the class of greater interest, may perform poorly in predictions due to models being biased toward the majority class. Oversampling methods such as SMOTE and SMOTEENN (SMOTE + Edited Nearest Neighbors) are employed to increase the number of instances in the minority class to balance the dataset. Class distributions in datasets can be balanced using the Random Oversampling (ROS) technique. ROS is simple and non-heuristic since it does not require making thoughtful decisions. To balance the training model, ROS essentially duplicates samples from the minority class. However, this can occasionally lead to overfitting, particularly for minority classes [22, 23]. ROS replicates its observations at random, sometimes with replacement, to increase the number of instances in the minority class. Although this approach can make the dataset more balanced, it also makes the dataset larger, which could result in increased computational costs.

ALGORITHM 1: SMOTE

```
Input:
  M: Matrix containing minority class samples
  N: percentage mount of SMOTE
  k: Number of nearest neighbors
Output:
  Synthetic: Matrix containing synthetic minority
  class samples

  c ← Number of minority class samples in M
  numAttributes ← Number of attributes in M
  N ← N / 100
  distances, indices ← get_neighbors(M, k)
  Synthetic ← Empty matrix of size (N * c) *
     numAttributes
  synth_idx ← 0

  for i ← 1 to c:
    for j ← 1 to N:
        neighbor ← Random integer between 1 and k
        diff ← M[indices[i, neighbur]] - M[i]
        gap ← Random float between 0 and 1
        Synthetic[synth_idx] ← M[i] + gap * diff
        synth_idx ← synth_idx + 1

Output Synthetic
```

1) SMOTE

Datasets from real-world situations typically include a higher percentage of normal than abnormal or notable cases. Misclassifying abnormal instances as normal presents great challenges. In contrast to simply focusing on undersampling the majority class, an effective strategy combines the oversampling of the minority class with the undersampling of the majority class to improve the efficacy of classifiers [24]. SMOTE creates synthetic samples for the underrepresented class. SMOTE can be used to generate fake instances of uncommon news topics in situations such as Arabic fake news detection, where these topics may be less common. This makes it easier for algorithms to be trained to identify and categorize these minority cases. More research is needed to fully understand the specific application of SMOTE in the context of Arabic fake news detection, as it is not covered in published studies. Algorithm 1 describes the SMOTE algorithm, which produces synthetic samples to balance class distribution given a matrix M containing minority class samples.

2) SMOTEENN (SMOTE + Edited Nearest Neighbors)

SMOTE and Edited Nearest Neighbors (ENN) are combined in the SMOTEENN technique [25]. This technique uses SMOTE to oversample the minority class to address class imbalance. Then, ENN is applied to eliminate noise and unclear samples. This novel method has greatly improved the performance of machine learning models, particularly in domains where data imbalance is a common problem. SMOTE creates synthetic samples for the minority class by interpolating between current instances and their neighbors. The final dataset is subjected to an ENN application to eliminate data that, given their neighborhood, are likely to be noisy or incorrectly classified. An instance is removed as noise if its class differs from the majority of its nearest neighbors.

ALGORITHM 2: SMOTEENN

```
Input:
  D: Dataset with minority and majority classes
  k: Number of nearest neighbors (default k=3 for
    ENN)
  Target_Proportion: The desired ratio for class
    balance
Output:
  D_Balanced: Balanced dataset after applying
    SMOTE-ENN

Procedure:
# SMOTE Process:
  While the proportion of minority class in D <
    Target_Proportion:
    Randomly select an instance x_min from the
      minority class in D
    Find k nearest neighbors of x_min within the
      minority class
    For each neighbor x_neighbor:
      Generate a random number rand_num between 0
        and 1
      Create synthetic instance x_synthetic =
        x_min + rand_num * (x_neighbor - x_min)
      Add x_synthetic to D

# ENN Process:
  For each instance x in D:
    Find k nearest neighbors of x in D
    Determine the majority class among neighbors
    If the class of x ≠ majority class of neighbors:
      Remove x from D
  Return D as D'
```

IV. CLASSIFIER IMPLEMENTATION

This study selected classifiers based on their demonstrated effectiveness in a variety of prediction tasks. The selected models accommodate a wide range of data characteristics and task requirements, spanning both conventional machine learning and sophisticated deep learning techniques. XGB excels at classification tasks in terms of speed and performance. XGBoost is a popular model for many different problems because of its effectiveness in managing sparse data and large datasets. It also optimizes computational resources and predictive accuracy using the gradient boosting framework. RF is a powerful and adaptable classifier that builds several decision trees during training and outputs the class that is the mean of the classes of the individual trees. CNNs have demonstrated efficacy in text analysis by exploiting their capacity to extract hierarchical features and patterns. By using their convolutional layers to find important features in word or character sequences, they are adept at identifying and deciphering complex structures found in text data. BiGRU processes data in both forward and backward directions, improving the standard GRU framework by capturing information from future and past contexts. This model excels when dealing with sequential data, where context from both directions is essential for precise predictions. BiLSTM networks extend the capabilities of LSTM in comprehending sequence data context by processing data in forward and backward directions. BiLSTM models work especially well for tasks such as language processing and time-series analysis, where precise predictions depend on knowledge from historical and future data points. The CNN-LSTM hybrid combines the spatial feature extraction capabilities of CNN with the long-term dependency capture expertise of LSTM. This combination works especially well in tasks such as video analysis or complex sequence prediction, where accurate predictions depend on local features and their temporal evolution. The CNN-BiGRU model combines the spatial pattern recognition skills of CNNs with the bidirectional sequence processing abilities of BiGRUs. This fusion is perfect for complex tasks that require nuanced interpretation of sequential and spatial data because it enables detailed data analysis where understanding the immediate details as well as the broader context is crucial.

V. EXPERIMENTATION AND RESULTS

The models were trained on the two datasets. Class weights, random under-sampling, SMOTE, and SMOTEENN were used to handle class imbalance. The performance of the models was evaluated using accuracy, AUC, precision, recall, and F1-score.

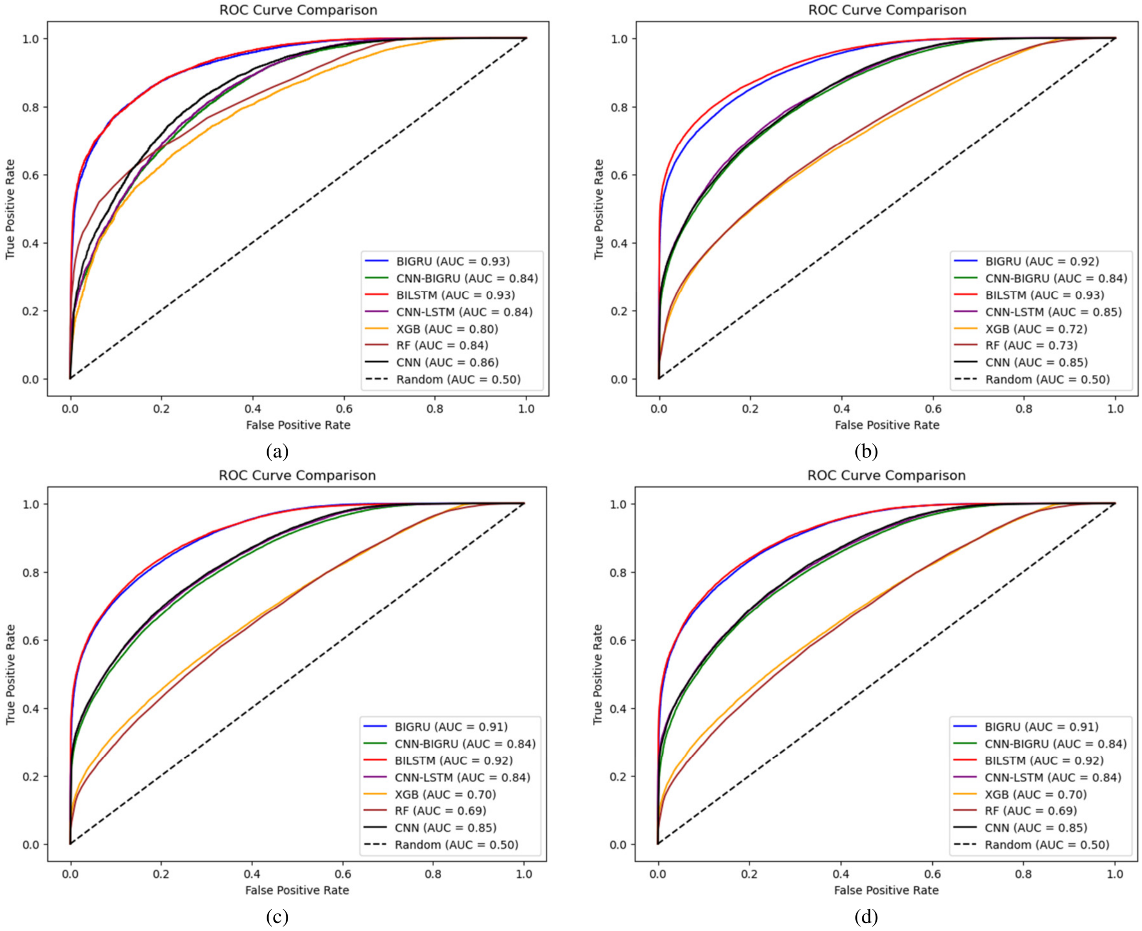
A. *Performance on AFND (Imbalanced Data)*

Machine learning models were trained on the AFND to assess class imbalance. Then, class balancing techniques were utilized to improve model performance, and Table II and Figure 4 display the results. Using SMOTEENN, XGBoost showed an increase in accuracy from 0.64 to 0.74 and in AUC from 0.69 to 0.80. SMOTEENN significantly improved RF, increasing accuracy from 0.64 to 0.75 and AUC from 0.69 to 0.84. CNN exhibited the highest AUC of 0.84 and an accuracy of 0.74 when no particular handling technique was employed, demonstrating its resilience to class imbalance. The robustness of CNN was enhanced when using SMOTE, as it achieved 0.79 accuracy and 0.86 AUC. BiGRU and BiLSTM outperformed the other models when utilizing SMOTEENN, obtaining the highest accuracy of 0.85 and an AUC of 0.93. Additionally, the CNN-LSTM and CNN-BiGRU hybrid models responded favorably to SMOTEENN, peaking at 0.78 accuracy and 0.84 AUC. These findings highlight the ability of bidirectional recurrent neural networks to handle complex dependencies and sequential data, making the BiGRU and BiLSTM models especially well-suited for the subtleties of textual fake news detection.

TABLE II. COMPARATIVE PERFORMANCE METRICS MODELS WITH VARIOUS CLASS IMBALANCE HANDLING TECHNIQUES

Model	Metric	Baseline	Class Weights	Under-sampling	SMOTE	SMOTEENN
XGB	Accuracy	0.64	0.65	0.65	0.65	0.74
	AUC	0.69	0.70	0.70	0.72	0.80
	Precision	0.65	0.65	0.65	0.65	0.73
	Recall	0.64	0.65	0.65	0.65	0.74
	F1	0.63	0.63	0.63	0.65	0.72
RF	Accuracy	0.64	0.63	0.63	0.65	0.75
	AUC	0.69	0.69	0.69	0.73	0.84
	Precision	0.64	0.64	0.64	0.65	0.78
	Recall	0.64	0.63	0.63	0.65	0.75
	F1	0.62	0.62	0.62	0.65	0.72
CNN	Accuracy	0.74	0.75	0.75	0.75	0.79
	AUC	0.84	0.85	0.85	0.85	0.86
	Precision	0.74	0.75	0.75	0.75	0.79
	Recall	0.75	0.75	0.75	0.75	0.80
	F1	0.74	0.75	0.75	0.75	0.79
BIGRU	Accuracy	0.82	0.82	0.82	0.83	0.85
	AUC	0.91	0.91	0.91	0.92	0.93
	Precision	0.82	0.82	0.82	0.83	0.85
	Recall	0.82	0.82	0.82	0.83	0.85
	F1	0.82	0.82	0.82	0.83	0.85
BILSTM	Accuracy	0.81	0.82	0.82	0.84	0.85
	AUC	0.91	0.92	0.92	0.93	0.93
	Precision	0.81	0.82	0.82	0.84	0.85
	Recall	0.81	0.82	0.82	0.84	0.85
	F1	0.81	0.82	0.82	0.84	0.85
CNN-LSTM	Accuracy	0.75	0.75	0.75	0.75	0.78
	AUC	0.84	0.84	0.84	0.85	0.84
	Precision	0.75	0.75	0.75	0.75	0.78
	Recall	0.75	0.75	0.75	0.75	0.78
	F1	0.75	0.75	0.75	0.75	0.78
CNN-BIGRU	Accuracy	0.74	0.74	0.74	0.75	0.78
	AUC	0.84	0.84	0.84	0.84	0.84
	Precision	0.74	0.74	0.74	0.75	0.77
	Recall	0.74	0.74	0.74	0.75	0.78
	F1	0.74	0.74	0.74	0.75	0.77

The experimental results provide informative insights into the performance of different models and strategies for handling class imbalance on AFND. BiGRU models show exceptional performance, stressing their suitability for text-related tasks, particularly in the context of Arabic language processing, where contextual understanding is critical. Almost all models tested performed best when handling imbalanced data with SMOTEENN. Real-world news can be distinguished from fake news by using temporal and contextual information, which is where recurrent neural architectures, such as BiGRU and BiLSTM, excel over conventional techniques. These results motivate further research into more complex ensemble methods and hybrid models to take advantage of various strategies and push the boundaries of fake news detection even further. SMOTEENN seems to be essential for maximizing performance in imbalanced datasets, since it combines oversampling the minority class with removing noisy instances from the majority class. It should be noted that CNN outperformed traditional machine learning methods and was less dependent on imbalance handling techniques. This might be explained by the fact that it can extract important features from various levels of abstraction, which is especially helpful for tasks involving text classification. The CNN-LSTM and CNN-BiGRU hybrid models performed better with SMOTEENN, but not as well as the standalone recurrent models. This could imply that although adding value to CNNs by combining them with recurrent layers, further fine-tuning of the particular architecture is necessary to fully utilize the benefits of both neural network types. Although the models performed differently depending on the class imbalances present in AFND, it is important to compare these results with their performance on a balanced dataset.



(a) (b) (c) (d)

Fig. 4. ROC curve comparison for the six classifiers on AFND using different class balancing techniques: (a) SMOTEENN, (b) SMOTE, (c) Undersampling, and (d) Class Weights.

B. Performance on ArCOV-19 (Balanced Dataset)

Table III reveals the effectiveness of various class balance handling techniques and the models examined on the ArCOV-19 dataset. XGBoost performed best with SMOTEENN, achieving the highest scores in accuracy (0.91) and AUC (0.97). RF achieved its best results when using SMOTEENN, with an accuracy of 0.88 and an AUC of 0.95. Furthermore, CNN performed well in all handling methods, achieving the highest AUC (0.98) when utilizing SMOTEENN. BiGRU showed a remarkable improvement with SMOTEENN, achieving 0.95 accuracy and 0.98 AUC. BiLSTM also reached its highest performance with SMOTEENN, with 0.93 accuracy and 0.98 AUC. The CNN-LSTM and CNN-BiGRU hybrid models reflected similar trends, where SMOTEENN provided the best results. However, CNN-BiGRU achieved 0.92 accuracy and 0.96 AUC, and CNN-LSTM achieved 0.90 accuracy and 0.97 AUC. XGBoost demonstrated a trade-off between sensitivity and precision, as evidenced by its high true positive rate in identifying misinformation with notable false positives. The balanced ArCOV-19 dataset was used to analyze misinformation detection and both the BiLSTM and BiGRU models exhibited strong classification abilities. As a measure of how well the BiLSTM model detects misinformation, it produced a high number of true positives (12,309) and a moderate number of false positives (1,682). Similarly, the BiGRU model demonstrated excellent precision, managing a similar number of false positives (2,039) and a marginally higher number of true positives (12,756). These findings emphasize how well recurrent neural networks can distinguish between true and fake information, as both models disclosed a good balance between recall and precision.

TABLE III. PERFORMANCE METRICS OF MODELS ON ARCOV-19 WITH VARIOUS CLASS BALANCE HANDLING TECHNIQUES

Model	**Metric**	**Baseline**	**Class weights**	**Under-sampling**	**SMOTE**	**SMOTEENN**
XGB	Accuracy	0.79	0.85	0.85	0.81	0.91
	AUC	0.87	0.92	0.92	0.90	0.97
	Precision	0.79	0.85	0.85	0.81	0.91
	Recall	0.79	0.85	0.85	0.81	0.91
	F1	0.79	0.85	0.85	0.81	0.91
RF	Accuracy	0.78	0.83	0.83	0.80	0.88
	AUC	0.88	0.92	0.92	0.89	0.95
	Precision	0.78	0.83	0.83	0.80	0.88
	Recall	0.78	0.83	0.83	0.80	0.88
	F1	0.78	0.83	0.83	0.80	0.88
CNN	Accuracy	0.85	0.88	0.89	0.88	0.91
	AUC	0.93	0.95	0.96	0.92	0.98
	Precision	0.85	0.88	0.89	0.88	0.91
	Recall	0.85	0.88	0.89	0.88	0.91
	F1	0.85	0.88	0.89	0.88	0.91
BIGRU	Accuracy	0.87	0.87	0.87	0.87	0.95
	AUC	0.94	0.94	0.95	0.93	0.98
	Precision	0.87	0.87	0.87	0.87	0.95
	Recall	0.87	0.87	0.87	0.87	0.95
	F1	0.87	0.87	0.87	0.87	0.95
BILSTM	Accuracy	0.86	0.88	0.88	0.88	0.93
	AUC	0.94	0.94	0.94	0.93	0.98
	Precision	0.87	0.88	0.88	0.88	0.93
	Recall	0.86	0.88	0.88	0.88	0.93
	F1	0.86	0.88	0.88	0.88	0.93
CNN-LSTM	Accuracy	0.86	0.88	0.88	0.88	0.90
	AUC	0.94	0.96	0.95	0.93	0.97
	Precision	0.86	0.88	0.88	0.88	0.90
	Recall	0.86	0.88	0.88	0.88	0.90
	F1	0.86	0.88	0.88	0.88	0.90
CNN-BIGRU	Accuracy	0.86	0.89	0.88	0.89	0.92
	AUC	0.93	0.96	0.96	0.92	0.96
	Precision	0.86	0.89	0.88	0.89	0.92
	Recall	0.86	0.89	0.88	0.89	0.92
	F1	0.86	0.89	0.88	0.89	0.92

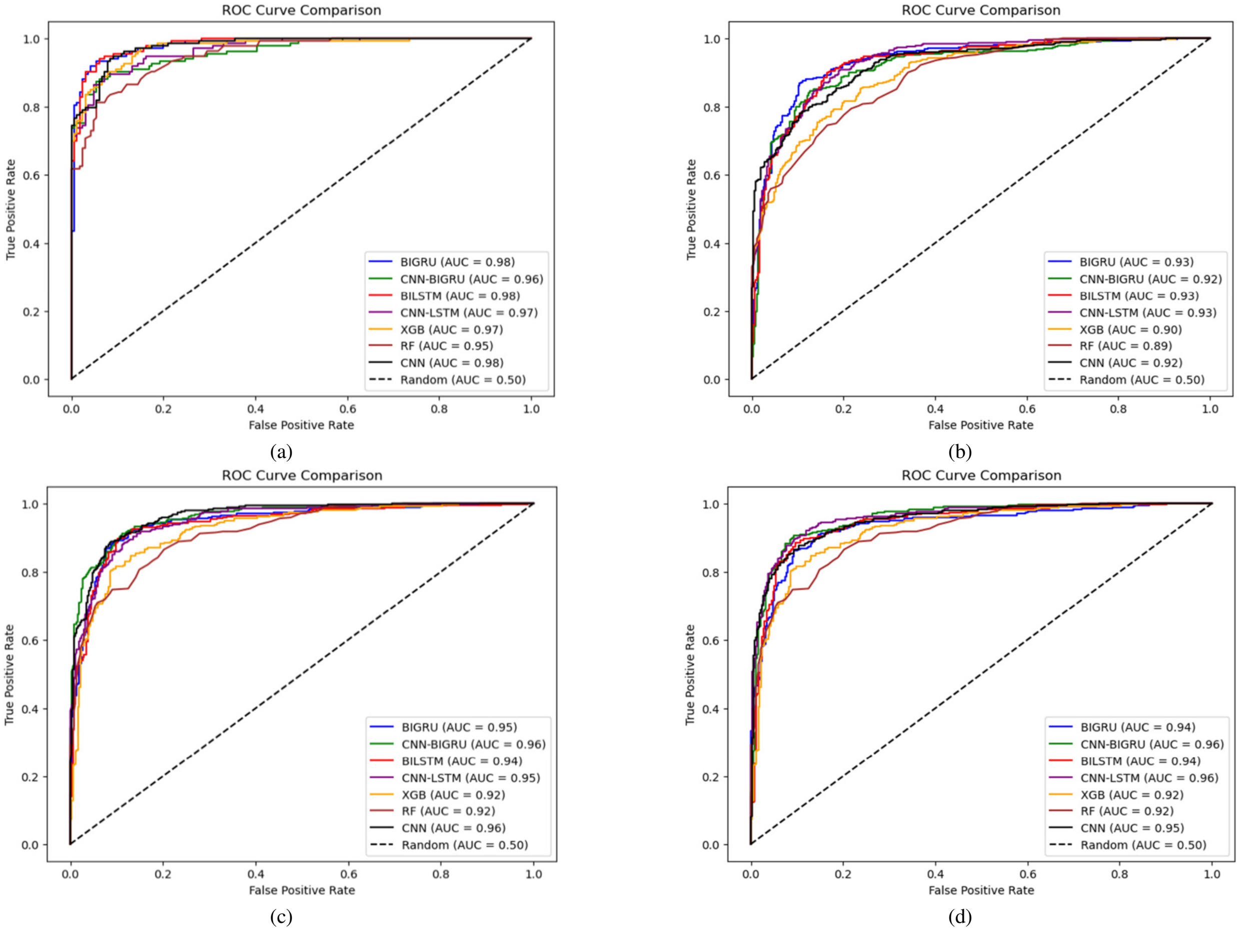
Figure 5 presents a detailed comparative analysis that highlights the performance of different machine learning models when subjected to different balancing techniques. The results on the balanced ArCOV-19 dataset show that SMOTEENN consistently improves both the performance of conventional machine learning models and sophisticated neural networks. This indicates that by improving the quality of the training data through adding synthetic samples and removing overlapping samples, SMOTEENN not only addresses class imbalance, but also adds to model robustness in balanced scenarios. The results suggest that SMOTEENN is the most effective technique for handling datasets with complex class distributions, as it was the most effective method in all models. Neural network models, especially the recurrent ones (BiGRU and BiLSTM) and their hybrid versions (CNN-LSTM, CNN-BiGRU), demonstrated notable performance advantages. Recurrent and hybrid models excel at capturing temporal and contextual nuances within the text, which is crucial to effectively distinguishing between true and fake news. The superiority of these models in handling complex datasets indicates their potential for deployment in fake news detection systems, where the ability to parse and understand subtle cues and discrepancies in news articles can be crucial.

Both datasets showed that neural network models and their hybrid forms consistently outperformed traditional machine learning models (XGB, RF) when improved by SMOTEENN. The neural network models were particularly robust across different balancing techniques but displayed optimal performance with SMOTEENN. The models exhibited significant variation in their performance on AFND depending on the balancing technique used, suggesting that the dataset is sensitive to how class imbalance is addressed. Despite the balanced nature of the ArCOV-19 dataset, models benefited from utilizing SMOTEENN, indicating that synthetic sample generation and noise reduction are beneficial even in datasets without severe class imbalances. In the ArCOV-19 dataset, CNN models disclosed excellent stability in a variety of approaches, with class weights and undersampling only slightly improving performance. This implies that the built-in feature extraction capabilities of CNNs are sufficiently resilient to small changes in class distribution. SMOTEENN consistently exhibits the highest improvement across all models and metrics in both datasets, highlighting its potency in handling imbalanced data. Although the imbalanced dataset demonstrated improvements with balancing techniques, the balanced dataset showed more noticeable improvements, especially in metrics such as F1 and AUC, indicating that the models perform better when predicting instances of the minority class in the balanced context. When deploying data balancing techniques, deep learning models (BIGRU, BILSTM, CNN-LSTM, and CNN-BIGRU) typically show more noticeable improvements, especially in the balanced dataset, demonstrating their sensitivity and flexibility to changes in data distribution.

This study builds on existing research in the field of fake news detection by focusing on the less-explored area of Arabic-language content. Although data imbalance techniques have been extensively studied in fake news detection in English, there are few comparable efforts for content in Arabic. By filling this gap, this study advances knowledge on fake news detection across linguistic contexts and offers insightful information on efficient data balancing strategies to enhance model performance in Arabic fake news detection. The use of CNN-BIGRU with SMOTEENN displays notable enhancements, expanding on the approaches applied in previous studies. Even though the dataset initially does not seem to be very unbalanced, using SMOTEENN significantly improves some metrics for these models. This implies that there may be some underlying complexity in the dataset that SMOTEENN helps to reduce, allowing the models to more accurately identify and categorize the minority class. This is essential in situations such as diagnosing rare diseases or detecting fraud.

VI. CONCLUSION AND FUTURE WORK

This study addressed the critical challenge of dataset imbalance in Arabic fake news detection, offering valuable insights into the effectiveness of various data balancing techniques. The former contributes to the field of Arabic fake news detection through demonstrating significant improvements in model performance and filling a literature gap. This study shows how data balancing techniques can significantly enhance the performance of machine learning models. With a consistent improvement in all important performance metrics across different models, SMOTEENN remained the most effective technique. For languages and applications where minority class detection precision and recall are critical, the present study emphasizes the need to use suitable data balancing techniques to counter the difficulties caused by imbalanced datasets. This study advances the field of detecting fake news while also adding to the larger discussion of how to make machine-learning applications reliable and robust in situations where they have a significant social impact. In general, the use of SMOTEENN demonstrated its effectiveness in addressing class imbalance in the context of Arabic fake news detection, especially with XGBoost, RF, BiGRU, and BiLSTM. The results indicate that SMOTE and SMOTEENN are particularly effective in improving the performance of machine learning models on the AFDN.



(a) (b)

(c) (d)

Fig. 5. Comparisons of ROC curves between various class balancing strategies for classifiers using the ArCOV-19 dataset: (a) SMOTEENN, (b) SMOTE, (c) Undersampling, and (d) Class Weights.

Future studies could investigate the creation of fresh or enhanced oversampling strategies suited to particular characteristics of unbalanced text data. Furthermore, the comparative efficacy of deep learning models, in particular BiGRU and BiLSTM, in managing imbalanced datasets requires more research to fully comprehend the mechanisms underlying their superior performance. According to the findings, deep learning models, especially those that combine convolutional and recurrent layers, offer notably improved performance over traditional machine learning models that offer a stable baseline. These models are benefited from the rich feature representations that augmented data in a balanced dataset can teach them. Subsequent research efforts should delve into the assimilation of these methods with increasingly sophisticated model architectures and their suitability for employment in other languages and domains where dataset imbalance is a pervasive issue.

REFERENCES

[1] L. Alsudias and P. Rayson, "COVID-19 and Arabic Twitter: How can Arab World Governments and Public Health Organizations Learn from Social Media?," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, Apr. 2020, [Online]. Available: https://aclanthology.org/2020.nlpcovid19-acl.16.

[2] B. Ahmed, G. Ali, A. Hussain, A. Baseer, and J. Ahmed, "Analysis of Text Feature Extractors using Deep Learning on Fake News," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 7001–7005, Apr. 2021, https://doi.org/10.48084/etasr.4069.

[3] S. Alqurashi, B. Hamoui, A. Alashaikh, A. Alhindi, and E. Alanazi, "Eating Garlic Prevents COVID-19 Infection: Detecting Misinformation on the Arabic Content of Twitter," *arXiv.org*, Jan. 09, 2021. https://arxiv.org/abs/2101.05626v1.

[4] K. M. Fouad, S. F. Sabbeh, and W. Medhat, "Arabic fake news detection using deep learning," *Computers, Materials and Continua*, vol. 71, no. 2, pp. 3647–3665, 2022, https://doi.org/10.32604/cmc.2022.021449.

[5] S. Alyoubi, M. Kalkatawi, and F. Abukhodair, "The Detection of Fake News in Arabic Tweets Using Deep Learning," *Applied Sciences*, vol. 13, no. 14, Jan. 2023, Art. no. 8209, https://doi.org/10.3390/app13148209.

[6] S. Bhattacharjee, S. Maity, and S. Chatterjee, "Addressing Class Imbalance in Fake News Detection with Latent Space Resampling," in *Computational Intelligence in Pattern Recognition*, Kolkata, India, 2023, pp. 427–438, https://doi.org/10.1007/978-981-99-3734-9_35.

[7] A. Khalil, M. Jarrah, and M. Aldwairi, "Hybrid Neural Network Models for Detecting Fake News Articles," *Human-Centric Intelligent Systems*, vol. 4, no. 1, pp. 136–146, Mar. 2024, https://doi.org/10.1007/s44230-023-00055-x.

[8] B. S. Arkok and A. M. Zeki, "Classification of Quranic topics based on imbalanced classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, pp. 678–687, May 2021, https://doi.org/10.11591/ijeecs.v22.i2.pp678-687.

[9] S. Al-Azani and E. S. M. El-Alfy, "Using Word Embedding and Ensemble Learning for Highly Imbalanced Data Sentiment Analysis in Short Arabic Text," *Procedia Computer Science*, vol. 109, pp. 359–366, Jan. 2017, https://doi.org/10.1016/j.procs.2017.05.365.

[10] P. Jeatrakul, K. W. Wong, and C. C. Fung, "Classification of Imbalanced Data by Combining the Complementary Neural Network and SMOTE Algorithm," in *Neural Information Processing. Models and Applications*, Sydney, Australia, 2010, pp. 152–159, https://doi.org/10.1007/978-3-642-17534-3_19.

[11] A. B. Nassif, A. Elnagar, O. Elgendy, and Y. Afadar, "Arabic fake news detection based on deep contextualized embedding models," *Neural Computing and Applications*, vol. 34, no. 18, pp. 16019–16032, Sep. 2022, https://doi.org/10.1007/s00521-022-07206-4.

[12] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "ArCOV-19: The First Arabic COVID-19 Twitter Dataset with Propagation Networks," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Virtual), Dec. 2021, pp. 82–91. [Online]. Available: https://aclanthology.org/2021.wanlp-1.9.

[13] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "ArCOV19-Rumors: Arabic COVID-19 Twitter Dataset for Misinformation Detection," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Virtual), Dec. 2021, pp. 72–81. [Online]. Available: https://aclanthology.org/2021.wanlp-1.8.

[14] M. S. Hadj Ameur and H. Aliane, "AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News & Hate Speech Detection Dataset," *Procedia Computer Science*, vol. 189, pp. 232–241, Jan. 2021, https://doi.org/10.1016/j.procs.2021.05.086.

[15] F. Alam *et al.*, "Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic, Aug. 2021, pp. 611–649, https://doi.org/10.18653/v1/2021.findings-emnlp.56.

[16] H. Mubarak and S. Hassan, "ArCorona: Analyzing Arabic Tweets in the Early Days of Coronavirus (COVID-19) Pandemic," in *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, online, Dec. 2021. [Online]. Available: https://aclanthology.org/2021.louhi-1.1.

[17] A. Khalil, M. Jarrah, M. Aldwairi, and M. Jaradat, "AFND: Arabic fake news dataset for the detection and classification of articles credibility," *Data in Brief*, vol. 42, Jun. 2022, Art. no. 108141, https://doi.org/10.1016/j.dib.2022.108141.

[18] A. Khalil, M. Jarrah, M. Aldwairi, and Y. Jararweh, "Detecting Arabic Fake News Using Machine Learning," in *2021 Second International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, Tartu, Estonia, Nov. 2021, pp. 171–177, https://doi.org/10.1109/IDSTA53674.2021.9660811.

[19] A. R. Mahlous and A. Al-Laith, "Fake News Detection in Arabic Tweets during the COVID-19 Pandemic," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021, https://doi.org/10.14569/IJACSA.2021.0120691.

[20] F. M. U. Baran, L. S. A. Alzughaybi, M. A. S. Bajafar, M. N. M. Alsaedi, T. F. H. Serdar, and O. M. N. Mirza, "Etiqa'a: An Android Mobile Application for Monitoring Teen's Private Messages on WhatsApp to Detect Harmful/Inappropriate Words in Arabic using Machine Learning," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12012–12019, Dec. 2023, https://doi.org/10.48084/etasr.6174.

[21] T. Feizi, M. H. Moattar, and H. Tabatabaee, "A multi-manifold learning based instance weighting and under-sampling for imbalanced data classification problems," *Journal of Big Data*, vol. 10, no. 1, Oct. 2023, Art. no. 153, https://doi.org/10.1186/s40537-023-00832-2.

[22] M. Azadbakht, C. S. Fraser, and K. Khoshelham, "Synergy of sampling techniques and ensemble classifiers for classification of urban environments using full-waveform LiDAR data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 73, pp. 277–291, Dec. 2018, https://doi.org/10.1016/j.jag.2018.06.009.

[23] T. Hasanin, T. M. Khoshgoftaar, J. L. Leevy, and R. A. Bauder, "Severely imbalanced Big Data challenges: investigating data sampling approaches," *Journal of Big Data*, vol. 6, no. 1, Nov. 2019, Art. no. 107, https://doi.org/10.1186/s40537-019-0274-4.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, https://doi.org/10.1613/jair.953.

[25] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, Mar. 2004, https://doi.org/10.1145/1007730.1007735.

AUTHORS PROFILE

Eman Aljohani is a leading computer science researcher at Taibah University, specializing in artificial intelligence and data mining. With a Ph.D. from the University of York, she has made significant contributions through her research, particularly in the areas of Arabic text classification and neural network-based ensemble learning. Dr. Aljohani's research interests are broad, encompassing fields, such as healthcare, big data analytics, and AI applications in various domains. Her work, notable for its innovative approaches in these areas, has been featured in prestigious conferences and journals, showcasing her dedication to advancing technology and its practical applications across a range of fields.